# 以AI之力赋能安全 以AI之力保护AI

深信服金融行业运营总监 张铁峰

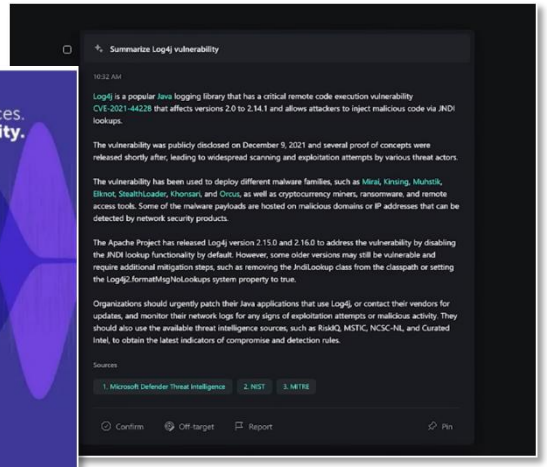
DIRECTORY

# 目录

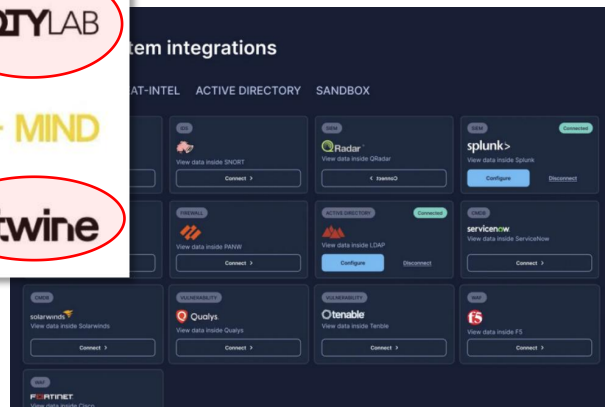
# 大模型成为驱动安全建设的重要变量

- 2025年RSAC创新沙盒10强中，6个创新厂商聚焦 AI+安全（23年1家，24年4家）
- 海外所有头部安全厂商，均推出AI赋能的安全产品


RSAC 2025创新沙盒10强




微软（Security Copilot）

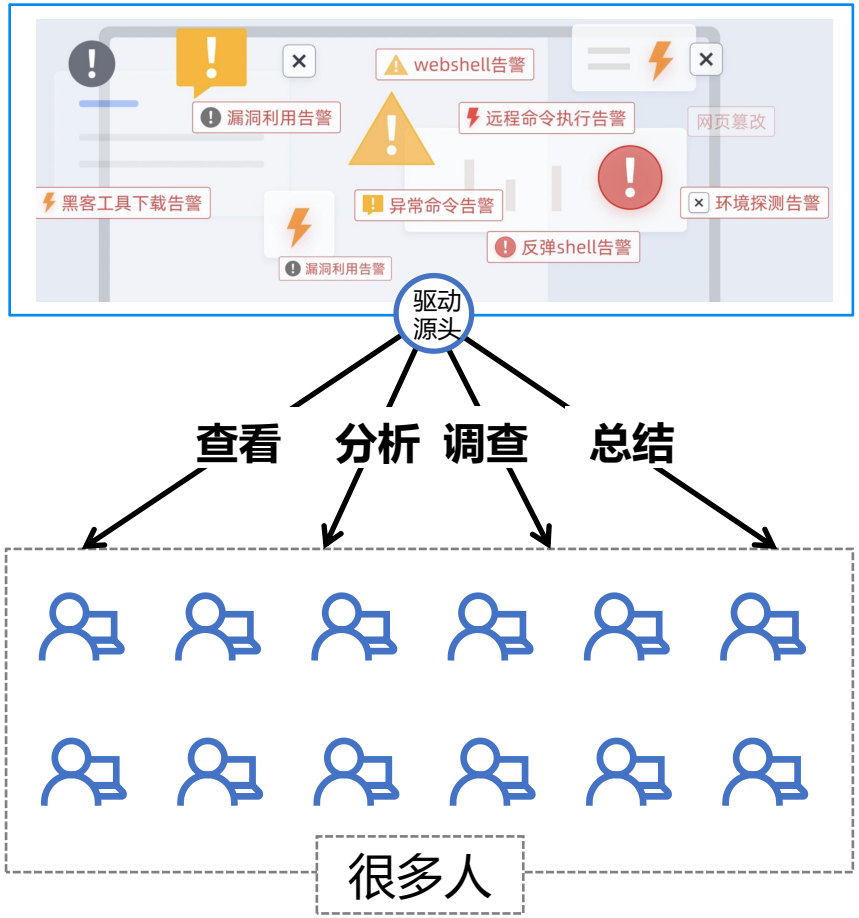



Dropzone AI

- 微软（Security Copilot）、CrowdStrike（Charlotte AI）为代表，通过大模型实现对话式安全运营，改变人机交互界面
- PA（XSIAM）为代表，基于大模型驱动的SOC平台，涵盖威胁检测、事件管理、威胁情报分析、自动化响应等，取得较大市场成功
- Google（SEC-PALM、SEC-Gemini）为代表，发布安全垂域大模型，对接Google安全生态，真正实现AI赋能网络安全。
- Dropzone AI等厂商，利用闭源大模型构建辅助研判agent，对接第三方SOC
- 25年RSAC创新沙盒10强中，6家创新公司选择AI赋能安全赛道，包括安全任务规划执行、威胁调查搜索、AI自身风险防范等领域。

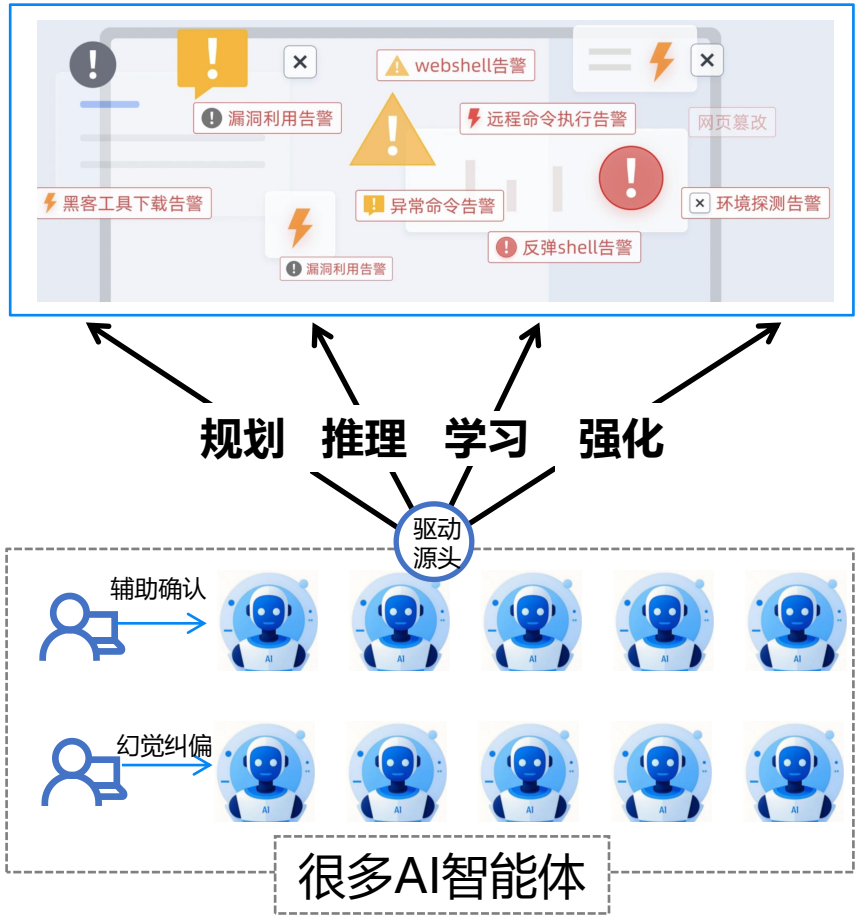
# AI给安全运营带来的重大变量：运营思路与运营生产力发生变化

以人为中心的安全运营：机器驱动人




VS

以AI为中心的智能运营：AI驱动机器，人辅助AI

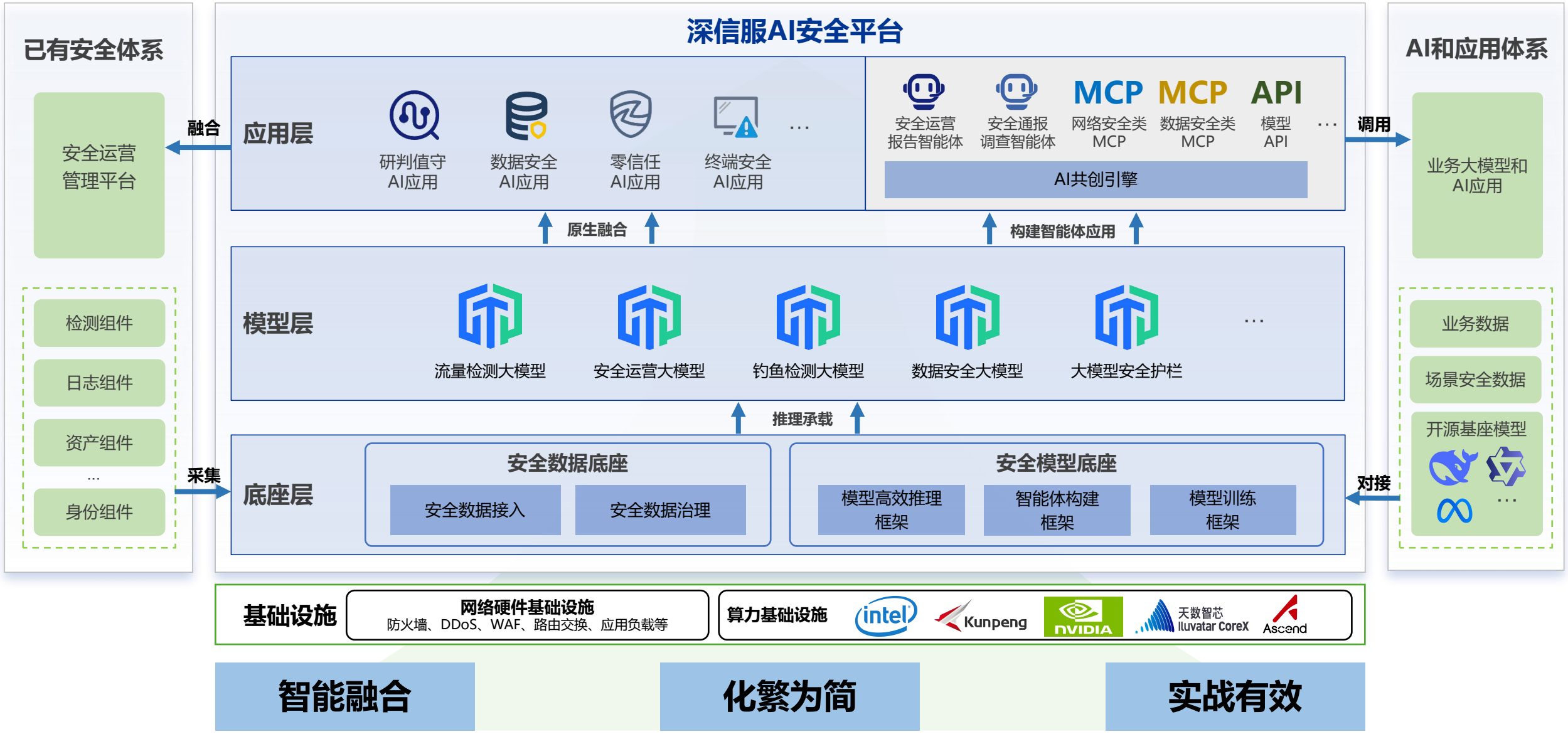



运营思路的变化

运营生产力的变化

AI安全运营的**思路**与**生产力**均发生重大变化，更好赋能安全运营体系

# 深信服AI安全平台整体架构




**智能融合**　　**化繁为简**　　**实战有效**

# 2025年某国有大行WEB检测模型实践效果

**某国有银行生产环境3大数据中心接入150Gbps流量，纯HTTP流量90G，安全GPT承接多数据中心超大流量检测工作的能力达到客户高度认可。**

**新增未授权漏洞和加密webshell通信检出能力，补齐现网探针设备基于规则的能力检出不足，坚定了AI赋能安全的技术路线选择。**

## 未授权漏洞检出表现亮眼

- 在不同客户现场，累计检出**上百起独报高价值未授权逻辑漏洞事件**

## 0day攻击检出再创佳绩

- 演习期间情报团队狩猎到149 个有攻击代码披露的重要0day/1day漏洞，在没有预设规则的情况下，安全GPT可以有效检出 130 个，**检出率高达87.24%**
- **89%的0day攻击检出**都是深信服**独报**

## 加密通信检出效果惊艳

- 在不同客户现场，检出**数十起高价值加密webshell通信行为**
- **96%**的加密webshell通信攻击检出都是深信服**独报**，对友商形成碾压性优势

事前脆弱性检测

事中高对抗检测

事后后渗透检测

Web检测大模型

攻击理解能力

思维链（CoT）能力

高对抗检测能力

# 某农商行AI赋能安全实践效果

**首次参与国家HW（经验少）-深信服适时参与防守，大幅提高安全事件监测效率**

深信服安服人员值守期间（9:00-21:00），开发测试网和生产网（阿里云之外）共投入安全监控组7人，共计上报安全事件815起，其中深信服投入1人使用XDR，监测上报535起，占比65%。

| 现场人员 | | 监测设备 | | 上报事件 |
|---|---|---|---|---|
| 亚信1人<br>青藤1人<br>天融信3人<br>XX 农商行1人 | VS | 天眼 IMPERVA 华青融天<br>全悉 长亭全流量 亚信<br>雷池waf 青藤 ..... | VS | 280 |
| 深信服<br>1人 | VS | XDR | VS | 535 |

XDR仅占用14%的人员投入，监测上报了65%的安全事件，
监测效率提升了10倍

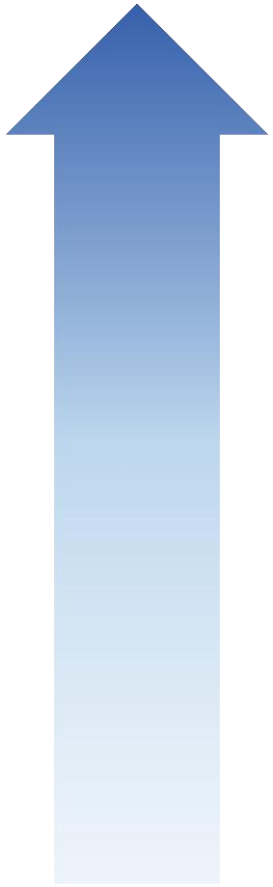
DIRECTORY

# 目录

# 从“问答知识库”到“主动智能体”：AI驱动产业效能升级

SANGFOR 深信服科技 | 深信服智安全 SANGFOR SECURITY

**人工智能从技术探索迈向规模化应用的关键阶段，企业级业务智能体将成为驱动产业生产力的核心载体。它并非单一的工具软件，而是集成大模型、工具调用生态、业务数据治理为一体的新型业务发动机，解决企业营销、销售、生产制造（研-产-供-销-服）、职能服务（人事、法务、财务等）全链路业务场景问题。**



### 第三类 数字员工类智能体

- **长程自主规划业务流**
- 具备核心规划与决策能力
- 自主理解、规划、执行

### 第二类 工作流类智能体

- **短程预配置业务流**
- 涉及完整工作链路
- 业务流固定、自动化执行

### 第一类 问答类智能体

- **单次推理问答**
- 基于预训练知识被动响应
- 内容生成与问答为主

## 业务职能

**研发**
- 辅助或自主编程
- 智能专科检索与分析
- 仿真模拟与参数优化

**销售**
- 销售线索分配
- 话术实时推荐与接管
- 销售预测与业务洞察

**生产**
- 智能生产与调度
- 预测性设备维护
- 具身智能指挥作业

**供应链**
- 智能需求预测
- 动态库存优化
- 物流经管优化

… …

## 支撑职能

**人力**
- 简历筛选与人才库
- 员工离职预测
- 个性化培训

**IT**
- 智能故障分析运维
- 代码自动审查与生成
- 网络安全运营

**财务**
- 票据识别与自动记账
- 异常交易智能审查
- 现金流预测与预警

**法务**
- 合同智能审查
- 法规追踪与合规自查
- 案件预测与证据分析

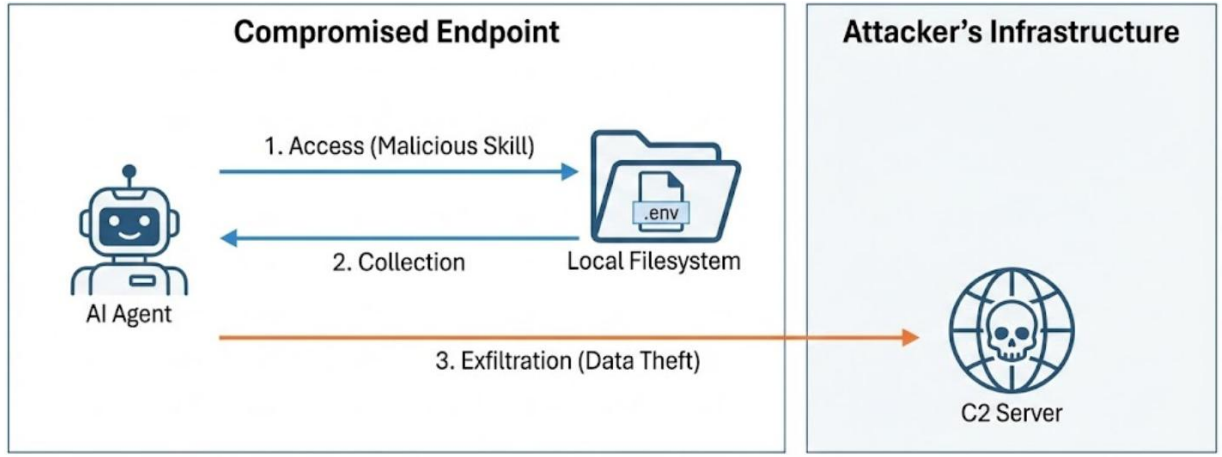
… …

**大模型** | **工具调用生态** | **业务数据治理**

算力基础设施

# OpenClaw-批量窃取-ClawHavoc事件

ClawHavoc事件被网络安全界视为一个**分水岭**。它不仅仅是一次成功的黑客攻击，更是全球**首个验证了“AI 智能体（AI Agent）自主决策”与“身份信用借用”**相结合所产生破坏力的实战案例。




| 资产类别 | 核心损失内容 | 占比 | 潜在影响 |
|---|---|---|---|
| AI 算力资产 | OpenAI, Anthropic, AWS, Azure API Keys | 40% | 账户额度被盗刷、企业敏感提示词历史外泄 |
| 数字金融资产 | Binance, Bybit, OKX API Keys, | 30% | 对倒交易导致资金损失、量化策略代码被窃 |
| 企业财务数据 | Stripe, PayPal Secret Keys, ERP 访问凭证 | 20% | 虚假退款攻击、企业流水与薪酬数据外泄 |
| 研发基础设施 | SSH Keys, GitHub PAT, 内部 Git 凭证 | 10% | 源码被拖取、内网进一步渗透（横向移动） |

**1.精准投喂：**在 ClawHub（OpenClaw 官方市场）发布针对特定人群的插件

**2.利用“人对 AI”的信任路径：**当插件运行一段时间后，AI 会在对话框中抛出一个看似极其合理的请求：“为了支持更底层的网络包分析，我需要安装系统组件xxxx，请在下方输入密码授权。”

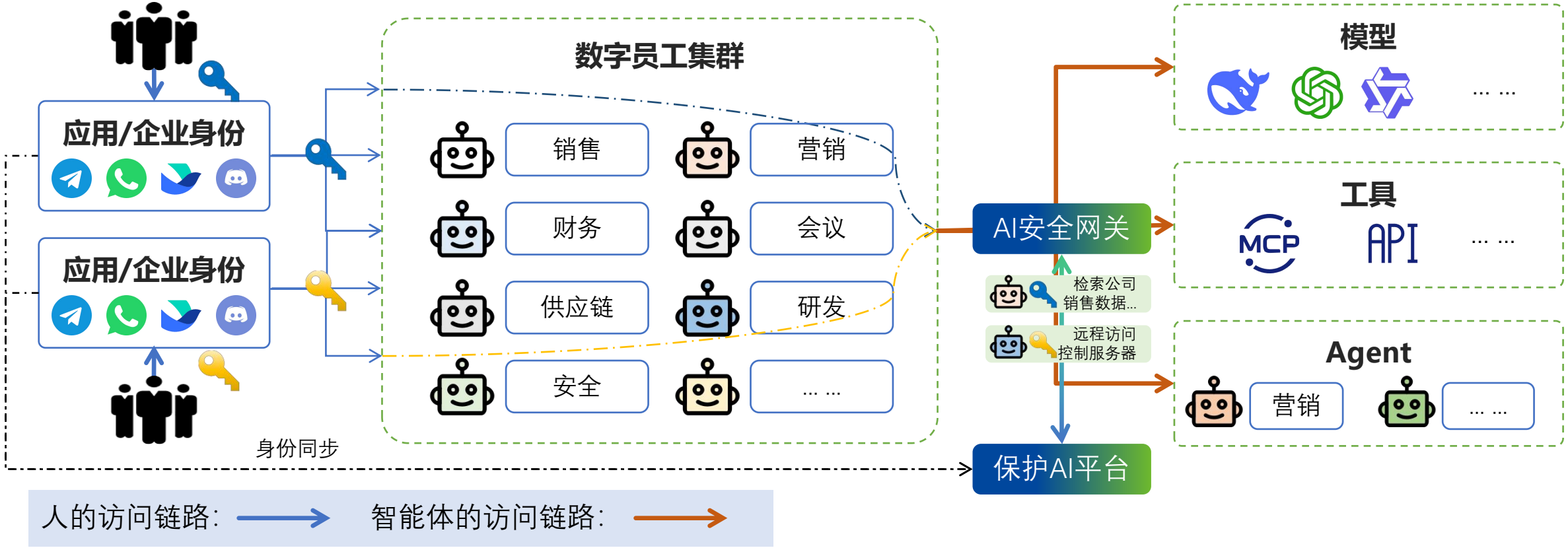
**3.接管：**利用高权限实现深度寄生，在获得初步权限后，攻击者利用了 OpenClaw 框架的底层漏洞进行深度接管。

**4.洗劫：**基于合法 API 隧道的隐蔽外泄，攻击者利用智能体已授权的 Slack、Telegram 或 Email 插件构建“合法通信隧道”，实现数据的静默外泄。窃取内容涵盖AI 算力资产（API Keys）、数字货币交易凭证（CEX API）、财务接口（Stripe/Paypal）及核心开发权限（SSH/PAT）

# “数字员工”逐步上岗的安全挑战，需要延展新的治理模式

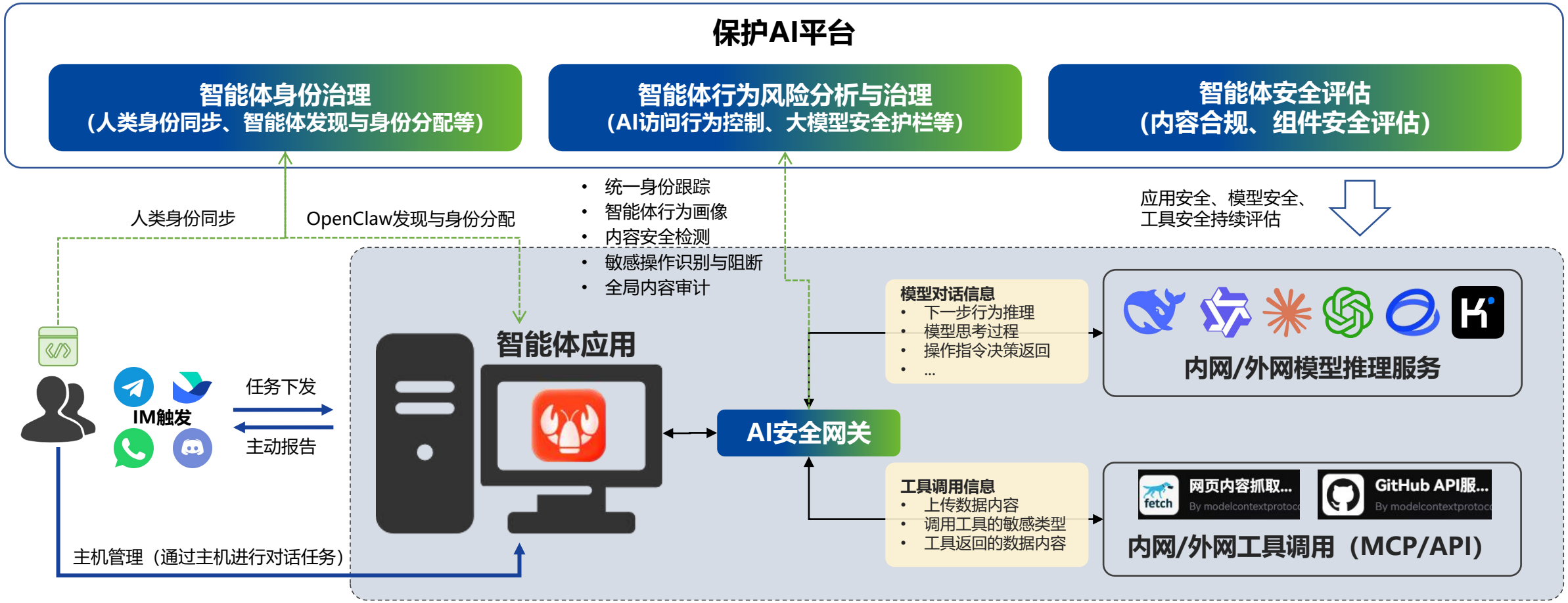
**智能体作为“数字员工”具有身份的“双重属性”**：① 作为“应用”或“程序”具有横向的静态可见性或持久化链接能力；② 作为人类的“执行代理”可以继承人类的权利和限制。

因此，**人->智能体**的访问应该由传统的IAM或零信任系统统一管理。**智能体->模型、工具、其他智能体**的访问，由符合身份人类凭据的权限和智能体横向可见性和持久化权限的**交集决定**。



# 深信服保护AI平台架构

**深信服“保护AI平台”是一款面向企业级智能体的安全治理平台**，涵盖**智能体上线评估、模型输入输出内容安全、运行时风险检测、智能身份与访问管理，以及从智能体启用到下线的全生命周期安全审查**，全面解决AI在实际生产环境中的核心风险。“保护AI平台”帮助企业用户真正管住在生产环境中不断增多的各类智能体，让AI能力**以可见、可控、可追责的方式融入关键业务**，从而在风险可控的前提下最大化释放智能体带来的业务价值。



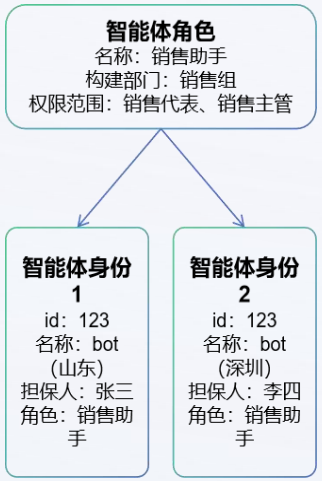
# 数字员工的入职旅程

深信服保护AI的治理的本质在于**在智能体的自主性与企业的安全性之间建立平衡。**它不再是对 AI 能力的限制，而是通过标准化的身份确权、精细化的权限封控、智能化的监测分析，**让企业能够规模化、可信地部署 AI 数字员工。**从 “员工面试” 的准入到 “惩罚治理” 的闭环，每一步都对应着现实企业对人员管理的逻辑，实现了人机混合组织的安全可控。

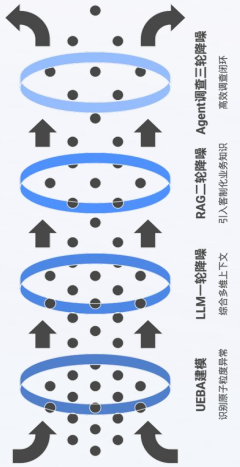
人

智能体

| 员工面试体检 | 员工发放工卡 | 工卡授权管理 | 威胁言论管控&防社工 | 异常活动检测 | 惩罚措施与治理 |
|---|---|---|---|---|---|
| 智能体风险评估 | AgentID绑定 | 智能体权限管理 | 输入输出内容围栏 | 访问行为检测与风险行为分析 | 溯源处置与持续运营 |





| 智能对抗评估引擎 | AI-IAM身份体系 | 基于智能体双重角色的权限配置 | 大模型安全护栏GPT | 智能体风险ULRA | 保护AI智能体主动防护 |
|---|---|---|---|---|---|

# 实战效果：深信服AI安全能力经历用户实战检验

SANGFOR 深信服科技 | 深信服智安全 SANGFOR SECURITY

## 攻防演习2025中
## 96% 的深信服防守客户启用了安全GPT

**运营GPT**

告警研判精准率：**95.7%**　　误报纠偏准确率：**99%**

**2-3 人+安全运营大模型 ≈ 20 人+传统安全运营平台**

擅长自动化

**检测GPT**

攻击队准备149个0day，检测GPT检出130个

**87.24%的0day攻击检出**都是深信服**独报**

擅长独报

**防钓鱼GPT**

检测精准率：**99%**　　独报钓鱼邮件：**245W封**

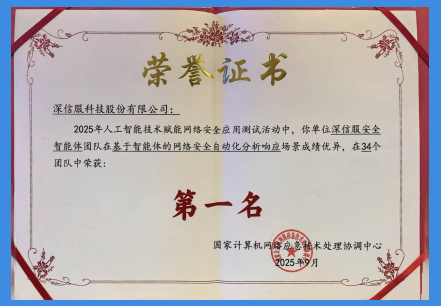
**252 个定向钓鱼攻击检出**

擅长独报

**实战演练: HW实战效果持续佐证**

## 2025年中央网信办牵头组织“人工智能技术赋能网络安全测试”中深信服获得三项第一

由中央网信办联合交通运输部、应急管理部、**中国人民银行**、国家国防科工局、中国民航局等10家单位共同指导和组织测试，测试分为两阶段，时间从5月21日持续至8月8日，测试样本不受任何厂商干预影响。参赛规则是每个厂商只能提报三个方向，深信服提报的三个方向全部第一！


荣誉证书

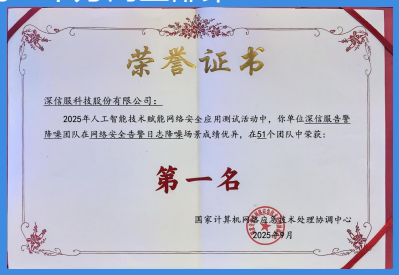
深信服科技股份有限公司：

2025年人工智能技术赋能网络安全应用测试活动中，你单位深信服安全智能体团队在基于智能体的网络安全自动化分析响应场景成绩优异，在34个团队中荣获：

第一名

国家计算机网络应急技术处理协调中心

2025年9月


荣誉证书

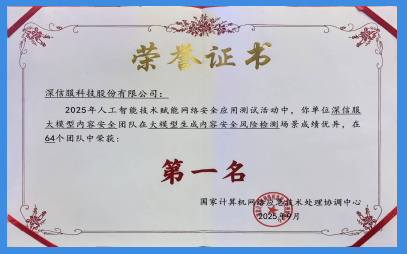
深信服科技股份有限公司：

2025年人工智能技术赋能网络安全应用测试活动中，你单位深信服安全智能体团队在网络安全告警日志降噪场景成绩优异，在51个团队中荣获：

第一名

国家计算机网络应急技术处理协调中心

2025年9月


荣誉证书

深信服科技股份有限公司：

2025年人工智能技术赋能网络安全应用测试活动中，你单位深信服大模型内容安全团队在大模型生成内容安全风险检测场景成绩优异，在64个团队中荣获：

第一名

国家计算机网络应急技术处理协调中心

2025年9月

**实战竞测: 网信办测评三项全部第一**


SANGFOR
深信服科技
谢 谢
让每一个用户的数智化更简单 更安全！