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

Open Agent Ecosystem Roundtable · COSCon'26

The Ceiling of Agents Is Trust, Not Intelligence



Speaker: Bai Jingyuan

CEO, QuantumNous · Co-founder, New API

New API | qtnous 锟腾

Democratizing Intelligence

Our Place in the Ecosystem

Agents & apps

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

Local gateways

AstrLink (OURS) · magpie · CC Switch · Claude Code Router

Gateways & relays

New API (OURS) · AxonHub · sub2api · CLIProxyAPI · Metapi Go · LiteLLM · OpenRouter · Higress · Kong · Cloudflare

Model providers

OpenAI · Anthropic · Google · xAI · DeepSeek · Qwen · Kimi · Zhipu

Our foundation

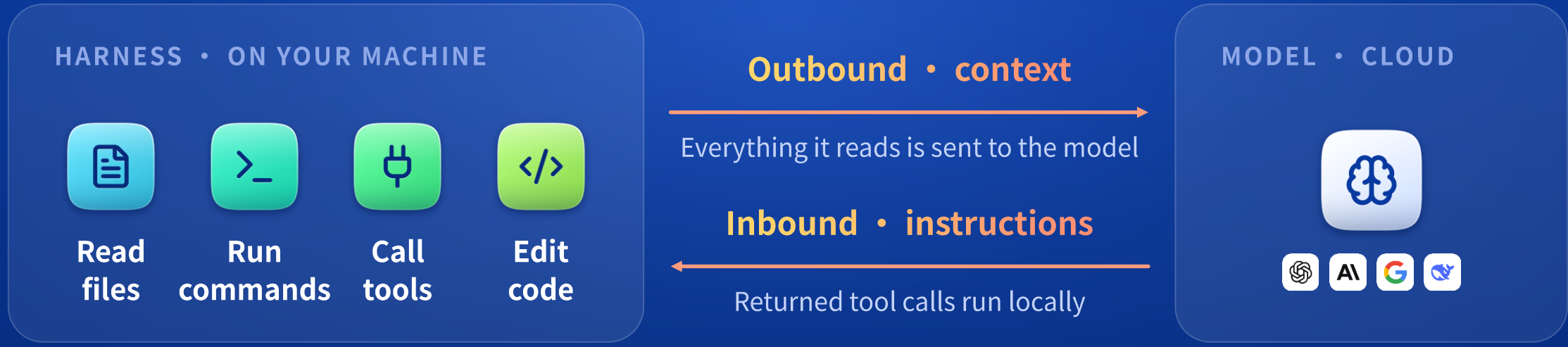
Used by badged projects

- **RelayKit** — Protocol bridge
- **moego** — Gateway JSON
- **moejs** — JS plugin runtime
- **AstrLink Guard** — Finds sensitive data

Sources: each project's go.mod, source and README (2026-11). Badged projects use our components directly; the rest are representative, not exhaustive

Agent = Model + Harness

The harness is the layer around the model that lets it act on your machine.



Who sees what goes out? Can what comes back be trusted?

Want great models, fear data leaks.

Cloud models

- ✓ Most capable, pay as you go
- ✗ **Code, docs and keys go out**

Local models

- ✓ Data stays on your machine
- ✗ **GPUs and operations are costly**

A THIRD WAY

Keep cloud models; redact secrets locally first.

AstrLink Guard

On-device redaction model

Redact locally, before it leaves.



WHAT THE AGENT SENDS

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

WHAT EVERY HOP READS

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

On the way back, placeholders are restored locally

Whatever Guard detects, **every hop sees as a placeholder.**

Source: AstrLink README (2026-11). The placeholder format is illustrative

AstrLink Guard

Detects keys, emails, phone numbers and 7 more types on device

97.46	93.54	90.58
General text	Code & logs	Coding tools

F1 • SELF-TESTED

Runs on device

36 MiB

INT8 quantized size;
runs offline on CPU, no GPU

Open, embeddable

Apache-2.0

Open weights;
any product can embed it

Accurate

2.4 %

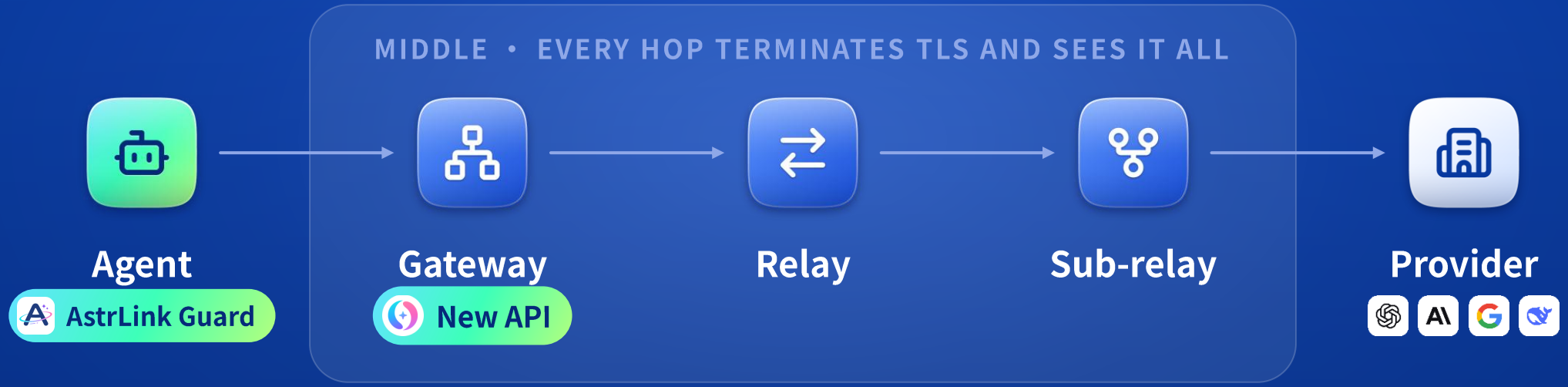
Credential miss rate on
2,158 test credentials

Used by AstrLink · magpie · Godwit · Claude Web · NexTerm

Product first: **privacy as a building block anyone can afford.**

Data: AstrLink Guard model card 0.1.3 (2026-11), evaluated on the project's own benchmarks; adoption per each project's source code

One layer is not enough.



9 / 428
relays injected malicious code into the tool calls they returned

401 / 440
sessions ran with auto-approve, executing whatever came back

99 credentials
leaked in these sessions

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

Five Layers of AI Traffic Security

		QUESTION IT ANSWERS	MEANS	WHO DOES IT
L5	**Instruction integrity**	Who sent it? Was it altered?	Sign · attest · verify	No standard yet
L4	**Execution**	How far may the agent go?	Sandbox · scopes · approval	
L3	**Data**	What can be sent out?	Detect · redact on device	AstrLink Guard
L2	**Access**	Who can call, and how much?	Auth · quotas · audit	New API
L1	**Transport**	Is anyone eavesdropping?	TLS encryption	Widely deployed

One party can secure four layers; **the top needs the whole chain.**

The layering is our team's synthesis (2026-11)

No one can build this layer alone.

01 · PROVIDER

Sign

"Who said this?"

Providers sign tool calls; agents check the public key

Email	DKIM · RFC 6376
Agents	No standard yet

02 · EVERY HOP

Declare

"Who changed what?"

Each hop declares its edits; undeclared changes fail

Email	ARC · RFC 8617
Agents	No standard yet

03 · AGENT

Verify

"Safe to auto-run?"

Auto-run once verified; otherwise ask a human

Email	DMARC · RFC 9989
Agents	No standard yet

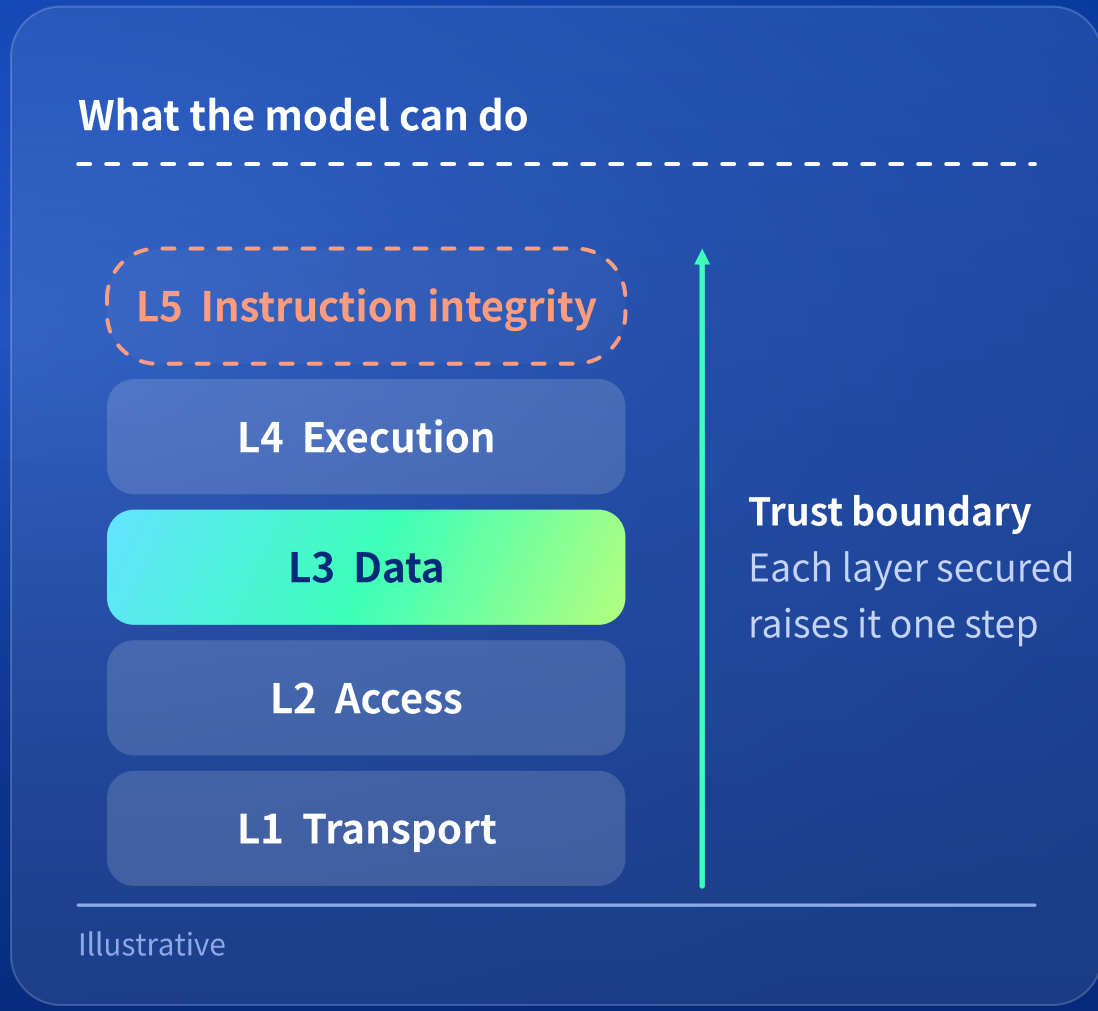
Cross-org trust needs open standards. **Let's write this one together.**

References: arXiv:2604.08407 Appendix C · RFC 6376 DKIM · RFC 8617 ARC · RFC 9989 DMARC

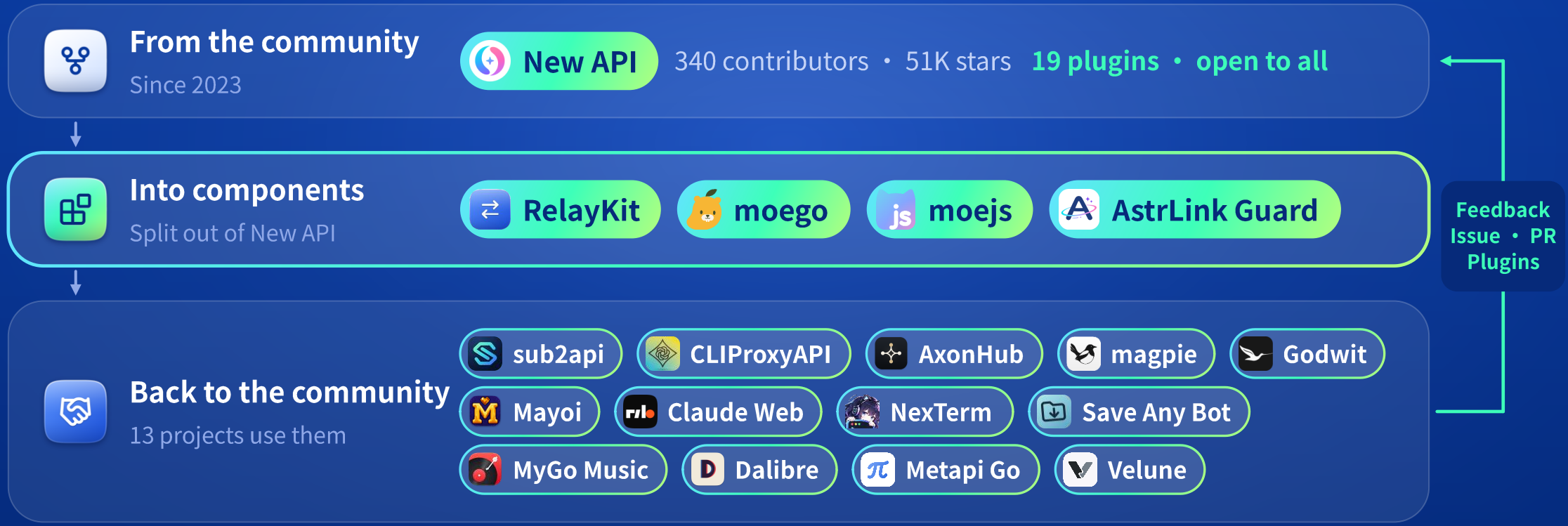
Security earns trust.

Each layer we secure lets us trust the agent with a little more. The closer trust gets to what the model can do, the more value agents deliver.

The ceiling of agents is trust, not intelligence.



From the community, for the community.



In the agent era, **trust is for the open-source community to define.**

Sources: GitHub contributor stats, index.json of new-api-plugins, each project's go.mod and source code (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

Bai Jingyuan

CEO, QuantumNous • Co-founder, New API

New API | QuantumNous 锟腾

GitHub WeChat X RedNote

