可信大模型分控策略网关

用户名

admin

密码

登录


DC
DP
HDMI
USB
ETH0
ETH1
ETH2
ETH3

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

大模型网关运营看板 系统状态

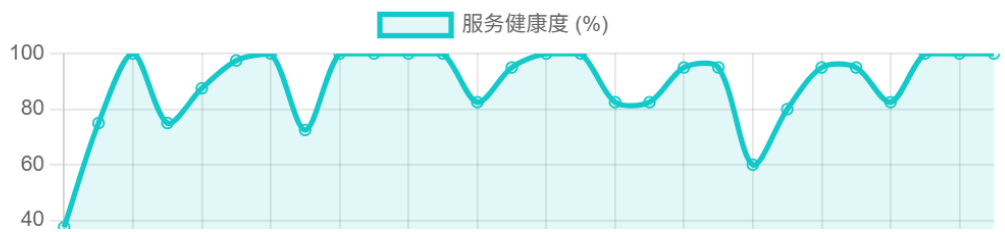
大模型网关运营看板 (近7天统计数据)

近 1 小时 近 24 小时 近 7 天 自动刷新：开启

总请求次数	消耗 Token 总计	安全与准入拦截数	Token 超额告警	当前连接数
196 次	1,094,537 输入: 757,235 | 输出: 337,302	25 次	29 次	1 个活跃连接

活跃用户数	服务健康度
6 人	100%

AI服务质量





系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

实时长尾系数 P95/Avg 前五模型

服务商-模型 (比值越高抖动越大)

1	UCloud2-glm-5.3	1.52x (P95:5463ms)
2	UCloud2-qwen3.6-35b-a3b	1.16x (P95:415ms)
3	阿里云百炼-deepseek-v4-pro	1.00x (P95:1071ms)
4	UCloud2-deepseek-v4-flash-0731	1.00x (P95:2015ms)
5	UCloud2-mimo-v2.6-flash	1.00x (P95:5342ms)

实时 P95 延迟前五模型

服务商-模型 (首字延迟 ms)

1	UCloud2-glm-5.3-flash	7490 ms
2	UCloud2-glm-5.3	5463 ms
3	UCloud2-mimo-v2.6-flash	5342 ms
4	UCloud2-deepseek-v4-flash-0731	2015 ms
5	UCloud2-deepseek-v4.1-flash-1017	1975 ms

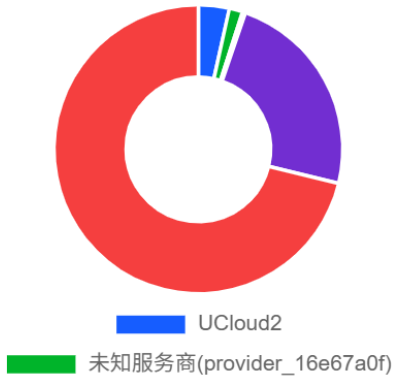
24小时服务健康度倒数前五模型

服务商-模型 (健康度越低越靠前)

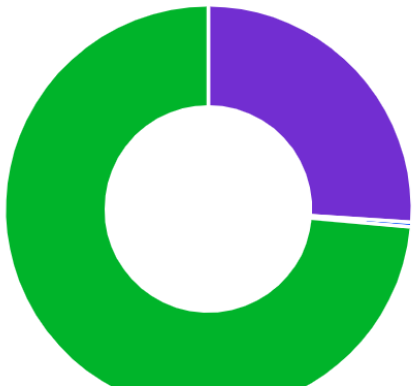
1	博伦智汇-deepseek-v4.1-flash-1017	57.1%
2	硅基流动-deepseek-v4.1-flash-1017	57.1%
3	天鸿智算-deepseek-v4-flash-vision-exp	64.3%
4	天鸿智算-deepseek-v4.1-flash	67.9%
5	天鸿智算-deepseek-v4.1-flash-1017	67.9%

近7天资源与配额占用分布

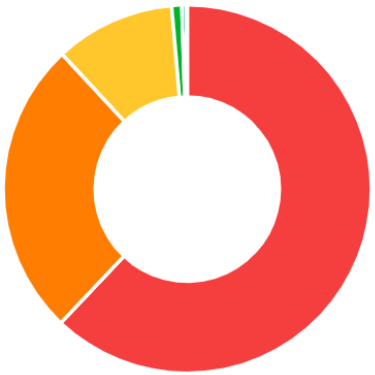
服务商 Token 消耗占比



工作组 Token 消耗占比



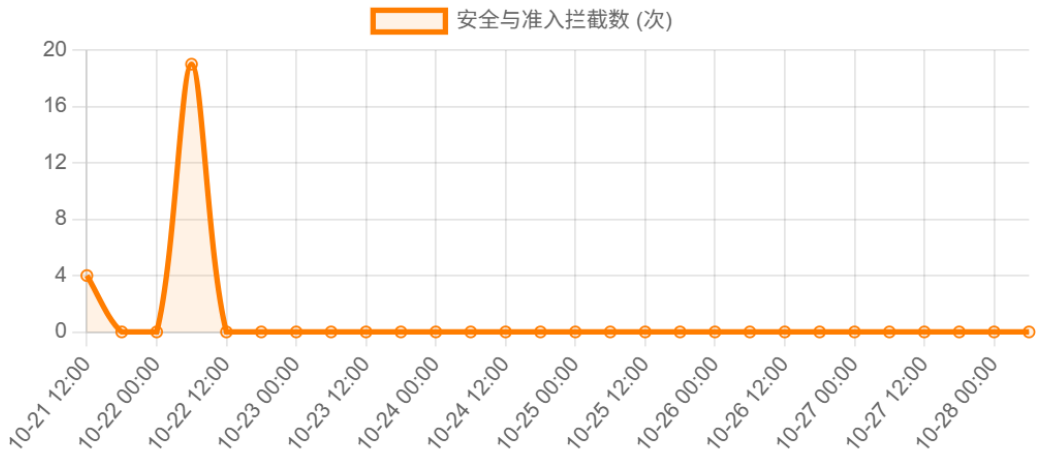
前10用户 Token 消耗占比

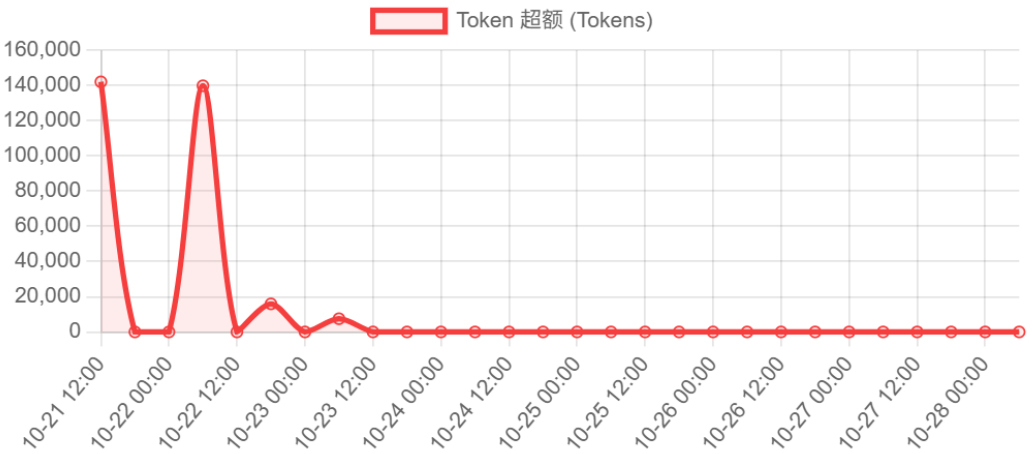


系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

近7天风控趋势

总计安全与准入拦截数: **23** 次 总计超额 Token: **305,160**





近7天实时报警

一般报警 重要报警 紧急报警

报警类型	详细信息概要	触发时间	状态
远程日志	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default [Connect)，当前未启用磁盘缓存保护，审计日志已丢弃	2026-10-28 09:46:31	未确认
远程日志	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default [Connect)，当前未启用磁盘缓存保护，审计日志已丢弃	2026-10-28 09:44:30	未确认
远程日志	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default [Connect)，当前未启用磁盘缓存保护，审计日志已丢弃	2026-10-28 09:43:15	未确认

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

大模型网关运营看板 系统状态

系统硬件状态

实时硬件与底层网络监控

运行时间： 0天 20时 22分

CPU 温度： 32.0 ℃

0.0%
CPU 使用率

22.4%
内存使用率
已用427M / 总1913M

0%
存储使用率
已用8M / 总11382M

网口链路状态：


WAN
1000M


LAN1


LAN2


LAN3

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统
账号与安全

账号与安全 管理员操作日志

账号安全与防护（当前账号：admin）

修改当前登录账号 (admin) 密码

新密码

如需修改密码请在此输入 (至少6位)

确认新密码

再次输入新密码

防爆破安全控制

允许连续输错次数

0

设置为 0 表示不生效（禁用锁定策略）；设置大于 0 时达到次数将锁定对应来源 IP

触发锁定后限制时长 (分钟)

15

锁定时间内禁止该来源 IP 登录

HTTPS 加密传输

☐ **启用 HTTPS 加密访问**

启用后系统将通过加密通道（HTTPS）提供 Web 管理服务；取消勾选将恢复标准 HTTP 访问

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统
账号与安全

默认 HTTPS 端口为 443

保存并应用配置

管理员账号管理（仅 admin 主账号可见）

➕ 添加新管理员账号：

管理员账号： 如：operator01　初始登录密码： 至少 6 位字符

授权可访问的功能模块：

☑ 系统概览 ☑ AI网关 ☑ AI服务商管理 ☑ 内网用户管理 ☑ AI安全 ☑ token策略 ☑ 统计报表 ☑ 网络配置 ☑ 定向加速 ☑ 智能组网
☑ 网络诊断 ☑ 设备信息 ☑ 重启系统

添加管理员

管理员账号	重置密码 (选填)	授权访问的功能模块	操作
user1 创建于: 2026-10-22 11:32:16	不修改请留空 (至少6位)	☑ 系统概览 ☑ AI网关 ☑ AI服务商管理 ☑ 内网用户管理 ☑ AI安全 ☑ token策略 ☑ 统计报表 ☑ 网络配置 ☑ 定向加速 ☑ 智能组网 ☑ 网络诊断 ☑ 设备信息 ☑ 重启系统	删除账号

保存并应用配置

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统
账号与安全

账号与安全 管理员操作日志

管理员操作日志（日志保存时间：1 年 (365天)）

导出操作日志 (CSV)

管理员： 全部管理员 开始时间： 2026-09-28 10:19 结束时间： 2026-10-28 10:19 每页显示： 20条 筛选查询

操作时间	管理员	页面模块	修改的功能详情	来源 IP
2026-10-28 10:10:29	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 134
2026-10-28 10:09:31	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 131
2026-10-28 10:05:54	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 131
2026-10-28 10:03:02	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 131
2026-10-28 09:57:03	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 131
2026-10-28 09:56:05	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 131
2026-10-28 09:55:11	admin	系统登录	管理员成功登录网关管理后台	192. 168. 10. 131
2026-10-28 09:53:43	admin	API网关	保存远程安全日志审计与磁盘缓存配置 (开关:关闭)	192. 168. 10. 131
2026-10-28 09:53:37	admin	API网关	保存远程安全日志审计与磁盘缓存配置 (开关:关闭)	192. 168. 10. 131

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

AI 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：

http://0.0.0.0:3000 保存配置 停止服务

提示：网关支持全自动识别OpenAI / Claude / Gemini 请求协议

数据保存间隔配置： 30 分钟 (控制数据的聚合采样及报表输出颗粒度)

数据最长保存时间： 1 年 (365天 / 默认) 保存配置

远程安全日志审计

远程日志审计开关

关闭推送

审计粒度模式

完全记录日常所有对话提问与回答全文

远端 HTTP Webhook 接收地址

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

远程安全日志审计

远程日志审计开关

关闭推送

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

如 http://192.168.1.100:8080/api/logs

Webhook 鉴权 Token (可选)

远端校验请求头 X-Audit-Token

磁盘缓存（网络异常或远端不可达时落盘保护）

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)

如 /overlay/aiapi/audit_cache

磁盘缓存最大空间限制 (MB)

50

保存审计与缓存配置

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

故障切换与负载策略

负载模式: 负载均衡（在支持目标模型的多个服务商间随机均摊流量） **故障处置与探活周期:** 连接异常临时禁用 1800 秒

首字超时(TTFT)与重试: 15 秒 / 重试 1次 **QoS劣化检测:** 启用 (P95 TTFT ≥ 基线2.5倍判异常)

保存策略配置

说明：1. **故障处置与探活周期**：选择「连接异常临时禁用」时，异常节点在设定秒数内暂停调度，期满后由后台自动探活恢复；选择「异常不禁用」时，节点异常依然标记异常状态、触发报警并在设定秒数后由后台自动探活恢复，但不隔离后续请求的正常路由；

2. **负载模式**控制内网客户端勾选了多个服务商时的选路方式（若处于**主备模式**，可通过服务商列表的 ▲ / ▼ 按钮调整主备优先级顺序；模型池内部可在「调度编排」中独立配置为负载均衡模式或按权重排序的主备模式）。

第三方 AI 服务提供商（上游）

一键清零统计

服务商名称 例如：硅基流动 | 接口协议规范 OpenAI 兼容协议 | 第三方 API Base URL 如：https://api.siliconflow.cn | 第三方真实 API Key (可选) 无鉴权可留空 | 添加上游

优先级排序	服务商 (协议)	API URL	真实 API Key	用量与词元(Tokens)	操作
1 ▲ ▼	**天鸿智算** OPENAI	https://api.thhtcloud.com	sk-**********KVwu 查看	总请求: 3 次 总 Tokens: 0 输入: 0 输出: 0	限定模型 重置 删除
	硅基流动		sk-**********mxif 查	总请求: 0 次	

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

故障切换与负载策略

负载模式： 主备模式（按服务商列表优先级顺序首选，故障顺延切换） **故障处置与探活周期：** 异常不禁用（仅标记报警并保持探活） 1800 秒

保存策略配置

说明：1. **故障处置与探活周期**：选择「…不禁用」时，节点异常依然标记异常状态、触发报警并在设定秒数后由后台自动探活恢复，但不隔离后续请求的正常路由；
2. **负载模式**控制内网客户端勾选了…序；模型池内部始终按池编排权重负载均衡）。

第三方 AI 服务提供商（上游）

一键清零统计

服务商名称 例如：硅基流动 接口协… Op… Key (可选) 添加上游

优先级排序	服务商 (协议)			用量与词元(Tokens)	操作
1 ▲ ▼	天鸿智算 连接异常(未禁用) OPENAI	https://api.thhtcloud.c…	sk-**********KVwu 查看	总请求: 0 次 总 Tokens: 0 输入: 0 输出: 0	限定模型 重置 删除
2 ▲ ▼	UCloud2 OPENAI	https://api.modelverse.…	cEl**********1d0e 查看	总请求: 3 次 总 Tokens: 35,228 输入: 33,683 输出: 1,545	限定模型 重置 删除
3 ▲ ▼	博伦智汇 OPENAI	https://max.booleanint.…	sk-**********4LVl 查看	总请求: 0 次 总 Tokens: 0 输入: 0 输出: 0	限定模型 重置 删除

限定模型设定 - 天鸿智算

保存限定模型 ×

点击模型切换限定状态（**蓝色为启用**，灰色为停用，全灰为不限定模型）：

- deepseek-v4-flash
- deepseek-v4-flash-0731
- deepseek-v4-pro
- deepseek-v4-pro-0813
- deepseek-v4.1-flash
- doubao-seedance-1-5-pro-251215
- doubao-seedance-2-0-260128
- doubao-seedance-2-5-260628
- doubao-seedream-4.5
- glm-5.3
- kimi-k3
- kling-v3-omni
- MiniMax-M3
- qwen3.7-plus

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

8 ▲ ▼ **UCloud3(与2同key)** OPENAI https://api.modelverse.cn cEl**********1d0e 查看 总请求: 37 次 总 Tokens: 271,912 输入: 237,407 输出: 34,505 限定模型 重置 删除

模型池管理 模型名称映射

模型池名称 例如：旗舰GPT池 模型名 (可输入或下拉选择) 手动输入或选择映射模型名 池调度模式 负载均衡模式 新建模型池

模型池名称	匹配模型名	上游路由状态	操作
111	glm-5.2	主备模式 已挂载 0 个服务商 (0 正常)	调度编排 删除池
test	deepseek-v4.1-flash	负载均衡 已挂载 1 个服务商 (1 正常) 今日请求: 0 次 | 今日Token: 0	调度编排 删除池
GLM模型池	glm-5.3	负载均衡 已挂载 0 个服务商 (0 正常)	调度编排 删除池
测试组模型池	deepseek-v4.1-flash	负载均衡 已挂载 2 个服务商 (2 正常) 今日请求: 0 次 | 今日Token: 0	调度编排 删除池
混合模型	glm-deepseek	负载均衡 已挂载 2 个服务商 (2 正常) 今日请求: 0 次 | 今日Token: 0	调度编排 删除池

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

8 ▲ ▼ UCloud3(与2同key) OPENAI https://api.modelverse.cn cEl**********1d0e 查看 总请求: 37 次 总 Tokens: 271,912 输入: 237,407 输出: 34,505 限定模型 重置 删除

模型池管理 模型名称映射

说明：配置映射后，网关将自动把客户端请求的模型名透明重定向为上游真实模型名（支持映射单个或多个目标模型，且自动同步至「模型池管理」的模型名下拉列表）。 刷新全量上游模型库

添加 / 修改模型映射规则（支持关键字跨服务商检索勾选 + 手动输入）：

检索上游真实模型： 输入模型关键字 (如 deepseek / qwen / glm / clau 遍历服务商匹配相似模型

客户端请求模型名 (Alias)： 手动输入新名称，或下拉拉取已有映射修改

→

映射到的上游真实模型名 (可勾选上方候选或手动输入，多个用逗号分隔)： 如：deepseek-ai/DeepSeek-V3, qwen-max

添加/修改到映射列表

客户端请求模型名 (源模型)	映射	上游真实模型名 (目标模型，支持多个用逗号分隔)	操作
glm-deepseek	→	qwen3.8-max, Qwen/Qwen3.8-27B	修改 移除
666	→	deepseek-v4.1-flash-1017, deepseek-v4-pro-0813, deepseek-v4.1-flash, deepseek-v4-flash-vis	修改 移除

保存模型名称映射配置

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

模型池调度编排：111 (目标匹配: glm-5.2)

智能扫描匹配服务商

当前模型池负载模式： 主备模式（按权重高低排列主备优先级，同权重按顺序优先）

主备模式规则：通过权重排列主备访问顺序，**权重越高越优先**；若权重相同，则按服务商列表顺序（#序号越小越优先）排列。

当前实时主备访问顺序：暂未勾选池内服务商

加入池	上游服务商 (列表顺序)	主备权重 (1-10, 越大越优先)	每日Token限额(0为不限)	今日连接数	今日已用 Token
☐	#1 天鸿智算 正常 模型: glm-5.2 | P95: 待采样 OPENAI	1	0	0 次	0
☐	#2 硅基流动 正常 模型: glm-5.2 | P95: 待采样 OPENAI	1	0	0 次	0
☐	#3 博伦智汇 正常 模型: glm-5.2 | P95: 待采样 OPENAI	1	0	0 次	0
☐	#4 阿里云百炼 正常 模型: glm-5.2 | P95: 待采样 OPENAI	1	0	0 次	0
☐	#5 UCloud2 正常 模型: glm-5.2 | P95: 待采样	1	0	0 次	0

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

内网用户 API Key

一键清零展示累计数据

接口协议规范: 自动匹配 | 接入模式: 自动识别 | 绑定目标 (可勾选多个): ☆ 111 | 所属工作组: 默认 | 自定义/生成的 Key: 1888c98502d6a6051ab653b3e3f9c596 | 组成员: 例如：张三 | 有效期至(留空长期): 年 /月/日

添加分发 Key

提示：1. **接口协议规范**：选「自动匹配」时网关根据请求特征自动识别协议；若选择具体协议（OpenAI / Claude / Gemini），则严格按所选协议匹配解析；
2. **接入模式**：选「自动识别」时根据 User-Agent 自动判断；若选择「客户端」或「程序调用」，则严格按所选模式返回 HTTP 200 对话气泡或标准 HTTP 错误状态码；
3. **有效期 / 状态**：留空（或点击「设为长期」）表示长期有效；设定日期后，到期次日自动标红并禁用该 Key 访问，修改后请点击下方「保存用户配置」。

工作组：默认 (9 个账号)　组累计请求：393　组累计 Tokens：1,772,806 (输入: 1,652,362 输出: 120,444)

内网分发 Key (客户端)	接口协议	接入模式	映射目标 (可多选池/商)	所属工作组	组成员	有效期 / 状态	累计展示数据	操作
b262bb**********3a5 查看	OpenAI 兼容	自动识别	天鸿智算	默认	小茜2	年 /月/日 长期有效 设为长期	累计请求: 2 次 累计 Token: 0 输入: 0 输出: 0	重 删
dfe31d**********9b1 查看	OpenAI 兼容	自动识别	☆ 111	默认	小茜	年 /月/日 长期有效 设为长期	累计请求: 3 次 累计 Token: 0 输入: 0 输出: 0	重 删
59865a**********410 查看	OpenAI 兼容	客户端 (气泡接	☆ 666	默认	老王	年 /月/日 长期有效 设为长期	累计请求: 263 次 累计 Token: 1,253,032 输入: 1,161,478 输出: 91,554	重 删

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

内网分发 Key (客户端)	接口协议	接入模式	映射目标 (可多选池/商)	所属工作组	组成员	有效期 / 状态	累计展示数据	操作
0dadef**********65f 查看	OpenAI 兼容	自动识别	UCloud3(与2同key)	默认	cj	年 /月/日 长期有效 设为长期	累计 Token: 49 输入: 17 输出: 32	重 删

工作组：张三组 (1 个账号) 组累计请求：135 组累计 Tokens：942,361 (输入: 904,379 输出: 37,982)

内网分发 Key (客户端)	接口协议	接入模式	映射目标 (可多选池/商)	所属工作组	组成员	有效期 / 状态	累计展示数据	操作
442318**********ca9 查看	OpenAI 兼容	自动识别	阿里云百炼	张三组	张三	年 /月/日 长期有效 设为长期	累计请求: 135 次 累计 Token: 942,361 输入: 904,379 输出: 37,982	重 删

工作组：调研组 (1 个账号) 组累计请求：2 组累计 Tokens：4,038 (输入: 752 输出: 3,286)

内网分发 Key (客户端)	接口协议	接入模式	映射目标 (可多选池/商)	所属工作组	组成员	有效期 / 状态	累计展示数据	操作
e17e73**********29f 查看	自动匹配	自动识别	UCloud2	调研组	老李头	2026/11/30 有效 设为长期	累计请求: 2 次 累计 Token: 4,038 输入: 752 输出: 3,286	重 删

工作组：测试组 (1 个账号) 组累计请求：31 组累计 Tokens：299,546 (输入: 253,492 输出: 46,054)

内网分发 Key (客户端)	接口协议	接入模式	映射目标 (可多选池/商)	所属工作组	组成员	有效期 / 状态	累计展示数据	操作
f7559e**********c62 查看	自动匹配	自动识别	☆ 混合模型	测试组	李世民	年 /月/日 长期有效 设为长期	累计请求: 31 次 累计 Token: 299,546 输入: 253,492 输出: 46,054	重 删

保存用户配置

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

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	9 人	默认组不可删
张三组	1 人	删除组
调研组	1 人	删除组
测试组	1 人	删除组

保存工作组配置

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: 10,261 上限: 1000000 个	已用: 12,021,078 上限: 0 个	最高已用: 5,253 上限: 1000000 个	最高已用: 10,340,392 上限: 0 个

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

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 个	已用: 942,361 上限: 0 个	移除策略
黄玮	默认	已用: 0 上限: 0 个	已用: 0 上限: 0 个	移除策略
老王	默认	已用: 0 上限: 30000 个	已用: 1,160,961 上限: 0 个	移除策略
李世民	测试组	已用: 0 上限: 20000 个	已用: 299,546 上限: 30000 个	移除策略

说明：配置了组成员专属配额策略后，将忽略所属工作组中的「组成员单日/单月配额」（填 0 表示不限制个人上限），但仍受工作组「全组每日/每月总配额」限制。修改后请点击下方保存按钮。

保存组成员配额策略

系统概览

AI网关

- AI服务商管理

- 内网用户管理

- Token策略

- AI安全

- 统计报表

网络配置

定向加速

智能组网

网络诊断

设备信息

重启系统

	默认	上限: 30000 个	上限: 0 个	移除策略
李世民	测试组	已用: 0 上限: 20000 个	已用: 299,546 上限: 30000 个	移除策略

说明：配置了组成员专属配额策略后，将忽略所属工作组中的「组成员单日/单月配额」（填 0 表示不限制个人上限），但仍受工作组「全组每日/每月总配额」限制。修改后请点击下方保存按钮。

保存组成员配额策略

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

系统概览

AI网关

- AI服务商管理

- 内网用户管理

- Token策略

- AI安全

- 统计报表

网络配置

定向加速

智能组网

网络诊断

设备信息

重启系统

双向安全围栏与 DLP 动态脱敏防护

输入与输出检测独立管控

输入方向: ☑ 启用提问阻断 ☑ 启用提问动态脱敏

输出方向: ☐ 启用回答熔断 ☐ 启用回答动态脱敏

内置隐私数据识别规则 (DLP)

启用此规则

☑ 二代身份证号 ☑ 手机号码 ☑ 银行卡号 ☑ 企业统一信用代码 ☑ 电子邮箱

☑ API 密钥凭证 (sk-/AKIA/AIza/ghp_) ☑ JWT 认证令牌 ☑ 数据库连接串 ☑ 私钥证书块 (PEM/RSA)

☑ 内网 IPv4 地址 ☑ 互联网 IPv4 地址 ☑ MAC 物理地址

处置模式: 动态脱敏模式（自动将隐私信息替换为 *** 掩码后放行

严格阻断模式 - 敏感词与正则列表

启用此表

命中此表立即拦截提问或熔断回答流（一行一条，支持 regex:正则表达式 或 regex:/表达式/g）

```
违禁词
涉密红头文件
regex:DROP\s+TABLE
```

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

动态脱敏模式 - 敏感词与正则列表 启用此表

命中此表自动替换为 *** 掩码并继续正常请求大模型（一行一条，支持 regex:表达式、regex:/表达式/g 或 regex:表达式=>替换词）

```
内部代号A
客户真实姓名
regex:password\s*[:=]\s*\S+
```

严格阻断模式触发后的自定义拦截提示内容

⚠ 您的提问或 AI 的回答包含不合规内容，已被系统拦截。

保存安全围栏配置

提示词注入（Prompt Injection）与越狱攻击预警

越狱指令与提示词注入特征库（AC自动机极速匹配） 关闭越狱特征阻断

已预置常用高频越狱特征词（如 DAN、ignore previous instructions、系统提示词套取等），支持自行增删或添加 regex:表达式。命中后将触发拦截并记录「越狱攻击」报警日志。

```
do anything now
dan mode
```

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

⚠ 检测到疑似提示词注入或越狱攻击指令，请求已被网关安全引擎阻断。

保存越狱防护配置

组成员接入安全（IP 与设备 MAC 地址准入控制）

添加组成员安全设备绑定（支持通过最近连接记录自动获取 IP 与 MAC）：

选择或输入组成员： 如：张三 🔍 自动获取最近连接IP/MAC 绑定访问 IP： 自动获取或手动输入IP 绑定设备 MAC： 自动获取或手动输入MAC

+ 添加到绑定列表

接入安全验证开关： 启用设备 IP/MAC 准入校验 **未授权设备拦截提示：** ⚠ 您使用了未授权的设备进行访问，请联系管理员。

组成员	所属工作组	允许访问的 IP 列表 (逗号分隔)	允许访问的设备 MAC 列表 (逗号分隔)	操作
李四	默认	192.168.10.131	66:0c:78:05:30:2d	移除绑定

说明：当不知道组成员设备的实际 IP 或 MAC 时，只需选择或输入组成员名称并点击「🔍 自动获取最近连接IP/MAC」，系统将自动读取该成员最近的连接日志并反查 ARP/DHCP 表为您一键填好。

保存组成员接入安全配置

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

说明：当不知道组成员设备的实际 IP 或 MAC 时，只需选择或输入组成员名称并点击「自动获取最近连接IP/MAC」，系统将自动读取该成员最近的连接日志并反查 ARP/DHCP 表为您一键填好。

保存组成员接入安全配置

瞬时速率限制与频控（RPM / TPM / 并发数）

优先级规则：系统严格按照 **指定组成员优先级最高 > 所属工作组其次 > 全体默认最低（默认 RPM: 60 / TPM: 20000 / 并发: 5路）** 独立计算每分钟最大请求数（RPM）、每分钟最大词元数（TPM）与同时在线最大并发请求数（并发限制）。填 0 表示该层级不限制。

瞬时频控总开关： 启用瞬时速率与并发限制

全体成员默认 RPM： 60 次/分钟 **全体成员默认 TPM：** 20000 Tokens/分钟

全体成员默认并发： 5 路同时请求

按工作组设置组内成员每分钟频控与并发限制（优先级次之）：

工作组名称	组内单人最大请求 (RPM)	组内单人最大词元 (TPM)	组内单人最大并发 (路)
默认	0 次/分钟	0 Tokens/分钟	0 路并发
张三组	0 次/分钟	0 Tokens/分钟	0 路并发

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

传输层与通信防重放攻击（HMAC 动态签名与 Nonce 防重放）

防重放与签名校验总开关： 关闭传输层防重放控制 **时间戳与 Nonce 防重放窗口 (秒)：** 300

当前勾选赋予模式： 勾选成员 = 高安全 HMAC 签名模式 (; 全选所有成员 取消全选

说明：可按工作组整组勾选或手动单选组成员；**标准模式**仅校验静态 Bearer Token（兼容 Cursor / Chatbox 等所有第三方软件）；**高安全 HMAC 模式**强制校验 X-Timestamp / X-Nonce / X-Signature 动态签名防抓包重放。

工作组：默认 (9 人) 本组全设为 HMAC 模式 | 本组全设为标准模式

小茜2 标准模式 | 小茜 标准模式 | 老王 标准模式 | 666 标准模式 | 6661 标准模式
李四 标准模式 | 777 标准模式 | 金豆豆 标准模式 | cj 标准模式

工作组：张三组 (1 人) 本组全设为 HMAC 模式 | 本组全设为标准模式

张三 标准模式

工作组：调研组 (1 人) 本组全设为 HMAC 模式 | 本组全设为标准模式

老李头 标准模式

工作组：测试组 (1 人) 本组全设为 HMAC 模式 | 本组全设为标准模式

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

工作组：调研组 (1 人)　本组全设为 HMAC 模式 | 本组全设为标准模式

老李头 标准模式

工作组：测试组 (1 人)　本组全设为 HMAC 模式 | 本组全设为标准模式

李世民 标准模式

保存通信防重放配置

请求体大小与 JSON 嵌套深度硬限制（Payload Guard）

Payload Guard 开关	设定模式	单次请求体最大限制 (MB)	JSON 最大嵌套深度 (层)
关闭硬限制防护	自定义	10	32

说明：超限将分别触发报警类型 **PG-01**（大小超限）与 **PG-02**（深度超限）；当 JSON 最大嵌套深度设为 **0 层** 时将**禁用 JSON 解析**（适用于纯二进制流或大体积音视频上传场景）。

保存 Payload Guard 配置

系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

192.168.10.181

图表统计 报警日志

统计周期： 今天 昨天 本月 近7天 近30天 自定义： 2026/10/28 00:00 至 2026/10/28 23:59 图表生成 下载报表 (CSV)

点击下方任意「业务关注场景卡片」即可一键切换对应分析图表： 当前视图：Token 消耗与成本趋势

Token 消耗趋势 成本与用量
14,299 Tokens
总请求 21 次 | 输入 5,239 / 输出 9,...

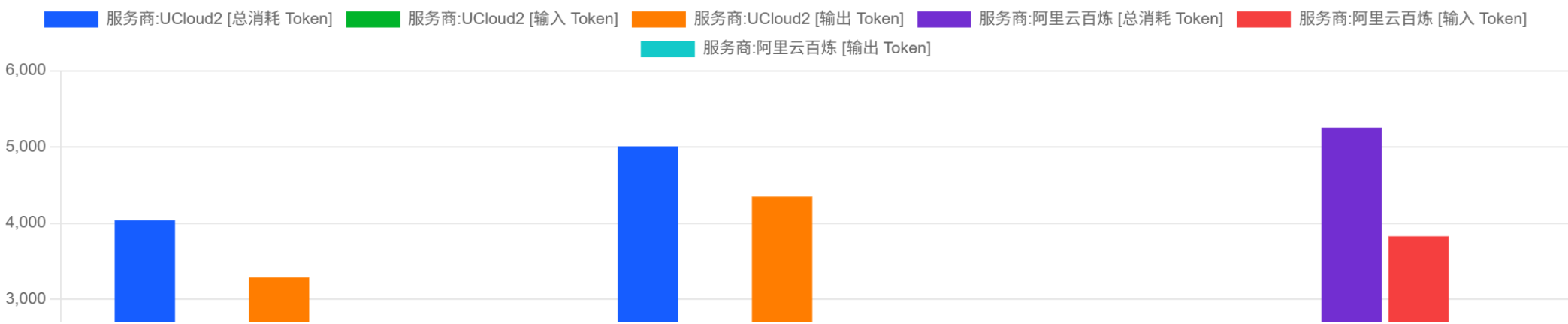
服务商质量 (QoS) TTFT与长尾
1549 ms 平均首字
峰值P95: 7490ms (UCloud2-glm-5...

服务商状态与恢复 可用性审计
0 次 异常事件
全局异常 0 | 模型异常 0 | 自动恢复 0

团队与成员排行 组织分摊
2 个 活跃工作组
Top消耗: (默认) 777

安全风控与超额 合规与配额
0 次 安全拦截
Token 配额超限告警: 0 次

聚合视角： 按服务商 按服务商+模型 按工作组+成员 图表样式： 柱状图 曲线图 环形占比 展开高级多维自定义筛选 (全部维度)



系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

Token 消耗趋势 成本与用量

14,299 Tokens

总请求 21 次 | 输入 5,239 / 输出 9,...

服务商质量 (QoS) TTFT与长尾

1549 ms 平均首字

峰值P95: 7490ms (UCloud2-glm-5...

服务商状态与恢复 可用性审计

0 次 异常事件

全局异常 0 | 模型异常 0 | 自动恢复 0

团队与成员排行 组织分摊

2 个 活跃工作组

Top消耗: (默认) 777

安全风控与超额 合规与配额

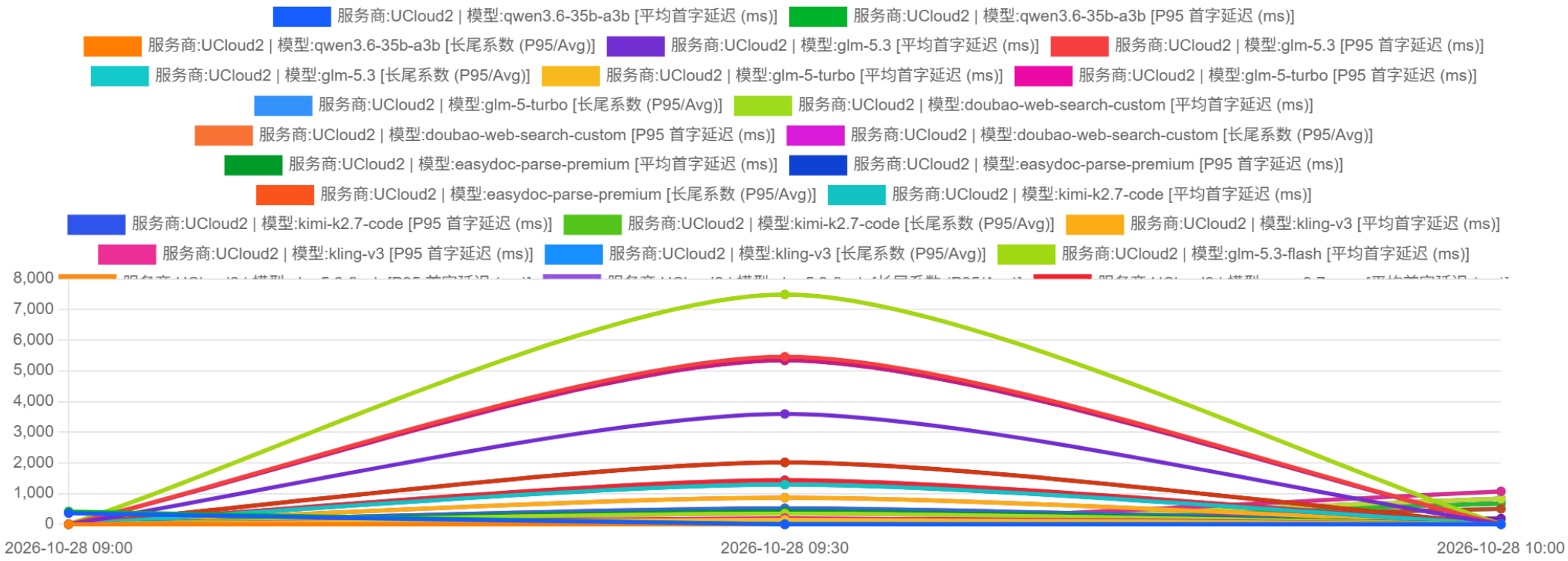
0 次 安全拦截

Token 配额超限告警: 0 次

聚合视角： 按服务商 | 按服务商+模型 | 按工作组+成员

图表样式： 柱状图 | 曲线图 | 环形占比

展开高级多维自定义筛选 (全部维度)



系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

图表统计 报警日志

统计周期： 今天 昨天 本月 近7天 近30天 自定义： 2026/10/21 10:56 至 2026/10/28 10:56 图表生成 下载报表 (CSV)

点击下方任意「业务关注场景卡片」即可一键切换对应分析图表： 当前视图：Token 消耗与成本趋势

Token 消耗趋势 成本与用量
983,233 Tokens
总请求 175 次 | 输入 655,994 / 输...

服务商质量 (QoS) TTFT与长尾
1177 ms 平均首字
峰值P95: 7490ms (UCloud2-glm-5...

服务商状态与恢复 可用性审计
271 次 异常事件
全局异常 44 | 模型异常 227 | 自动...

团队与成员排行 组织分摊
3 个 活跃工作组
Top消耗: (默认) cj

安全风控与超额 合规与配额
19 次 安全拦截
Token 配额超限告警: 20 次

聚合视角： 按服务商 按服务商+模型 按工作组+成员 图表样式： 柱状图 曲线图 环形占比 展开高级多维自定义筛选 (全部维度)



系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

统计周期： 今天 昨天 本月 近7天 近30天 自定义： 2026/09/28 10:57 至 2026/10/28 10:57 图表生成 下载报表 (CSV)

点击下方任意「业务关注场景卡片」即可一键切换对应分析图表：

当前视图：Token 消耗与成本趋势

Token 消耗趋势 成本与用量
13,359,094 Tokens
总请求 812 次 | 输入 12,803,815 / ...

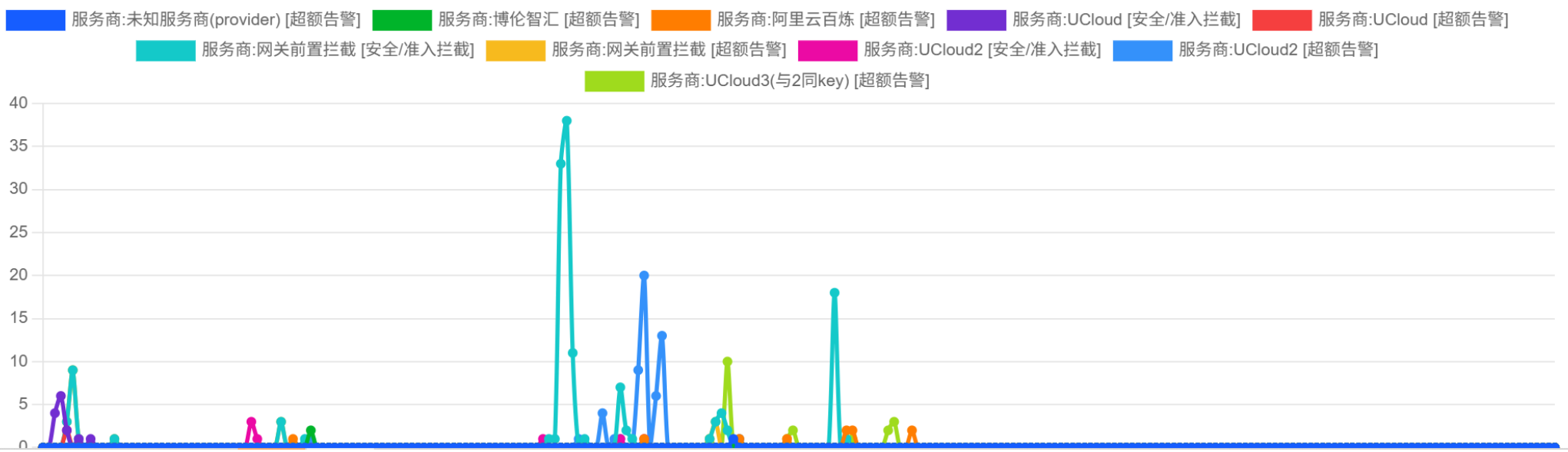
服务商质量 (QoS) TTFT与长尾
1182 ms 平均首字
峰值P95: 81999ms (UCloud2-默认...

服务商状态与恢复 可用性审计
390 次 异常事件
全局异常 163 | 模型异常 227 | 自动...

团队与成员排行 组织分摊
5 个 活跃工作组
Top消耗: (默认) cj

安全风控与超额 合规与配额
163 次 安全拦截
Token 配额超限告警: 113 次

聚合视角： 按服务商 按服务商+模型 按工作组+成员 图表样式： 柱状图 曲线图 环形占比 展开高级多维自定义筛选 (全部维度)



系统概览
AI网关
- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

图表统计 报警日志

报警日志

一般报警 重要报警 紧急报警

报警等级设置 确认选中项 导出日志 (CSV)

开始时间: 2026/01/01 08:00

结束时间: 2026/10/28 10:58

报警类型 (默认全显):

- 服务商
- 敏感词
- 越狱攻击
- Token配额

每页显示: 20条

筛选查询

	报警类型	报警等级	内容概要	产生时间	确认时间	确认人
	远程日志	重要报警	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default [Connect)，当前未启用磁盘缓存保护，审计日志已丢弃	2026-10-28 09:46:31	-	-
	远程日志	重要报警	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default [Connect)，当前未启用磁盘缓存保护，审计日志已丢弃	2026-10-28 09:44:30	-	-
	远程日志	重要报警	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default [Connect)，当前未启用磁盘缓存保护，审计日志已丢弃	2026-10-28 09:43:15	-	-
	远程日志	重要报警	远程日志服务器 [http://172.16.1.49:3000/webhuuk/test] 推送失败 (Cannot connect to host 172.16.1.49:3000 ssl:default	2026-10-28 09:42:01	-	-

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

LAN 局域网配置

IP 地址

192.168.137.1

子网掩码

255.255.255.0

DHCP 服务

开启

WAN 上网配置

IP/mask	192.168.10.181/24
网关	192.168.10.1
DNS	114.114.114.114 1.2.4.8

上网协议

静态 IP

IP 地址

192.168.10.181

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

定向加速管理

服务开关

当前运行：已停止

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

分流总节点

选择后会同时应用到TCP和UDP协议

加速模式

分流模式(默认)

分流模式只走指定规则流量，全局模式所有流量走代理节点

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 保存并应用

当前定向出口流量

设备信息

重启系统

选择后会同时应用到TCP和UDP协议

加速模式

分流模式(默认)

分流模式只走指定规则流量，全局模式所有流量走代理节点

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 保存并应用

当前定向出口流量

获取失败：连接接口失败 ('NoneType' object has no attribute 'get')

当前定向出口IP

IP：58.216.171.202

城市：常州

省份：江苏省

国家：中国

运营商：AS4134 CHINANET BACKBONE

查看更多

第三方平台数据仅供参考

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

服务状态

❌ 服务已停止

配置参数

设备编码

84e2a618582d55f128a7d2885107eaf0

可信设备ID

84e2a618582d55f128a7d2885107eaf0@box.netbooster

连接密码

••••••••

保存并应用

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

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

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

- AI服务商管理
- 内网用户管理
- Token策略
- AI安全
- 统计报表
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

在线检查新版本

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

AI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**

系统概览
AI网关
网络配置
定向加速
智能组网
网络诊断
设备信息
重启系统

重启系统

重启网关设备将导致当前的网络连接短暂中断（预计需要 1 分钟）。
请确认是否继续操作?

立即重启

可信大模型分控策略网关

边缘EDGE MLMtoken系列

简称：tokenBoss盒子

联系人：陈珂 博士 13775099862

常州万联云网信息技术有限公司 官网：www.macrounion.com