开源社 kaiyuanshe | Datawhale | HOAIF 杭州市开源人工智能基金会

开源下一代 Open Source: The Next Generation | COSCon'26 第十一届中国开源年会

开放 Agent 生态圆桌 · COSCon'26

Agent 的上限
不是智能，是信任



演讲嘉宾：白景元

锟腾科技 CEO · New API 联合创始人

New API | qtnous 锟腾

Democratizing Intelligence　让智能普惠每一个人

生态全景：我们的位置，谁在用

Agent 与应用

Godwit · Claude Web · NexTerm · Save Any Bot · MyGo Music · Mayoi · Dalibre · Velune · Claude Code · Codex · Cursor · Gemini CLI · OpenClaw · OpenCode · Cline · OpenHands · Cherry Studio · DeepChat · Lody · Raft · Memoh · Nowledge Mem · Dify · Open WebUI

本地网关

AstrLink 我们的项目 · magpie · CC Switch · Claude Code Router

网关与中转

New API 我们的项目 · AxonHub · sub2api · CLIProxyAPI · Metapi Go · LiteLLM · OpenRouter · Higress · Kong · Cloudflare

模型厂商

OpenAI · Anthropic · Google · xAI · DeepSeek · 通义千问 · Kimi · 智谱

我们开源的底座

高亮的项目都在用这些组件

- **RelayKit** — 模型协议互转
- **moego** — 网关专用 JSON 库
- **moejs** — JS 插件运行时
- **AstrLink Guard** — 识别敏感信息的模型

来源：各项目的 go.mod、源码与 README（2026-11）。带徽标的项目直接采用了我们的组件；其余是各层的代表项目，未穷举

Agent = 模型 + Harness

Harness 是包在模型外面的那一层，让模型能在你的机器上动手。

HARNESS · 在你的机器上

读文件　跑命令　调工具　改代码

发出去 · 上下文

读到的内容，都会随请求发给模型

收回来 · 指令

模型返回的 tool call，会在本机执行

模型 · 在厂商的服务器上

发出去的，谁看得到？　收回来的，可不可信？

想用好模型，又怕资料外泄。

用第三方模型

- ✓ 能力强，按量付费
- ✗ **代码、文档、密钥随上下文发出去**

本地部署模型

- ✓ 资料留在自己的机器上
- ✗ **显卡和运维的成本高**

第三条路

模型照常用云上的，敏感内容先在本机换掉。

AstrLink Guard

在本机运行的隐私脱敏模型

发给模型之前，先在本机脱敏。



Agent 发出的

```
OPENAI_API_KEY=sk-proj-8Hq2…Zx4T
email: user@example.com
```

中间层和模型读到的（示意）

```
OPENAI_API_KEY=<API_KEY_1>
email: <EMAIL_1>
```

响应回来，占位符在本机换回原文

识别出的敏感内容，**中间层和模型读到的都是占位符。**

来源：AstrLink README（2026-11）。占位符的写法是示意

AstrLink Guard

在本机识别密钥、邮箱、手机号等 10 类敏感信息的模型

97.46	93.54	90.58
通用文本	代码与日志	编程工具请求

F1 · 自测基准

在本机运行

36 MiB

INT8 量化后的大小，
CPU 离线运行，不依赖显卡

开源，可集成

Apache-2.0

权重开源，
任何产品都可以集成

识别准确

2.4 %

凭证漏检率，
自测集共 2,158 个凭证

已经接入 AstrLink · magpie · Godwit · Claude Web · NexTerm

产品导向：**把隐私保护做成谁都用得起的基础组件。**

数据：AstrLink Guard 模型卡 0.1.3（2026-11），评测为项目自测基准；接入情况见各项目源码

只守住这一层，远远不够。



9 / 428
个中转在返回的 tool call 里
注入了恶意代码

401 / 440
个会话开着自动批准，
收到什么就执行什么

99 份凭证
在这些会话里泄露

来源：Liu et al., Your Agent Is Mine, ACM CCS 2026, arXiv:2604.08407，§5.2、§5.3

AI 流量安全五层模型

		回答的问题	手段	谁在做
L5	**指令可信**	指令是谁发的，被改过吗?	签名 · 声明 · 验证	尚无标准
L4	**执行安全**	Agent 可以做到哪一步?	沙箱 · 权限 · 人工确认	
L3	**数据安全**	哪些内容可以发出去?	本机识别 · 脱敏	AstrLink Guard
L2	**访问安全**	谁能调用，能用多少?	鉴权 · 配额 · 审计	New API
L1	**传输安全**	链路上有没有人偷听?	TLS 加密	已经普及

下面四层，一方就能做到；**最上面一层，要整条链路一起做。**

分层为本团队的归纳（2026-11）

这一层，一家做不成。

01 · 模型厂商

源头签名

「这句话是谁说的」

厂商对 tool call 签名，
Agent 用公钥验证出处

邮件的先例 DKIM · RFC 6376

Agent 调用 尚无标准

02 · 网关与中转

逐跳声明

「谁经手过，改了什么」

每一跳声明自己改动的字段，
没声明的改动通不过验证

邮件的先例 ARC · RFC 8617

Agent 调用 尚无标准

03 · AGENT

端侧验证

「敢不敢自动执行」

验证通过才自动执行，
验不过就交回给人确认

邮件的先例 DMARC · RFC 9989

Agent 调用 尚无标准

跨组织的信任，要靠开放标准；**这份规范，想和大家一起定。**

参考：arXiv:2604.08407 附录 C · RFC 6376 DKIM · RFC 8617 ARC · RFC 9989 DMARC

守住安全，就守住了信任。

每守住一层，敢交给 Agent 的事就多一些；
信任的边界越接近模型的能力，
Agent 的价值才发挥得出来。

Agent 的上限不是智能，是信任。



示意图

取之于社区，贡献于社区。



Agent 时代的信任，**应该由开源社区一起定义。**

来源：GitHub 贡献者统计、new-api-plugins 的 index.json、各项目的 go.mod 与源码（2026-11）

开源社 kaiyuanshe | Datawhale | HOAIF 杭州市开源人工智能基金会

开源下一代 Open Source: The Next Generation | COSCon'26 第十一届中国开源年会

THANKS

白景元

锟腾科技 CEO · New API 联合创始人

New API | qtnous 锟腾



GitHub 公众号 X 小红书