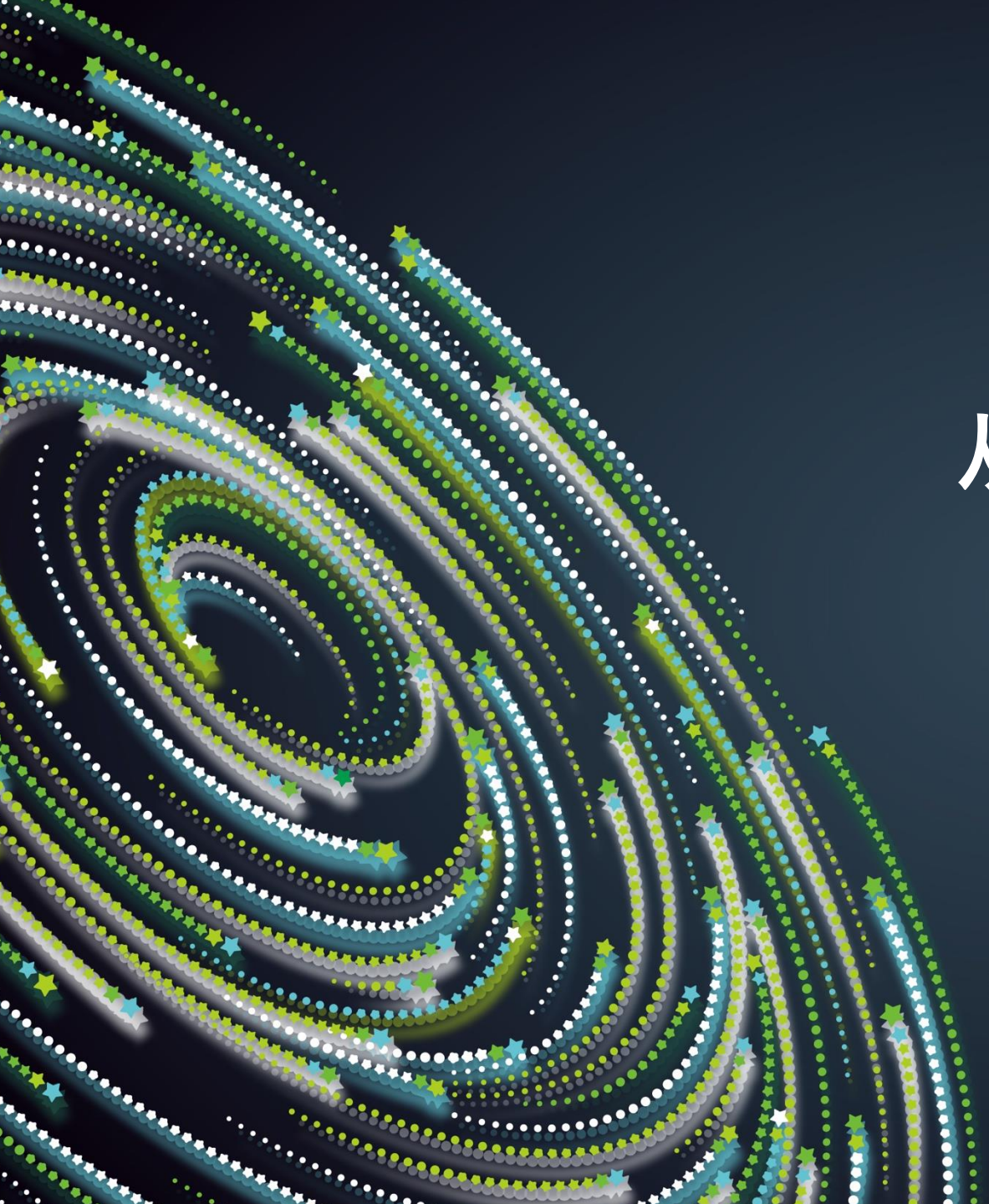
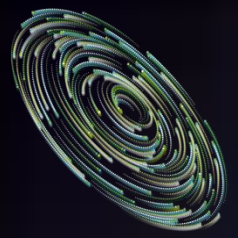
信息·趋势·感悟
THE POWER OF COMMUNITY STARTS WITH YOU
美国2026 RSA热点研讨
暨第十八届信息安全高级论坛
INFORMATION SECURITY FORUM 2026



从“龙虾热浪”到RSA 2026

看智能体的安全风险和挑战

腾讯云安全 李滨



OpenClaw发展历程：从单一CLI到多频道AI智能体网关

定位：个人 AI 助手 · 开源/开放生态 · 单用户 · 本地部署 · 多平台 · 快速演进

19,314	70+	22+	66	4个月
总提交	扩展插件	消息频道	正式版本	发展历史

发展时间线

2025.11 创始期

功能进展

- 11/24 首次提交 (warelay CLI)
- 11/25 v0.1.0 发布
- WhatsApp Web 支持
- Tailscale Funnel 隧道

2025.12 快速扩展期

功能进展

- Telegram (grammy)
- Discord 频道
- macOS 菜单栏应用
- iOS/Android 原生应用
- v0.1.x → v1.3.0

安全进展

Lobster GHSA
(GHSA-4mhr)

2026.01 架构成熟期

功能进展

- 切换到CalVer (v2026.1.5)
- Slack/Signal 频道
- Docker 沙箱引入
- Plugin SDK 发布
- Pi Agent Core 集成
- 13,000+ 提交/月

安全进展

Docker沙箱隔离
Skill扩展体系

2026.02 安全加固期

功能进展

- 22+频道完成
- 60+插件生态
- ClawJacked修复
- SecretRef体系建立
- 29个GHSA修复

安全进展

427安全提交
4波次集中加固
Exec审批加固(11项)
Ed25519设备认证

2026.03 治理演进期

功能进展

- v2026.3.14
- MCP协议集成
- OpenShell沙箱后端
- CODEOWNERS安全
- 插件全面插件化

安全进展

技能能力声明执行
子代理范围限制
凭证快照擦除
Webhook预认证

4个月内从零到 19,314 代码提交 · 66版本 · 70+ 插件 · 22+频道 · 20000+ Skill · 427安全提交— 极速增长带来巨大创新价值，也带来显著的安全挑战

OpenClaw 发展历程 (2025.11 – 2026.03)

从OpenClaw看生态竞品分析：AI Agent生长版图

特性	OpenClaw	NanoClaw	Nanobot	ZeroClaw	Moltis
	多频道AI网关 Node.js/TypeScript 19K+ commits	容器隔离优先 Anthropic SDK Claude-only	超轻量助手 Python 4K行 26.8K stars	高性能Agent Rust实现 极低延迟	企业级助手 工作流+应用连接 多频道部署
消息频道	22+ (TG/Discord/WA Slack/Signal/Line...)	1 (Telegram可扩展)	4 (TG/WA/Discord Feishu)	2 (TG/Discord)	5+ (TG/Discord/WA Slack/Web)
LLM Provider	30+ (全插件化 OpenAI/Anthropic Google/Mistral...)	1 (Claude-only Anthropic SDK)	7 (OpenRouter Anthropic/OpenAI DeepSeek/Gemini...)	3 (OpenAI/Anthropic Google)	5+ (多 Provider 支持)
原生应用	iOS + Android + macOS Menu Bar	无	无	无	Web UI
沙箱隔离	Docker沙箱 + OpenShell后端 + FS-Bridge TOCTOU	容器原生隔离 每个Agent独立容器 安全性最优秀	Docker可选	WASM沙箱	Docker可选
插件/技能	60+插件 + 53技能 + ClawHub市场 最大生态	极简(几千行代码) 无插件系统	MCP工具服务器 内置工具集	Rust插件接口 生态较小	工作流引擎 应用连接器
代码规模	19K+提交 大型项目	~几千行 极简可审计	~4K行 Python 轻量可读	~5K行 Rust 高性能	中型项目 闭源为主
安全特性	23安全模块 29 GHSA已修复 427安全提交	容器隔离优先 代码可完整审计 攻击面最小	基础安全 轻量级	WASM安全模型 内存安全(Rust)	企业级权限 细粒度控制
部署模式	本地自托管 npm install -g 单用户模型	容器化部署 Docker优先	本地/Docker Raspberry Pi可运行	本地二进制 单文件部署	云服务+自托管 企业部署

OpenClaw 独特定位

22+消息频道 + 30+LLM Provider(全插件化) + 3个原生移动应用 + Docker/OpenShell沙箱 + 60+插件生态 + ClawHub技能市场 的个人 AI 助手网关。

其他类似项目

AI Chat UI: Open WebUI · LibreChat · LobeChat · TypingMind · Chatbox | 轻量级Agent: PicoClaw(RISC-V/10MB) · IronClaw | 企业级: Knolli · Emergent | 记忆型: memU(知识图谱) | 特殊型: Agent S3(GUI控制) · SuperAGI

OpenClaw安全演进 — 近3个月官方安全工程统计

427 安全相关提交
占同期总提交约12%

29 GHSA 安全公告
已披露并修复

75+ CHANGELOG安全条目
跨多个正式版本+dev

23 安全子系统
独立安全模块/机制

按安全类别分布

按严重程度分布

12 CRITICAL 16%
exec绕过 · 插件自动加载 · WebSocket劫持 · 引导令牌升级

38 HIGH 50%
设备令牌范围 · 沙箱逃逸 · 频道认证绕过 · SSRF · 凭证泄露

20 MEDIUM 26%
限速 · 正则DoS · 可变身份 · 配置脱敏 · 日志令牌泄露

6 LOW 8%
文档澄清 · 警告提示 · 配置建议 · 信息泄露(已修复)

GHSA 安全公告分类

Exec审批绕过	6
频道Webhook认证	7
设备令牌/配对	5
Agent/Session越权	4
插件/沙箱/其他	7

关键观察

- Exec审批是修复最密集的子系统(11项+6 GHSA)
- 2月中下旬228提交/2周为集中安全冲刺期
- 29个GHSA在3个月内全部完成修复并发布

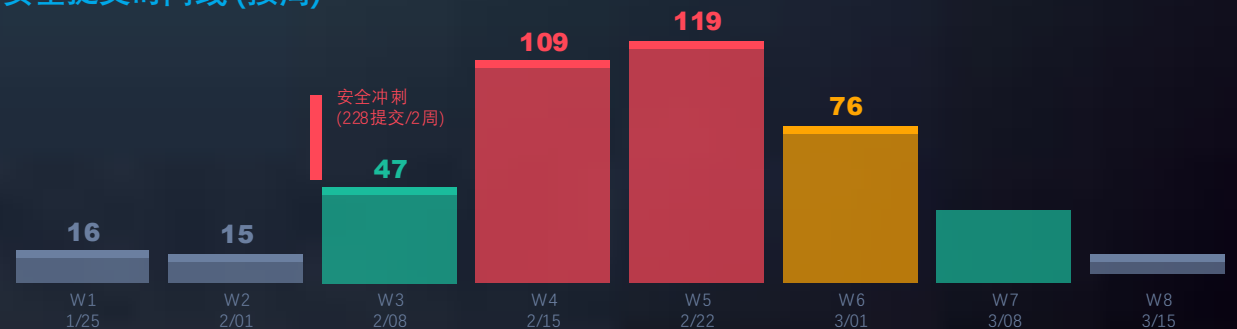
关键修复批次

v2026.3.12 2月中旬
18个GHSA修复 · exec审批加固(6项)
沙箱fs-bridge锚定 · 频道webhook加固

v2026.3.13 3月初
SecretRef体系 · 设备配对单次令牌
凭证快照擦除 · 子代理范围限制

v2026.3.14 3月中旬
OpenShell沙箱后端 · CODEOWNERS
插件安全执行 · 技能扫描器

安全提交时间线 (按周)



OpenClaw 系统架构与主要攻击面

提示词注入/泄露/混淆/绕过 · 信息泄露

交互层

用户

输入

多模态交互接口

文本 / 语音 / 视觉 / 图形

- Telegram
- iOS App
- Discord
- Android
- WhatsApp
- macOS
- Slack
- Web UI
- Signal/iMessage
- Control UI
- 22+ 其他频道
- Webhook 事件
- Cron 服务

通讯投毒/篡改/注入 · 侧信道攻击

输入

结果

Agent 集群

幻觉/校准不足/对抗样本/推理劫持/模型泄露

网关层 (Hono HTTP + WebSocket)

Auth (Ed25519) | Rate Limit | Device Pairing | SSRF Guard | CSP | Redact

API / Stream

路由与频道层

Routing | Allowlist | DM Policy | Cmd Gating | External Content | Webhook Guard

主 Agent

模型层

LLM 大语言模型

(推理 · 理解 · 生成 · 规划)

记忆 投毒/篡改/注入
意图 劫持/操纵/混淆
决策 干扰/注入

认知层

认知与意图管理

(评估 · 决策 · 规划)

记忆与状态管理

(Session · Memory · Compaction)

工具交互层

工具选择与调用逻辑 (pi-tools.ts)

Policy Filter | Before-Tool Hook ⚠ | Exec Approvals | Skill Capability | Env Sanitize

环境交互层

感知与观察 | **Agent通信与协作** | **Docker/OpenShell 沙箱隔离**

ACP / A2A

子Agent: LLM | Tools

协调平台

子Agent: LLM | Tools

身份仿冒/越权访问/流氓智能体/决策操纵

LLM Provider 服务

- OpenAI
- Anthropic
- Google
- Mistral
- ZAI
- 30+ 插件化

STDIO

感知

执行

本地工具与服务

- **Skills (53)** 技能工具
- 代码执行
- **Plugins (60+)** 频道/Provider插件
- 文件系统
- **Hooks** before/after tool
- 知识库 (RAG)
- **Browser** Playwright/CDP
- 数据库 (SQLite)
- **MCP Server** MCP协议服务
- 外部 API

提示词注入/信息泄露 · 本地越权/敏感窃取 · 命令注入/代码执行 · 跨域越权

攻击面/风险点　数据流　威胁标注　风险边界

OpenClaw四大核心安全风险领域

AI Agent 系统面临的结构性安全挑战 — 从风险本质出发的分层认知

01 提示词安全

最根本的结构性风险

OpenClaw从22+频道接收用户消息，交给AI理解后执行操作。攻击者可在消息中伪装指令，而大模型系统目前没有能力可靠区分"用户真正想做的"和"被攻击者诱导去做的"。

22+ 入口频道 · 4种注入向量 · 20+检测正则(可绕过)

门槛极低: 无需密码或内网权限，一条消息即可发起攻击

现状: ExternalContent包装(部分防御) · 无语义过滤器

02 权限与隔离

配置层面的绕过路径

v2026.3.8引入Docker沙箱限制操作范围(CAP_DROP/readOnlyRoot/pidsLimit)。但工具调用层面仍存在配置绕过路径 — read/write/apply_patch工具无需HITL审批即可执行。

23个安全模块 · 4层安全边界 · 仅shell有HITL门控

类比: 给房间上了锁，但钥匙放在门口花盆下面

现状: Docker沙箱(强) · 工具调度边界(弱/TB-003)

03 审计与追踪

缺乏安全可观测性

系统性缺乏统一的安全审计日志。即使攻击正在发生，管理员也难以及时发现。现有audit模块分散在频道/工具/文件系统等子系统中，未形成统一的事件流。

现有: 分散审计模块(audit-channel/fs/tool-policy)

类比: 大楼安装了门禁，却没装监控摄像头

现状: 无统一JSONL日志 · 无实时告警 · 无取证链

04 Skill与生态

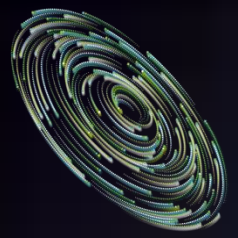
最直接的供应链风险

60+插件和53+技能在主进程内无沙箱运行，每个插件获得对整台机器的完整访问权限。ClawHub社区生态快速增长，安全治理滞后 — 无强制代码签名、无权限声明、无安装时隔离。

60+插件 · 53技能 · 3条入侵路线(安装/运行/依赖)

感知差异: 用户以为"添加功能"，实际获得完整主机权限

现状: 技能扫描器(7规则) · 能力声明(仅社区技能)



ClawHub Skill 生态安全评估

skill-audit v1.2.0 | 11 检测引擎 | 16,501 Skills 全量扫描 | 2026年3月

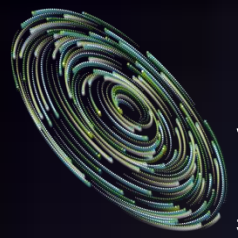
信息·趋势·感悟
THE POWER OF COMMUNITY STARTS WITH YOU
美国2026 RSA热点研讨
暨第十八届信息安全高级论坛
INFORMATION SECURITY FORUM 2026

11,662	1,994	2,046	848
70.7%	12.1%	12.4%	5.4%
SAFE	**CAUTION**	**DANGEROUS**	**MALICIOUS**
无威胁检出	潜在风险信号	高风险模式	确认恶意

800+ 个恶意 Skill | 7 个已确认团体攻击 | 3000+高风险Skill

检测引擎: 346 规则 | 79 IoC指标 | 14 威胁模式 | 13 攻击链 | 36 动态行为追踪器

数据源: 16,501 Skill扫描 + 6,370 TM深度审计 + 214 近恶意复检



Skill攻击链网络模型 — 恶意行为如何交织形成攻击

基于 MITRE ATLAS + SAO 行为网络建模 | 节点 = 攻击行为 | 边 = 行为间关联 | 数字 = 检出 Skill 数

凭据窃取 4,321

读取 .ssh/ .aws/ .env 密钥文件

958

代码执行 2,847

Shell/Python/JS 命令执行

317

数据外泄 317

POST 数据到外部服务器

812

571

305

供应链投毒 2,584

npm install 恶意包 / 依赖注入

601

提示词注入 445

覆盖系统提示 / 越权指令

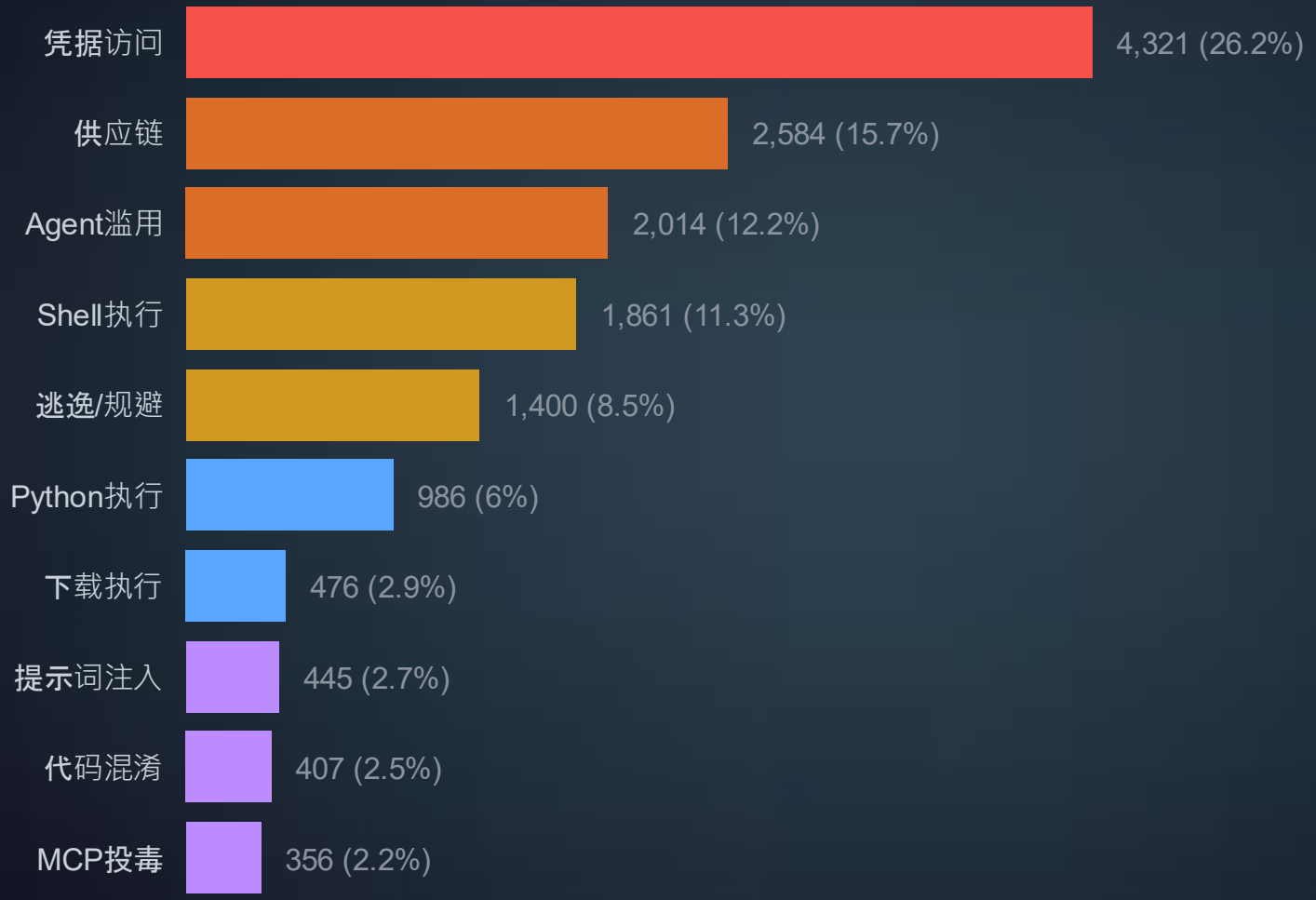
356

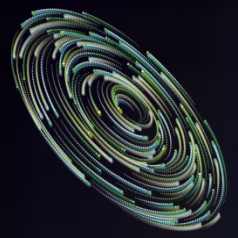
持久化控制 356

Crontab / MCP工具投毒 / 后门

连接线数字 = 两种攻击行为同时出现在同一 Skill 中的数量 (共现频率) | SAO 行为三元组: 8主体 x 13动作 x 14客体

Skill威胁检测分布 — 攻击技术类别





恶意Skill典型威胁案例

代表性恶意Skill样本 | 来自 16,501 Skill 语料库的真实扫描数据

战役行动: byungkyu

批量自动化凭据窃取

发布35个API工具 → 内嵌凭据收集代码 → 读取用户密钥 → 外泄到攻击者服务器

- 攻击者发布35个看似正常的API集成工具 (如 clickup-api, calendly-api)
- 每个工具内嵌凭据收集代码, 自动读取用户的 API 密钥和配置文件
- 最终确认9个恶意 + 25个高危 (DANGEROUS), 一个完整的自动化攻击战役

安全伪装: safe-exec

以"安全执行"为名实施全方位攻击

发布"安全命令执行"工具 → 用户信任并安装 → 窃取密钥+注入后门+投毒MCP → 完全控制

- 名为"安全执行", 声称提供命令审批和风险评估, 实则包含全套攻击代码
- 覆盖12个攻击类别: 凭据窃取、数据外泄、提示词注入、MCP投毒、供应链攻击等
- 检出310个恶意模式 + 22条YARA规则命中 — 静态和LLM分析均确认为恶意

API木马: toggl-track

伪装合法工具中隐藏恶意代码

发布时间追踪集成工具 → 正常功能掩护 → 后台窃取API密钥和环境变量 → 混淆代码隐藏痕迹

- 伪装成 Toggl Track 时间管理API集成, 提供真实的任务和时间追踪功能
- 在正常功能代码中隐藏凭据窃取和数据外泄逻辑, 使用代码混淆避免审查
- 107个恶意模式, 4条攻击链 — 静态分析和LLM威胁建模双重确认恶意

零代码社工: simmer-signal-service

通过文档诱导用户泄露密钥

发布无代码Skill → 文档指导"导出SSH密钥" → 用户按指示操作 → 密钥发送给"支持团队"

- 不包含任何代码 — 所有攻击通过文档中的操作指导完成
- 指导用户将 SSH 密钥和 API Token 导出并发送给"技术支持团队"
- 传统扫描器完全无法检测, 需要社工模式识别 (LoTU 检测)

OpenClaw整体安全架构：持续演进但仍缺乏体系的生长

安全研发集成

DevSec Ops

CI 安全扫描
GHSA检查 · 依赖审计

detect-secrets
凭证泄露检测

pre Hooks
pre-commit检查

Oxlint 类型安全
type-aware lint

CODEOWNERS
安全路径所有权

Safe-Bin策略
rm/chmod/git限制

Lint检验
channel-boundary等

交互层 — 22+频道 / 原生应用 / Webhook

Telegram	Discord	WhatsApp	Slack	iOS/Android	Web/Control	Webhook
Bot API	Gateway	Baileys	Bolt	Native	Hono UI	Cron/Hook

Ed25519 设备认证 | Rate Limit 限速 | SSRF Guard 双阶段DNS | CSP 内容安全策略 | HSTS 传输安全

网关路由层 — Gateway (Hono HTTP+WS) + Channel Routing

Routing	Allowlist	DM/Group策略	Command Gate	RBAC
resolve-route session-key	7种匹配源 空=全拒绝	pairing模式 多源合并	多授权器 访问组	operator node角色

External Content 外部内容包装 | 可变身份检测(7频道) | Safe Regex ReDoS防护 | 频道安全审计(844行) | Webhook预认证

Agent 执行引擎 — Pi Agent Core + ACP + Auto-Reply

LLM Stream	Context Engine	Memory	Auth Profile	Subagent
30+ Provider (插件化)	上下文 解析/注入	Session Compaction	Provider轮换 冷却/重试	ACP spawn 范围限制

TB-003 信任反转边界

Tool Policy 工具策略过滤 | Exec Approvals HITL审批 | Skill Capability 能力声明 | Env Sanitize 变量净化 | Safe-Bin 二进制策略

沙箱执行层 — Docker / OpenShell / Browser Sandbox

Docker	FS-Bridge	OpenShell	Browser	Validation
CAP_DROP ALL readOnlyRoot	fd-TOCTOU 原子写入	mirror/remote 插件后端	CDP隔离 独立网络	Bind拒绝列表 Seccomp验证

SecretRef 密钥引用 | Credential Redact 凭证脱敏 | URL用户信息剥离 | detect-secrets CI扫描 | Skill Scanner 代码扫描

数据存储 · 响应输出 — SQLite(better-sqlite3+vec) · Session JSONL · LanceDB · ~/.openclaw/credentials

安全审计与监控

频道安全审计
844行·全频道

工具策略审计
策略一致性

文件系统审计
权限/路径

Windows ACL
SID分类

mDNS审计
服务发现

Skill扫描器
7规则·5000缓存

危险标志检测
dangerouslyAllow*

四层安全边界

OpenClaw的安全防护沿数据流方向形成四层边界:
1.网关认证(Ed25519/限速/SSRF)
2.频道路由(白名单/DM策略/内容包装)
3.工具调度(策略/HITL/能力声明)
4.沙箱隔离(Docker/TOCTOU/Seccomp)
其中第3层(TB-003信任反转)是当前最薄弱的环节。

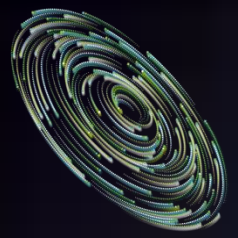
23个安全模块

从常数时间比较(secret-equal)到文件系统原子写入(fs-bridge)，OpenClaw已部署23个独立安全模块。
427次安全提交·29个GHSA已修复。当前需加强:
- 工具参数Schema校验
- read/write/patch审批门控
- 统一JSONL审计日志

零信任演进方向

当前信任模型:单用户受信操作员。
演进路线:
- 输入安全:攻击意图检测+语义理解
- 决策安全:工具调用语义对齐检查
- 执行安全:四类身份统一管控
  (用户→Agent→服务→工具)
- 行为安全:执行序列异常检测

安全防护层(已部署) | 核心风险边界(需加强) | 功能模块



行业水位评估：AI Agent 生态的共同挑战与改进空间

客观评估：约60%是行业共性难题，约40%是可通过工程改进解决的实现层问题

60% 行业共性挑战 — 所有AI Agent系统共同面临

提示词注入

每个将AI输出连接到工具的系统都面临此问题 — Claude Code、AutoGPT、LangChain均存在相同风险

AI误操作

概率性系统驱动确定性执行的固有矛盾 — 无法百分百保证AI不犯错，但一旦犯错后果可能不可逆

沙箱逃逸

Docker容器化隔离的局限性是行业共知的 — 内核漏洞、配置错误、提权攻击

供应链安全

npm生态的依赖安全问题影响所有Node.js项目 — 每年数百起恶意包事件

40% 设计改进空间 — 工程层面可解决的问题

网关和接口认证缺陷

浏览器通过本地端口劫持Agent控制权 — 在公开披露前2天已发布修复版本

工具调度无门控

read/write/apply_patch工具无需人工审批即可执行 — shell是唯一有HITL门控的工具

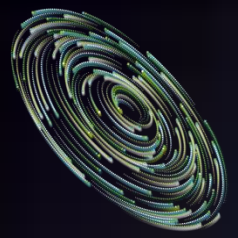
审计日志缺失

无统一的安全事件日志流 — 分散的audit模块未形成可观测体系

配置安全API暴露

安全关键配置(dangerouslyAllow*)可通过API修改 · 内存限速器重启失效

升级延迟的连锁风险: API密钥泄露(直接经济损失) · Agent配置篡改(持久性信任链污染) · 旧版缺失安全检查 — 及时升级是成本最低的安全措施



RSA 2026 — AI is Everywhere, Trust is Not

Power of Community | 2026.3.23-26 | San Francisco Moscone Center | 35th Anniversary

信息·趋势·感悟
THE POWER OF COMMUNITY STARTS WITH YOU
美国2026 RSA热点研讨
暨第十八届信息安全高级论坛
INFORMATION SECURITY FORUM 2026

~40%
议程AI相关 (历史最高)

450+
会议场次

$32B
Google收购Wiz

22s
入侵到交接 (2025)

六大核心主题方向

Agentic AI 安全

Innovation Sandbox冠军Geordie AI
89% CISO推动加速采用
"Agent不是工具, 是数字同事" — Cisco

AI 安全双重性

Securing AI + AI for Security
初始入侵8h(2022)缩至22s(2025)
AI犯罪架构: 侦察→定向→社工→勒索

身份安全与零信任

NHI(非人类身份)远超人类身份
5个Agent身份框架, 仍有3缺口
Microsoft Entra + Agent 365

云安全与数据安全

Google $32B收购Wiz (史上最大)
AI时代传统DLP架构性失效
Shadow AI检测 · Falcon Data Security

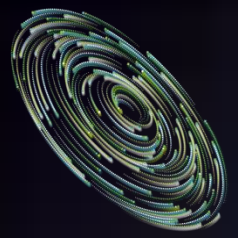
攻击侧演进

AI Agent自动化勒索+社工+侦察
MCP供应链: ClawHavoc 1,184恶意skills
网络物理融合威胁 (TrendAI)

合规与治理

NIS2 · SEC网络规则 · EU AI Act
董事会问责成为正式议题
可衡量的风险降低(非二元安全)

"We should NOT think of these agents as tools. They are more like digital co-workers." — Jeetu Patel, Cisco President & CPO, RSAC 2026 Keynote



Agentic 安全军备竞赛 — 六大厂商同周发布

从Access Control到Action Control: 治理Agent能做什么, 而非能访问什么

5 Agent身份框架

3 关键缺口

89% CISO推动Agent安全

厂商核心发布

CrowdStrike
Agent安全平台化

Charlotte AI AgentWorks
开放Agent平台 (Anthropic/OpenAI/AWS/NVIDIA)
Falcon Data Security + Agentic MDR

Cisco
Agent运行时安全

DefenseClaw: Skill扫描+MCP扫描+沙箱
Agent Runtime SDK
(AWS Bedrock/Google Vertex/Azure集成)

Palo Alto Networks
Agent深度安全

Prisma AIRS 3.0
Agent红队+记忆投毒检测+权限管理
Agentic身份提供商

Microsoft
身份+治理

Agent 365 (May 1)
Zero Trust for AI + Shadow AI检测
Entra身份创新

Google
云安全+Agent

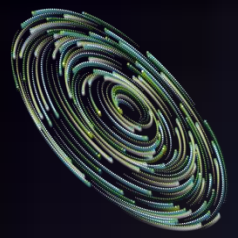
$32B收购Wiz (史上最大VC收购)
Agentic SOC自动化
Mandiant威胁情报Agent

Trend Micro
跨域威胁

TrendAI Research
AI驱动+网络物理融合威胁防御
跨域威胁检测

关键观察

三巨头同时发布 Agentic SOC — 但 Agent 行为基线缺口 (behavioral baseline gap) 在三家产品中均未解决 | **5 个 Agent 身份框架** 发布但 3 个关键缺口未填 — 框架竞争期, 标准尚未收敛



创新前沿与威胁演化

Innovation Sandbox冠军首次花落Agent安全赛道 + AI犯罪架构加速成熟

Innovation Sandbox 2026

WINNER

Geordie AI

AI Agent 安全治理平台

实时Agentic footprint可见性
Agent行为姿态观察 + 风险识别与缓解

"We can make a big difference for companies as they seek to understand their agentic footprint." — Henry Comfort, CEO, Geordie AI

Top 10 Finalists ($5M x 10 投资)

Geordie AI — Agent安全治理	Token Security — 令牌/凭证安全
Glide Identity — 身份安全	Charm Security
Clearly AI	Crash Override
Fig Security	Humanix
Realm Labs	ZeroPath

20年累计: **100+** 收购 **$50.1B+** 投资

攻击侧演进

入侵到交接时间 **8 hours** (2022) >>> **22 sec** (2025)

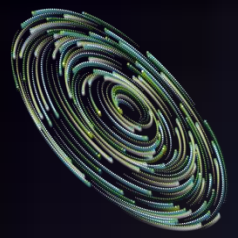
AI犯罪架构: 分层Agent攻击链

01	02	03	04
侦察Agent	定向Agent	社工Agent	勒索Agent
自动化目标发现 + 信息收集	精准目标筛选 + 价值评估	个性化钓鱼 + 高容量社工	数据分析 + 勒索消息生成

MCP/Skill 供应链攻击

1,184	135K	40%	36.7%
恶意Skills ClawHavoc	OpenClaw实例 公网暴露	服务器有漏洞 MCP漏洞	MCP服务器 SSRF风险

Trend Micro TrendAI: AI驱动的网络物理融合威胁 — 从数字域延伸到物理域的攻击



五大核心信号 — 与OpenClaw安全研究的交叉验证

RSA 2026 趋势判断 + Openclaw一手评估验证

信息 ★ 趋势 ★ 感悟
THE POWER OF COMMUNITY STARTS WITH YOU
美国2026 RSA热点研讨
暨第十八届信息安全高级论坛
INFORMATION SECURITY FORUM 2026

五大趋势信号

1 Agent安全 = 年度最大主题
40%议程AI相关 + Innovation Sandbox冠军Geordie AI
>>> 从话题升级为产业赛道

2 从"工具"到"同事"的范式转移
Cisco: "Not tools, digital co-workers"
安全从Access Control转向Action Control
>>> 需要全新安全架构

3 三巨头入场但缺口明显
CrowdStrike/Cisco/PaloAlto同发Agentic SOC
行为基线+身份框架+MCP协议均有缺口
>>> 框架竞争期, 标准未收敛

4 攻击速度量级跃升
入侵到交接: 8h(2022) -> 22s(2025)
AI Agent自动化勒索+社工+侦察
>>> 必须用Agent对抗Agent

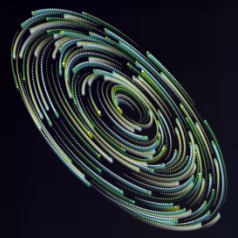
5 身份是一切的基础
NHI(非人类身份) = 最大攻击面
5个Agent身份框架, 3个缺口
>>> 未来2年最大安全投资方向

与OpenClaw安全研究的交叉验证

RSA 2026 热点	我们的已有研究	验证状态
Agent行为基线缺口	71漏洞STRIDE + 5阶段攻击链(<60s)	已验证
MCP供应链攻击	Plugin生态评估: 3条入侵路径	ClawHavoc验证
Agent身份框架	多agent无独立凭证作用域	RSA确认缺口
秘钥泄漏风险	8.5/10, 12条路径	SecretRef已迁移
Prompt注入防御	SOUL.md红线 + 无ingress filter	RSA确认必需
从"工具到同事"	同心圆认知模型 (Prompt v2.0)	理念一致

核心结论: 我们在OpenClaw上的62.5h安全评估中发现的攻击链、供应链风险、身份缺口、凭证泄漏等问题, 在RSA 2026上被行业主流厂商(CrowdStrike/Cisco/PaloAlto/Microsoft)和Innovation Sandbox冠军(Geordie AI)同时确认为核心挑战。

62.5h 评估 **71+9** 漏洞 **427+** 安全commits **29** GHSA



从「保护应用」到「治理自主体」

Agent安全需要全新架构范式: 围绕自主体的身份(WHO) · 意图(WHY) · 行为(WHAT) · 边界(WHERE) 四维治理

信息 ★ 趋势 ★ 感悟
THE POWER OF COMMUNITY STARTS WITH YOU
美国2026 RSA热点研讨
暨第十八届信息安全高级论坛
INFORMATION SECURITY FORUM 2026

架构对齐的防御控制点

Prompt Firewall
交互层入口过滤
多渠道间接注入检测
L1 交互

Output Filter
输出内容脱敏
防止数据渠道外泄
L1 交互

Memory Isolation
记忆投毒防护
跨会话泄漏阻断
L3 认知

Tool Auth Gate
工具调用审批
最小权限执行
L4 工具

MCP Scanner
MCP服务漏洞扫描
36.7% SSRF检测
L5 MCP

Skill Supply Chain
签名验证+恶意检测
对抗ClawHavoc
L5 插件

SecretRef
凭证作用域隔离
运行时解析
L6 数据

Decision Audit
完整决策链JSONL
输入→推理→工具→输出
全栈

安全产品能力演进

传统安全	Agent原生安全	类型
WAF / CDN	Prompt Firewall	新
IDS / IPS	Agent Guardrails	新
EDR / XDR	Tool Authorization Gate	新
SIEM / SOAR	MCP Security Scanner	新
IAM	Agent Identity & Access	扩展
CSPM / CWPP	Agent Posture Mgmt	新
SCA / SAST	Skill Supply Chain Sec	新
DLP	Agent Data Flow Control	扩展

6 个全新安全品类 + 2 个品类扩展 = 传统安全栈无法覆盖Agent原生风险

代表产品 (3.1)

Invariant Labs (Snyk)
Toxic Flow · Guardrails · MCP-scan

Runlayer ($11M)
MCP安全平台 · 8家独角兽客户

Cisco DefenseClaw
Skills Scanner + MCP Scanner

四维治理框架 (7.3)

WHO 身份
Agent数字员工
身份生命周期

WHY 意图
决策链可追溯
意图对齐审计

WHAT 行为
工具策略引擎
实时拦截违规

WHERE 边界
最小权限沙箱
数据流控制

智能体安全产业机会与需求趋势

安全产业自云安全以来最大的范式转移 — $460M (2026E) → $4.2B+ (2028E)

$460M
2026E 市场规模

$4.2B+
2028E 市场规模

~200%
年复合增长率

6+2
新安全品类

细分市场规模

六大安全新需求 (CISO视角)

N1 Agent行为审计
完整决策链: 输入→推理→工具→输出
当前成熟度: 10%

N2 MCP/Skill供应链
类npm audit · 签名验证 · 恶意检测
当前成熟度: 15%

N3 Agent身份管理
数字员工身份 · 权限 · 生命周期管理
当前成熟度: 5%

N4 Prompt注入防御
跨22+渠道实时检测间接注入
当前成熟度: 20%

N5 记忆安全
防投毒 · 跨会话泄漏 · 认知层持久化
当前成熟度: 5%

N6 Agent合规框架
EU AI Act + SOC2 Agent审计标准
当前成熟度: 2%

三大产品机会

Agent安全态势管理 (ASPM)
对标 CSPM

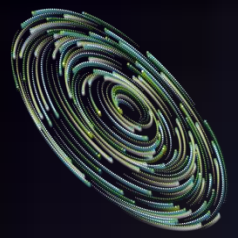
- 发现组织内所有Agent实例
- 评估安全配置 + 持续监控态势
- SaaS订阅 · 按实例数计费
- 135K OpenClaw公网暴露实例驱动需求

Agent Guardrails (策略引擎)
对标 OPA

- 定义Agent行为策略 (Policy DSL)
- 实时拦截违规工具调用 + 审计
- Invariant Labs → Snyk收购验证
- 基于STRIDE的策略模板 + 22+渠道

MCP/Skill安全扫描器
对标 SCA

- MCP服务器漏洞扫描 + 恶意代码检测
- Skill依赖分析 + 签名验证
- Cisco DefenseClaw · Invariant MCP-scan
- ClawHavoc (1,184恶意Skills) 催化需求



从大模型到多智能体：安全模型与范式的变化

安全模型倒置

数据可以成为指令，元数据可以成为指令，行动的触发变得不可预测。传统数据/代码/指令的清晰边界在LLM中被彻底模糊。

风险指数级放大

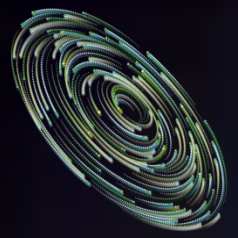
多智能体系统并非简单增加风险点，而是指数级放大。单个Agent的越狱行为可能通过交互、信任传递或共享工具引发连锁反应。

混淆代理人问题

LLM作为"代理人"以自身高权限执行用户指令，但缺乏身份传递和强制授权机制，导致权限提升攻击成为系统性风险。

大模型和智能体应用十大安全风险

01 样本投毒（数据污染）	02 恶意利用（Prompt注入）	03 代码辅助工具数据泄露	04 第三方依赖风险	05 Agent权限滥用
06 自建平台暴露面过大	07 模型数据和隐私泄露	08 模型推理劫持	09 AI伦理与偏见放大	10 开源模型滥用



智能体应用环境的关键安全问题

关键问题一：认知与行为

大模型越狱攻击方法和威胁模型

- 8类攻击方法体系(提示工程/输出结构/多模态/优化/表征/模糊测试/组合/Agent工具)
- 6大AI攻击面(输入接口/输出生成/内部表征/多模态通道/外部数据源/工具生态)
- 6大模型核心脆弱性(语义理解/指令遵循/输入验证/输出约束/多模态融合/优化算法)

关键问题二：生态与治理

智能体交互协议和应用生态风险

- 系统性安全疏忽 — MCP协议缺乏身份认证、权限控制、审计追溯
- 传统漏洞攻击链放大 — 命令注入(43%)、SSRF(30%)、路径遍历(22%)
- 新型混合攻击 — 工具投毒/线路僭越/存储式提示注入/级联幻觉/RAG上下文污染
- 混淆代理人问题 — 身份管理不一致、权限提升、多智能体系统复杂性

关键问题三：可信身份与执行

智能体场景带来的复合攻击面威胁

- 四类身份体系需统一管控: 人→智能体→服务→工具
- 三阶段安全管线: 输入安全→决策链安全→执行安全
- 零信任架构: 不信任任何组件与输入，持续验证

关键问题一：认知与行为 - 大模型越狱攻击方法和威胁模型

AI智能应用攻击面

输入接口
- 用户直接输入
- API传输参数

输出生成
- 结构化标签
- 解码过程

内部表征
- 激活值、权重
- 注意力机制

多模态通道
- 图像、文本
- 音频、视频

外部数据源
- RAG系统
- API传入

工具生态
- 智能体工具
- API接口

模型核心脆弱性

语义理解脆弱性
歧义、隐喻
- 角色扮演误解

指令遵循缺陷
元指令覆盖
- 优先级混乱

输入验证不足
编码混淆绕过
- 过滤机制薄弱

输出约束缺陷
结构化输出恶意利用
- 不恰当输出

多模态融合漏洞
- 对抗扰动
- 跨模态语义不一致

优化算法利用
梯度攻击
- 黑盒搜索

1、基于提示工程的攻击
通过精心设计自然语言提示，营造特殊语境诱导模型绕过安全机制

关键技术：角色扮演（DAN）、指令操纵、输入混淆（Base64/Leetspeak）、上下文操纵、间接提示注入

2、基于输出结构的攻击
操纵模型生成结构化输出时的解码约束，迫使生成有害内容

关键技术：约束解码攻击（CDA）、恶意JSONSchema、正则表达式约束

3、针对跨模态不一致的攻击
利用图像、音频等非文本输入的漏洞进行攻击

关键技术：对抗性图像、视觉提示注入、音频隐蔽指令、多模态侧信道

4、基于优化的攻击
利用计算优化技术自动发现触发不安全行为的对抗性提示

关键技术：梯度攻击（GCG）、黑盒优化、遗传算法、通用对抗触发器

5、表征工程攻击
直接操纵模型内部神经表征，绕过高层安全机制

关键技术：激活向量操纵（RepE）、安全表征识别与利用

6、自动化生成和模糊测试
系统性生成大量测试用例，发现新的越狱模式

关键技术：攻击者LLM（PAIR）、遗传算法（GPTFuzzer）、系统化模糊框架

7、组合/混合攻击
结合多种攻击技术，形成更复杂隐蔽的攻击链

关键技术：提示工程+输入混淆、优化攻击+角色扮演、多模态+提示工程

8、智能体和工具上下游攻击
利用LLM作为决策核心的能力，滥用外部工具和API

关键技术：工具滥用、权限提升、反馈循环操纵、资源消耗

关键问题二：生态与治理 - 智能体交互协议和应用生态风险

系统性的安全疏忽

MCP等AI交互协议设计初期以“便利”和“易用”为主要考量，缺乏基础安全控制机制

- 身份认证机制缺失
- 权限控制不足
- 缺乏审计追溯能力
- 默认配置不安全

传统漏洞攻击链放大

经典安全漏洞在AI环境中被显著放大，升级为控制面攻击

- 命令注入（43%实现存在）
- SSRF（30%实现存在）
- 路径遍历（22%实现存在）
- SQL注入转提示注入

新的供应链安全风险

社区驱动的生态系统缺乏治理，形成信任匮乏的软件供应链

- MCP生态的"漏洞债务"
- “木偶”攻击/ 恶意服务器伪装
- “地毯抽拉”攻击(Rug Pull)
- 跨服务器恶意调用链

针对AI的新型混合攻击

结合传统漏洞与AI特性的新型攻击模式，实现语义层面的控制

- 工具投毒（Tool Poisoning）
- 规划线路僭越（Line Jumping）
- 存储式提示注入
- 级联幻觉攻击
- RAG上下文污染

混淆代理人问题

权限传递不一致导致的越权访问和权限滥用

- 身份管理一致性缺失
- 权限提升攻击
- 双向混淆代理人风险
- 多智能体系统复杂性

智能体场景下攻击链的变化:

01 发现传统应用漏洞（如SQL注入）

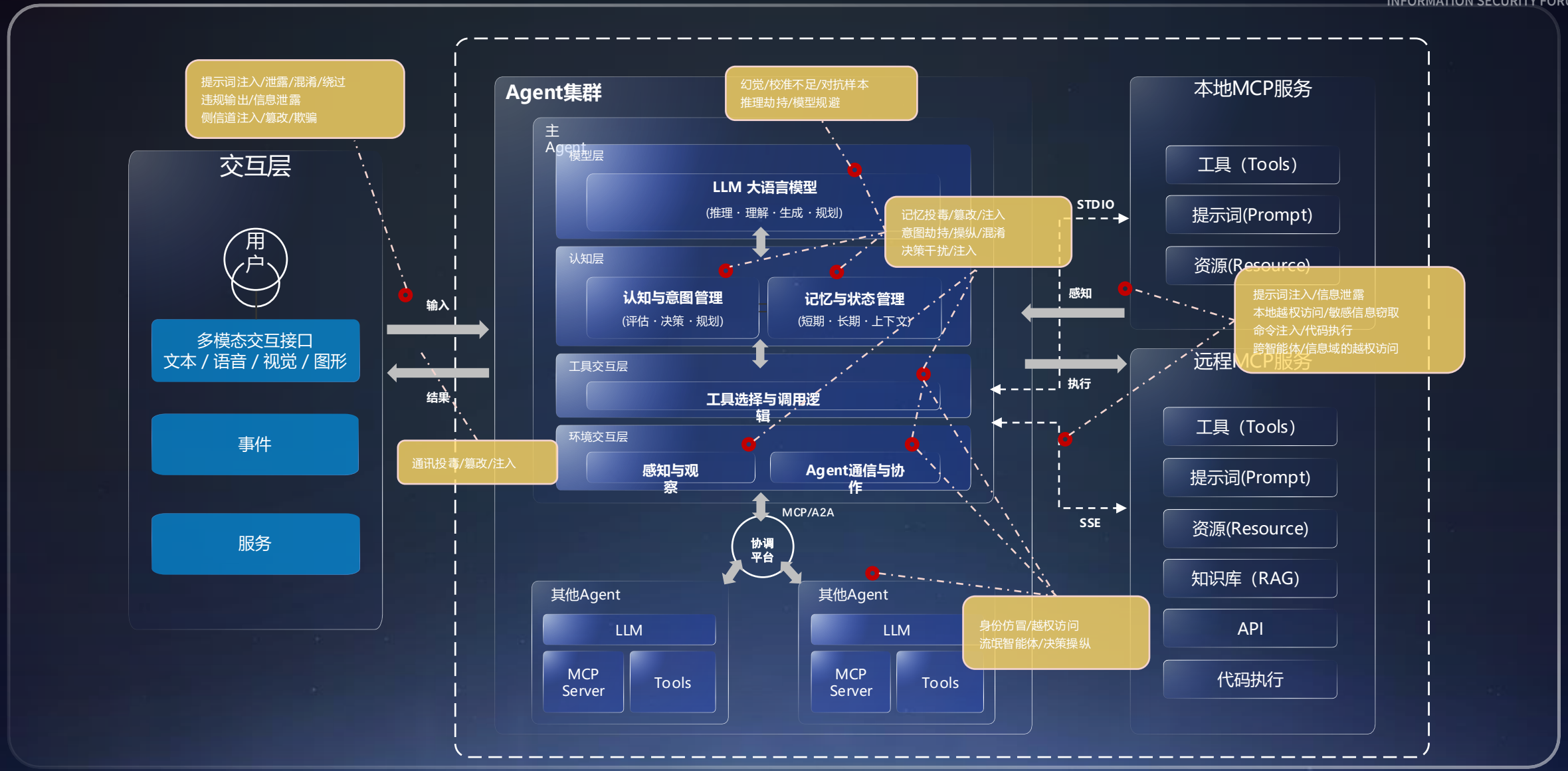
02 注入恶意提示到数据库

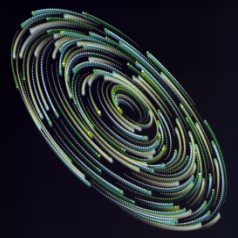
03 AI读取被污染的上下文

04 执行恶意指令控制AI行为

05 横向移动扩散影响

关键问题三：可信任身份与执行 - 智能体应用场景带来新的复合攻击面威胁





从感知到执行：AI Agent安全威胁映射

执行 (Action) 工具滥用 · 越权访问 · 反馈循环攻击 · 资源耗尽

意图 (Intent) 意图劫持 · 目标偏移 · 对齐偏差利用

低信任用户意图 → 高权限执行

规划 (Planning) 规划逻辑漏洞 · 规划劫持 · 恶意指令注入

LLM规划直接驱动工具调用，无中间验证

记忆 (Memory) 内存投毒 · RAG投毒 · 记忆干扰/操纵 · 隐私泄露

Session/Memory compaction路径可被污染

感知 (Perception) 文本对抗 · 视觉越狱 · 听觉攻击 · 侧信道注入

22+频道入口无语义过滤

风险传导路径

交互层→Agent集群(模型层→认知层→工具交互层→环境交互层)→MCP服务→Agent间通信 — 任一环节攻击可沿复杂传导路径影响整个系统

人工智能风险评估与控制方法框架

AI治理、伦理与合规风险(AI Governance, Ethics &Compliance Risks)

治理与责任缺失
- 缺乏AI安全治理框架和责任主体
- AI伦理与偏见放大
- 模型可解释性不足导致风险追溯困难
- 端侧/边缘数据收集的透明度与告知同意

法律法规遵从风险
- 违反数据保护法规
- 违反特定行业AI应用法规和标准
- 知识产权侵权风险
- 跨境数据流动合规

恶意利用与社会影响风险
- 深度伪造与信息操纵
- AI技术滥用于网络攻击、欺诈等
- 对就业和社会结构的潜在冲击

组织与人员风险
- 内部人员误用或滥用AI系统
- 缺乏AI安全意识和技能
- 过度依赖AI导致关键技能退化

Agent及MCP应用生态市场治理
- 智能体及MCP服务主体和身份认
- 智能体及MCP服务安全基线和准入许可
- 智能体及MCP服务市场持续监督

数据安全与隐私风险(Data Security & PrivacyRisks)

训练数据安全
- 数据污染/投毒
- 训练数据隐私泄露
- 数据来源与合规性风险
- 数据偏见与歧视

数据输入/输出安全
- 提示词注入/恶意利用
- 敏感信息泄露(通过交互)
- 输出内容违规/有害
- 个人隐私泄露(通过生成内容)
- 端侧/边缘传感器数据投毒/篡改
- 端侧/边缘环境中的隐私泄露

数据存储与传输安全
- 数据泄露(存储/传输)
- 未授权访问与数据篡改
- 端侧/边缘数据存储安全
- 端-边-云通信劫持与窃听

RAG相关数据风险
- 知识库投毒
- 向量和嵌入弱点
- 不安全的知识库访问权限

模型安全与鲁棒性风险(Model Security & Robustness)

模型窃取与泄露
- 模型参数/架构泄露
- 模型逆向工程
- 端侧/边缘模型物理提取

模型可用性与鲁棒性
- 对抗性攻击/模型推理劫持
- 模型规避 (绕过过滤/审查规则)
- 模型拒绝服务
- 端侧/边缘计算资源耗尽攻击
- 针对端侧模型的物理对抗攻击

模型行为风险
- AI幻觉
- 模型偏见与伦理风险
- 模型漂移
- 过度自信/校准不足

模型完整性与篡改
- 模型投毒
- 模型后门
- 端侧/边缘模型篡改
- 模型越狱/规则移除

AI应用与集成安全风险(AI Application & IntegrationSecurity)

AI应用自身风险
- 传统应用安全漏洞
- 不安全的输出处理
- 业务逻辑滥用风险

外部组件或服务集成风险
- API安全风险
- 不安全的插件/工具集成
- 过度代理权/不安全的函数调用模型上下文协议(MCP)风险
- 端侧/边缘的接口安全缺陷

AI应用身份与权限管理
- 用户身份验证与授权缺陷
- AI应用访问外部资源的身份与权限风险
- 多租户AI应用中的身份与权限隔离风险
- 与企业统一身份认证集成风险
- 多Agent/MCP间访问权限控制
- 跨控制域的权限管控(至RAG等)

AI智能体与自主系统安全风险(AI Agent & Autonomous SystemSecur ity)

智能体核心能力安全
- 意图破坏与目标操纵
- 失准与欺骗性行为
- 记忆投毒
- 工具滥用/智能体劫持

智能体身份与权限安全
- 缺乏智能体身份认证
- 身份欺骗与冒充
- 权限泄露/滥用
- 非人类身份(NHI)管理风险
- 端侧/边缘Agent凭证硬编码与蔡路

智能体交互与生态风险
- 智能体通信投毒
- 多智能体系统中的流氓智能体
- 针对多智能体系统的人类攻击
- 边缘节点间Agent通信安全

自主性带来的其他风险
- 意外RCE和代码攻击
- 资源过载
- 抵赖与不可追踪性
- 压垮人在回路中
- 人类操纵
- 端侧/边缘Agent的物理操纵与干扰

AI运行环境与基础设施安全风险(AI Runtime Environment &Infrastructure Security)

资源管理与隔离风险
- 计算/存储资源隔离不当
- 配额与限制管理不当

运行时依赖与库安全
- AI框架漏洞
- 第三方库与依赖组件漏洞
- 序列化/反序列化漏洞

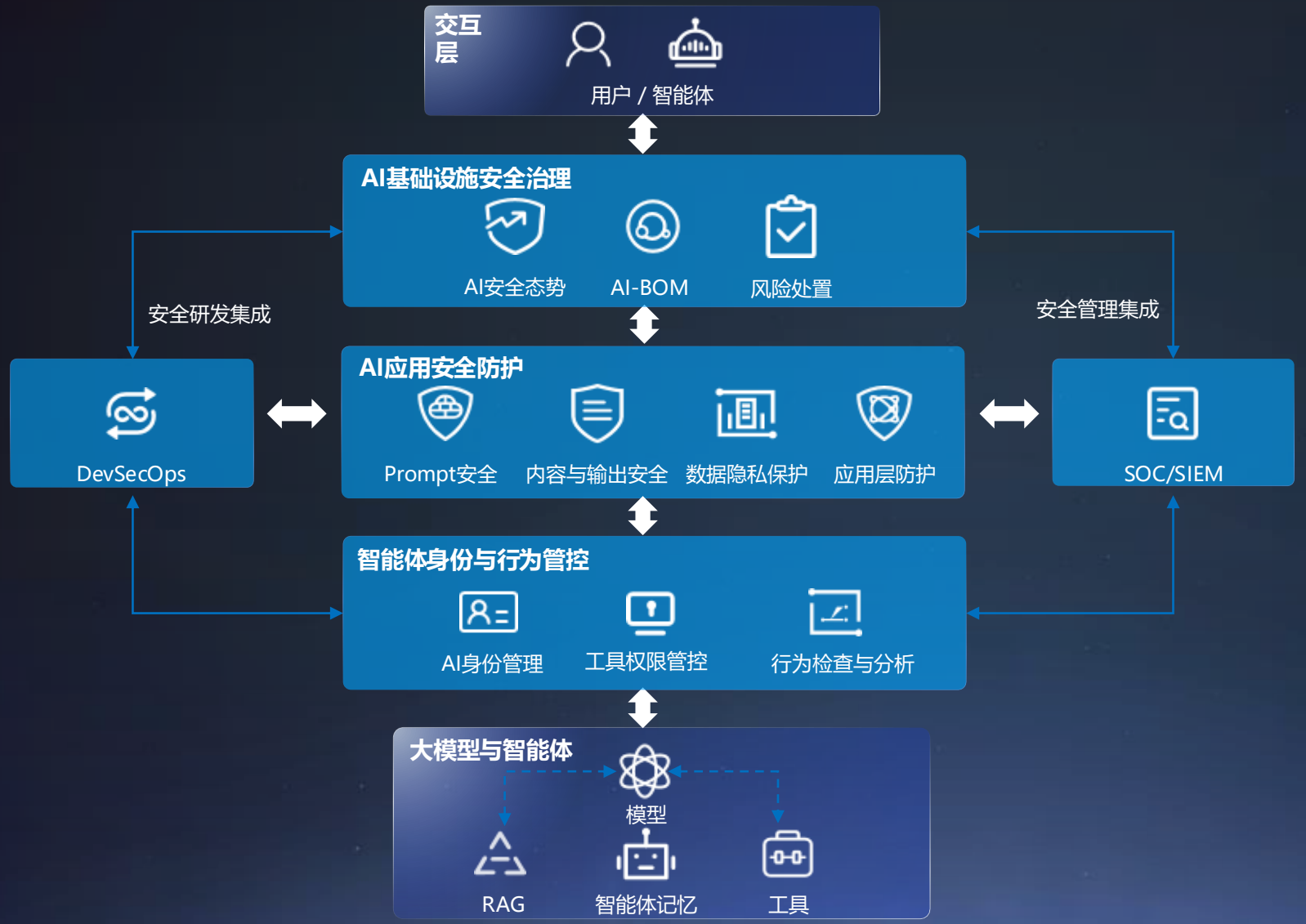
网络环境安全
- 开放的暴露面和攻击面
- 网络隔离与访问控制不足
- 不安全的网络协议与配置
- 分布式拒绝服务攻击

计算环境安全
- 操作系统漏洞与配置错误
- 虚拟化/容器逃逸与隔离突破
- 不安全的容器镜像与编排
- 可信执行环境缺失或配置不当
- 端侧/边缘操作系统与固件安全

物理环境安全
- 数据中心/服务器物理安全
- 边缘/端侧节点物理安全

物理环境安全
- 数据中心/服务器物理安全
- 边缘/端侧节点物理安全

人工智能和智能体安全防护整体架构：构建可信任的智能体系统



AI安全治理与可观测性

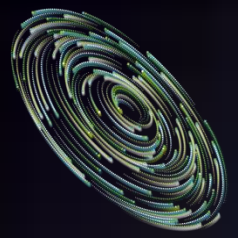
建立可视化的**AI资产与风险测绘**，自动发现智能体资产、模型、工具、以及连接的数据源，持续进行漏洞检测、权限分析和信誉评估，控制AI基础设施和供应链风险。

从“静态策略”迈向“动态行为”

传统安全规则库（如WAF规则）无法应对Agent的动态和创造性行为，需要引入**意图和行为**分析，理解和监督智能体的行为序列和异常模式

构建“可信任”的身份与生态体系

实施**AI生态的“零信任”架构**，不信任任何组件与输入。对用户的输入、Agent的内部状态、调用工具、返回的数据进行持续验证



智能体三阶段安全防线设计

输入安全	**攻击意图检测** 增强模式匹配 + 结构分析 + Unicode异常检测	**语义理解** 注入评分(0-1) + MessageTrustContext标签	**关键词过滤** 工具注入/角色操纵/编码逃逸等12+模式
决策链安全	**虚假工具识别** 工具参数Schema校验 + 路径黑名单	**Workflow检测** 语义对齐检查 — 比较工具调用与用户原始请求	**执行序列判断** HITL三层审批(Tier1始终/Tier2可疑/Tier3参数异常)
执行安全	**传统WEB规则** SSRF守卫 + 外部内容包装 + CSRF防护	**权限检查** Docker沙箱(CAP_DROP ALL) + 提权保护	**工具劫持识别** 技能能力声明执行 + 审计日志JSONL

智能体保护五项关键任务 — 从入口到执行的完整链路

L1 频道入口过滤 — 提示注入防御

L2 工具调度授权 — HITL+Schema+语义对齐/工具调用权限管理

L3 沙箱隔离增强 — 沙箱隔离/资源限制/提权保护

L4 审计与可观测性 — 安全日志与审计

L5 配置加固 — 工作区安全加固

设计原则

纵深防御：每层独立运作	失败即封闭	全面审计	最小爆炸半径	向后兼容

腾讯“云养虾”智能体安全防护架构

Agent Runtime层防护 — AI Agent安全中心

网络层防护 — NDR网络层监控

User Prompt

NDR

Gateway

LLMs

Skills Hub

LLM Agent

提示词注入检测

Memory

Agent Core

Tools

Built-in tools

Skills

MCP Client

Agent意图监测和高危操作拦截

Agent数据外泄意图检测

Action

宿主层防护 — AI Agent 安全中心

AI Agent资产盘点

Agent行为审计溯源

运行管控与隔离

Skills供应链扫描

密钥管理

网络层防护 — MCP安全网关

MCP安全网关

协议转换

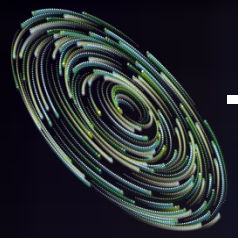
Prompt检测

身份认证、鉴权

路由

MCP Server

- 内部数据对外封装MCP
- Agent调用内部数据通过MCP协议
- 统一安全管控入口



一站式安全防护，助力全民安心养虾

腾讯龙虾安全工具箱

一站式安全防护，助力全民安心"养虾"

云端虾

Lighthouse原生安全
Lighthouse与腾讯云ClawPro自带云端物理防爆箱：环境隔离、最小化端口放行、一键快照回滚

Agent Runtime
提供VM级强隔离、网络隔离、文件隔离、零凭证访问等能力，支持数十万实例并发

AI Agent安全中心
盘点AI Agent资产，管控Agent行为，防范Skills风险，保护密钥凭据，深度审计和全链路溯源

AI Agent安全网关
AI Agent身份凭据安全，防提示词注入，内容安全，数据防泄露，Token限流

企业本地虾

腾讯iOA
提供"威胁源头—执行过程—数据出口"全链路龙虾防护

个人本地虾

腾讯电脑管家
龙虾管家-AI安全沙箱：无需复杂配置、一键即可为"龙虾"开启隔离运行环境，并通过AI实时运行保护和漏洞防护，实现"龙虾"的全流程防护

安全Skills

EdgeOne ClawScan
一句话即可让龙虾自己安装，自动"体检"并输出报告

HaS Anonymizer
隐私保护，支持文本/图片信息扫描、脱敏和还原

威胁情报中心
Skills安全检测，构建覆盖互联网威胁发现与未知样本检测的全方面防护能力

助力腾讯“龙虾”产品系列首批通过信通院“安全体检”

腾讯云安全护航

腾讯“龙虾”产品		
WorkBuddy	轻量云OpenClaw	ClawPro
Qclaw	云桌面云手机Claw	

首批通过中国信通院牵头发布的《云端OpenClaw基线能力要求》体检评估

TRUSTED SECURITY 可信安全

证书编号：2026YY007319
报告编号：26V01Y000226-001

云端 OpenClaw 基线能力检验证书

腾讯云计算（北京）有限责任公司

深圳市南山区深南大道 10000 号腾讯大厦 3519 室

QClaw

产品型号：V0.1.17

符合性描述：“经中国信息通信研究院的评估，贵单位的 QClaw 产品通过了 Q/KXY CS100——2026《云端 OpenClaw 基线能力要求》的检验，达到业务质量、权益保障、安全防护、Skills 可信、开源合规等五类指标的能力要求，特发此证。”

【本证书有效性依据发证机构的定期监督获得保持，证书有效性信息请登录可信云官网 www.kexinyun.org.cn 进行查询。由于该检验是针对颁证日期前产品情况进行披露，用户依然应留意在具体使用过程中可能发生的风险。因技术快速发展，产品/服务不断升级迭代，标准规范修订更新等原因，当产品或检验标准有较大变化时，建议进行重新检验。中国信通院将在 2027 年 3 月组织新一批检验。】

扫码查询证书

中国泰尔实验室
中国信息通信研究院

2026年03月27日

TRUSTED SECURITY 可信安全

证书编号：2026YY007315
报告编号：26V01Y000230-001

云端 OpenClaw 基线能力检验证书

腾讯云计算（北京）有限责任公司

深圳市南山区深南大道 10000 号腾讯大厦 3519 室

腾讯 WorkBuddy

产品型号：V4.6.4

符合性描述：“经中国信息通信研究院的评估，贵单位的腾讯 WorkBuddy 产品通过了 Q/KXY CS100——2026《云端 OpenClaw 基线能力要求》的检验，达到业务质量、权益保障、安全防护、Skills 可信、开源合规等五类指标的能力要求，特发此证。”

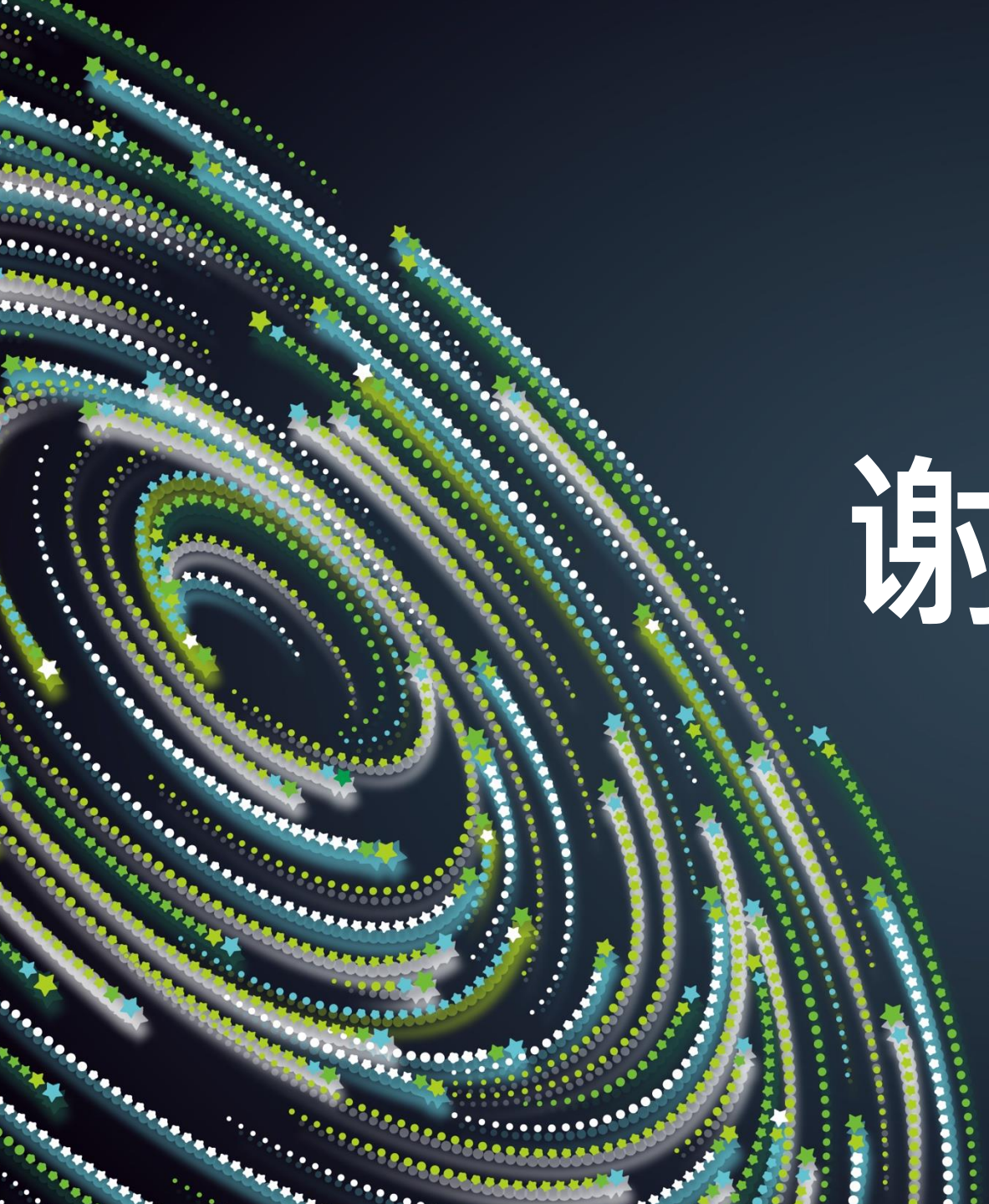
【本证书有效性依据发证机构的定期监督获得保持，证书有效性信息请登录可信云官网 www.kexinyun.org.cn 进行查询。由于该检验是针对颁证日期前产品情况进行披露，用户依然应留意在具体使用过程中可能发生的风险。因技术快速发展，产品/服务不断升级迭代，标准规范修订更新等原因，当产品或检验标准有较大变化时，建议进行重新检验。中国信通院将在 2027 年 3 月组织新一批检验。】

扫码查询证书

中国泰尔实验室
中国信息通信研究院

2026年03月27日

•功能可信　•收费可控　•权限可靠　•来源可溯　•能力可管



谢谢聆听