

云网融合 Edge GLMToken 系列

Trusted Api-Token 可信分控策略网关 产品介绍

简称：边缘 TokenBoss 盒子——「云网融合大模型 api-token 可信分控策略网关」

一句话简单概括：

市面上绝大多数大模型 API 聚合与 Token 分发平台工厂是纯软件，仅提供公有云侧的大模型 API 中转聚合服务，更多面向个人用户；没有解决云网融合环境下企业级用户侧 api 可信管控和多席位 AI 应用客户端的 Token 管理需求；

万联云网推出的边缘 TokenBoss 盒子是「边缘侧 api-token 管控策略网关盒子，内置大模型 API-token 分控 Boss 管理功能，客户端坐席的 api 调用/策略日志，以及 SD-WAN 定向访问加速功能」，一体化边缘侧交付，主打云网融合 API-token 分控+token 流量审计+Ai 专网接入，面向企业级 to B 边缘 Ai 安全可控场景。

一、“边缘 TokenBoss 盒子”（云网融合可信大模型 api-token 分控策略网关）的产品描述：

1. 部署形态：分两级，运营级大模型 api 聚合中间服务平台（软件系统），边缘 TokenBoss 盒子（硬件网关）

- 运营级大模型 api 聚合中间服务平台——大模型 api to token 聚合分发与调度管理系统（开源二开）
- 边缘 TokenBoss 盒子，本地部署在客户机房/集团数据中心/园区边缘

算力 token 节点/办事处或协同研发的第三方内网

BOSS 功能内嵌在 api-token 管控网关内，不是独立系统。网关含各家大模型 API 管理、管控 token 用度与审计、token 用量坐席统计能力等基础服务。

2. 定位：云网融合 API-token 2 User 接入分控功能

- 不只是 API 聚合转发，更是云网融合接入网关 + AI API-token 管控平台二合一。管控对象不仅仅是对各家的 API Key，还包含终端坐席（PC/手机 AI 助手）的 Ai 工作台的可信 api 接入身份、链路权限、token 流量管理、密钥、api 调用审计、token 用量清单。
- 管理员账号（运维运营）和政企大模型账户、使用者坐席账号（自定义）两层隔离。

3. 网络底座：跨境访问优化服务+SD-WAN Overlay 加速功能

- 原生支持 4G/5G、WiFi6、千兆 LAN，内置 SD-WAN，Overlay 隧道加速。模型调用流量可以走专网/运营商专线，而不再是裸公网。（网关定制）

4. 管控边界：token 分发层 + 大模型 API 层双层可信管控

- 网络侧：终端坐席 Ai 大模型 API 身份 key 管理（二次内控）、链路 QoS、隧道加密、接入权限
- AI 调用侧：模型 API 密钥统一托管、路由分发、token 计量、全链路 api 审计、token 分帐户、API 应用调用明细

审计是端到端：哪个终端、走哪条专网链路、调用哪个模型、消耗多少 token，日志全部在边缘盒子留存，可本地留存审计数据，满足数据不出域要求。

5. 密钥与数据可信

- Key 管理在边缘网关侧，可自主配置密钥不上公有云；模型请求流量可在专网隧道内传输，敏感业务流量不必暴露在裸公网。这是与纯公有云 token 工厂最大的分水岭。

6. 目标客户场景：这类私有化 token 工厂，集团 Ai 服务商、算力大模型运营商/Ai 专网政企客户

面向大中型集团 AI 新基建服务，广电、电信、移动的 ToB 边缘政企客户。客户需求不是简单“凑一堆大模型 API”，而是需要一套敏捷的“本地私有化 AI 大模型应用+云网融合 api-token 分级管控+可审计可溯源可计量的 Boss 工具” AI 服务系统。

二、边缘 TokenBoss 盒子 api 网关核心面板介绍

云网融合接入网关

系统概览

网络配置

定向加速

智能组网

API网关

网络诊断

设备信息

API 网关服务状态

当前网关运行状态: 运行中

内网发布 URL / 监听地址

http://0.0.0.0:23333

保存地址

停止服务

1. 配置第三方 AI 服务提供商 (上游)

服务商名称

例如: CherryIn AI

第三方 API Base URL

例如: https://express-ent-admin.cherryin.net/v1

第三方真实 API Key

输入上游给的真实 API Key

添加 AI 服务提供商

服务商	API URL	真实 API Key	用量与词元(Tokens)	操作
aitest	https://express-ent-admin.cherryin.net/v1	oZK*****rsZL 显示	总请求: 4 次 总 Tokens: 5,769	重置 删除

2. 分发内网用户 API Key (下游)

绑定目标 AI 服务商:

自定义/生成的 Key:

使用人/部门备注:

aitest

c49b4b373db514680ee5aad12a14cf93

例如: 张三 / 研发部

添加分发 Key

内网分发 Key (客户端使用)	映射服务商	使用者备注	用量与词元(Tokens)	操作
606290*****b130 显示	aitest	老王 保存	请求次数: 4 次 消耗 Token: 5,769	重置 删除
df8af9*****fdc6 显示	aitest	杨洋 保存	请求次数: 0 次 消耗 Token: 0	重置 删除

1、页面两大核心模块（上游接入 + 下游分发）

1) 配置第三方 AI 服务提供商【上游：聚合多家大模型】

- 录入上游大模型厂商信息：服务商名称、上游 BaseURL、上游真实 API Key，添加后系统托管上游原始密钥。

- 列表展示：记录上游地址、脱敏原始 Key、总请求次数、总 Token 消耗。

作用：把多家上游大模型 API 统一接入到本地边缘网关，盒子替客户统一管理各家大模型原生密钥。

✅ 关键安全点：原始上游 Key 只保存在本地盒子内，不对外下发给终端用户。

2) 分发内网用户 API Key 【下游：给内部终端/人员分发代理密钥】

- 选择绑定的上游服务商，由网关自动生成代理 Key（下发给内网用户）
- 填写使用人/部门备注，实现按人/按部门分配
- 列表展示：下发给客户终端的内网分发 Key、映射绑定的上游模型、使用人、独立用量统计（请求次数+消耗 Token），支持单独查看/删除。

核心逻辑：密钥代理映射

用户（老王、杨工）拿到的是盒子生成的代理 Key，拿不到上游真实大模型密钥；终端请求全部经过这个本地网关中转，网关做 Token 计量、审计、权限控制。

- 顶部 API 网关状态

网关本地监听内网地址 `http://0.0.0.0:2333`，在内网提供统一接入端点；终端 PC/手机 AI 助手，只访问这个本地网关地址，由网关代转请求到上游大模型。

- 自主可控 API-TOKEN

- 审计日志、密钥、Token 用量可以本地留存，不需要全部上传到外部云端。

- 密钥代理隔离：上游原始密钥锁在本地盒子，下发给员工/终端的是代理映射 Key。终端永远看不到真实上游 API Key，解决集团/政企场景密钥外泄风险。
- 内网私有化接入：终端坐席的 AI 助手只访问内网网关地址，不需要终端直接访问公网大模型接口；内网统一出口，便于安全管控。

其它产品功能和使用说明详见《产品手册》

联系人： 万联云网 400 0277590 陈珂 博士：13775099862

万联云网官网：www.macrounion.com