《AI 重塑网络安全：网络安全智能化产品与市场报告》

从安全助手到安全智能体：能力边界、应用路径与代表厂商

数说安全

2026 年 4 月

目 录

报告概览

一、报告范围与定位

本报告聚焦于 **AI for Security**（AI 赋能网络安全）领域，研究人工智能技术如何提升网络安全防御、检测、响应和治理能力。报告**不涵盖** Security for AI（AI 安全）相关内容，即不讨论大模型自身的安全风险、对抗攻击、提示注入等议题。

本报告围绕五个核心研究问题展开：一是 AI 技术如何重塑安全运营（SecOps）工作流；二是从安全助手（Copilot）到安全智能体（Agent）的技术演进路径是什么；三是产业界和学术界在 AI for Security 领域的最新进展与差距在哪里；四是不同应用场景下 AI 能力的边界与局限性如何；五是代表性厂商的技术路线、产品形态与商业模式有何异同。

二、目标读者

本报告面向四类专业人群。一是 **CISO/安全负责人**，为战略决策提供 AI 技术趋势判断、投资优先级建议、供应商评估框架。二是 **SecOps 团队**，为一线安全运营人员提供 AI 工具选型指南、工作流改造参考、技能转型方向。三是**产品负责人**，为安全产品开发团队提供 AI 能力集成路线图、功能设计参考、差异化定位思路。四是**投研分析师**，为投资机构和战略研究团队提供市场规模预测、竞争格局分析、技术成熟度评估。

三、相对 2024 版的核心增量

本报告是 2024 年版报告的全面升级，主要增量包括四个方面。

一是**新增学术研究章节。**系统梳理 2023-2026 年 AI for Security 领域的顶会论文（NDSS、S&P、USENIX Security 等），对比学术界与产业界的技术路线差异，识别学术成果向产品转化的典型路径与障碍。

二是**海外厂商覆盖扩展。**从 2024 版的 10 家扩展至 17 家国际代表厂商，新增欧洲、以色列新兴安全 AI 创业公司，深度分析微软、Google、CrowdStrike 等大厂 AI 战略调整。

三是 **Agent 化趋势深入研究。**涵盖从 Copilot（辅助决策）到 Agent（自主执行）的能力演进、多 Agent 协作框架在 SOC 场景的应用案例，以及人机协同的新型工作流设计模式。

四是**评测与治理新增专题。**包括 AI 安全产品的效果评测方法学、幻觉/误报/责任归属等治理挑战，以及国内外监管政策对 AI 应用的影响。

四、报告结构导览

本报告共包含 **12 个正文章节**和 **1 个附录。**

章节	**标题**	**核心内容**
Ch0	概览	报告范围、读者、方法、结构
Ch1	关键发现	十三项核心发现，涵盖市场、技术、产业格局
Ch2	行动建议	面向甲方、乙方、产学研的可操作建议
Ch3	战略假设	2026-2028 技术/市场/监管/人才/数据假设

Ch4	市场定义	AI for Security 边界、子市场切分、与既有产品关系
Ch5	市场描述	历史脉络（专家系统→大模型）、工程闭环视角
Ch6	市场走向	趋势、驱动力、阻碍因素、投融资动态
Ch7	学术研究前沿	9 大方向 100+篇顶会论文综述、产业启示
Ch8	常见能力清单	Must-Have/Standard/Advanced 分级、决策树
Ch9	差异化能力与选型评估	8 维度评估框架、40 个 RFI/RFP 问题
Ch10	代表厂商	国内 27 家+海外 17 家，共 44 家厂商画像
Ch11	市场落地建议	五层架构、PoC→运营化→规模化路线图
Ch12	案例	案例覆盖多大场景，量化收益数据（待补充）
附录	术语表、数据源、参考文献	30+术语中英对照、调研方法说明

五、调研方法与数据来源

本报告采用混合研究方法（Mixed Methods），结合定量与定性数据，数据来源分为四个维度。

一是国内企业调研。问卷调查方面，66 家国内安全厂商完成结构化问卷（数说安全调研口径），涵盖公司类型、客户行业、产品形态、AI 整体成熟度、技术路线、客户驱动力、趋势方向、主要阻碍、量化收益、审批审计、部署方式、未来规划等 60 余个维度。深度访谈方面，先后完成 11 场头部厂商深度 Briefing（各 1.5-2 小时，多数为问卷样本企业的同步深度交流），包括技术演示、路线图讨论、案例分享。综合画像累计覆盖国内代表性厂商约 70 家（以问卷样本为主，补充公开材料与公司发布会口径）。

二是**国际公开资料。**包括 17 家海外厂商的产品文档、技术白皮书、财报电话会议，Gartner、Forrester、IDC 等分析机构报告，以及厂商官方博客和技术峰会演讲视频。

三是**学术论文研究。**覆盖 2023-2026 年 NDSS、IEEE S&P、USENIX Security、CCS、RAID 等顶会论文和 arXiv 预印本平台 AI for Security 方向论文，以及与 3 位高校安全实验室 PI 的邮件交流。

四是**用户侧验证。**包括 5 家企业 CISO/SOC 负责人的非正式访谈、安全社区舆情分析，以及公开的 POC 测试报告和用户评价。

数据采集时间窗口：问卷与访谈为 2025 年 11 月至 2026 年 3 月，文献检索截止至 2026 年 2 月，市场数据引用 2025Q4 最新数据。

六、阅读指南

快速阅读路径（30 分钟）：Ch1（关键发现）→ Ch2（行动建议）→ Ch10（代表厂商，浏览）。

技术人员路径（2-3 小时）：Ch4（市场定义）→ Ch5（市场描述）→ Ch7（学术研究前沿）→ Ch8（能力清单）→ Ch9（选型评估）。

管理者路径（1.5-2 小时）：Ch1（关键发现）→ Ch3（战略假设）→ Ch6（市场走向）→ Ch10（代表厂商）→ Ch11（落地建议）→ Ch2（行动建议）。

投资者路径（1.5 小时）：Ch1（关键发现）→ Ch3（战略假设）→ Ch6（市场走向）→ Ch10（代表厂商）→ Ch12（案例）。

七、报告更新计划

本报告计划年度更新，2027 年 2 月发布 2027 版持续追踪技术与市场变化。同时通过公众号/知识星球发布季度观察笔记。欢迎读者通过邮件/Issue 提供厂商补充、案例分享、纠错建议。

报告版本：v1.0

发布日期：2026 年 4 月

作者单位：数说安全研究院（北京赛博英杰科技有限公司旗下）

联系方式：ssaq@geniuscybertech.com

第一章：关键发现

1. AI 大模型公司对网络安全公司造成压力

2026 年 2 月 20 日，Anthropic 正式发布 Claude Code Security，以推理驱动的代码漏洞扫描能力直接进入代码安全市场。消息落地当日，CrowdStrike（CRWD）股价下跌约

8%，Cloudflare 下跌 8.1%，Zscaler 下跌 5.5%，SailPoint 下跌 9.4%，Okta 下跌 9.2%，Global X Cybersecurity ETF 下跌约 9%。仅隔 14 天，OpenAI 跟进发布 Codex Security。全球前两大 AI 实验室在半个月内先后进入代码安全这一传统安全厂商的核心业务领域，这在行业历史上尚属首次。2026 年 3 月下旬，Anthropic 内部文档因 CMS 配置错误意外泄露，曝光了其更下一代模型 Claude Mythos 的能力边界——文件显示 Mythos 被描述为"迄今最强的网络安全模型"，能够自动发现主流安全产品未能检测到的 0day 漏洞利用路径，并可直接对抗 EDR、SIEM 等传统安全产品的检测逻辑。消息传出当日，CrowdStrike 再度下跌约 7%，Palo Alto Networks 下跌 6%，Zscaler 下跌 4.5%，iShares Cybersecurity ETF 下跌 4.5%。

传统 SAST（静态应用安全测试）工具依赖预定义规则库，对跨文件业务逻辑漏洞和需要深层推理的复杂路径漏洞几乎无能为力——这类漏洞正是真实攻击中最常见的入口。推理驱动的大模型可以理解跨组件语义关联，自主生成假设并验证漏洞利用路径，在检测能力维度上形成对传统工具的降维打击。然而，独立测评机构 Checkmarx Zero 的评估同时指出 Claude Code Security 的实际准确率存在争议（8 个发现中仅 2 个为真阳性），Snyk 则提出"检测易、修复难"的协同路线——传统安全情报数据库与 AI 推理能力结合才能形成完整修复闭环。这意味着市场的短期反应（股价下跌）蕴含了对"AI 替代"的过度定价，但中长期（12-24 个月）的结构性替代压力是真实存在的。

对国内安全厂商而言，这一事件的影响是双向的。一方面，直接冲击代码安全赛道的传统 SAST 玩家，海云安 AI 白盒、长亭慧鉴、云起无垠等产品需要在 12-24 个月内明确 AI 战略，否则面临被 AI 原生工具取代的压力。另一方面，悬镜安全、安恒信息已敏锐捕捉到

这一机会——悬镜以 SAST 误报率降至 2%以下为卖点，并发布多模态 AIST 平台；安恒信息以"中国版 Cloud Code"为定位加速布局。业内出现的"传统 SAST 产品只有 1-2 年生存周期"判断，虽有警示价值，但也存在一定的舆论放大成分——实际替代进程取决于私有化部署效果、客户切换成本和监管合规要求的综合约束。

这一事件的产业意义远超一款新产品的发布。首先，它验证了通用大模型的推理能力已足以胜任高度专业化的安全分析任务，无需漫长的垂域专项训练。其次，它以资本市场的剧烈反应证明，即便是头部传统安全厂商，也无法对 AI 颠覆免疫。第三，它为国内安全厂商设定了明确的能力参照系——本轮调研中，长亭科技和悬镜安全均将 Claude Code Security 作为代码安全能力的对标基准，国内最优模型的差距被量化为约 20%。

2. 市场进入爆发增长期，年均增速约 19-24%

AI 驱动的安全运营市场正经历前所未有的增长。据 Grand View Research 估算，2024 年全球 AI 网络安全市场规模约 253.5 亿美元，预计 2030 年达到 937.5 亿美元，CAGR 为 24.4%；Fortune Business Insights 的预测更为乐观，2025 年市场规模为 340.9 亿美元，CAGR 约 21.7%；Precedence Research 则给出 2025 年 296.4 亿美元、CAGR 18.9%的预测。综合多家机构数据，**2025 年全球市场规模约在 250-340 亿美元区间，年均增长率约 19-24%。**这一增速远超传统网络安全市场的平均增长水平，标志着 AI for Security 从概念验证阶段进入规模化商用阶段。

投融资数据进一步印证了市场热度。7AI（Agentic Security 方向）从出隐身到$7 亿估值仅用 10 个月，融资总额达$1.66 亿；Dropzone AI 实现 11 倍 ARR 增长。更广泛地看，网络安全行业整体进入大额并购期——Google 以$320 亿收购 Wiz（云安全）、Palo Alto 以

$250 亿收购 CyberArk（身份安全）、Cisco 以$280 亿整合 Splunk（安全运营），虽然这些交易并非都直接围绕 AI，但安全平台的整合为 AI 能力的落地提供了数据和场景基础。

3. AI SOC / Agentic SOC 已进入美国网络安全运营厂商的主流产品路线图

多家分析机构与厂商口径均预计，2026—2028 年 AI/Agent 能力在 SOC 的渗透将显著提升。其中 Gartner 在公开沟通中给出的可核验口径包括：到 2028 年至少 15%的日常工作决策将由 agentic AI 自主完成、约 33%的企业软件应用将内嵌 agentic AI、超过 50%的企业将部署 AI security platforms 以保护其 AI 投资。需要说明的是，业内常被引用的"AI SOC Agent 渗透率从 2025 年 5%升至 2028 年 70%"这一具体数字目前缺乏可公开核验的原始出处，公开口径之间差异较大，本报告不将其作为定量基准，仅作为方向性参考。这一预测基于两个核心判断：一是 AI 技术在安全场景的成熟度快速提升，二是企业面对告警疲劳和人力短缺的刚性需求无法通过传统方式解决。

2026 年 4 月 RSAC 大会进一步印证了这一方向：AI SOC / Agentic SOC 已从差异化亮点演进为美国网络安全运营产品提供商的主流产品路线图与头部厂商共识。从展会现场看，几乎所有主流 SOC 产品厂商均在展台核心位置展示 AI SOC / Agentic SOC 能力——Palo Alto Networks（Cortex XSIAM 的 Agentic SOC 演进）、CrowdStrike（Charlotte AI + Agentic SOC Workflows）、Microsoft（Security Copilot + Sentinel 的 Agentic 能力）、SentinelOne（Purple AI Athena）、Torq（HyperSOC 2o 的 MCP 原生编排）、Splunk（Cisco 整合后的 Agentic 路线）、Exabeam / LogRhythm / Securonix 等 SIEM 厂商均将 AI Agent 列为下一代产品的主推方向。与此同时，Dropzone AI、7AI、Prophet Security、Radiant Security、Simbian、Intezer 等新兴专项厂商在本届 RSAC 期间批量发布或升级 Multi-Agent

SOC 平台。需要指出的是，主流厂商集中发布、方向基本成型，并不等同于"已成为事实标配"——Gartner 自身的公开口径同时提示，相当比例的 agentic AI 项目仍可能因 ROI、治理与可靠性问题被取消或回撤，行业整体仍处于试错期。本报告将这一阶段刻画为"头部厂商共识 + 主流产品路线图 + 客户加速选型期"，而非"已落定的行业标准"。

这一趋势将深刻改变安全行业的人才结构、产品形态和服务模式。从"先锋实验"向"主流产品路线图"的演进，意味着 AI SOC / Agentic SOC 正在成为头部厂商不可回避的能力建设方向，缺位该能力的 SOC 类产品将在客户选型中承受越来越大的压力——但行业最终的渗透率与节奏，仍取决于 2026—2028 年间数据质量、治理框架与 ROI 验证的成熟度。

4. 头部加速分化，中腰部仍在追赶

对 66 家国内网络安全企业的结构化问卷调研（含 11 场深度 Briefing）显示，头部厂商与中腰部厂商之间的分化正在加速扩大，市场呈现出"两个世界"并行的格局。

一是头部厂商已全面进入产品化乃至规模商业化阶段。安恒信息 2025 年 AI 相关收入超过 2 亿元、订阅客户逾 1000 个；启明星辰交付 AI 项目 35 个、AI 版客单价从 20 万元翻倍至 40-50 万元；亚信安全 AIXDR 联动防御系统在一年内落地 60+客户；深信服 MSS 覆盖 3500 家企业客户，正在将 T1 研判员从 50 人削减至 5 人，并已斩获 8 个海外百万美金大单。这些数据彻底打破了"AI for Security 叫好不叫座"的刻板印象。

二是中腰部厂商仍以功能嵌入和试点验证为主。以华云安公开披露的口径为例，其政企客户群体中绝大多数要求纯内网部署（厂商自述约 95%，样本范围限于该公司客户），AI 以 Copilot 形态嵌入现有产品，尚无独立 AI 产品线；更多中小安全厂商仍停留在将大模型接入已有工具的阶段，缺乏系统性的 AI 战略。

三是造成分化的核心是数据底座、工程化能力和客户生态的差异，而非模型本身。头部厂商普遍具备 15 年以上的安全运营数据积累（深信服反馈），或在垂直客群中形成了深度绑定（启明星辰与中国移动、亚信安全与运营商生态）。这种结构性优势短期内难以被追赶者复制。这种分化格局预计将在 2026-2027 年进一步固化，形成"强者愈强"的马太效应。

5. 从 Copilot 到 Agent：正在发生的范式革命

2025-2026 年最重大的变化是 AI 安全工具从"辅助式 Copilot"向"自主式 Agent"的范式转变。Copilot 模式下，AI 充当分析师的"智能助手"，提供建议但决策权在人；Agent 模式下，AI 可以在明确边界内自主执行调查、响应、修复等完整工作流，人类角色转变为监督者和策略制定者。

这一转变的技术支撑是多 Agent 架构的成熟。学术界已发表 100+篇顶会论文探索多 Agent 协同机制，SecBench 等评测基准的涌现为 Agent 能力提供了客观度量标准。海外领先企业如 Microsoft、CrowdStrike 以及 Dropzone AI 等新兴公司已推出成熟的 Agent 产品。国内企业中，深信服明确提出"人是 AI 的一个环节"的激进理念——不是 AI 辅助人，而是人成为 AI 工作流的审核环节；知道创宇的渗透测试团队要求"每个工程师必须使用 AI，否则跟不上进度"。这代表了行业最前沿的人机关系重构。

6. 告警疲劳、人力缺口、MTTR 是驱动 AI 应用的三大刚需

本轮 66 家受访企业的问卷反馈显示，约四分之三受访企业（50 家/66 家）明确提到"告警疲劳"是推动 AI 应用的首要驱动力。现代安全工具每天产生数千乃至数万条告警，但真正需要人工处理的高危事件可能不足 1%，分析师淹没在噪音中无法聚焦真正威胁。约半数

受访企业（35 家/66 家）反馈面临严重的人力缺口，安全人才的供需失衡在短期内无解。约六成受访企业（40 家/66 家）将压缩平均响应时间（MTTR）作为核心 KPI，传统人工流程难以满足快速响应要求。

这三大刚需构成了 AI for Security 的"铁三角"需求基础。任何 AI 安全产品如果无法在这三个维度之一产生可量化价值，将难以获得客户买单。

7. 量化收益已经显现：头部客户实现工时下降 50%-83%

尽管整体产品化率不高，但先行者已开始报告显著的量化收益。根据厂商访谈与公司发布会口径，5 家受访企业自述实现工时下降超过 50%，其中知道创宇披露通过 AI Agent 将安全分析工时降低 83%，美创科技公开披露其 AI 风险监测智能体在已投产节点的检出率达到 95%。上述均为厂商自述/客户案例口径，尚缺独立第三方审计数据，但仍可作为方向性证据：在数据质量可控、场景边界清晰的条件下，AI 已具备实质性替代人工重复劳动的潜力。

然而需要注意的是，这些成功案例多集中在头部企业或特定场景（如告警降噪、漏洞检测），尚未实现全场景、全流程的覆盖。中小企业由于数据积累不足、工程能力受限，难以复制这些成功经验，这也是中腰部厂商仍在追赶的重要原因。

8. 数据质量是最大阻碍，私有化与云端效果"差一个数量级"

66 家受访企业中，约六成（40 家）将数据可用性与质量列为 AI 应用的最大阻碍。安全数据普遍存在标注不足、格式不统一、噪音比高的问题，而大模型的效果严重依赖训练和微调数据的质量。火山引擎在访谈中的描述具有代表性：私有化部署与云端大模型的效

果差距"不止一个数量级"。云端模型拥有海量数据训练和持续迭代能力，而私有化部署受限于单一客户的数据规模和更新频率，效果大打折扣。

这一现实迫使厂商和客户重新思考部署模式。对于数据敏感度极高的金融、政务等行业，私有化是刚性要求，但需要接受效果妥协；对于数据敏感度较低的场景，云端服务或混合部署可能是更优选择。约三分之一受访企业（24 家/66 家）提到评测体系缺失，也与数据质量问题密切相关——没有高质量的测试集，就无法客观评估模型能力。

9. 技术路线分化：自训大模型收益递减，工程化壁垒崛起

技术路线调研显示，开源大模型+RAG（检索增强生成）+规则融合已成为主流选择，仅烽台科技和绿盟科技等少数企业选择自研大模型。这一分化反映了厂商对成本、效果和时间窗口的不同权衡。

然而，本轮多场深度 Briefing 揭示了一个更重要的方向性转变：**安全垂域大模型微调的边际收益正在快速递减**，这一判断已在受访厂商中形成高度共识。启明星辰坦言"现在调用第三方有时候更好"；深信服明确表示 32B 以上模型不再微调，直接使用开源基座；长亭科技的战略判断是"只做小/中模型，不做 100B 以上"；360 保留 14B 自训模型仅用于精度敏感场景，通用任务全面转向开源。开源基座模型（千问、DeepSeek 系列）的通用能力已足以覆盖大部分场景，进一步专项微调的回报已不足以支撑投入。

这一共识指向一个关键的竞争格局转变：竞争壁垒正从"谁的模型更强"转向"谁的工程化更扎实"。具体而言，一是语料生产平台的质量，深信服 15 年安全运营积累的语料库构成实质壁垒；二是数据底座的规模，360 百 PB 级安全数据在垂域精调场景仍不可替代；三

是工具链的 MCP 化程度，360、深信服等已率先将核心能力 API/MCP 化，快速响应行业接口标准化的趋势。

值得强调的是，在威胁检测这一核心场景中，小模型（0.1B-0.85B 参数级别）并非"退而求其次"的选择，而是技术刚需。实时流量分析、终端检测等场景要求毫秒级延迟和数万 QPS 吞吐量，大模型根本无法满足。无论大厂还是创业公司，都需要构建"分层模型架构"：小模型做实时检测、中等模型做告警研判、大模型做交互分析。

10. 竞争格局重构：头部格局扩容，国内外差距依然显著

本轮调研后，国内头部梯队已从"三家领先"扩容至更丰富的竞争格局。**第一梯队**（深度验证+规模商业化）包括深信服、360 数字安全、安恒信息、启明星辰四家，均已在 AI 营收规模、客户数量或运营效率方面取得可量化的规模化成果。**第二梯队**（技术领先但规模待验证）涵盖绿盟科技、知道创宇、海云安、长亭科技、青藤云安全等，在特定赛道或技术维度具备差异化优势。**第三梯队**（细分赛道专精）包括悬镜安全（AI 原生安全）、亚信安全（智能体安全）、未来智安（XDR+MCP 原生）等具备赛道先发优势的专精型企业。

启明星辰以"1+1+N"架构（统一大模型+AIDK 智能体框架+N 条产线智能体）推进全产线 AI 化，AI 版客单价翻倍、中国移动战略级绑定构成独特壁垒。亚信安全依托并购亚信科技形成的"安全+数智+连接"三合一战略，以及在成都世运会实战验证的 MTDR 从小时级压缩至 1 分钟的效果，正在构建运营商生态内的差异化优势。

然而，与海外巨头 Microsoft、Google、CrowdStrike 相比，国内企业在技术深度、生态整合和全球化能力上仍有明显差距。长亭科技、悬镜安全在代码安全场景的独立测试均量

化了这一差距：国内最优模型与 Claude 的能力差距约为 20%，私有化场景下差距则更大。智谱唐杰将这一差距描述为"至少半个代差"。海外市场"巨头平台化+创业公司场景化"的双轨竞争格局在国内尚未形成，这既是挑战也是战略窗口。

11. 人才与评测双缺失，成为产业化瓶颈

66 家受访企业中，约三分之一（21 家）明确提到人才短缺是核心阻碍，安全+AI 的复合型人才极度稀缺。传统安全人才缺乏 AI 工程能力，AI 人才又不了解安全场景和威胁机理，这种结构性错配短期内难以通过市场自发调节解决。值得注意的是，国内安全行业面临的并非简单的"总量缺口"，而是结构性错配——网安专业大规模扩招后，初级安全人才就业已开始出现困难，但 AI+安全的高级复合型人才仍一将难求。

同样有 36%的企业提到评测体系缺失，不同厂商对"AI 安全能力"的定义和度量标准各异，客户难以客观比较和选型。SecBench 等学术评测基准的出现是积极信号，但尚未形成行业统一标准。缺乏标准化评测，会导致"演示型成功"横行——厂商针对特定演示场景过度优化，实际部署效果大打折扣。建立覆盖检测、调查、响应、预测等全流程的标准化评测体系，是产业从混战走向成熟的必经之路。

12. MCP/A2A 协议与 OpenClaw 等运行时协同推动安全产品开放化

本轮多场 Briefing 中，有 5 家厂商明确披露了 MCP/A2A 协议化或 OpenClaw 等运行时集成战略，且态度惊人一致：不将其视为外部威胁，而是主动将自身核心能力以 MCP 协议对外开放，成为可被调用的开放能力节点。这一现象标志着安全行业正在经历从"封闭平台竞争"向"开放能力生态"的结构性转型。

各厂商的布局路径各有侧重。360 的策略最为清晰——将所有安全检测、情报、响应能力全面 MCP 化，中石油和上海公安已通过 OpenClaw 直接调用 360 的安全能力，这代表了"能力即服务"的最前沿实践；深信服将检测大模型和运营大模型全部 API/MCP 化，与主流开发环境集成；启明星辰的 AIDK 框架同时支持 MCP 协议和 A2A 协议，强制全产线统一接口标准；安恒信息已推动产品 MCP 化改造并建立 MCP 平台；华云安的智能体平台也已支持 MCP 协议。

这一趋势的深层逻辑是：随着大型企业的安全运营体系日益复杂，单一厂商的"全栈闭环"模式已无法满足客户对最优能力组合的需求。MCP 作为事实标准接口，使得客户可以像调用 API 一样灵活调用来自不同厂商的安全能力，安全产品的竞争维度从"谁的平台功能更全"转向"谁的专项能力更强、接口更标准"。对于中小安全厂商而言，MCP 化不是选择题而是生存题——成为开放生态中的专项能力提供者，可能比继续维护封闭平台更具商业可持续性。

13. 智能体安全（Security for AI Agent）正在成为独立赛道

随着 AI 智能体在企业内部的快速渗透，如何保护智能体本身的安全、如何防范智能体被滥用或劫持，正在从学术讨论演变为有真实产品竞争的独立市场赛道。本轮调研发现，多家厂商正从不同角度切入这一新兴赛道，但各自的技术深度和商业化进度差异显著。

亚信安全郑鑫提出的三层智能体安全框架代表了国内思考最深的系统性方案。框架的逻辑是：在生态互信层，通过类似 DNS 机制的智能体枢纽实现统一身份注册和发现；在组织治理层，解决人、应用、智能体三类主体的混合身份信任传递问题；在纵深防护层，以意图驱动访问控制（IBC）追踪智能体意图是否偏离初始任务——这实际上是将传统的零

信任理念延伸到了 AI Agent 的动态行为层面。深信服则从运营角度切入，将于 5 月底发布智能体身份治理和内容合规网关产品。悬镜安全以供应链安全为切入点，推出模型血缘图谱（识别模型微调和量化溯源）和 MCP/Skills 投毒检测能力。长亭科技则提供 Security for AI 评估服务，2026 年预计贡献 1000-2000 万营收。

然而，这一赛道面临早期市场特有的两重挑战。其一，需求端尚未形成购买共识，企业客户对"智能体安全"的理解和预算分配仍在摸索阶段，短期内难以支撑大规模商业化；其二，供给端已出现恶性价格竞争的苗头——长亭科技披露，字节、百度、蚂蚁等互联网大厂正以 20 万元报价抢占原本价值 300 万元的 Security for AI 评估项目，专业安全厂商的回报周期面临被严重拉长的风险。这一新兴赛道能否形成健康的商业生态，将在很大程度上取决于头部买方市场能否建立基于质量而非价格的采购决策机制。

核心洞察：AI for Security 市场已从"是否应用"全面转向"如何应用"，并在头部厂商中进一步演进至"如何规模化"。技术可行性已被多个头部案例充分验证，头部厂商正在建立难以追赶的工程化壁垒。与此同时，AI 巨头入局代码安全、MCP 开放化趋势和智能体安全赛道的出现，预示着市场竞争的边界和规则正在被重新定义。未来 24 个月将是产业格局定型的关键窗口期：能够同时完成规模化商业化、MCP 能力开放、以及 Security for AI 布局三件事的企业，才有望在下一轮竞争中占据制高点。

第二章：行动建议

基于对 AI for Security 市场现状与趋势的分析，我们针对甲方安全团队、乙方产品厂商、产学研机构三个不同视角，提出可操作的行动建议。

一、甲方视角：CISO 与 SecOps 团队的实践路径

1.1 分阶段试点路线图：从告警降噪开始的 90 天/180 天计划

第一阶段（0-90 天）：告警降噪与验证

不要试图一步到位构建 AI 驱动的完整 SOC。从最痛的点切入——告警降噪。选择单一告警源（如 SIEM、WAF 或 IDS）作为试点场景，目标是将告警数量降低 30-50%，同时保持零漏报。

具体步骤分为四个阶段：一是第 1-2 周盘点现有告警数据，统计过去 3 个月的告警量、误报率、平均处理时长；二是第 3-4 周选择 2-3 家供应商进行 POC 测试，要求在真实数据上运行而非演示数据；三是第 5-8 周小范围试运行，AI 处理结果由人工复核，建立"AI 判断-人工验证-反馈标注"闭环；四是第 9-12 周量化评估效果（告警降噪率、误报率、漏报率、分析师工时节省），决定是否扩大范围。

成功标志包括三方面：告警数量下降≥30%且未发生漏报事件；分析师平均每日处理告警时间下降≥2 小时；团队对 AI 判断的信任度达到"愿意自动化处理低风险告警"的程度。

第二阶段（90-180 天）：扩展场景与流程整合

在第一阶段成功的基础上，将 AI 能力扩展到 2-3 个相关场景，并开始整合到现有工作流。一是威胁狩猎辅助，利用 AI 从海量日志中发现异常模式，为威胁狩猎提供线索；二是

事件调查加速，将 AI 应用于安全事件的初步调查，自动收集关联日志、资产信息、历史事件；三是漏洞优先级排序，基于资产重要性、可利用性、威胁情报，AI 辅助确定漏洞修复优先级。

成功标志是 AI 应用场景从 1 个扩展到 3 个，MTTR（平均响应时间）下降≥20%，并建立起"AI 建议-人工决策-执行反馈"的标准化流程。

1.2 选型三大陷阱：如何避免"买家秀"与"卖家秀"的差距

陷阱一：演示型成功——完美演示背后的数据陷阱。 供应商演示时 AI 表现完美，但实际部署效果大打折扣，原因在于演示数据经过精心清洗和标注，而真实数据质量远低于此。规避策略：一是坚持要求在你的真实数据上进行 POC 测试，而非供应商提供的示例数据；二是测试数据应覆盖至少 3 个月跨度，包含正常流量和已知攻击案例；三是关注"边缘案例"处理能力——问供应商"遇到从未见过的攻击类型怎么办"；四是要求提供现有客户的量化指标（工时节省、误报率等），并核实其真实性。

陷阱二：全场景覆盖承诺——从检测到响应的"大而全"陷阱。 供应商承诺 AI 可以覆盖检测、分析、响应、预测全流程，但实际能力可能仅在某一环节成熟。规避策略：一是要求供应商明确列出哪些场景已产品化、哪些在试点、哪些是路线图；二是优先选择在 1-2 个场景做到极致的产品，而非什么都能做但都不精的产品；三是查看产品更新日志，判断研发重心在哪里——频繁更新的模块才是成熟模块；四是与现有客户沟通，了解他们实际在用哪些功能、哪些功能被搁置。

陷阱三：自研大模型噱头——"自主可控"背后的效果与成本权衡。 供应商宣称"自研大模型"以彰显技术实力和数据安全，但要看具体场景，实际效果未必如开源模型+微调。

规避策略：一是关注实际效果而非技术路线——不管黑猫白猫，抓到老鼠就是好猫；二是对比测试，要求供应商与主流开源模型在相同任务上 PK；三是评估私有化部署的真实成本（算力、存储、运维、升级迭代的 TCO）；四是如果选择私有化部署，确保合同中包含"效果不达标可切换云端服务"的条款。

1.3 数据准备先行：AI 应用成功的前置条件

在引入 AI 工具之前，先对自身数据进行健康检查。需要回答四个关键问题：一是数据完整性，日志收集覆盖率是否≥80%，关键资产的日志是否完整；二是数据标准化，不同来源的日志格式是否统一，时间戳是否同步；三是数据标注，是否有历史事件的标注数据，至少需要 100+标注案例作为训练集；四是数据可访问性，AI 系统能否实时访问所需数据，API 对接是否完善。

如果数据健康度低于 60 分，建议优先投入 3-6 个月进行数据治理，再引入 AI，否则"垃圾进，垃圾出"。快速启动路径为：第 1 个月集中解决日志收集完整性问题，确保关键资产 100%覆盖；第 2 个月建立数据标注机制，每周标注 10-20 个典型案例（攻击/误报/正常）；第 3 个月搭建数据湖或数据仓库，统一日志格式，建立快速查询能力。

1.4 权限审计不能省：AI Agent 的边界与红线

AI Agent 具备自主执行能力，意味着它可以"代表你"执行操作。这带来效率提升，也引入新的风险。必须建立四方面的权限控制机制：一是最小权限原则，AI Agent 仅能访问完成任务所需的最小数据和系统权限；二是操作白名单，明确列出 AI 可以自动执行的操作（如查询日志、生成报告）和必须人工审批的操作（如阻断 IP、隔离主机）；三是操作日

志与回溯，所有 AI 执行的操作必须留存日志，支持事后审计和回滚；四是异常行为监控，监控 AI 的行为模式，如突然大量查询、访问敏感数据等异常行为应触发告警。

推荐分阶段放权：第一阶段（0-90 天）AI 仅具备"建议权"，所有操作需人工确认；第二阶段（90-180 天）对低风险操作（如生成周报、基础告警分类）授予自动执行权限；第三阶段（180 天+）基于风险评分体系，AI 可在明确边界内自主响应。需要强调的红线是，修改生产系统配置、删除或修改原始日志数据、对外发送敏感数据、执行不可逆的封禁/隔离操作，这四类操作永远不能完全自动化（除非有强制人工复核机制）。

1.5 人才培养：从"安全分析师"到"AI 安全工程师"

AI 的引入不是为了裁员，而是让分析师从重复劳动中解放，聚焦高价值工作。但这需要技能升级。安全团队的技能升级分为三个阶段：初级阶段（AI 使用者），学会使用 AI 工具的基本功能，理解 AI 判断的逻辑并能识别误判，掌握 prompt 工程技巧；中级阶段（AI 调优者），学习基础的机器学习概念，能够参与 AI 模型的标注、测试、评估工作，理解 RAG、微调等技术；高级阶段（AI 架构师），设计 AI 与现有安全工具的集成架构，制定 AI 应用的策略、边界和评估体系，探索新的 AI 应用场景。

培养建议方面：一是每季度组织 1-2 次 AI 技术培训（外部专家或供应商）；二是建立"AI 创新小组"，鼓励团队成员尝试新工具和场景；三是将"AI 工具熟练度"纳入绩效考核，激励主动学习；四是与高校或培训机构合作，定制"安全+AI"复合型课程。

二、乙方视角：产品厂商的战略选择

2.1 产品进化路线：从 Copilot 到 Agent 的四个阶段

AI 安全产品的演进呈现清晰的四阶段路径。

阶段 1：智能助手（Copilot 1.0）。 核心能力是自然语言交互、告警摘要、知识库问答，交互模式为人提问、AI 回答，决策权 100%在人。技术门槛较低，基于通用大模型+RAG 即可实现，市场定位为降低使用门槛，提升分析师效率 10-20%。代表产品为早期的 SecurityGPT 及各家"AI 助手"功能。

阶段 2：深度助手（Copilot 2.0）。 核心能力升级为告警分析、威胁关联、调查建议、响应手册推荐，AI 主动分析并给出结构化建议，人做决策。技术门槛提升，需要领域知识注入（安全知识图谱、攻击链模型等），市场定位为显著提升分析质量，效率提升 30-50%。代表产品为 Microsoft Security Copilot、Splunk AI Assistant。

阶段 3：半自主 Agent（Agent 1.0）。 核心能力是在明确边界内自主执行标准化工作流（如 L1 告警处理、日常巡检），AI 自主执行，关键节点人工审批，例外情况升级。技术门槛较高，需要多 Agent 协作、工作流编排、外部工具调用（Function Calling），效率提升 50-70%。代表产品为 CrowdStrike Charlotte AI、Dropzone AI。

阶段 4：全自主 Agent（Agent 2.0）。 核心能力是端到端自主处理复杂安全事件，自主学习和优化策略，AI 独立运作，人类监督和审计。技术门槛最高，需要强化学习、持续学习、可解释 AI、高可靠性工程。目前处于研究阶段，尚无成熟产品。

给厂商的建议是：2025-2026 年的竞争焦点是从 Copilot 2.0 向 Agent 1.0 的跨越；不要跳跃式发展——Copilot 能力不扎实的情况下做 Agent 必然失败；明确告知客户当前处于哪个阶段，不要过度承诺；在 1-2 个场景率先实现 Agent 1.0，建立标杆案例，再横向扩展。

2.2 评测与治理：差异化竞争的关键战场

在开源大模型普及的背景下，单纯的"我也能做 AI"不再是竞争优势。评测与治理能力将成为差异化方向。

评测能力分为三个层次。L1 是结果评测，评估 AI 输出的准确率、召回率、误报率，与人工基线对比证明 AI 优于人工——客户最关心"你的 AI 到底准不准"，但多数厂商说不清楚。L2 是过程评测，评估 AI 的推理过程是否符合安全逻辑（可解释性），识别 AI 的知识盲区和高风险误判场景——从"黑盒 AI"到"透明 AI"，建立客户信任。L3 是持续评测，建立自动化评测流程，每次模型更新都运行评测，监控生产环境中 AI 的性能退化（数据漂移、对抗样本等）——能够持续保证 AI 质量的厂商才能获得长期订单。

治理能力的构建涵盖四个维度：一是数据治理，帮助客户建立数据质量评估、清洗、标注的流程和工具；二是模型治理，提供模型版本管理、A/B 测试、灰度发布能力；三是风险治理，建立 AI 误判应急响应机制、红蓝对抗测试、对抗样本检测；四是合规治理，满足数据安全法、个人信息保护法等合规要求，提供审计日志。

行动建议：成立专门的"AI 评测与治理"团队，不隶属于产品开发部门；借鉴学术界的 SecBench 等评测基准，建立自己的评测集；将评测报告作为销售材料，公开透明地展示 AI 能力边界；提供"评测即服务"，帮助客户评估不同供应商的 AI 产品。

2.3 小模型在安全检测场景的刚性需求

一个常见的误解是将小模型视为"中小厂商的廉价替代品"。事实上，**在威胁检测这一核心场景中，小模型不是退而求其次的选择，而是技术刚需**——无论企业规模大小。

威胁检测场景对延迟和吞吐量有极高要求。实时流量分析需要在毫秒级完成判断，EDR/NDR 每秒需要处理数千到数万条事件。大模型（数十 B 乃至百 B 参数）的推理延迟通常在秒级，根本无法满足这类场景。具体来看：一是延迟约束，网络流量检测、恶意文件实时扫描等场景要求推理延迟<10ms，只有 0.1B-0.85B 级别的模型才能达到；二是吞吐量要求，大型企业 SOC 每秒处理数万事件，需要单卡支撑数万 QPS，大模型根本跑不动；三是边缘部署，终端/网关/IoT 设备上的检测任务算力受限，只能部署超小模型；四是成本效率，即使在云端，对海量日志逐条调用大模型，推理成本也完全不可接受。

头部安全厂商（CrowdStrike、SentinelOne 等）的核心检测引擎正是基于轻量级模型。这不是因为它们缺乏资源部署大模型，而是因为场景决定了技术选择。由此形成了**"分层模型架构"**：实时检测层必须用小模型（0.1B-0.85B），保证延迟和吞吐；深度分析层可以用中等模型（7B-32B），对小模型标记的可疑事件做二次研判；交互辅助层可以用大模型（70B+或 API），支撑安全分析师的自然语言交互和报告生成。不同层级用不同规模的模型，匹配各自的延迟、精度、成本需求。

对于分析研判类场景（非实时检测），小模型结合 RAG 是兼顾效果与效率的方案。通过检索最新威胁情报和内部知识库弥补知识缺陷，在生成答案前检索相关上下文减少幻觉，知识库更新即时生效无需重新训练模型，检索到的文档可作为答案的"引用来源"增强可信度。

实践路径方面：一是按场景选择模型规模（实时检测用 0.1B-0.85B，告警研判用 1B-32B，交互分析用 13B-70B 或云端 API），选择标准是延迟优先，在满足延迟约束的前提下最大化精度；二是构建领域知识库用于 RAG 增强，包括威胁情报库、内部安全知识库、合

规标准文档，使用向量数据库（如 Milvus、Qdrant）支持语义检索；三是针对性微调，实时检测模型用大规模标注数据进行任务专用微调，分析研判模型用 LoRA/QLoRA 在安全分析师的真实案例上微调；四是工程化部署，实时层用 ONNX Runtime、TensorRT 等推理加速并做 INT4/INT8 量化，分析层用 vLLM 等高吞吐推理框架配合 RAG pipeline。

给厂商的建议是：不要简单地"大模型 or 小模型"二选一，而是设计分层模型架构；将小模型检测能力作为核心竞争力打造，这是产品化的关键差异点；大模型用于提升用户体验，但不要让它成为检测引擎的瓶颈；关注模型蒸馏技术，用大模型的能力"教"小模型，持续提升小模型的检测精度。

2.4 开放生态 vs 封闭平台：战略选择的分水岭

封闭平台策略的逻辑是构建"大而全"的一体化安全 AI 平台，从数据采集、分析到响应全覆盖，优势在于用户体验统一、数据流转顺畅、厂商掌控全链条，但风险是开发周期长、客户锁定困难、难以适应快速变化的 AI 技术。开放生态策略的逻辑是提供标准化 API 和插件机制，允许第三方扩展功能，优势是快速响应长尾需求、形成网络效应、降低自身研发压力，但挑战在于质量控制难度大、生态建设需要长期投入。

对头部厂商的建议是采用"核心封闭+边缘开放"策略：核心能力（大模型、数据平台、Agent 引擎）自主掌控，边缘功能（特定场景的插件、行业定制化）开放给生态。对中小厂商的建议是专注做好 1-2 个场景的"最佳插件"而非自建平台：适配主流安全平台的 API，参与头部厂商的生态计划，降低客户的迁移成本，用效果而非锁定获得黏性。

开放生态的最佳实践包括提供详尽的 API 文档和 SDK、建立开发者社区和激励计划、定期举办 Hackathon 发掘优秀插件、建立插件认证机制确保安全性和质量。

三、产学研视角：构建产业长期竞争力

3.1 安全 AI 评测基准标准化：从学术成果到行业共识

当前面临的核心问题是：不同厂商对"AI 安全能力"的定义和度量标准各异，客户难以横向比较；学术界已有 SecBench 等评测基准，但行业采用度低，产学脱节；缺乏权威第三方评测机构，厂商自说自话，公信力不足。

为此提出三方面行动建议。一是建立行业标准评测基准，由行业协会（如中国网络安全产业联盟、信息安全测评中心）牵头，联合头部厂商、高校、研究机构，共同制定任务分类标准（将 AI 安全能力细分为告警分类、威胁检测、事件调查、响应建议、漏洞分析等）、评测数据集（覆盖常见攻击类型和正常行为）、评测指标体系（准确率、召回率、误报率、响应时延、可解释性、鲁棒性等多维度）和评测流程规范。二是成立第三方评测认证机构，独立于厂商，定期发布评测报告（类似汽车碰撞测试），提供"AI 安全能力认证"成为客户选型参考，建立"评测沙箱"供厂商自主提交测试。三是推动学术成果转化，将 SecBench、CyBench 等学术评测基准引入工业界，鼓励高校开源最新研究成果，设立"产学研联合实验室"解决评测中的技术难题。

预期 3 年内形成国家标准或行业标准，建立权威的 AI 安全能力排行榜每半年更新，使客户选型有客观依据，促进良性竞争。

3.2 数据共享机制：破解"数据孤岛"困境

核心矛盾在于：AI 效果严重依赖大规模高质量训练数据，但安全数据高度敏感、企业不愿对外共享，而单一企业的数据规模和多样性又不足以训练高质量模型。

务实的解决方向有三个。一是脱敏数据交易平台，建立安全数据交易平台，企业可出售脱敏后的数据、购买所需数据，建立基于数据规模/质量/稀缺性的定价机制，确保合规审查和使用记录防止滥用。二是众包标注机制，将安全数据标注任务众包给安全从业者，平台提供标注工具和质量控制机制，标注者获得报酬同时提升自身技能，适用于威胁情报标注、恶意样本分类等任务。三是开源安全数据集建设，由行业协会或国家机构牵头构建公开的代表性安全数据集（脱敏处理），参考计算机视觉领域 ImageNet 的成功经验，但需正视安全数据的特殊性——攻击手法快速演变，数据集需持续更新。

3.3 安全 AI 人才培养体系：填补结构性人才缺口

当前现状是传统安全人才缺乏 AI 技能，AI 人才不懂安全场景，高校课程设置滞后，企业内训成本高、周期长。需要清醒认识到，在 AI 技术以月为单位快速迭代的当下，传统的"在安全专业中增加几门 AI 课程"的思路已经不够——等本科四年读完，AI 技术早已面目全非。

行动建议分为两个层面。一是高校教育范式变革：核心理念是 **AI Embedded**，即从本科阶段起就将 AI 融入网络安全教育的全流程，而非简单叠加 AI 课程。安全数据分析、威胁检测、漏洞挖掘、应急响应等每一门专业课都应以 AI 为底座重新设计——用 AI 做实验、用 AI 写报告、用 AI 辅助攻防，让学生从入学起就建立"AI 是基本工具"的认知范式。研究生阶段则聚焦 AI for Security 的前沿研究方向（多 Agent 安全运营、对抗机器学习、AI 安全评测等），实践项目与企业合作解决真实问题。师资方面鼓励企业一线专家到高校任教，将实战经验引入课堂。

二是在职从业者的理念升级：对于已经工作的安全从业者，重点不在具体技术技能的补课，而在于**理念层面的根本转变**——理解 AI 如何重塑安全运营的工作范式，从"手工分析"思维转向"人机协同"思维，从"规则驱动"转向"数据驱动+AI 增强"。头部厂商和行业协会应组织面向 CISO 和一线团队的短期密集培训，帮助从业者建立对 AI 能力边界的正确认知，掌握 AI 工具的选型评估方法，理解从 Copilot 到 Agent 的演进路径对组织架构和人员配置的影响。竞赛方面，举办"AI 安全对抗赛"（用 AI 检测隐蔽攻击、AI Agent 自动化渗透测试等）是检验和提升实战能力的有效方式。

总结：协同演进，共建 AI 安全新生态

AI for Security 不是某一方的独角戏，而需要甲方、乙方、产学研的协同演进。甲方应大胆试点、小步快跑，用真实需求牵引技术进步，同时做好数据准备和人才储备。乙方应聚焦场景、打磨产品，从 Copilot 稳步向 Agent 演进，构建开放生态而非封闭壁垒。产学研应制定标准、共享数据、培养人才、建设基础设施，为产业提供公共品。

未来三年是 AI for Security 从"概念验证"到"规模落地"的关键窗口期。谁能在这个窗口期解决数据质量、评测标准、人才供给等核心瓶颈，谁就能在多家机构所预期的"2026—2028 年 AI/Agent 能力在 SOC 加速渗透"的窗口期内占据主导地位。

第三章 战略假设

时间范围与适用说明

本章战略假设覆盖 **2026-2028 年** 时间窗口（3 年中期视野），为后续章节的趋势判断、市场预测、技术路线分析提供基准前提。假设基于当前已观察到的技术轨迹、市场信号和政策动向，但不排除黑天鹅事件导致的路径偏移。

本报告所有预测性结论均建立在以下假设成立的前提下。若关键假设被证伪（如大模型能力增长停滞、重大监管政策转向），相关结论需相应调整。

一、技术假设

假设 1.1：大模型能力持续提升

在推理能力方面，到 2028 年主流大模型在复杂逻辑推理任务上的表现预计接近人类安全分析师中位水平（如 MITRE ATT&CK 战术链推理、多步攻击溯源）。在领域适配方面，安全垂直领域模型（如 CodeLlama Security、SecGPT 系列）通过持续 SFT 和 RLHF，将在威胁情报解读、日志分析、策略生成等任务上显著超越通用模型。在成本方面，推理成本按年均 40-50%速率下降（美元/百万 Token），使实时全流量分析和大规模日志 AI 处理在经济上可行。

假设 1.2：Agent 框架趋于成熟

标准化方面，LangGraph、AutoGen、CrewAI 等多 Agent 框架预计在 2026-2027 年完成工业级稳定性验证，形成事实标准。工具调用可靠性方面，工具调用成功率将从当前 70-80%提升至 90%+，幻觉率下降至个位数（在结构化任务中）。在人机协同模式方面，预计

形成"Agent 执行+人类审批关键节点"的成熟工作流模板，支持自动化率 60-80%的 SOC 运营场景。

假设 1.3：多模态能力普及

预计到 2028 年，AI 可准确解析安全态势大屏、网络拓扑图、漏洞截图，辅助非技术管理者理解安全事件。视觉+语言模型结合将实现代码补丁自动审查、配置文件漏洞扫描（准确率 85%+）。时序数据方面，原生支持网络流量时序模式识别和用户行为异常检测（UEBA 场景）。

假设 1.4：本地化部署技术成熟

在小模型性能方面，7B-13B 参数的安全专用模型在特定任务上预计达到 GPT-4-turbo 70-80%能力，可部署在企业私有云。量化与推理优化方面，INT4/INT8 量化、投机采样等技术使单张 4090/A800 可支撑 10+ QPS 的生产负载。在部署架构方面，云端大模型（复杂推理）+本地小模型（实时检测）的混合架构将成为主流部署形态。

二、市场假设

假设 2.1：市场规模增长轨迹

全球 AI for Security 市场 2025 年规模约**$25-34B**，年均增长率约 **19-24%**。三家主要研究机构的预测数据为：Grand View Research 估算 2024 年$25.4B、2030 年$93.8B（CAGR 24.4%）；Fortune Business Insights 预测 2025 年$34.1B、2034 年$213.2B（CAGR 21.7%）；Precedence Research 给出 2025 年$29.6B、2035 年$167.8B（CAGR 18.9%）。增长主要由安全人才短缺、攻击复杂度上升、监管合规压力和 AI 成本下降等因素驱动。

中国市场方面，预计占全球份额从 2025 年的 12%提升至 2028 年的 15-18%，本土化需求是核心驱动力。政务、金融、运营商等重点行业 AI 安全产品渗透率预计从当前 5-8%提升至 25-30%。

假设 2.2：预算再分配趋势

企业安全预算中 AI 相关支出（工具采购+人才培养+数据治理）占比预计从 2025 年的 8-12%提升至 2028 年的 20-25%。传统 SIEM/IDS 等产品预算增速将放缓（CAGR 3-5%），预算向 AI 原生安全产品迁移。

假设 2.3：买家行为变化

预计出现三方面变化：一是 POC 周期延长，从平均 3-6 个月延长至 6-12 个月，买家更关注 AI 效果可证性、误报控制和长期 ROI；二是多厂商组合，企业倾向于"最佳品种（Best-of-Breed）"策略，同时采购 2-3 家 AI 安全产品而非依赖单一平台；三是订阅模式主导，按用户数/数据量的 SaaS 订阅取代传统 License，占新签合同的 70%+。

三、监管假设

假设 3.1：中国数据安全法与 AI 治理政策持续收紧

在数据出境限制方面，《数据安全法》《个人信息保护法》执法力度加强，安全日志、威胁情报等敏感数据禁止出境训练海外模型。在可解释性要求方面，关键信息基础设施运营者使用 AI 安全产品时必须确保决策可审计、可追溯。

假设 3.2：中国政府和国企市场中私有化部署仍是主流

需要区分不同市场的差异：在美国等西方国家，SaaS 类安全 AI 服务仍然是主流交付模式，企业对云端部署的接受度较高。但在中国，由于政府对保密性的高度重视，对政府

机关、国有企业、关键信息基础设施运营者有明确的"数据不出域"合规性要求，这使得**中国政企市场中私有化部署仍是主流**。政务、金融、能源等行业客户中 80%+要求 AI 模型与数据全部部署在自有环境（私有云/本地）。混合云仅在互联网、零售等数据敏感度较低的行业缓慢渗透。

假设 3.3：国际监管碎片化

欧盟 AI Act 对高风险 AI 系统（包括关键基础设施安全）的合规要求将在 2026-2027 年逐步生效，影响欧洲市场准入。美国方面以 NIST AI RMF、CISA 指南等软性框架为主导，短期内无强制性 AI 安全产品认证。

四、人才假设

假设 4.1：安全人才市场结构性错配加剧

(ISC)²、Cybersecurity Ventures 等机构发布的"全球缺口 350 万""中国缺口 95 万"等数据需审慎看待，其发布方多为人才培养利益相关方（培训机构、认证机构、高校），存在夸大市场需求的动机。中国现实情况是，网络安全专业经历大规模扩招后，应届毕业生就业已开始出现困难，基础岗位供大于求。真正稀缺的是既懂安全攻防又懂 AI 工程的高端复合型人才，这与"总量缺口"是完全不同的问题。大量初级安全人才找不到工作，同时企业招不到能驾驭 AI 安全工具的高级人才——这是技能结构问题，不是总量问题。

假设 4.2：AI 自动化将重塑安全岗位结构

AI 将大量替代 L1/L2 级别的告警分析、日志审查等重复性工作，进一步压缩初级安全岗位的需求。安全分析师工作重心将从"处理告警"转向"训练 AI、审核 AI 决策、处理 AI 无法解决的复杂案例"。安全 AI 工程师、AI 红队测试员等新职位需求增长，但总量远小于被

替代的初级岗位。净效应是 AI 可能使安全行业整体人力需求下降而非上升，但对剩余岗位的技能要求显著提高。

假设 4.3：教育体系与市场需求脱节

高校网络安全专业课程仍以传统攻防为主，普遍缺乏 AI+安全交叉课程。大量扩招培养的是市场已不再急需的初级人才，而企业真正需要的 AI 安全复合型人才高校尚无力培养。企业内训和在线课程成为高级技能补充的主要渠道。

五、数据假设

假设 5.1：高质量安全数据仍是稀缺资源

在标注瓶颈方面，安全告警数据需要资深分析师标注（成本$50-200/小时），导致高质量训练集获取困难。在数据孤岛方面，企业间威胁情报、攻击样本共享意愿低（竞争与合规顾虑），限制模型泛化能力。

假设 5.2：数据飞轮效应显现

头部厂商（拥有大量客户部署）积累的反馈数据形成训练飞轮，模型能力持续拉开与中小厂商差距。

六、假设风险与应对

假设项	主要风险	若不成立的影响	监控指标
大模型能力持续提升	技术瓶颈、训练数据枯竭	AI 产品价值受限，增长放缓	GPT-5/6 性能评测、顶会论文数量
Agent 框架成熟	幻觉、可靠性问题未解决	自动化应用延后 2-3 年	工具调用成功率、生产事故率

市场高增长	经济衰退、IT 预算削减	市场规模达成率<70%	Gartner 季度追踪报告
中国政企私有化部署主导	无法使用最先进大模型能力，本地算力受限导致安全能力发挥受限	私有化客户 AI 安全能力与云端差距拉大，厂商需在本地化效果优化上加大投入	私有化 vs 云端效果对比、国产算力进展
人才结构性错配加剧	AI 工具降低技能门槛	复合型人才供给改善	企业招聘数据、高校就业率
数据稀缺性	开源数据集爆发	降低数据壁垒，加剧竞争	Hugging Face/GitHub 发布量

七、假设更新机制

本章假设按年度审查（每年 2 月随报告更新一并审查假设有效性），并设触发式修正机制——若出现重大技术突破（如 AGI 实现）、政策变化（如 AI 禁令）、市场震荡（如头部厂商倒闭），即启动假设修正。欢迎读者通过 ssaq@geniuscybertech.com 提供假设挑战与替代情景建议。

本章假设的哲学：我们相信未来是由当下的技术轨迹与市场力量塑造的，但保持对"意外"的敏感。假设不是预言，而是帮助我们在不确定性中做出更好决策的工具。

第四章 AI for Security 市场定义

> **核心观点**：AI for Security 市场正在经历从"能力插件"到"智能中枢"的范式转变。市场边界不再由技术实现方式划分,而是由业务价值链的完整性和自主决策能力来界定。本章明确划定市场边界、子市场结构,并解析 AI 如何重构既有安全产品的边界。

4.1 AI for Security 市场定义

4.1.1 核心定义

AI for Security(人工智能赋能网络安全)是指利用人工智能技术（包括机器学习、深度学习、大语言模型、生成式 AI 等）提升网络安全防护、检测、响应、运营、治理等能力的软件、服务和平台的集合。其核心价值在于**从人工规则驱动转向数据和知识驱动,从人工决策转向人机协同乃至自主决策。**

4.1.2 市场边界："包含"与"不包含"

包含范围

本报告所界定的 AI for Security 市场**包含**以下场景和能力：

一是**威胁检测与响应**，涵盖 AI 驱动的异常检测（端点、网络、云、身份等）、威胁情报自动化分析与关联、自动化事件调查与根因分析，以及自主或辅助响应处置（封禁、隔离、修复等）。二是**安全运营智能化**，包括 AI SOC Analyst（智能安全分析师）、告警降噪与优先级排序、自然语言查询与威胁猎捕，以及安全知识库与专家经验沉淀等核心能力。三是**数据安全治理**，覆盖智能化数据分类分级、敏感数据识别与脱敏、数据泄露风险检测，以及数据安全态势管理（DSPM）等场景。四是**攻防对抗能力**，主要体现在自动化

渗透测试（AI Red Team）、漏洞挖掘与代码审计，以及攻击路径推演与仿真等方面。五是**开发安全与供应链**，涉及 AI 驱动的代码安全扫描、开源组件漏洞检测，以及 AI 生成代码的安全验证等能力。六是**邮件与通信安全**，包括钓鱼邮件识别与 BEC（商业邮件入侵）检测、邮件溯源与归因等功能。七是**认知安全与鉴伪**，聚焦于深度伪造检测（Deepfake Detection）、虚假信息识别与内容溯源等新兴场景。八是**安全策略智能化**，涵盖策略冲突检测与优化、安全配置基线智能生成等能力。九是 **AI 自身安全治理**（Security for AI 的防护侧），包含 AI 资产发现与治理、影子 AI 检测、AI 运行时保护（Runtime Protection）、提示注入与模型投毒等攻击检测等内容。需要说明的是，虽然该领域属于"Security for AI"范畴，但从安全产品提供商视角看，其实现技术与 AI for Security 重叠，故纳入本报告范畴。

不包含范围

本报告**不包含**以下内容：

首先是**纯 IT 运维类 AI 应用**，即不涉及安全场景的 AIOps、日志分析、性能监控等。需要说明的是，若 AI 能力同时服务于安全和运维（如 SIEM/SOAR 同时用于故障和安全事件），则计入本市场。其次是**非 AI 技术的传统安全产品**，即纯规则引擎、签名库、静态策略的安全产品。若产品中 AI 能力占核心价值链比重<20%，或仅作为边缘辅助功能，则不计入。第三是 **AI 基础设施与算力服务**，包括 GPU 云、训练平台、MLOps 等通用 AI 基础设施。若专为安全场景定制（如安全专用训练平台、安全数据标注服务），则计入。第四是 **AI 技术本身的研发攻防**，即对抗样本生成、模型提取等攻击技术研发（非防御侧）。"Security for AI"中的攻击技术研究不计入，但防御产品计入。第五是**通用 BI/数据分析产**

品，即非安全场景的数据可视化、报表生成等。安全数据的智能分析和可视化计入本市场范畴。

4.1.3 三种能力定位的区分

AI for Security 市场中的厂商能力可分为三类：

能力定位	**定义**	**典型特征**	**代表厂商**
平台化赋能	为多产品/多场景提供统一 AI 能力,形成 AI 中枢或 AI 服务层	• 支持多场景复用 • 统一知识库/工具库 • 跨产品调度	微软 Security Copilot Google Gemini for Security 火山引擎 360 安恒信息 深信服 金睛云华 绿盟科技
单点插件	针对单一产品或场景嵌入 AI 能力,不具备跨产品复用性	• 与现有产品深度耦合 • 解决特定痛点 • 快速见效	Abnormal Security（邮件安全） Snyk（开发安全） 美创科技（数据安全） 安恒信息 悬镜安全 灵云数科
基座模型服务	提供可被安全厂商或客户调用的 AI 推理/训练能力,类似"安全 AI 中间件"	• 以 API/SDK 形式输出 • 模型可定制训练 • 算力按需调用	Anthropic OpenAI

三者关系：**平台化赋能**是大厂和综合性厂商的战略选择，追求规模效应和生态控制；**单点插件**是专业厂商的快速切入路径，依赖场景深度和产品粘性；**基座模型服务**是介于两者之间的能力输出形态，可作为独立业务，也可作为平台化赋能的服务化延伸。

4.1.4 与"Security for AI"的边界

AI for Security 与 Security for AI（AI 自身安全）存在交集但核心关注点不同：

维度	**AI for Security**	**Security for AI**
核心目标	用 AI 保护信息系统和数据	保护 AI 系统自身不被攻击
防护对象	传统 IT 资产、网络、应用、数据	AI 模型、训练数据、推理服务
威胁类型	传统网络攻击、数据泄露、恶意代码	对抗样本、模型投毒、提示注入、模型窃取
技术手段	机器学习检测、大模型推理、Agent 自动化	鲁棒性训练、输入过滤、模型水印、运行时防护
交集领域	**AI 安全治理**：既需要 AI 能力检测 AI 威胁,又需保护 AI 资产安全	同左

本报告立场：将"AI 安全治理"（如 AI 资产发现、影子 AI 检测、AI 运行时保护）纳入 AI for Security 范畴，理由有三：一是这些能力的提供方是安全厂商（非 AI 研发方）；二是技术实现与其他 AI 安全能力重叠（如威胁检测、行为分析）；三是客户采购决策方通常是 CISO（非 AI 部门）。

4.2 子市场切分

4.2.1 按场景簇切分

根据业务价值链和技术实现的相似性，AI for Security 市场可划分为以下 10 大场景簇（代表厂商仅作为场景定义的锚点，国内厂商分场景完整清单与能力要点详见 10.10 节，国际厂商深度画像详见 10.1–10.5 节）：

1. 威胁检测（Threat Detection & Response）

定义：利用 AI 技术识别已知和未知威胁,覆盖端点（EDR）、网络（NDR）、云（CWPP/CSPM）、身份（ITDR）等多个维度的异常行为检测和自动化响应。

技术特征：在技术实现上，该场景主要依赖无监督学习建立基线行为模型、实时流式数据处理与异常检测、攻击链拼接与攻击意图推理，以及自动化响应编排（封禁、隔离、取证）等核心能力。

代表厂商：国际以 CrowdStrike（EDR 领导者）、Darktrace（自学习 NDR）、Vectra AI（NDR Leader）为代表；国内以 360 数字安全、深信服、安恒信息、金睛云华为头部厂商。完整国内厂商清单见 10.10 节。

关键指标：衡量该场景价值的关键指标包括检测准确率（Precision）与召回率（Recall）、误报率（False Positive Rate）下降幅度，以及平均检测时间（MTTD, Mean Time to Detect）。

2. 安全运营（Security Operations, SecOps）

定义：将 AI 应用于 SOC 日常运营,实现告警降噪、事件调查自动化、威胁情报关联、安全知识沉淀等,核心是"AI SOC Analyst"。

技术特征：该场景的技术特征体现为大语言模型驱动的自然语言交互、RAG/TAG 增强的知识库检索、Multi-Agent 协同的调查与处置，以及闭环工作流（调查→判定→处置→复盘）。

代表厂商：国际以 Dropzone AI、Microsoft Security Copilot、Google Security Operations（Chronicle/Gemini）为代表；国内以深信服、360 数字安全、奇安信、启明星辰、安恒信息、绿盟科技为头部厂商。国内分场景全量厂商清单见 10.10 节。

关键指标：该场景的核心评估指标包括告警处置自动化率、平均响应时间（MTTR, Mean Time to Respond）缩短幅度，以及 SOC 分析师工时节省比例。

3. 数据安全（Data Security Governance）

定义：利用 AI 自动化识别敏感数据、分类分级、检测异常访问和泄露风险,覆盖数据发现、分类、脱敏、DLP、DSPM 等全生命周期。

技术特征：技术实现层面，该场景依托自然语言处理识别敏感字段（PII、商业机密等）、多模态识别（文本+图像+音频）、基于行为基线的异常访问检测，以及智能脱敏与数据合成等能力。

代表厂商：国际以 Cyera、BigID、Varonis 为代表；国内以美创科技、炼石网络、明朝万达、闪捷信息为代表。完整国内厂商清单见 10.10 节。

关键指标：衡量该场景效果的关键指标包括敏感数据自动识别准确率、数据分类分级效率提升，以及数据泄露事件的事前拦截率。

4. 渗透测试与漏洞挖掘**（Penetration Testing & Vulnerability Research）**

定义：利用 AI 自动化执行渗透测试流程、挖掘 0day 漏洞、生成漏洞利用代码,降低攻防对抗的人力依赖。

技术特征：该场景的核心技术能力体现为大模型生成攻击脚本与 Payload、强化学习优化攻击路径、模糊测试（Fuzzing）智能化，以及攻击面自动化探测与映射。

代表厂商：国际以 XBOW（自主渗透）、Horizon3.ai NodeZero、Pentera 为代表；国内以长亭科技、云起无垠、矢安科技、华云安为代表。完整国内厂商清单见 10.10 节。

关键指标：该场景的关键评估指标包括 0day 发现数量、渗透测试覆盖范围，以及测试效率（单位时间测试用例数）。

5. 软件供应链安全（Software Supply Chain Security）

定义：在开发阶段嵌入 AI 安全扫描,覆盖代码审计、开源组件漏洞检测、AI 生成代码验证、软件供应链风险分析等。

技术特征：技术实现上，该场景依托代码语义理解与漏洞模式匹配、Agent 自动生成修复代码、开源组件 SCA（Software Composition Analysis），以及软件物料清单（SBOM）智能分析等能力。

代表厂商：国际以 Snyk（AST Leader）、GitHub Advanced Security（Copilot Autofix）、Semgrep 为代表；国内以悬镜安全、奇安信（代码卫士）、孝道科技为代表。完整国内厂商清单见 10.10 节。

关键指标：衡量该场景价值的关键指标包括千行代码漏洞率（Bugs per KLOC）下降幅度、修复建议采纳率，以及开发流程中的安全扫描覆盖率。

6. 邮件安全（Email Security）

定义：利用 AI 检测钓鱼邮件、BEC 攻击、恶意附件、邮件溯源等,保护企业通信渠道安全。

技术特征：该场景的技术特征体现为行为 AI 学习正常通信模式、多模态检测（文本+图像+链接）、实时拦截与用户安全培训，以及邮件归因与威胁情报关联。

代表厂商：国际以 Abnormal Security（邮件行为 AI）、Proofpoint、Microsoft Defender for Office 365 为代表；国内以深信服、广东盈世/CACTER、知其安为代表。完整国内厂商清单见 10.10 节。

关键指标：该场景的核心评估指标包括钓鱼邮件拦截率、误报率（合法邮件被拦截比例），以及 BEC 攻击检测准确率。

7. 鉴伪/认知安全（Deepfake Detection & Cognitive Security）

定义：利用 AI 检测深度伪造内容（Deepfake）、虚假信息、合成媒体,保护组织免受认知操纵和信息战攻击。

技术特征：技术实现层面，该场景依托生成对抗网络（GAN）检测、多模态一致性验证、数字水印与内容溯源，以及虚假信息传播链分析等能力。

代表厂商：国际以 Reality Defender、Sensity AI、Pindrop 为代表；国内以云弈科技、摄星科技、任子行为代表。完整国内厂商清单见 10.10 节。

关键指标：衡量该场景效果的关键指标包括 Deepfake 检测准确率、误判率（真实内容被误判为伪造），以及响应时效（从生成到检测的时间窗口）。

8. 工控/OT 安全（Industrial Control / OT Security）

定义：面向工业控制系统（ICS/SCADA）、OT 网络与车联网、物联网等融合场景的 AI 驱动威胁检测、资产测绘与合规审计，兼容国产化算力与信创要求。

技术特征：该场景的核心技术能力体现为小样本/无监督异常检测、工控协议语义解析（Modbus、IEC 61850、OPC UA 等）、边侧轻量模型推理（毫秒级延迟），以及视觉 AI 辅助的现场安全监测。

代表厂商：国际以 Claroty、Nozomi Networks、Dragos 为代表；国内以烽台科技、威努特、长扬科技、六方云、万物安全（IoT/车联网延伸）为代表。完整国内厂商清单见 10.10 节。

关键指标：衡量该场景价值的关键指标包括工控协议深度解析覆盖率、现场部署端到端延迟、误报率（生产环境对误报高度敏感），以及资产发现与测绘完整度。

9. 云安全与 WAAP（Cloud Security & WAAP）

定义：面向公有云/私有云工作负载与 Web 应用、API、边缘接入的 AI 驱动防护，覆盖 WAF、Bot 管理、API 安全、DDoS 防御，以及针对 LLM 应用的 Prompt Injection 与数据泄漏防护（WAAP for LLM）。

技术特征：该场景的核心技术能力体现为大流量实时在线学习与特征提取、API 行为建模与异常调用检测、Bot 指纹与流量基线、以及面向 LLM 应用层的提示词攻击识别与语义护栏。

代表厂商：国际以 Cloudflare、Akamai、Imperva 为代表；国内以瑞数信息（WAAP for LLM）、网宿安全（全球边缘 AI 防护）、国舜股份、厦门快快网络为代表。完整国内厂商清单见 10.10 节。

关键指标：衡量该场景价值的关键指标包括 0day 拦截率、LLM Prompt Injection 拦截率、Bot 识别准确率，以及防护开销与业务时延增幅。

10. 安全策略管理（Security Policy Management）

定义：利用 AI 自动化生成、优化、审计安全策略（防火墙规则、访问控制策略、加密策略等）,减少策略冲突和错误配置。

技术特征：该场景的核心技术能力体现为策略冲突自动检测、基于业务意图的策略生成、零信任策略动态调整，以及策略合规性自动审计。

代表厂商：国际以 Tufin、AlgoSec、FireMon 为代表；国内以蔷薇灵动（微隔离）、安博通为代表，整体厂商数量较少（该场景 10.10 节未单列）。

关键指标：该场景的关键评估指标包括策略冲突自动发现率、策略优化后的性能提升，以及配置错误导致的安全事件下降幅度。

补充说明：本节采用"定义 + 技术特征 + 代表厂商（旗舰锚点）+ 关键指标"四要素界定 10 大场景簇，代表厂商仅作为场景定义的锚点；国内厂商在各场景的完整分布与能力要点详见 10.10 节，国际厂商深度画像见 10.1–10.5 节。本轮国内调研还观察到反欺诈情报（如威胁猎人）、攻防推演与安全教育（如丈八网络、永信至诚）、AI 安全治理（Security for AI，如 HiddenLayer、长亭科技守元；严格意义上属于 Security for AI 边界，详见 4.1.4）三类相邻赛道，已在 10.10 中作为独立类别列示，不在本节作为一级场景簇重复定义。

4.2.2 按交付形态切分

从用户交互和价值交付方式看,AI for Security 可划分为 **3 种交付形态**：

1. Chatbot/Copilot（对话式助手）*

定义：以自然语言交互为主,辅助安全人员完成查询、分析、决策等任务,**决策权在人。**

典型场景：该形态的典型应用包括威胁情报查询（"最近有哪些针对金融行业的勒索软件？"）、告警解读（"这条告警是真阳性还是误报？"），以及策略建议（"如何配置防火墙规则拦截此类攻击？"）。

代表产品：代表性产品包括 Microsoft Security Copilot、Google Gemini for Security Operations、SentinelOne Purple AI、CrowdStrike Charlotte AI、Palo Alto Cortex Copilot、IBM QRadar AI Assistant 等国际产品，以及国内的 360（BrainGPT/安全大模型 Copilot）、

深信服（安全 GPT Copilot）、奇安信（Q-GPT）、启明星辰（安星/观星）、安恒信息（恒脑）、绿盟（风云卫 Copilot）、海云安、天懋信息、摄星科技、石犀科技等。

优势与局限：该形态的优势在于降低技能门槛、提升人员效率、决策透明可控。局限性则体现在仍需人工介入，无法实现全自动化。

2. Agent 智能体（自主执行）*

定义：具备工具调用和决策能力的智能体,可**自主完成调查、处置、修复**等闭环任务,**决策权部分或全部在 AI。**

典型场景：该形态的典型应用场景包括 AI SOC Analyst（自动调查 100%的告警，生成处置建议或直接执行低风险处置）、AI Detection Engineer（根据威胁情报自动调优检测规则），以及 AI Threat Hunter（主动搜索潜在威胁并生成猎捕报告）。

代表产品：代表性产品包括 Dropzone AI（AI SOC Analyst）、7AI、Prophet Security、Torq（HyperSOC）、CrowdStrike Agentic Security Platform、Microsoft Security Copilot（Agent 模式）、Radiant Security、Simbian、Qevlar AI 等国际产品，以及国内的深信服（MSS+安全 GPT Agent）、360（安全运营智能体平台）、奇安信（AISOC）、启明星辰（AIDK 智能体框架）、亚信安全（AIXDR）、未来智安（XDR+Agent）、青藤云安全（无相 AI）、知道创宇、绿盟科技、天懋信息、瀛云科技、摄星科技、炼石网络、美创科技、宁数安全、石犀科技等。

核心能力：该形态的核心能力体现为三个方面。一是**工具编排**（Tool Orchestration），能够调用查询、封禁、工单、取证等工具；二是**闭环处置**（Closed-Loop Remediation），实

现从检测到修复的全流程自动化；三是**审批审计**（Approval & Audit），关键动作需人审批，全链路可追溯。

优势与局限：该形态的优势在于大幅降低人工工时（30-80%）、缩短 MTTR、提升处置闭环率。局限性则在于需要严格的权限控制和审计机制，防止误操作和越权。

3. 内嵌增强（Embedded AI）*

定义：将 AI 能力嵌入既有安全产品的功能模块,**用户无感知或轻感知**,直接提升产品原有能力。

典型场景：该形态的典型应用包括 EDR 产品内嵌 AI 检测引擎以提升未知威胁检测率、SIEM 自动化告警降噪和事件关联，以及 DLP 产品智能识别敏感数据。

代表产品：代表性产品包括 CrowdStrike Falcon（内嵌 Charlotte AI）、Darktrace（自学习威胁检测）、SentinelOne Singularity（Purple AI 内嵌）、Palo Alto Cortex（Precision AI）、Microsoft Defender（内嵌 AI）、Fortinet FortiAI、Check Point ThreatCloud AI 等国际产品，以及国内的深信服（EDR/NDR 内嵌 AI）、奇安信（天擎/天眼）、360（终端安全内嵌）、安恒信息（明御+恒脑）、亚信安全（XDR 内嵌）、启明星辰（天清系列内嵌）、美创科技、广东盈世、烽台科技、天懋信息、云弈科技、石犀科技等。

优势与局限：该形态的优势在于无需改变用户习惯、部署成本低、价值立竿见影。局限性则体现在可见度低，用户难以感知 AI 价值，定价溢价困难。

4.2.3 子市场规模与增速（2025-2030 预测）

综合 Grand View Research、Fortune Business Insights、Precedence Research 等机构数据，2025 年全球 AI 网络安全市场总规模约$26-36B，CAGR 约 20-24%。以下按场景簇拆分为估算值（注：子市场拆分为作者基于总量和各领域相对权重的推算，非机构直接发布）：

场景簇	**2025 规模（估算）**	**CAGR（估算）**	**成熟度**
威胁检测与响应	$8-10B	15-20%	成熟期
网络安全运营	$6-8B	24-28%	快速增长期
数据安全治理	$5-6B	20-24%	增长期
开发安全与供应链	$3-4B	22-26%	增长期
邮件与通信安全	$2-3B	16-20%	成熟期
渗透测试与漏洞挖掘	$1-2B	26-30%	早期
认知安全与鉴伪	$0.5-1B	28-35%	萌芽期
安全策略管理	$1-2B	20-24%	增长期
合计	**$26-36B**	**20-24%**	-

数据来源：Grand View Research（2024 基年$25.4B, CAGR 24.4%，按此推算 2025 约 $31.6B）、Fortune Business Insights（2025, $34.1B, CAGR 21.7%）、Precedence Research（2025, $29.6B, CAGR 18.9%）。子市场拆分为基于行业结构的估算。

关键洞察：从市场数据分析可以看出，**SecOps（网络安全运营）在大体量赛道中增速最快**，Agentic AI 驱动的 SOC 自动化成为最热赛道；**威胁检测规模最大**，传统优势领域，但边际增速放缓；**认知安全与鉴伪潜力最大**，虽规模小，但增速>30%，Deepfake 等新型威胁推动需求。

2030 年终值推算：按合计 20-24% CAGR 复合 5 年推算，2030 年全球 AI 网络安全市场预计达到$65-105B 区间，中位约$85B，相比 2025 年市场规模实现约 2.5-3 倍扩张。

4.3 AI 能力如何重构既有安全产品边界

AI 技术的引入不仅是功能增强,更是**产品边界的重构**和**价值链的重组**。以下分析 AI 如何改变主流安全产品的定位与能力范围。

4.3.1 从安全信息与事件管理（SIEM）到 AI 驱动的安全编排与响应（AI-Driven SOAR/XDR）

传统 SIEM 痛点：传统 SIEM 面临三大核心痛点。一是告警泛滥、误报率高（Gartner 数据显示，SOC 每日收到 11000+告警，其中 99%为误报）；二是需人工编写复杂查询语句（SPL、KQL 等）；三是缺乏自动化响应能力。

AI 重构路径:

首先是**自然语言查询**，实现从"编写 SPL"到"用自然语言提问"的转变。典型案例包括 Microsoft Sentinel + Security Copilot、Google Security Operations + Gemini 等，效果显著，查询效率提升 60-80%，降低技能门槛。

其次是**告警降噪与自动分类**，从"人工逐条研判"转向"AI 自动聚类和优先级排序"。Dropzone AI 宣称可自动调查 100%告警，只将高优先级呈现给人类，效果为误报率下降 30-90%（国内厂商数据）。

第三是**自动化响应编排**（SIEM → SOAR 融合），从"生成报告"转向"直接执行处置动作"。Torq HyperSOC 2o 宣称可自主关闭 90% Tier-1 告警，效果为 MTTR 从天级缩短到分钟级。

边界重构结果：从产品边界重构的结果看，**SIEM 不再是纯"检测平台"**，而是**"检测+响应+编排"的融合体**；**SIEM 与 SOAR 的边界模糊**，AI 能力使两者功能趋同；**新品类诞生**，出现"AI SOC Analyst" 或 "Agentic SIEM"等新产品类型。

4.3.2 从安全编排自动化响应（SOAR）到超自动化（HyperAutomation）

传统 SOAR 痛点：传统 SOAR 存在三大痛点。一是 Playbook 编写复杂，需要专业技能；二是缺乏灵活性，无法应对未知场景；三是部署周期长（12-18 个月才见效，如 Valvoline 案例）。

AI 重构路径：

首先是**自动生成 Playbook**，从"人工编写 YAML/Python"转向"自然语言描述需求，AI 生成工作流"。Torq 宣称可在 48 小时内实现 ROI（传统 SOAR 需 12-18 个月）。

其次是 **Multi-Agent 协同**，从"单一工作流"转向"多个专业化 Agent 协作"。Torq HyperSOC 2o 原生支持 MCP（Model Context Protocol），多 Agent 间可互操作。

第三是**动态调整响应策略**，从"静态 Playbook"转向"根据攻击演进实时调整"，基于强化学习优化响应路径。

边界重构结果：从产品边界重构来看，**SOAR 概念被"Hyperautomation"取代**，Torq 等厂商明确宣称"SOAR 已死"；**从编排工具到自主决策系统**，不再依赖人类定义的规则；**新品类诞生**，出现"Agentic SOAR" 或 "HyperSOC"等新产品形态。

4.3.3 从扩展检测与响应（XDR）到 AI 原生统一检测（AI-Native Unified Detection）

传统 XDR 痛点：传统 XDR 面临三大核心挑战。一是跨数据源关联依赖人工规则；二是缺乏对云原生、AI 工作负载的支持；三是检测逻辑无法快速适应新攻击手法。

AI 重构路径：

首先是**跨域智能关联**，从"规则引擎"转向"大模型推理攻击链"。典型案例包括 CrowdStrike Charlotte AI、SentinelOne Purple AI 等。

其次是 **AI 工作负载保护整合**，从"传统 IT 资产"扩展到"AI 模型+推理服务+数据管道"。案例包括 Palo Alto Networks AI Runtime Security、HiddenLayer AISec Platform 等。

第三是**自适应检测逻辑**，从"威胁情报更新检测规则"转向"AI 自学习新攻击模式"。Darktrace Enterprise Immune System（企业免疫系统）是该方向的典型代表。

边界重构结果：从产品边界重构的角度看，**XDR 不再局限于 Endpoint+Network+Email**，而是**"IT+OT+IoT+AI"的全域防护**；**从检测平台到自适应防御系统**，具备自我进化能力；**新品类诞生**，出现"AI-Native XDR" 或 "Extended AI Security"等新产品类别。

4.3.4 从终端/网络检测与响应（EDR/NDR）到自主响应（Autonomous Response）

传统 EDR/NDR 痛点：传统 EDR/NDR 存在三大痛点。一是检测到威胁后仍需人工研判和响应；二是对未知威胁（0day）检测能力弱；三是缺乏跨端点的协同防御。

AI 重构路径：

首先是**自主响应能力**，从"生成告警"转向"自动隔离、修复、回滚"。Darktrace Autonomous Response 可在分钟级自动阻断威胁。国内案例包括天懋信息、炼石网络等实现关键动作审批+自动执行。

其次是**未知威胁检测**，从"签名库匹配"转向"行为基线+异常检测"。Vectra AI 专注攻击者行为（而非单纯异常），误报率低。

第三是**协同防御**，从"单点防御"转向"全网联动"。CrowdStrike Falcon 平台跨 2 亿+端点共享威胁情报。

边界重构结果：从产品边界重构来看，**EDR/NDR 从"检测工具"升级为"自主防御系统"**；**响应能力成为核心竞争力**，不仅要"发现威胁"，更要"自动阻断"；**新品类诞生**，出现"Autonomous EDR/NDR"等新产品形态。

4.3.5 从云原生应用保护/云安全态势管理（CNAPP/CSPM）到 AI 驱动的云安全态势（AI-Driven Cloud Security Posture）

传统 CNAPP/CSPM 痛点：传统 CNAPP/CSPM 面临三大挑战。一是配置检查依赖人工编写规则；二是漏洞优先级排序不准确；三是修复建议不可执行。

AI 重构路径：

首先是**智能配置基线**，从"人工定义合规规则"转向"AI 学习最佳实践"。Palo Alto Prisma Cloud 内嵌 AI 配置检查是典型案例。

其次是**风险评分与优先级**，从"CVSS 静态评分"转向"结合业务上下文的动态评分"。Tenable Exposure Management 被 Gartner 称为"AI 驱动暴露评估标杆"。

第三是**自动化修复**，从"生成修复建议"转向"Agent 自动执行修复"。Furl（Agent 驱动的安全修复 startup，2026 年完成$10M 融资）是该方向的代表。

边界重构结果：从产品边界重构的角度看，**CNAPP/CSPM 从"检查工具"升级为"持续自愈系统"**；**从态势可见到主动治理**，不仅展示问题，更要自动解决；**新品类诞生**，出现"Agentic Exposure Management"等新产品类型。

4.3.6 数据安全态势管理（DSPM）到 AI 驱动的数据发现与保护（AI-Powered Data Discovery & Protection）

传统 DSPM 痛点：传统 DSPM 存在三大核心痛点。一是敏感数据识别依赖正则表达式和关键词；二是无法识别非结构化数据中的敏感信息；三是缺乏跨数据源的统一治理。

AI 重构路径：

首先是**智能数据分类分级**，从"关键词匹配"转向"NLP 语义理解+多模态识别"。炼石网络、美创科技、海云安等国内厂商已实现智能分类分级，效果为准确率提升 30-50%，效率提升 60-80%。

其次是**跨数据源统一治理**，从"单一数据库"扩展到"结构化+非结构化+多云"。Tenable DSPM 覆盖传统 IT+云+AI 资产。

第三是**智能脱敏与数据合成**，从"规则脱敏"转向"保留数据特征的智能脱敏"。炼石网络实现智能脱敏，美创科技支持数据合成。

边界重构结果：从产品边界重构来看，**DSPM 从"数据发现"扩展到"数据全生命周期治理"**；**与 DLP、数据库审计、脱敏产品边界模糊**；**新品类诞生**，出现"AI-Driven Data Security Fabric"等新产品形态。

4.3.7 跨产品边界融合趋势

AI 能力的引入打破了传统产品的技术壁垒,出现**平台化整合**趋势：

融合方向	原产品边界	AI 驱动的新边界	代表厂商
SIEM + SOAR + XDR	三类独立产品	统一的"AI SOC 平台"	Dropzone AI、7AI、Prophet Security；奇安信 AISOC、深信服安全 GPT、启明星辰泰合、绿盟智慧安全运营平台
EDR + NDR + ITDR	按检测位置划分	统一的"行为检测平台"	CrowdStrike、Microsoft Defender；奇安信天擎、360 终端安全、深信服 EDR、安天智甲
CNAPP + DSPM + CWPP	按防护对象划分	统一的"云安全中枢"	Palo Alto Prisma Cloud、Wiz；阿里云安全中心、

			腾讯云安全、青藤云安全、华为云 HSS
DLP + DSPM + 数据库审计	按功能划分	统一的"数据安全治理平台"	Cyera、BigID、Varonis；炼石网络、美创科技、安恒信息、闪捷信息

融合背后的逻辑：产品融合的驱动因素包括四个方面。一是**数据共享需求**，AI 模型训练需要跨产品的数据融合；二是**能力复用**，统一的 AI 引擎可服务于多个场景；三是**客户采购偏好**，"单一供应商"策略降低集成成本；四是**技术可行性**，大模型的泛化能力支持跨场景应用。

4.4 本章小结

本章明确界定了 AI for Security 市场的三大核心要素：

一是**市场边界**方面，以"是否利用 AI 技术提升安全能力"为核心判断标准，包含但不限于威胁检测、SecOps、数据安全、攻防对抗、开发安全、邮件安全、认知安全、策略管理等 8 大场景簇。与 Security for AI 的交集领域（AI 安全治理）纳入本市场范畴。

二是**能力定位**方面，市场中的厂商可分为"平台化赋能"、"单点插件"、"基座模型服务"三类，分别对应不同的战略选择和商业模式。

三是**产品重构**方面，AI 能力的引入正在重构 SIEM、SOAR、XDR、EDR/NDR、CNAPP/CSPM、DSPM 等既有产品的边界，核心方向是**从检测到响应**、**从辅助到自主**、**从单点到平台**。

第五章 AI for Security 市场描述

> **核心观点**：AI for Security 不是横空出世的新物种,而是网络安全与人工智能交织演进 40 余年的必然结果。从专家系统的"规则智能",到机器学习的"数据智能",再到大模型时代的"知识智能",技术范式的每一次跃迁都在解决上一代的痛点,同时催生新的挑战。2026 年,市场正在经历从"能力堆砌"到"工程闭环"的关键转折——AI 不再是孤立的检测引擎或聊天机器人,而是具备自主决策、工具编排、审计追溯能力的智能体（Agent）。本章将追溯这一演进历程,解析大模型带来的范式突破,并阐明 2026 年工程实践中的核心矛盾与破局方向。

5.1 历史脉络：从专家系统到大模型

5.1.1 专家系统时代（1980s-1990s）：规则驱动的"知识工程"

技术特征：**核心思想**：将安全专家的经验编码为"IF-THEN"规则,构建知识库驱动的推理引擎。

代表技术：该时代的代表技术主要包括三类。一是 IDES（Intrusion Detection Expert System, 1987），由 SRI International 开发的基于规则的入侵检测系统,被视为 IDS 的鼻祖；二是 P-BEST（1989），用于 Unix 系统的专家系统,通过规则匹配检测异常行为；三是知识表示方法,即规则库（Rule Base）加推理引擎（Inference Engine）的组合架构。

典型应用：专家系统在安全领域的典型应用主要体现在三个方向。首先是签名库匹配,代表产品如 Snort（1998 年发布,基于规则引擎）；其次是漏洞扫描,如 ISS（Internet Security Systems,1994 年创立,基于漏洞规则库）；第三是防火墙策略,如 Checkpoint FireWall-1（1994 年,基于规则引擎）。

局限性：专家系统时代存在四大主要局限性。一是规则爆炸问题,攻击手法快速演进导致规则库维护成本高；二是无法应对未知威胁,只能检测已知攻击模式,对 0day 攻击无能为力；三是误报率高,规则过于死板,缺乏上下文理解；四是对专家的依赖性强,规则编写需要安全专家参与,难以规模化。

历史意义：专家系统时代具有三方面的重要历史意义。首先,它奠定了"知识驱动"的安全理念；其次,形成了威胁情报（规则库）的雏形；第三,为后续机器学习方法提供了对比基准。

5.1.2 机器学习时代（2000s-2015）：数据驱动的"统计智能"

技术演进路径：**早期探索（2000-2010）**：这一阶段的探索主要集中在两大方向。一是异常检测,基于统计模型（如高斯分布、聚类算法）检测偏离基线的行为,典型案例包括 2003 年 Symantec 推出基于启发式的反病毒引擎,以及 2005 年 Arbor Networks 的 DDoS 检测引入机器学习；二是垃圾邮件过滤,贝叶斯分类器（Naive Bayes）成为标配,如 2004 年 Gmail 推出基于贝叶斯的垃圾邮件过滤,准确率超过 99%。

商业化浪潮（2010-2015）：这一时期机器学习技术开始大规模商业化应用,主要体现在两大领域。一是端点行为分析,机器学习取代签名库成为 EDR 核心,典型案例包括 2011 年 CrowdStrike 创立,首次提出"Endpoint Detection and Response"概念,核心技术是行为分析加机器学习,以及 2013 年 Carbon Black（现 VMware Carbon Black）推出 CB Response,基于进程行为建模；二是网络流量分析,如 2013 年 Darktrace 创立,提出"企业免疫系统"概念,使用无监督学习建立网络行为基线,2014 年 Vectra Networks（现 Vectra AI）推出 AI 驱动的 NDR 产品。

核心技术栈：机器学习时代的核心技术包括三大类。一是监督学习,随机森林（Random Forest）、支持向量机（SVM）、逻辑回归被广泛用于恶意软件分类；二是无监督学习,K-Means、DBSCAN 被用于异常检测和攻击者行为聚类；三是特征工程,人工提取特征（如文件哈希、API 调用序列、网络流特征）成为关键瓶颈。

代表性厂商与产品

时间	厂商/产品	核心技术	市场影响
2009	**360 QVM（奇虎 360）**	机器学习驱动的启发式反病毒引擎（QVM，Qihoo Virtual Machine）	2010 年集成于"360 安全卫士"，中国率先将机器学习引擎规模化部署于亿级终端
2011	**CrowdStrike Falcon**	行为分析+机器学习	定义 EDR 市场,2019 年 IPO,市值超$600 亿（2025）
2013	**Darktrace**	无监督学习的异常检测	定义"企业免疫系统"概念,2021 年 IPO
2013	**Carbon Black**	进程行为建模	2019 年被 VMware 以$21 亿收购
2014	**Vectra AI**	攻击者行为检测（非单纯异常）	2025 年被评为 NDR 领域 Leader（Gartner, GigaOm）
2014	**Cylance**	纯 AI 驱动的反病毒引擎	2019 年被 BlackBerry 以 $14 亿收购

局限性（为何未能彻底解决问题）：机器学习时代虽然取得显著进展,但仍面临五大核心局限性。一是特征工程瓶颈,需要安全专家和数据科学家协作提取特征,如恶意软件检测需手工提取 API 调用序列、字符串特征等,工作量巨大；二是泛化能力弱,模型对训练数据分布高度敏感,攻击者稍作变形即可绕过,案例如 Cylance 的 AI 引擎曾被证明可通过在恶意软件中嵌入良性字符串绕过检测；三是可解释性不足,"黑盒"模型难以向安全人员解释决策逻辑,导致误报处理困难,SOC 人员对 AI 结果信任度低；四是依赖标注数据,监督学习需要大量标注样本（恶意/正常）,新型攻击（APT、0day）缺乏训练数据；五是实时性挑战,复杂模型推理耗时长,难以满足毫秒级响应需求,早期机器学习引擎在端点运行时 CPU 占用率可达 20-30%。

历史意义：机器学习时代具有三方面的重要历史意义。首先,证明了"数据驱动"优于"规则驱动"的技术路线；其次,培育了一批 AI 原生安全公司（CrowdStrike, Darktrace 等）；第三,为深度学习时代积累了数据资产和应用场景。

5.1.3 深度学习时代（2015-2020）：表征学习的"感知智能"

技术突破：**核心变化**：从人工特征工程转向端到端学习（End-to-End Learning）。

关键技术：深度学习时代的关键技术主要包括四类。一是卷积神经网络（CNN）,应用于恶意软件可视化检测,如将 PE 文件转换为灰度图,用 CNN 识别恶意软件家族（2016 年起流行）；二是循环神经网络（RNN/LSTM）,应用于日志序列异常检测、网络流量时序分析,案例如 2017 年 IBM Watson for Cyber Security 引入 LSTM 分析威胁情报；三是 Transformer 初探,2017 年 Transformer 论文发布,但在安全领域应用有限（主要用于 NLP 预处理）；四是

生成对抗网络（GAN）,用于对抗样本生成和异常检测,2018 年多篇学术论文探索 GAN 在 IDS 中的应用。

商业化进展：**新兴应用**：深度学习在安全领域催生了多个新兴应用方向。一是恶意软件家族分类,深度学习将准确率从 85%提升至 95%以上；二是钓鱼邮件检测,如 2019 年 Abnormal Security 创立,宣称使用"行为 AI"（实为深度学习加传统 ML 混合）检测 BEC 攻击；三是威胁情报自动化,如 2015 年 Recorded Future 创立,使用 NLP 加深度学习从海量文本提取威胁情报；四是漏洞挖掘辅助,如 2019 年 Mayhem（ForAllSecure 公司）赢得 DARPA Cyber Grand Challenge,使用深度强化学习自动化漏洞挖掘。

平台化趋势：深度学习开始嵌入云安全平台成为普遍趋势。案例包括 2018 年 Microsoft Azure Security Center 引入机器学习检测异常活动,以及 2019 年 AWS GuardDuty 升级,使用深度学习检测威胁。

局限性（为何仍未成为主流）：深度学习虽然技术先进,但仍面临四大局限性导致未成为主流。一是训练成本高,深度学习需要 GPU 算力和海量数据,中小厂商难以负担（国内厂商普遍反馈算力成本是主要阻碍）；二是可解释性更差,相比传统 ML,"黑盒"程度加深,监管和审计需求难以满足（如金融、医疗行业）；三是对抗攻击风险,对抗样本（Adversarial Examples）攻击兴起,如 2019 年研究证明可通过微小扰动绕过恶意软件检测模型；四是泛化到新场景困难,模型在训练数据分布内表现优异,但对未见过的攻击手法泛化能力有限,案例如疫情期间远程办公场景激增,原有网络行为模型失效,误报率飙升。

历史意义：深度学习时代具有三方面的重要历史意义。首先,证明了端到端学习的可行性；其次,为大模型时代的 Transformer 架构铺路；第三,暴露了"纯技术驱动"的不足——缺乏业务知识和工程闭环。

5.1.4 知识图谱时代（2018-2022）：结构化知识的"推理智能"

技术特点：**核心理念**：将安全领域知识（威胁情报、攻击技战法、资产关系等）结构化为图谱,利用图算法进行推理。

关键技术：知识图谱时代的关键技术主要包括两类。一是 ATT&CK 框架（2018 年 MITRE 正式发布）,将攻击者行为标准化为战术（Tactics）和技术（Techniques）,成为威胁检测和情报共享的事实标准；二是图神经网络（GNN）,用于攻击路径推演、横向移动检测,如 2020 年多篇学术论文探索 GNN 在 APT 检测中的应用。

商业化进展：**ATT&CK 驱动的产品创新**：ATT&CK 框架推动了三大类产品创新。一是威胁情报平台,如 Recorded Future、ThreatConnect 将情报与 ATT&CK 框架关联；二是 SIEM/SOAR 增强,如 2019 年 Splunk 推出 ATT&CK 应用,自动将告警映射到 ATT&CK 技术；三是红蓝对抗平台,如 2020 年 AttackIQ 推出基于 ATT&CK 的自动化红队测试平台。

知识图谱应用：知识图谱技术在安全领域的应用主要体现在两个方向。一是资产关系分析,如 JupiterOne（2018 年创立）构建云资产知识图谱,用图查询语言（Gremlin）分析攻击面；二是攻击链重构,如 SentinelOne 的 Storyline 技术（2019 年推出）自动构建攻击事件的因果链。

国内实践：国内厂商在知识图谱方面的实践主要包括两类。一是知道创宇、绿盟科技等厂商将 ATT&CK 框架融入威胁情报和检测规则；二是炼石网络、美创科技等数据安全厂商构建数据资产知识图谱。

局限性：知识图谱技术存在三大核心局限性。一是知识构建成本高,人工维护知识图谱工作量巨大,且知识老化快（新攻击手法层出不穷）；二是覆盖不完整,ATT&CK 主要覆盖企业 IT 环境,对云原生、OT、IoT、AI 工作负载支持有限,2025 年 ATT&CK 才新增"Cloud"和"ICS"矩阵；三是推理能力受限,基于图的推理依赖预定义规则,难以应对攻击者的创新战术。

历史意义：知识图谱时代具有三方面的重要历史意义。首先,将安全知识从"隐性"转为"显性"；其次,为大模型时代的 RAG（检索增强生成）奠定基础；第三,证明了"知识加推理"的价值。

5.1.5 大模型时代（2022-至今）：生成式 AI 的"认知智能"

技术奇点：ChatGPT 的震撼：**2022 年 11 月 30 日**：OpenAI 发布 ChatGPT,标志着生成式 AI（Generative AI）进入大众视野。

对安全行业的冲击：ChatGPT 的发布对安全行业产生了双向冲击。一方面是攻击者武器化,2023 年起,ChatGPT 被用于生成钓鱼邮件、恶意代码、社工脚本,2023 年 CrowdStrike 报告显示使用 ChatGPT 生成的钓鱼邮件点击率提升 30%；另一方面是防御者工具化,2023 年起,安全厂商竞相发布基于大模型的安全助手,时间线上 2023 年 3 月 Microsoft 发布 Security Copilot（基于 GPT-4）,2023 年 5 月 Google 宣布 Sec-PaLM（安全专用大模型）,2023 年下半年 CrowdStrike、Palo Alto、SentinelOne 等相继推出 GenAI 能力。

技术范式突破：**1. 自然语言理解与生成**

突破点：安全人员可用自然语言与系统交互,无需学习查询语言。

典型应用：自然语言理解与生成在安全领域的典型应用主要包括三个方向。一是自然语言查询威胁,如 Microsoft Security Copilot 可接受"过去 24 小时内有哪些失败的登录尝试?"并自动生成 KQL 查询,SentinelOne Purple AI 支持"Show me all endpoints with critical vulnerabilities in production"；二是告警解读与摘要,如 Dropzone AI 可将数百条原始日志自动生成人类可读的事件摘要,天懋信息、摄星科技、炼石网络等厂商实现了告警摘要功能；三是报告自动生成,如 Trellix Wise 可自动生成调查报告包括攻击时间线、影响范围、处置建议,烽台科技、绿盟科技等实现了报告自动生成。

价值：自然语言理解与生成带来三方面价值。一是降低技能门槛,非专家也能完成复杂查询；二是提升效率,查询和分析时间从小时级降至分钟级；三是工时效益显著,国内厂商数据显示工时下降 10-50%（问卷汇总数据）。

2. 多任务学习与迁移能力

突破点：单一模型可处理多种安全任务,无需针对每个任务单独训练。

典型应用：多任务学习与迁移能力在安全领域的典型应用包括两个方向。一是跨场景知识迁移,如 Google Gemini 在威胁情报分析、代码审计、策略优化等场景间无缝切换,绿盟科技的 AI 平台支持威胁检测、SecOps、数据安全、渗透测试等多场景复用；二是 Few-Shot Learning,只需少量样本即可适应新攻击类型,如 Microsoft Security Copilot 可根据用户反馈快速调整检测逻辑。

价值：多任务能力带来三方面价值。一是降低训练成本,无需为每个场景单独标注数据；二是快速适应新威胁,攻击手法变化时无需重新训练；三是泛化能力提升,国内厂商反馈显示泛化能力提升但仍存在"幻觉"问题。

3. 推理与逻辑链构建

突破点：大模型具备一定的因果推理能力,可构建攻击链和处置逻辑。

典型应用：推理与逻辑链构建在安全领域的典型应用包括三个方向。一是攻击路径推演,如 CrowdStrike Charlotte AI 可推演攻击者的下一步行动,摄星科技、绿盟科技实现了攻击路径推演；二是根因分析（RCA）,如 SentinelOne Purple AI 可追溯告警的根因（如初始访问方式、权限提升路径）,Palo Alto XSIAM 提供了 AI 驱动的 RCA；三是威胁情报关联,如 Recorded Future 使用 LLM 自动关联多源情报生成可操作的威胁简报。

价值：推理与逻辑链构建带来三方面价值。一是提升检测深度,从"发现异常"到"理解攻击意图"；二是加速响应,根因明确后可直接处置,无需人工逐层排查；三是响应效率显著提升,国内厂商数据显示 MTTR 改善 10-50%（问卷汇总数据）。

4. 代码生成与自动化

突破点：大模型可生成查询语句、修复脚本、检测规则,甚至攻击代码（红队用）。

典型应用：代码生成与自动化在安全领域的典型应用包括三个方向。一是查询语句生成,如 Microsoft Security Copilot 可生成 KQL 查询,Google Security Operations 可生成 YARA-L 检测规则；二是修复代码生成,如 Snyk Agent Fix 可自动生成漏洞修复代码并提交 PR,Furl

（Agentic Remediation startup）可自动化执行漏洞修复；三是渗透测试脚本,如知道创宇的AI Pentester 可生成渗透测试脚本,海云安、绿盟科技实现了漏洞挖掘辅助。

价值：代码生成与自动化带来三方面价值。一是降低专业门槛,非开发人员也能生成代码；二是提升自动化率,从"生成建议"到"直接执行"；三是代码质量显著提升,国内厂商数据显示海云安宣称千行代码漏洞率下降 50%。

5. AI 驱动的攻防新范式

攻击侧：AI 驱动的攻击手段主要体现在三个方向。一是社工工程升级,如 2024 年起攻击者使用 ChatGPT 生成高度个性化的钓鱼邮件规避传统检测,以及 Deepfake 语音/视频用于CEO 欺诈（BEC 2.0）；二是漏洞挖掘自动化,如 2024 年 ZeroGPT（虚构案例,但类似工具已在暗网流传）可自动化寻找 0day 漏洞；三是对抗样本攻击,攻击者针对 AI 检测模型生成对抗样本,绕过检测。

防御侧：防御侧的 AI 能力主要体现在两个方向。一是 AI Red Team,Microsoft PyRIT、HiddenLayer AutoRTAI 等工具可自动化测试 AI 系统安全性,SuperClaw（2026 年 2 月开源）则是针对 AI Agent 的红队框架；二是 AI 运行时保护,HiddenLayer、Noma Security、Operant AI 等提供 AI 模型和 Agent 的运行时防护,Palo Alto AI Runtime Security 保护 AI 工作负载。

新平衡：攻防双方都在利用 AI,形成"AI vs AI"的新对抗格局。

大模型时代的代表性厂商（2023-2026）

时间	厂商/产品	核心能力	融资/估值

2023.3	**Microsoft Security Copilot**	基于 GPT-4 的安全助手，横跨全栈安全产品	微软内部产品
2023.5	**Google Gemini for Security**	原生多模态能力,整合 Chronicle 和 Mandiant	Google 内部产品
2023.5	**深信服安全 GPT**	国内首个安全 GPT 大模型，赋能 SIEM、XDR、SOC 场景，推出 AI 安全助手	深交所上市公司
2023.7	**360 安全大模型（奇虎 360)**	基于 360GPT，构建“智能安全大脑”，演进为 360 AISOC 安全运营智能体	深交所上市公司
2023.8	**奇安信 Q-GPT / QAX-GPT**	国内安全行业首个百亿级专有大模型，后升级为“奇安信 AISOC”平台	上交所科创板上市公司
2023.9	**安恒信息“恒脑”**	垂直领域安全大模型，赋能 AiLPHA 态势感知、AiGuard 数据安全等产品线	上交所科创板上市公司
2024.2	**Dropzone AI**	AI SOC Analyst,自主调查 100%告警	$37M B 轮（2025.7）
2024.6	**7AI**	Agentic SOC,动态推理能力	$130M A 轮（2025.12），估值$7 亿
2025.3	**Prophet Security**	Multi-Agent SOC 平台	$30M A 轮（2025 夏）

2025.6	**Torq HyperSOC 2o**	首个 MCP 原生安全编排平台	$140M D 轮（2025.1），估值$12 亿
2025.9	**Snyk AI Trust Platform**	AI 驱动的开发安全+AI 安全治理	上市公司
2026.1	**WitnessAI**	AI Agent 安全监控和治理	$58M 新轮（2026.1）

国内厂商进展：国内厂商在大模型时代的进展呈现四个梯队。第一梯队为头部安全厂商的安全大模型与 AI SOC 平台布局，代表包括深信服"安全 GPT"、奇安信"QAX-GPT/AISOC"、360"安全大模型/智能安全大脑"、安恒信息"恒脑"等，主攻 AI SOC、威胁检测、数据安全治理、终端安全等场景；第二梯队为产品化平台阶段（基于问卷汇总），包括海云安、烽台科技、绿盟科技、云弈科技共 4 家；第三梯队为规模交付阶段，包括天懋信息、美创科技共 2 家；第四梯队为试点交付阶段，包括瀛云科技、摄星科技、炼石网络、和利时、知道创宇、宁数安全、石犀科技、广东盈世共 8 家。

技术演进总结：| 时代 | 核心范式 | 代表技术 | 典型应用 | 局限性 |

专家系统	**规则驱动**	**IF-THEN 规则库**	**签名库匹配、Snort**	**无法应对未知威胁，规则维护成本高**
机器学习	数据驱动	随机森林、SVM	EDR 行为分析、垃圾邮件过滤	特征工程瓶颈,泛化能力弱,可解释性不足
深度学习	表征学习	CNN、RNN/LSTM	恶意软件分类、日志异常检测	训练成本高,可解释性更差,对抗攻击风险

知识图谱	知识推理	ATT&CK、GNN	攻击链重构、资产关系分析	知识构建成本高,覆盖不完整,推理能力受限
大模型	认知智能	Transformer、LLM	自然语言交互、代码生成、Agent 自主决策	幻觉/不确定性、算力成本、提示注入风险

演进逻辑：AI for Security 的技术演进遵循一个核心逻辑,即每一代技术都在解决上一代的核心痛点,但同时也引入了新的挑战。大模型时代的核心突破体现为两个维度的升级：从"感知"到"认知"，从"检测"到"决策"。

5.2 前大模型时代的痛点：为何 AI 安全难以规模化

尽管 AI 技术在安全领域应用已有 20 余年,但直到大模型时代之前,AI 安全产品的规模化应用仍面临诸多障碍。这些痛点不仅是技术问题,更是工程实践、组织能力、商业模式的综合挑战。

5.2.1 误报率困境：准确率与召回率的"不可能三角"

问题表现：**核心矛盾**：提升检测覆盖率（Recall）必然导致误报率（False Positive Rate）上升,降低误报则会漏报真实威胁。

数据支撑：多方数据揭示了误报率困境的严重程度。一是 Gartner 数据（2023）显示平均每个 SOC 每天收到 11000 以上告警,其中 99%为误报；二是 Ponemon Institute 调研（2022）发现 SOC 分析师每天花费 25%时间处理误报,导致告警疲劳（Alert Fatigue）；三是国内厂商数据（问卷汇总）反映误报率是客户驱动 AI 采购的首要痛点（10 家厂商提及）。

案例：典型的误报问题体现在多个场景。端点检测方面,CrowdStrike 早期版本曾因误报导致客户投诉,直到引入 Charlotte AI 才显著改善。云安全方面,传统 CSPM 工具对云配置检查过于严格,单一租户可产生数千条告警,其中大部分非安全风险。国内案例中,某金融客户部署传统 NDR 后,每日产生 3000 以上告警,SOC 团队疲于应付,最终放弃使用。

根因分析：误报率困境的根本原因主要有三个方面。一是缺乏业务上下文,传统 ML 模型只看行为特征,不理解业务逻辑,如研发人员深夜访问代码库是正常行为,却被误判为"异常访问"；二是阈值难以调优,阈值过高导致漏报,阈值过低导致误报,需要安全专家长期调优,成本高昂；三是攻击者对抗,攻击者通过"低慢"攻击（Low and Slow）绕过检测阈值,对抗样本攻击直接欺骗 ML 模型。

5.2.2 数据可用性与质量：AI 的"燃料危机"

问题表现：**数据缺失**：数据缺失问题体现在三个维度。一是标注数据稀缺,恶意样本（尤其是 APT、0day）极难获取,标注成本高昂；二是数据孤岛,客户数据无法共享,厂商难以积累跨客户的训练数据；三是国内厂商反馈（问卷汇总）显示,数据可用性/数据质量是最大技术阻碍（9 家厂商提及）。

数据质量问题：数据质量问题主要表现在三个方面。一是噪声数据,日志中包含大量无关信息,影响模型训练；二是标注错误,人工标注样本存在主观偏差；三是数据老化,攻击手法快速演进,历史数据迅速失效。

案例：典型案例充分说明了数据质量问题的严重性。Cylance 数据投毒事件（2019）中,研究人员证明可通过在恶意软件中嵌入良性字符串绕过 AI 检测,暴露训练数据质量问题。国内案例方面,某电力客户的工控网络环境独特,厂商通用模型效果差,需现场采集数据

重新训练,周期长达 6 个月。此外,国内问卷汇总中石犀科技反馈"客户现场的实测数据由于安全顾虑无法直接使用"。

根因分析：数据可用性与质量问题的根本原因主要包括三个方面。一是隐私与合规约束,GDPR、PIPL 等法规限制数据跨境和共享,客户担心数据泄露,不愿将日志上传到厂商云端；二是标注成本高,安全样本标注需要专家参与,单条样本标注成本可达$10-$100,APT 攻击样本极其稀缺,市场上甚至出现"样本交易"灰产；三是厂商能力不足,中小厂商缺乏数据清洗和治理能力,国内厂商问卷显示自评"安全数据"能力≥4 分的仅 5 家。

5.2.3 泛化能力弱：模型的"脆弱性"

问题表现：**场景迁移失效**：模型泛化能力弱首先体现在场景迁移失效,在 A 客户环境训练的模型,部署到 B 客户环境后效果急剧下降。典型案例是某 NDR 产品在互联网公司表现优异（误报率<5%）,但在制造业客户误报率飙升至 40%（工控协议与 Web 流量特征差异巨大）。

时间漂移：泛化能力弱还体现在时间维度上,模型在训练时表现良好,但随时间推移效果下降（Concept Drift）。典型案例是疫情期间远程办公激增,原有网络行为基线失效,误报率普遍上升 30-50%。

对抗攻击脆弱：模型对对抗攻击的脆弱性也暴露了泛化能力不足,攻击者可通过微小扰动绕过检测模型。案例如 2020 年研究证明可通过在恶意软件中添加无害字节绕过深度学习检测器。

根因分析：模型泛化能力弱的根本原因主要有三个方面。一是训练数据分布窄,模型在有限场景数据上训练,泛化到新场景能力差,深度学习模型尤其依赖"i.i.d."假设（独立同分布）,现实中难以满足；二是缺乏常识推理,传统 ML/DL 模型缺乏对攻击意图的理解,只能识别"表面特征",无法推理"深层逻辑"；三是模型更新困难,重新训练模型耗时长（数周到数月）,增量学习（Incremental Learning）在安全领域应用有限。

5.2.4 可解释性不足：AI 的"黑盒"困境

问题表现：**决策不透明**：可解释性不足首先体现在决策不透明,SOC 人员无法理解 AI 为何判定某行为为恶意,难以向管理层或客户解释。案例如某银行使用 AI 检测内部威胁,模型标记某高管为"高风险",但无法解释原因,导致误会和投诉。

审计与合规困难：可解释性不足还导致审计与合规困难,金融、医疗等强监管行业要求 AI 决策可追溯、可解释。案例如欧盟 GDPR 要求"算法决策的解释权"（Right to Explanation）,传统 ML 模型难以满足。

误报处理成本高：可解释性不足的第三个问题是误报处理成本高,SOC 人员无法判断 AI 告警是否可信,仍需人工逐条排查,导致 AI 价值大打折扣。

根因分析：可解释性不足的根本原因主要有三个方面。一是模型复杂度高,深度学习模型参数量达百万至亿级,内部机制难以解释,集成学习（如随机森林）的决策逻辑也难以直观呈现；二是可解释性与性能矛盾,简单模型（如决策树）可解释性好但性能差,复杂模型（如深度神经网络）性能好但不可解释；三是可解释 AI 技术不成熟,LIME、SHAP 等方法仅提供局部解释,无法解释全局决策逻辑,在安全领域应用有限。

5.2.5 实时性与性能成本：边缘计算的挑战

问题表现：**推理时延高**：实时性与性能成本问题首先体现在推理时延高,复杂深度学习模型推理耗时长,难以满足毫秒级响应需求。案例如某 EDR 产品的 DL 引擎单次文件扫描耗时 200-500ms,在高负载场景下 CPU 占用率达 30%,影响业务系统性能。

算力成本高：第二个问题是算力成本高,GPU 推理成本高昂,尤其在边缘设备（端点、IoT）部署困难。案例如某制造企业部署 AI 驱动的工控安全产品,需为每个边缘网关配备 GPU,成本增加 3 倍。

云端延迟：第三个问题是云端延迟,将数据上传到云端推理存在网络延迟,且存在隐私风险。案例如某医疗客户要求 AI 模型部署在本地,但本地算力不足,导致项目搁置。

国内厂商反馈（问卷汇总）：国内厂商的反馈印证了实时性与性能成本是主要技术阻碍（4 家厂商提及）,和利时反馈表示"主要用来完成非实时性问题"。

根因分析：实时性与性能成本问题的根本原因主要有三个方面。一是边缘算力受限,端点设备、IoT 设备算力有限,无法运行大型模型,工控环境对稳定性要求高,GPU 故障率高；二是模型压缩技术不成熟,量化、剪枝、蒸馏等技术在安全领域应用有限,压缩后模型性能下降明显；三是实时推理需求与成本矛盾,威胁检测要求毫秒级响应,但复杂模型推理慢,云端推理延迟高,边缘推理成本高。

5.2.6 AI 自身安全：攻击者的新目标

问题表现：**对抗攻击**：AI 自身安全问题首先体现在对抗攻击,攻击者通过对抗样本绕过 AI 检测模型。案例如 2019 年研究证明可通过在图像中添加不可见扰动绕过人脸识别,2020 年类似技术被应用于恶意软件检测。

模型窃取：第二个问题是模型窃取,攻击者通过 API 查询逆向推断模型参数。案例如 2021 年研究证明可通过数千次查询窃取商业 ML 模型。

数据投毒：第三个问题是数据投毒,攻击者在训练数据中注入恶意样本,污染模型。案例如 2018 年研究证明可通过投毒使垃圾邮件过滤器失效。

提示注入（大模型时代新增）：第四个问题是提示注入,这是大模型时代的新增风险,攻击者通过精心构造的提示词绕过 AI 安全策略。案例如 2023 年起,大量"越狱"提示词在社交媒体传播,ChatGPT 等模型多次被绕过。

国内厂商反馈（问卷汇总）：国内厂商对提示注入/模型安全已有警觉,有 2 家厂商将其视为新兴阻碍,摄星科技、宁数安全明确提及此风险。

根因分析：AI 自身安全问题的根本原因主要有三个方面。一是 AI 系统缺乏鲁棒性,模型对输入扰动高度敏感,缺乏对抗训练和鲁棒性验证；二是 AI 安全研究滞后,攻击技术发展快于防御技术,AI 安全测试标准缺失；三是供应链风险,开源模型和数据集可能被投毒,模型部署环境（容器、API）存在漏洞。

5.2.7 人才稀缺：复合型人才的"荒漠"

问题表现：**技能鸿沟**：人才稀缺问题首先体现在技能鸿沟,AI 技术（机器学习、深度学习）与安全领域知识（威胁情报、攻防技术）跨度大,既懂 AI 又懂安全的复合型人才极其稀缺。

数据支撑：人才稀缺的数据支撑主要来自两方面。一是国内厂商反馈（问卷汇总）显示,缺懂 AI 和 AI 开发的人是主要阻碍（5 家厂商提及）；二是关于(ISC)²等机构的"全球缺口 400 万"数据需审慎看待,发布方多为培训/认证利益相关方，存在夸大需求的动机,实际上中国网安专业大规模扩招后，基础岗位已出现供大于求，真正稀缺的是安全加 AI 复合型高端人才。

案例：典型案例反映了人才稀缺的现实困境。某安全厂商招聘"AI 安全工程师",要求同时具备 ML 算法、威胁建模、攻防实战经验,半年未招到合适人选。国内某金融客户希望自建 AI 安全能力,但数据科学家不懂安全,安全分析师不懂 AI,项目推进困难。

根因分析：人才稀缺问题的根本原因主要有三个方面。一是教育体系滞后,高校网络安全专业缺乏 AI 课程,AI 专业缺乏安全实战训练；二是经验积累慢,AI 安全需要长期实践经验（数年到十年）,人才培养周期长；三是薪资竞争,AI 人才被互联网大厂高薪挖走,安全行业薪资竞争力不足。

5.3 大模型带来的新可能性：从"感知"到"认知"

大语言模型（Large Language Models, LLMs）和生成式 AI（Generative AI）的出现,为解决前述痛点提供了新的技术路径。其核心突破在于从**"模式识别"升级为"认知推理"**,从**"单一任务"扩展到"多任务泛化"**。

5.3.1 自然语言理解：打破技能壁垒

技术突破：**核心能力**：大模型可理解和生成自然语言,使非专家也能完成复杂安全任务。

典型应用：自然语言理解与生成在安全领域的典型应用包括三个方向。

一是自然语言查询威胁。传统方式中,SOC 分析师需学习 SIEM 查询语言（如 Splunk SPL、Microsoft KQL、Google YARA-L），学习曲线陡峭；大模型方式则是用自然语言提问,AI 自动生成查询语句,如 Microsoft Security Copilot 可回应"Show me all failed login attempts in the past 24 hours from external IPs"自动生成 KQL 查询,Google Gemini for Security Operations 可根据"Create a YARA-L rule to detect ransomware behavior"自动生成检测规则。效果是查询效率提升 60-80%（微软内部数据），非专家也能完成 Tier-2 级任务。

二是告警解读与分类。传统方式中,SOC 分析师逐条阅读原始日志,判断真/假阳性,耗时长；大模型方式则是 AI 自动解读告警,生成人类可读的摘要和处置建议,如 Dropzone AI 可生成"This alert is a true positive. Attacker used stolen credentials to access sensitive database. Recommended action: Isolate the compromised account and initiate forensic investigation."的分析,天懋信息、摄星科技、炼石网络、烽台科技、石犀科技、绿盟科技等国内厂商实现了告警摘要功能。效果是告警分类准确率超过 90%（Dropzone AI 宣称），国内厂商数据显示人工工时下降 10-50%（问卷汇总）。

三是安全知识问答。传统方式中,SOC 人员查阅威胁情报、ATT&CK 框架、产品文档,耗时长；大模型方式则是直接提问,AI 从知识库检索并生成答案,如 CrowdStrike Charlotte AI 可回应"What is the latest ransomware targeting financial institutions?"自动检索威胁情报并总

结,天懋信息、摄星科技、海云安、和利时、知道创宇、宁数安全、石犀科技、绿盟科技、云弈科技等国内厂商实现了研判问答功能。效果是知识查询时间从分钟级降至秒级,降低对专家的依赖。

价值分析：自然语言理解与生成带来三方面重要价值。一是降低技能门槛,初级分析师也能完成高级任务,缓解人才短缺；二是提升工作效率,减少重复性劳动,专注于复杂决策；三是知识民主化,安全知识不再是少数专家的特权。

5.3.2 多任务能力：一个模型,多个场景

技术突破：**核心能力**：大模型通过预训练学习通用知识,可快速迁移到多个安全任务,无需针对每个任务单独训练。

典型应用：多任务学习与迁移能力在安全领域的典型应用包括三个方向。

一是跨场景知识迁移。Google Gemini 可在以下场景间无缝切换：威胁情报分析（从文本提取 IoC）、代码审计（识别漏洞）、策略优化（检测防火墙规则冲突）、日志分析（识别异常模式）。效果是单一模型替代多个专用模型,降低维护成本。

二是 Few-Shot Learning（少样本学习）。传统方式中,新攻击类型出现时,需收集大量样本（数千到数万）重新训练模型；大模型方式则是只需少量样本（几个到几十个）即可适应新威胁,如 Microsoft Security Copilot 可根据用户反馈（几次纠正）快速调整检测逻辑,Abnormal Security 的行为 AI 可在几天内学习新客户的邮件通信模式。效果是新威胁响应时间从周级缩短到天级或小时级,降低对标注数据的依赖。

三是多模态能力。大模型（如 GPT-4V、Gemini）可处理文本、图像、代码、日志等多种模态数据,如 Google Gemini 可分析网络拓扑图（图像）加日志（文本）加流量包（二进制）综合判断攻击路径,美创科技、烽台科技、广东盈世、绿盟科技等国内厂商实现了多模态分析（文本加图像加流量）。效果是单一模型替代多个专用模型（如图像分类器加文本分析器），提升检测准确率（多模态特征融合）。

价值分析：多任务能力带来三方面价值。一是降低训练成本,无需为每个场景单独标注数据和训练模型；二是快速适应新威胁,Few-Shot 能力使模型可快速学习新攻击手法；三是提升泛化能力,预训练的通用知识提升模型在新场景的表现。

5.3.3 推理逻辑：从"What"到"Why"和"How"

技术突破：**核心能力**：大模型具备一定的因果推理和逻辑链构建能力,可回答"为什么"和"如何做"。

典型应用：

推理与逻辑链构建在安全领域的典型应用包括四个方向。

一是**攻击链重构与攻击意图推理。**传统方式基于规则拼接攻击事件（如 MITRE ATT&CK 链）,但规则覆盖有限；大模型方式则通过推理能力自动构建完整攻击链并推断攻击者意图。案例如 CrowdStrike Charlotte AI 可分析多个告警后推断"Attacker likely conducting ransomware attack. Next steps: data exfiltration and encryption.",SentinelOne Storyline 加 LLM 可自动构建攻击事件的因果链（初始访问→权限提升→横向移动→数据窃取）,国内摄星科技、绿盟科技也实现了攻击路径推演。效果是提升检测深度,从"发现异常"到"理解攻击意图",并降低误报（基于上下文判断）。

二是**根因分析（Root Cause Analysis, RCA）**。传统方式下 SOC 分析师手工回溯日志寻找根因,耗时数小时到数天；大模型方式下 AI 自动追溯告警的根因并生成因果链。案例如 Palo Alto XSIAM 可自动分析告警根因（如"Initial access via phishing email → credential theft → lateral movement"）,Dropzone AI 的 OSCAR 框架（Observe → Scope → Context → Analyze → Resolve）实现了自动化 RCA 流程,国内炼石网络、知道创宇、绿盟科技也实现了 RCA 能力。效果是 MTTR 从天级缩短到分钟级,国内厂商数据显示 MTTR 改善 10-50%,知道创宇案例达 95%（问卷汇总）。

三是**威胁情报自动关联与归因。**传统方式下人工阅读多个情报源、手动关联 IoC（指标）、TTP（战术技术）、归因攻击组织；大模型方式下 AI 自动关联多源情报,推理攻击者身份和动机。案例如 Recorded Future 使用 LLM 自动从暗网、社交媒体、技术博客提取情报并关联攻击事件归因 APT 组织,Microsoft Security Copilot 自动关联 Defender 告警与威胁情报生成攻击归因报告。效果是情报分析效率提升 10 倍以上,归因准确率提升（结合多源证据）。

四是**可解释性增强。**核心突破在于大模型可用自然语言解释决策逻辑,缓解"黑盒"困境。案例如 Microsoft Security Copilot 对于每个告警提供"为何判定为威胁"的推理过程（如"This alert is triggered because: 1) External IP attempted login 50 times in 5 minutes (brute force pattern). 2) IP is listed in threat intelligence as malicious. 3) No prior legitimate access from this IP."）,国内炼石网络、知道创宇、绿盟科技等也实现了可解释性增强。效果是提升 SOC 人员对 AI 决策的信任度,满足审计与合规要求（如 GDPR 的"解释权"）。

价值分析：推理与逻辑链构建带来三方面价值。一是提升检测深度,从"表面特征"到"深层逻辑",降低误报；二是加速响应,根因明确后可直接处置,无需人工逐层排查；三是增强可信度,可解释性提升用户对 AI 的信任,促进规模化应用。

5.3.4 代码生成：从"建议"到"执行"

技术突破：**核心能力**：大模型可生成查询语句、修复脚本、检测规则、攻击代码,实现从"提出建议"到"直接执行"的跨越。

典型应用：代码生成与自动化在安全领域的典型应用包括四个方向。

一是查询语句自动生成,包括 Microsoft Security Copilot 将自然语言转换为 KQL 查询、Google Security Operations 生成 YARA-L 检测规则、Splunk AI Assistant 生成 SPL 查询,效果是查询编写时间从分钟级降至秒级。

二是漏洞修复代码生成,包括 Snyk Agent Fix 自动分析漏洞生成修复代码并提交 PR 到 GitHub/GitLab、Furl 自动化执行漏洞修复利用真实系统上下文生成精准修复方案,效果是开发人员修复漏洞时间从数小时降至数分钟,海云安数据显示千行代码漏洞率下降 50%（问卷汇总）。

三是渗透测试脚本生成,包括知道创宇 AI Pentester 根据目标系统特征自动生成渗透测试脚本、海云安和绿盟科技实现漏洞挖掘辅助自动生成 PoC（概念验证代码）,效果是渗透测试覆盖范围扩大,效率提升 3-5 倍。

四是自动化编排工作流生成,包括 Torq 用自然语言描述响应需求 AI 自动生成 SOAR 工作流（Playbook）、Microsoft Security Copilot 自动生成 PowerShell 脚本执行响应动作（如

隔离主机、重置密码）,效果是 SOAR 部署周期从 12-18 个月缩短到 48 小时（Torq 宣称）。

价值分析：代码生成与自动化带来三方面重要价值。一是降低专业门槛,非开发人员也能生成代码；二是提升自动化率,从"生成建议"到"直接执行",真正实现闭环；三是加速响应,自动化脚本执行响应动作,MTTR 从小时级降至分钟级。

5.3.5 AI 驱动的攻防新范式："对抗"升级为"协同"

攻击侧：AI 成为攻击者的武器

攻击侧的 AI 武器化主要体现在三个方面。一是社工工程自动化,包括钓鱼邮件生成和 Deepfake 社工,攻击者使用 ChatGPT 生成高度个性化、无语法错误的钓鱼邮件,2023 年 CrowdStrike 报告显示使用 ChatGPT 生成的钓鱼邮件点击率提升 30%,同时攻击者使用 AI 生成假语音/视频冒充 CEO 进行诈骗（BEC 2.0）,如 2024 年香港某公司被 Deepfake 视频诈骗 $25M（CNN 报道）；二是漏洞挖掘自动化,攻击者使用 LLM 分析开源代码自动化寻找 0day 漏洞,如 2024 年研究证明 GPT-4 可自动化利用 1-day 漏洞（CVE 发布后 24 小时内）；三是对抗样本攻击,攻击者针对 AI 检测模型生成对抗样本绕过检测,如 2023 年研究证明可通过微小扰动绕过恶意软件检测模型。

防御侧：AI 成为防御者的盾牌

防御侧的 AI 盾牌能力主要体现在三个层面。一是 AI Red Team 自动化,工具包括 Microsoft PyRIT、HiddenLayer AutoRTAI、SuperClaw 等,能自动化测试 AI 系统安全性,发现提示注入、数据泄露等风险,如 SuperClaw（2026 年 2 月开源）可针对 AI Agent 执行红队测

试,生成对抗场景；二是 AI 运行时保护,厂商包括 HiddenLayer、Noma Security、Operant AI、Palo Alto AI Runtime Security,能保护 AI 模型、推理服务、Agent 免受攻击,案例如 HiddenLayer AISec Platform 2.0（2025 年 4 月发布）提供 AI 资产全面可见性、影子 AI 检测、运行时防护,Operant AI MCP Gateway（2025 年中发布）保护 MCP 应用全面安全；三是自适应防御,AI 系统可根据攻击者行为动态调整防御策略,如 Darktrace Autonomous Response 实时学习网络行为自动阻断异常活动,绿盟科技的多智能体协同的自主调查、自主基线、自主值守能力（问卷汇总）。

新平衡："AI vs AI"的军备竞赛

AI 驱动的攻防新范式呈现出新的平衡态势。首先,攻防双方都在利用 AI,攻击者用 AI 生成攻击,防御者用 AI 检测和响应；其次,对抗不断升级,从"人 vs 人"演进到"人+AI vs 人+AI",再到"AI vs AI"；第三,从发展趋势看,未来网络安全将是 AI 系统间的对抗,人类角色转变为"监督者"和"决策者"。

5.4 从"能力提升"到"工程闭环"：2026 年新增视角

2023-2025 年,AI for Security 市场经历了"能力堆砌"阶段：厂商竞相发布 AI 功能（Chatbot、检测引擎、自动化工具）,但缺乏系统化的工程实践。进入 2026 年,市场开始关注**从"能力碎片"到"工程闭环"**的转变——AI 不再是孤立的功能模块,而是**具备自主决策、工具编排、审计追溯能力的智能体（Agent）。**

5.4.1 Agent 化：从问答到工具编排与自动处置

技术演进

Chatbot 时代（2023）：该阶段的核心能力是自然语言交互,回答问题,生成建议,但局限在于决策权在人,无法自动执行动作,典型案例是早期的 Microsoft Security Copilot 和 Google Gemini for Security。

Agent 时代（2025-2026）：该阶段的核心能力演进为自主决策加工具调用加闭环处置。技术架构包括四层：一是感知层,接收告警、日志、情报等输入；二是决策层,大模型推理,生成处置策略；三是执行层,调用工具（查询、封禁、工单、取证等）；四是反馈层,记录结果,沉淀知识,优化策略。

典型产品：Agent 时代的典型产品包括多个代表性厂商。一是 Dropzone AI 的 AI SOC Analyst,具备自主调查 100%告警、人类级推理能力；二是 7AI 的 Agentic SOC,具有动态推理能力,可消除误报,实现非人工安全运营任务自动化；三是 Prophet Security 的 Agentic AI SOC Platform,包含 AI SOC Analyst、AI Threat Hunter、AI Detection Advisor；四是 Torq HyperSOC 2o,为 Multi-Agent System,原生支持 MCP,可自主关闭 90% Tier-1 告警；五是国内厂商方面,天懋信息、瀛云科技、摄星科技、炼石网络、美创科技、知道创宇、宁数安全、石犀科技、绿盟科技等已实现相关能力（问卷汇总）。

工具编排（Tool Orchestration）：**核心能力**：Agent 可调用多种工具完成复杂任务。

常见工具类型（基于国内厂商问卷汇总）：

1. **查询工具**：搜索威胁情报、查询日志、检索知识库。
2. **响应工具**：封禁 IP、隔离主机、重置密码、阻断流量。
3. **协作工具**：创建工单、发送通知（邮件/IM）、触发审批流程。

4. **策略工具**：下发检测规则、更新防火墙策略、调整访问控制。
5. **取证工具**：采集日志、抓取内存镜像、提取文件样本。

技术实现：工具编排的技术实现主要包括三种方式。一是 Function Calling,大模型识别用户意图并调用预定义函数（如 OpenAI Function Calling）；二是 MCP（Model Context Protocol）,作为统一的工具调用协议支持多 Agent 协作,如 Torq HyperSOC 2o 原生支持 MCP,Operant AI MCP Gateway 保护 MCP 应用安全；三是 API Gateway,Agent 通过 API 调用既有安全产品（SIEM、EDR、防火墙等）。

案例：Dropzone AI 的 OSCAR 框架

1. **Observe（观察）**：接收告警"User X accessed sensitive database at 3 AM"。
2. **Scope（范围）**：查询用户 X 的历史行为、访问权限、数据库敏感度。
3. **Context（上下文）**：关联威胁情报（User X 账号是否被泄露？）、业务逻辑（是否有合理的业务需求？）。
4. **Analyze（分析）**：推理判断"User X 账号被盗,攻击者尝试数据窃取"。
5. **Resolve（解决）**：自动执行：①隔离 User X 账号;②创建高优先级工单;③通知 SOC 团队;④提取审计日志。

效果：调查时间从 40 分钟降至 3 分钟（Dropzone AI 宣称）,国内厂商数据显示人工工时下降 10-50%,MTTR 改善 10-50%（问卷汇总）。

自主处置的边界与风险

自主程度分级（基于国内厂商实践,问卷汇总）：

等级	定义	典型行为	审批要求
L0-仅建议	生成建议,不执行	提示"建议隔离主机",不自动执行	无审批
L1-写入工单	生成建议并写入工单/报告	自动创建工单,人工审批后执行	人工审批
L2-低风险自动	自动执行低风险动作,可回滚	自动封禁已知恶意IP,可撤销	部分审批
L3-全自动闭环	可闭环处置,含审批/审计/回滚	自动隔离主机→审批→执行→记录→复盘	关键动作审批

风险控制机制：自主处置的风险控制机制主要包括五个方面。一是权限最小化,Agent只能执行预授权动作,无法执行高危操作（如删除数据、关闭防火墙）；二是人机协同审批,关键动作（如隔离核心业务系统）需人工审批；三是可回滚,所有自动化动作可撤销（如解除封禁、恢复访问）；四是全链路审计,记录 Agent 的所有决策和动作,可追溯；五是沙箱测试,高风险动作先在沙箱环境模拟,验证无误后执行。

案例：炼石网络的闭环能力（问卷汇总）体现了风险控制的最佳实践。该方案可实现闭环处置（含审批/审计/回滚/复盘沉淀）,关键动作需审批并全链路记录,技术亮点是基于零信任模型的细粒度权限控制,实现权限最小化。

5.4.2 RAG/TAG + 知识库/工具库：从"能说"到"能做、能追溯"

RAG（Retrieval-Augmented Generation）：知识增强

核心思想：大模型推理前,先从知识库检索相关内容,增强回答的准确性和时效性。

技术架构：RAG 的技术架构包括三个核心层次。一是知识库构建,包含威胁情报库（IoC、TTP、攻击组织）、安全知识库（ATT&CK、CVE、合规标准）、企业知识库（资产清单、业务逻辑、历史事件）；二是检索引擎,包括向量检索（Embedding-based Search,将文本转换为向量计算相似度）、关键词检索（BM25 等传统方法）、混合检索（向量加关键词）；三是生成增强,大模型基于检索结果生成回答,避免"幻觉"。

典型应用：RAG 在安全领域的典型应用包括三个方向。一是威胁情报问答,如用户提问"最近有哪些针对金融行业的勒索软件？",RAG 流程先检索威胁情报库（最近 30 天+金融行业+勒索软件）,再由大模型总结生成回答；二是合规审计辅助,如用户提问"我们的数据库配置是否符合等保 2.0 要求？",RAG 流程先检索等保 2.0 标准与当前配置,对比分析后生成审计报告；三是历史事件回溯,如用户提问"去年类似的告警是如何处置的？",RAG 流程先检索历史工单,提取处置方案后生成参考建议。

国内厂商实践（问卷汇总）：国内厂商在 RAG 方面的实践已较为广泛,已实现 RAG 的厂商包括摄星科技、炼石网络、和利时、知道创宇、烽台科技、宁数安全、石犀科技、绿盟科技、海云安共 9 家,重点投入 RAG/TAG 知识库的厂商包括天懋信息、瀛云科技、摄星科技、炼石网络、和利时、知道创宇、绿盟科技共 7 家。

效果：RAG 技术带来三方面显著效果。一是减少幻觉,从知识库检索事实,降低大模型"编造"内容的风险；二是提升时效性,知识库实时更新（如最新威胁情报）,大模型无需重新训练；三是增强可信度,回答附带引用来源（如"根据 MITRE ATT&CK T1566.001..."）。

TAG（Tool-Augmented Generation）：工具增强

核心思想：大模型推理前或推理中,调用外部工具（API、数据库、计算引擎）获取实时数据或执行动作。

与 RAG 的区别：TAG 与 RAG 的核心区别在于,RAG 检索静态知识库（文本、文档），而 TAG 调用动态工具（实时查询、执行动作）。

典型工具类型：TAG 涉及的工具主要分为三类。一是查询工具,包括威胁情报 API（如 Recorded Future、AlienVault OTX）、SIEM 查询 API（如 Splunk、Microsoft Sentinel）、资产管理 API（如 CMDB）；二是计算工具,包括风险评分引擎（如 CVSS 计算器）、加解密工具（如 GPG）；三是执行工具,包括防火墙 API（封禁 IP）、EDR API（隔离主机）、工单系统 API（创建 Jira 工单）。

典型应用：

TAG 在安全领域的典型应用包括三个方向。一是实时威胁情报查询,如用户提问"这个 IP 地址是恶意的吗？",TAG 流程先调用威胁情报 API 查询,返回结果（恶意/良性+置信度），再生成回答；二是风险评分,如用户提问"这个漏洞的风险等级是多少？",TAG 流程先调用 CVSS 计算器,结合资产重要性,生成风险评分和处置建议；三是自动化响应,如 Agent 决策"需要封禁 IP 1.2.3.4",TAG 流程先调用防火墙 API 执行封禁,返回结果后记录到审计日志。

国内厂商实践（问卷汇总）：国内厂商在 TAG 方面的实践也在推进中,已实现 TAG 的厂商包括天懋信息、瀛云科技、摄星科技、炼石网络、美创科技、知道创宇、云弈科技共 7 家。

效果：TAG 技术带来三方面显著效果。一是实时性,获取最新数据（如当前网络流量、最新威胁情报）；二是可执行性,从"建议"到"执行",实现闭环；三是准确性,调用专业工具（如风险评分引擎）,比大模型直接推理更准确。

知识库与工具库的持续运营

挑战：知识库与工具库的持续运营面临三大挑战。一是知识老化,威胁情报、漏洞库、攻击手法快速演进,知识库需持续更新；二是数据质量,知识库中的错误或过时信息会误导大模型；三是工具适配,不同客户环境中的工具 API 不统一,需适配。

解决方案：解决方案主要包括三个方面。一是自动化更新,包括威胁情报自动同步（如 RSS 订阅、API 拉取）和漏洞库自动更新（如 NVD、CVE 订阅）；二是知识质量管理,通过人工审核加自动化校验（如一致性检查）以及版本控制（Git 管理知识库）确保质量；三是工具抽象层,定义统一的工具接口（如 MCP）,并采用适配器模式（Adapter Pattern）对接不同厂商工具。

案例：绿盟科技的 AI 平台（问卷汇总）定位为独立的安全体系智能中枢,提供开放式 API,可为客户已有平台、产品、服务系统进行智能化赋能,支持 RAG、TAG、SFT 微调、RL/偏好优化等全技术栈。

5.4.3 审计治理：权限最小化、人机协同审批、操作可审计

为何审计治理成为 2026 年焦点?

背景：审计治理成为 2026 年焦点源于三方面驱动力。一是监管要求,欧盟 AI Act（2024 年通过,2026 年全面实施）要求高风险 AI 系统必须可审计、可解释,中国《生成式人

工智能服务管理暂行办法》（2023 年）要求 AI 服务提供者确保内容安全、可追溯；二是客户担忧,企业担心 AI Agent 越权操作（如误删数据、误封禁关键业务）,SOC 团队担心 AI 决策不透明难以向管理层解释；三是安全事件,2025 年多起 AI Agent 误操作事件（如 ChatGPT Plugin 误删用户数据）引发关注,Gartner 警告 2026 年会有 Agentic AI 部署导致的公开泄露事件并有人因此失业。

权限最小化（Principle of Least Privilege）

核心思想：Agent 只能执行其职责范围内的最小必要权限动作。

实现机制：权限最小化的实现机制包括三个层面。一是角色分离,AI SOC Analyst 只能查询、分析、生成建议而不能执行封禁/隔离,AI Remediation Agent 可执行低风险修复（如补丁安装）但不能修改核心配置,AI Policy Manager 可生成策略建议但不能直接下发到生产环境；二是动作白名单,预定义 Agent 可调用的工具列表（如只能调用"查询威胁情报 API",不能调用"删除数据 API"）；三是动态权限控制,基于上下文动态授权（如工作时间 vs 非工作时间、开发环境 vs 生产环境）,并采用零信任模型确保每次动作都需重新验证权限。

案例：炼石网络的权限控制（问卷汇总）基于零信任模型实现细粒度权限控制,实现权限最小化,防止越权操作。

人机协同审批（Human-in-the-Loop）

核心思想：高风险动作需人工审批,低风险动作可自动执行。

风险分级（基于影响范围和可逆性）：

风险级别	动作类型	审批要求	案例
低风险	可逆、影响小	无需审批,自动执行	封禁已知恶意 IP、创建工单、发送通知
中风险	可逆、影响中等	事后审计,可撤销	隔离测试环境主机、重置普通用户密码
高风险	不可逆或影响大	事前人工审批	隔离生产系统、删除数据、修改核心策略

审批流程:

1. Agent 生成处置建议,标注风险级别。
2. 低风险动作：自动执行。
3. 高风险动作：推送审批请求（邮件/IM/工单系统）。
4. 人工审批：通过→执行;拒绝→记录原因,优化策略。

国内厂商实践（问卷汇总）：国内厂商在人机协同审批方面的实践已较为普遍。关键动作需审批的厂商包括天懋信息、瀛云科技、摄星科技、炼石网络、美创科技、海云安、和利时、知道创宇、烽台科技、石犀科技、绿盟科技、广东盈世共 12 家；全链路记录的厂商包括天懋信息、摄星科技、炼石网络、美创科技、海云安、知道创宇、烽台科技、石犀科技、绿盟科技、广东盈世、云弈科技共 11 家。

效果：人机协同审批带来三方面效果。一是降低误操作风险,人工把关高风险动作；二是提升信任度,客户知道"关键决策在人手中"；三是满足合规要求,满足监管对"人类监督"的要求。

操作可审计（Auditability）

核心思想：Agent 的所有决策和动作必须可追溯、可回放、可解释。

审计维度：操作可审计涉及四个核心维度。一是输入审计,记录 Agent 接收的输入（告警、日志、用户提问）；二是决策审计,记录推理过程（如"为何判定为恶意"）和调用的工具与参数（如"查询威胁情报 API,参数 IP=1.2.3.4"）；三是输出审计,记录生成的建议或执行的动作（如"封禁 IP 1.2.3.4"）和执行结果（成功/失败加错误信息）；四是反馈审计,记录人工审批结果（通过/拒绝加原因）和动作回滚（如"解除封禁"）。

技术实现：审计日志的技术实现包括三个层面。一是日志系统,采用结构化日志（JSON 格式）,包含时间戳、用户 ID、Agent ID、动作类型、参数、结果等信息；二是区块链（可选）,将审计日志写入区块链,确保不可篡改；三是可视化回放,将审计日志可视化为时间线,支持回放 Agent 的决策过程。

案例：知道创宇的审计能力（问卷汇总）体现了操作可审计的最佳实践。该方案实现了可闭环处置（含审批/审计/回滚/复盘沉淀），关键动作需审批并全链路记录，量化收益包括人工工时下降 30-50%、MTTR 改善超过 50%、处置闭环率提升超过 50%。

效果：操作可审计带来三方面显著效果。一是问责与追溯,出现问题时可快速定位责任（AI 错误 vs 人工失误）；二是合规要求,满足监管对"可审计"的要求（如金融行业的审计要求）；三是持续优化,分析审计日志识别 Agent 的弱点（如哪些决策经常被人工否决）,优化策略。

治理框架：从技术到组织

技术层：治理框架的技术层包括权限控制、审批流程、审计日志等（如前所述）。

流程层：治理框架的流程层包括三个核心方面。一是 AI 使用政策,明确哪些场景可使用 AI、哪些不可,定义高风险动作清单；二是应急预案,建立 AI 误操作时的回滚流程,设置人工接管机制（如 AI 故障时切换到人工模式）；三是定期审查,每季度审查 AI 决策质量（准确率、误报率）,每年审查权限配置（是否需要调整）。

组织层：治理框架的组织层包括三个核心角色与机制。一是 AI 治理委员会,由 CISO、法务、业务负责人组成,负责审批高风险 AI 应用（如自动化数据删除）；二是 AI 伦理官,负责监督 AI 系统的公平性、透明性、问责性；三是培训与意识,培训 SOC 团队理解 AI 决策逻辑,培训业务团队理解 AI 的能力与局限。

案例：Gartner 建议的 AI 治理框架（2025）涵盖四个核心要素。一是 AI 风险评估,评估 AI 系统的风险等级（低/中/高）；二是人类监督要求,高风险 AI 必须有人类监督；三是透明度要求,AI 决策必须可解释；四是问责机制,明确 AI 错误的责任归属。

5.5 本章小结

本章追溯了 AI for Security 从专家系统到大模型时代的 40 余年演进历程,解析了前大模型时代的 7 大痛点（误报率、数据质量、泛化能力、可解释性、实时性、AI 自身安全、人才稀缺）,阐明了大模型带来的 5 大突破（自然语言理解、多任务能力、推理逻辑、代码生成、攻防新范式）,并重点剖析了 2026 年市场从"能力堆砌"到"工程闭环"的关键转折。

核心洞察：本章分析揭示了以下五个核心洞察。

一是技术演进逻辑,每一代技术都在解决上一代的核心痛点,但也引入新的挑战,大模型时代的核心突破是从"感知"到"认知",从"检测"到"决策"。

二是工程闭环是 2026 年主题,市场不再满足于"AI 能检测威胁""AI 能回答问题",而是要求从检测到处置的全流程自动化,从建议到执行的可信闭环,从能力到治理的体系化。

三是 Agent 化是行业头部已形成共识的演进方向,多家分析机构均预计 2026—2028 年 AI/Agent 能力在 SOC 的渗透将显著提升（业内常被引用的"SOC AI Agent 渗透率 2025 年 5%、2028 年 70%"具体数字目前缺乏可公开核验的原始出处，本报告仅作为方向性参考）；但 Agent 化不是简单的"自动化",而是自主决策加工具编排加审计治理的系统工程。

四是人机协同是现实路径,至少到 2028 年,AI 的角色仍是"增强而非替代人力"（Gartner 观点）,权限最小化、人机协同审批、操作可审计是实现可信 AI 的关键。

五是市场分化加速,平台厂商（Microsoft、Google、Palo Alto 等）追求全栈整合,通过大额收购快速补全能力；专业厂商（Dropzone AI、7AI、Abnormal Security 等）聚焦细分场景,打造 Agent 化的深度能力；国内厂商正在经历从"试点交付"到"规模交付"的跨越,头部厂商（绿盟科技、知道创宇、海云安等）已具备产品化平台能力。

展望未来（2027-2030）：未来几年 AI for Security 的发展将呈现四个重要趋势。

一是 Agentic AI 成为标配,到 2027 年,Agent 能力将超过 Chatbot 成为主流交付形态（Gartner 预测）。

二是 Multi-Agent 协同普及,单一 Agent 演进为多 Agent 协作系统（如 SOC Analyst 加 Threat Hunter 加 Detection Engineer 协同工作）。

三是 AI 安全治理标准化,AI Act（欧盟）、NIST AI RMF（美国）等法规推动行业形成统一的治理标准。

四是 AI vs AI 对抗常态化,攻防双方都在利用 AI,网络安全进入"智能对抗"新阶段。

第六章 市场走向（趋势、驱动、阻碍）

6.1 主要驱动力

AI 赋能网络安全的市场快速增长，背后是多重驱动力的叠加作用。根据国际市场调研和国内数十家厂商的问卷反馈，我们识别出六大核心驱动因素。

6.1.1 告警疲劳（Alert Fatigue）与 SOC 效率瓶颈

告警疲劳已成为全球 SOC 团队最严峻的挑战之一。国内问卷数据显示，约四分之三受访企业（66 家样本中 50 家提及）将"告警疲劳"列为客户采用 AI 安全的首要驱动力。

国际案例验证：

从国际实践来看，AI 降噪效果已在多个真实场景得到验证。一是 Dropzone AI 的客户案例，印第安纳农业保险局（Indiana Farm Bureau Insurance）部署 AI SOC Analyst 后，实现"分钟级提供高保真告警"，有效过滤了大量低价值告警；二是火山引擎 Circle 的客户典型场景，呈现出"10 万条告警 → 8 万条降噪 → 100 条事件 → 20 条自动闭环 + 80 条人工处理"的处理路径，告警聚合率达 99%。

国内案例数据：

国内厂商在告警降噪方面也取得了显著进展。美创科技风险监测智能体已投产的 3 个客户节点，日均告警量控制在约 100 条，检出率 95%、准确率 95%，与传统规则引擎相比，检出率提升 30%。知道创宇的降噪效果范围为 30%-99%，差异源于客户环境复杂度，其中互联网暴露面大的客户降噪率可达 99%。

告警疲劳的根本原因在于传统 SIEM/XDR 系统高度依赖静态规则，误报率居高不下。AI 通过行为基线、图谱关联和上下文推理，能够实现从"告警洪流"到"可操作事件"的质变。

6.1.2 MTTR 压缩需求

企业对 Mean Time To Respond（平均响应时间）的压缩需求日益迫切。问卷数据显示，约六成受访企业（66 家样本中 40 家提及）将"缩短 MTTR/MTTD"列为核心驱动力。

量化收益数据（问卷汇总）：

企业	MTTR 改善幅度	具体场景
美创科技	>50%	数据安全风险监测
知道创宇	>50%	威胁检测与响应
绿盟科技	>50%	SecOps 全流程
炼石网络	30-50%	数据安全事件响应
烽台科技	10-30%	工控安全运营

国际标杆案例：

国际上已有多个成功案例验证了 MTTR 压缩的显著效果。其中,Vectra AI 在 Texas A&M System 的部署将威胁调查时间从"数天"缩短至"数分钟",年节省成本$700 万；Dropzone AI 的 OSCAR 框架则将标准调查流程从 40 分钟压缩至 3 分钟,MTTR 降低 90%。

技术实现路径：

MTTR 的有效压缩主要依赖三大技术路径。一是自动化 RCA（根因分析），AI 通过攻击链重构和多维度关联，快速定位根本原因，CrowdStrike Charlotte AI、SentinelOne Purple

AI 等产品在此方面表现突出；二是端到端溯源可视化，美创科技实现了"接口→应用→数据库→表字段"的完整链路呈现，突破了传统"登录-登出会话"的局限；三是预案智能化生成，AI 能够动态生成处置建议，而非依赖静态 Playbook，火山引擎 Circle、炼石网络等均采用了这一技术路径。

MTTR 的压缩不仅是技术优化，更关乎业务连续性。研究表明，每缩短 1 小时 MTTR，可为企业平均节省$10,000-$50,000 损失（取决于行业和业务规模）。

6.1.3 全球安全人力缺口

企业侧反馈：

问卷数据显示，约半数受访企业（66 家样本中 35 家提及）将"人力缺口"列为 AI 采用的核心驱动。

注：(ISC)²等机构发布的"全球缺口 400 万"数据需审慎引用——其发布方为培训/认证利益相关方，存在夸大需求的动机。

国内人才现状的真实图景：

根据 Briefing 访谈和问卷反馈，国内安全行业面临的不是简单的"总量缺口"，而是**结构性错配**：一是**基础岗位供大于求**，网安专业大规模扩招后，初级安全人才就业已开始出现困难；二是 AI+安全复合型人才极度稀缺，66 家受访企业中约三分之一（21 家）将"缺懂 AI 和安全的复合型人才"列为主要阻碍，这才是企业真正感受到的"缺口"。

AI 如何填补人力缺口：

案例 2：Dropzone AI 服务 MSSP

CBTS（托管安全服务商）部署 Dropzone AI 后，卸载 30-50%告警量，释放团队专注于威胁猎捕和客户服务，**在不扩大团队规模的前提下服务更多客户。**

人力缺口不仅是招聘难题，更是**成本压力。**AI 通过自动化低价值重复性工作（Tier-1 告警分类、日志标准化、报告生成等），使有限的安全团队能够聚焦高价值决策。

6.1.4 合规审计自动化需求

问卷数据：约三分之一受访企业（66 家样本中 23 家提及）将"合规审计压力"列为客户核心驱动。

政策推动力：

政策层面的推动力主要体现在三个方面。一是中国政务系统的明确要求，浙江省政务已明确要求"AI+风险监测"（美创科技反馈）；二是欧盟 AI Act（2026 年全面实施），要求高风险 AI 系统必须通过安全评估和持续监控；三是美国 EO 14110，联邦机构 AI 应用需进行 Red Team 测试。

AI 赋能合规审计的典型场景：

场景 2：合规报告智能化生成

问卷数据显示，66 家企业中，56 家（约 85%）已实现"AI 报告生成"能力，覆盖场景包括三大类：一是等保合规报告（绿盟科技、宁数安全）；二是密评密改报告（炼石网络）；三是数据安全风险评估报告（石犀科技、美创科技）。

合规审计的自动化不仅降低人工成本，更实现**持续合规**（Continuous Compliance），将传统"年度审计"转变为"实时监控+自动修正"。

6.1.5 新增驱动力：AI 巨头入局安全工具市场

网络安全行业正在经历一场来自产业链上游的降维冲击。2026 年 2 月，Anthropic 以 Claude Code Security 入局代码安全审计市场，14 天后 OpenAI 跟进发布 Codex Security——两家全球顶级 AI 实验室在半个月内先后进入传统安全厂商的核心业务领域，这在行业历史上尚属首次。其产品发布方式同样颠覆常规：以"研究预览"形式向 Enterprise 用户免费开放，绕开了传统安全工具的商业化路径，直接以能力示范冲击市场。

这一事件的产业意义远超一款新产品的发布。首先，它验证了通用大模型的推理能力已足以胜任高度专业化的安全分析任务，无需经过漫长的垂域专项训练；其次，它以资本市场的剧烈反应（CrowdStrike 单日跌幅 8-10%、全球网安 ETF 跌幅约 9%）证明，即便是头部传统安全厂商，也无法对 AI 颠覆免疫；第三，它为国内安全厂商设定了一个明确的能力参照系——本轮调研中，长亭科技和悬镜安全均将 Claude Code Security 作为代码安全能力的对标基准，国内最优模型的差距被量化为约 20%。

对市场走向的判断：AI 巨头入局安全工具不是"狼来了"式的恐慌，而是行业竞争坐标系的根本性重置。未来 2-3 年，以下几类安全产品面临最直接的替代压力：一是依赖规则库的传统 SAST/DAST 工具；二是以人工为主的渗透测试和代码审计服务；三是竞争力主要来自"接入大模型"而非深度安全 Know-How 的浅度 AI 安全产品。而具备深厚安全情报积累、工程化修复能力或垂直场景壁垒的厂商，则有望在与 AI 巨头的协同中找到生存与增长空间。

6.1.6 AI 驱动攻击的倒逼效应

Recorded Future《2026 年安全状态报告》警告："AI 驱动的验证失败——AI 放大欺骗、社工、身份滥用"。攻击者使用 AI 的速度远超防御方想象。

攻击端 AI 应用现状：

攻击者对 AI 的应用主要体现在三个方向。一是钓鱼邮件生成，AI 生成的钓鱼邮件成功率提升 40%（无语法错误、高度个性化）；二是漏洞自动化利用，知道创宇内部测试显示，AI 可自主完成"扫描→读文档→编写 Exploit→拿下"全流程；三是深度伪造攻击，已出现 AI 生成 CEO 语音指令转账的真实案例（香港某跨国公司被骗$2500 万）。

防御方必须跟进的逻辑：

> "如果攻击者用 AI 编写恶意代码需要 5 分钟，而防御方人工分析需要 5 小时，这场战争已无悬念。"（知道创宇观点）

Gartner 预测：到 2026 年底，因 AI 风险护栏不足导致的"AI 致死"法律诉讼将超过 2000 起。这一预测加剧了企业对 AI 安全防护的紧迫感。

6.1.7 国内客户驱动力综合画像（问卷数据）

基于受访企业的反馈，我们绘制出国内客户驱动力全景图：

驱动因素	提及率	典型行业	核心诉求
告警疲劳	71% (10/14)	政务、金融、运营商	从告警洪流到可操作事件

人力缺口	64% (9/14)	全行业	用 AI 替代初级分析师
缩短 MTTR	50% (7/14)	金融、能源	业务连续性保障
趋势压力	43% (6/14)	互联网、制造	同行竞争倒逼
合规审计	36% (5/14)	政务、金融、医疗	持续合规+降低审计成本
成本压力	36% (5/14)	制造、中小企业	降本增效

关键洞察：

从驱动力综合画像中可以得出三点核心洞察（基于 66 家问卷样本）。一是告警疲劳+MTTR 压缩是核心双驱动（分别约四分之三与约六成受访企业提及），人力缺口紧随其后（约半数）；二是合规与成本呈现矛盾状态，既要满足合规（约三分之一），又要降低成本（约三成），AI 成为平衡两者的关键；三是趋势压力相对次要，约四分之一受访企业受"同行已部署 AI"驱动，FOMO（Fear of Missing Out）心态。

6.2 主要趋势（深度分析）

趋势一：从 Copilot 到 Agent——工具编排与自动处置成为核心竞争力

这是 AI for Security 领域最重要的范式转变。**多家分析机构与厂商口径预期，2026—2028 年 Multi-Agent AI 在威胁检测与响应场景的渗透将显著提升；其中坊间常被引用的"2025 年 5%→2028 年 70%"这一具体数字目前缺乏可公开核验的原始出处，公开口径间差异较大，本报告仅作为方向性参考。**

6.2.1 演进路径：三阶段进化

阶段 1：Chatbot/Assistant（2023-2024）

这一阶段的特征是自然语言问答、报告生成、知识检索,代表性产品包括早期 Microsoft Security Copilot、绿盟科技安全助手。其主要局限在于**仅提供建议,不执行操作。**

阶段 2：Copilot（2024-2025）

这一阶段的特征是工具调用、脚本生成、辅助决策,代表性产品包括 Microsoft Security Copilot 生成 KQL 查询、CrowdStrike Charlotte AI 威胁搜索。相比上一阶段的进步在于**可生成可执行命令,但需人工确认。**

阶段 3：Agent 智能体（2025-2026）

这一阶段的特征包括自主推理、工具编排、闭环处置、多 agent 协作，代表性产品有 Dropzone AI（OSCAR 框架）、火山引擎 Circle、7AI，实现的突破在于**自主执行+审计回滚+复盘沉淀。**

问卷数据验证:

数据显示 Agent 能力已成为主流趋势（66 家问卷样本）。一方面，约八成受访企业（52 家/66 家）已交付 Agent 智能体；另一方面，约七成将"Copilot→Agent"演进列为 2026 年重点投入方向（48 家/66 家）。

6.2.2 国际案例：从单 agent 到 multi-agent system

Microsoft Security Copilot Agents（2025 年 3 月）

该产品实现了能力跃升，从单一 Copilot 升级为多专业化 agent 协作，包括威胁猎捕 agent（主动搜索未知威胁）、事件响应 agent（自动隔离、封禁、取证）、合规检查 agent（持续监控配置偏离）。在架构特点方面，Agent 之间可共享上下文、联动调用工具。

CrowdStrike Agentic SOC（2025 年秋季）

CrowdStrike 定义"Agentic SOC"概念，推出了三类专业化 agent：一是检测工程师 agent，可自动调优检测规则；二是事件响应自动化 agent，能够分钟级完成封禁/隔离；三是威胁猎捕 agent，基于 MITRE ATT&CK 主动搜索威胁。

Dropzone AI Multi-Agent System（2026 演进）

Dropzone AI 从单一 AI SOC Analyst 扩展为专业化 agent 团队，包括五个专业角色：威胁猎手（Threat Hunter）、检测工程师（Detection Engineer）、取证分析师（Forensic Analyst）、威胁情报分析师（Threat Intel Analyst）、安全数据架构师（Security Data Architect）。

核心洞察：Agent 化的本质是将"人类 SOC 团队的专业分工"映射到"AI agent 矩阵"，实现**"虚拟 SOC 团队"24×7 自主运营。**

6.2.3 国内案例：从助手到智能体的工程化闭环

绿盟科技风云卫（NSFGPT）：20+专业化 Agent 矩阵

绿盟科技风云卫是国内场景覆盖最广的 AI 安全运营智能体平台，采用 SecLLM（自研安全大模型）+DeepSeek 双基座架构，构建了四层智能体池：**场景化智能体**（钓鱼邮件检

测、挂马检测、敏感数据识别、勒索识别、漏洞管理）、**工作流智能体**（降噪预判、未知攻击检测、AI 7×24 自主值守、基线自动生成、攻击故事还原）、**通用智能体**（意图识别、长上下文处理、文档处理、语音转换）、**原子智能体**（降噪、检测、研判、推荐、调查、响应、报告）。通过四层 agent 协同，绿盟风云卫实现了 AI 7×24 自主值守的完整闭环：统一数据湖→威胁检测→告警归并与降噪（降噪率达 98%）→自动调查与关联→自动处置或升级人工决策→可视化报告。在 2024 年中央网信办组织的"AI 赋能网络安全应用测试"中，绿盟风云卫在告警日志降噪赛道排名第 2（18 家参评）、钓鱼邮件识别赛道排名第 4（16 家参评）。

悬镜安全灵脉 AI：AI 原生 SAST + 多模态 AIST

悬镜安全是国内 DevSecOps 赛道的 AI 原生代表，其灵脉 AI 开发安全卫士将传统 SAST 重构为"向量化代码索引+LLM 编排+控制流/数据流知识图谱"的 AI 原生架构。**核心 agent 能力**包括：AI 代码业务识别 agent（按业务类型分类缺陷并给出上下文修复建议）、AI 审计核验 agent（基于知识图谱+历史缺陷批量审计，一键标记误报）、AI 修复 agent（自动生成修复代码，准确率 90%，修复时间下降 90%）。悬镜覆盖 6000+缺陷检查项、30+编程语言，支持 GB/T 34943/34944/34946、GJB-8114、CERT、CWE、OWASP、MISRA、PCI-DSS 等多套合规标准，误报率从业界平均 40%+降至 5%以下。在多模态 AIST 平台层面，悬镜将 SAST、SCA、RASP、模型扫描、代码护栏、红队对抗、组件指纹、SBOM 统一整合，并原生对接 Jenkins、GitLab-CI、Gitee、Azure DevOps、Zadig、阿里云效等 CI/CD 流水线，成为"透明"嵌入研发流程的开发安全智能体。

奇安信 QAX-GPT + AISOC：数据规模驱动的全产线 AI 化

奇安信 QAX-GPT 是国内首批以"安全大模型"为产品品牌的自研模型（2023 年发布，截至 2026 年已完成三轮增量预训练），在其之上构建了 **AISOC 智能运营平台**，将大模型能力下沉到天眼（NDR）、天擎（EDR）、盘古石（取证响应）、代码卫士等全产线，形成"一个大模型+一个 Agent 框架+N 个场景 agent"的统一体系。核心能力包括：告警语义研判与自动聚合、钓鱼邮件与恶意样本识别、漏洞与基线合规检查、自然语言转查询（NL2SQL/NL2KQL）、威胁情报增强，以及依托奇安信威胁情报中心的 APT 组织画像与攻击链重构。其差异化在于依托大规模安全数据积累（每日处理千亿级日志/告警、60 万+样本、200 万+APT 攻击事件），将"数据规模优势"直接转化为模型的上下文理解和推理准确率，形成与国内其它厂商的代差。

6.2.4 工具编排能力对比（问卷数据）

工具类型	已实现企业数	典型场景
查询（搜索/情报）	12 家（86%）	威胁情报检索、历史事件查询
工单/通知	11 家（79%）	自动创建工单、飞书/企微推送
规则下发	8 家（57%）	防火墙策略、IPS 签名更新
封禁/隔离	5 家（36%）	IP 封禁、主机隔离、账号冻结
取证采集	3 家（21%）	内存镜像、流量抓包、日志留存

关键洞察：

从工具编排能力对比可以得出三点核心洞察（基于 66 家问卷样本）。一是查询类工具普及率最高（大多数厂商覆盖），反映 RAG/TAG 是基础能力；二是执行类工具（封禁/隔离/取证）普及率偏低（少数到约三分之一），说明真正的 Agent 闭环能力仍是稀缺；三是在审批审计机制方面，大多数受访企业要求关键动作需审批，约七成实现全链路记录。

6.2.5 Agent 化的核心挑战与应对

挑战 1：权限与审计（越权风险）

问卷数据显示，66 家受访企业中少数（10 家，约 15%）将"权限与审计"列为主要阻碍。针对这一挑战，业界主要采用三类应对方案：一是**分级授权**，低风险动作自动执行，高风险动作需人工审批（炼石网络、火山引擎）；二是**沙箱模拟**，先在沙箱环境执行，验证无误后下发生产环境（绿盟科技）；三是**可回滚机制**，所有操作可追溯、可撤销（知道创宇、广东盈世）。

挑战 2：工具生态成熟度

火山引擎工程师反馈：

> "01 系列产品（公安部第一研究所推出的一系列安防产品）生态封闭，API 不足，无法满足智能体调用。必须 FDE（前沿部署工程师）团队定制化适配，成本高昂。"

应对方向：

业界主要从三个方向应对工具生态成熟度不足的问题：一是推动安全设备厂商开放 API（行业共同努力）；二是采用 MCP（Model Context Protocol）标准化工具调用（火山引擎、Operant AI）；三是自建工具适配层（绿盟科技的"开放式 API"策略）。

挑战 3：多 agent 协同的复杂度

多 agent 协同的技术难点主要包括 agent 间上下文传递、冲突决策、任务分解等。针对这些难点，业界主要采用两类解决方案：一是**主从架构**，由主 agent 负责任务规划，子

agent 执行细分任务（火山引擎的主模型+专用模型）；二是**协作协议**，定义 agent 间通信标准（如 MCP、OpenAI Function Calling）。

趋势判断：

Torq 宣称"SOAR 已死，Hyperautomation 是未来"。这一宣言激进但方向正确——**传统基于 Playbook 的静态编排，将被基于 Agent 的动态推理取代。**多家分析机构对 2026—2028 年 AI/Agent 能力在 SOC 的快速渗透均给出方向性预期（"2028 年 70%"等具体数字尚缺可公开核验的原始出处，仅作方向性参考），即便去除该具体数字，方向性结论依然成立：不具备 Agent 能力的安全产品将在 3 年内承受越来越大的客户选型压力。

趋势二：RAG/TAG + 知识库/工具库——从"能说"到"能做、能追溯"

RAG（Retrieval-Augmented Generation）和 TAG（Tool-Augmented Generation）已成为 AI for Security 的标配技术。**问卷数据显示，66 家受访企业中约七成采用 RAG（46 家），约六成采用 TAG（39 家）。**

6.2.6 RAG 在安全场景的应用模式

技术原理回顾：

RAG 通过"检索+生成"两阶段流程，将外部知识库注入大模型推理过程。一是**检索阶段**，根据用户 query 从知识库中检索相关文档/片段；二是**生成阶段**，将检索结果与 query 拼接，输入大模型生成答案。

安全场景的独特挑战：

安全场景下的 RAG 应用面临三大独特挑战。一是**知识时效性**，威胁情报、漏洞库每日更新，知识库需实时同步；二是**多源异构**，需整合告警日志、威胁情报、工单记录、安全知识文档等；三是**检索精度**，安全领域容错率极低，误检索会导致误判。

典型应用场景（问卷汇总）：

场景	实现企业	知识库类型	核心价值
研判问答	12 家（86%）	威胁情报、历史工单、知识库	将 SOC 分析师经验沉淀为可复用知识
告警摘要	10 家（71%）	告警规则库、处置预案	语义化描述威胁，而非技术术语
报告生成	9 家（64%）	行业合规标准、模板库	自动化生成等保/密评/审计报告
知识沉淀	8 家（57%）	历史事件库、专家经验	将每次事件的处置过程反哺知识库

国际案例：Recorded Future Intelligence Graph

Recorded Future Intelligence Graph 的核心能力体现在三个方面：一是**数据规模**,每日处理 TB 级威胁数据,覆盖 75+语言；二是**知识图谱**,从开源、暗网、技术源实时采集威胁数据,AI 自动关联和优先级排序；三是**预测性分析**,基于历史数据和当前态势,预测未来威胁(如 APT 组织的下一步行动)。

国内案例：摄星科技安全知识库

摄星科技的安全知识库体系包括三个核心要素：一是**知识类型**,涵盖告警/日志、样本/沙箱、威胁情报、漏洞库、安全知识文档；二是**核心方案**,采用纯提示词、RAG、规则/图谱融合；三是**量化收益**,实现人工工时下降 10-30%,MTTR 改善 10-30%,误报下降 30-50%。

炼石网络多源数据融合

> "构建多源数据融合平台，整合公开漏洞库、行业威胁情报和客户本地数据，实现数据可用不可见的跨机构协作。"

技术优化方向：

业界在 RAG 技术优化上主要聚焦三个方向。一是**向量化检索优化**，采用混合检索（关键词+语义向量）提升召回率（美创科技、火山引擎）；二是**分层知识库**，构建通用知识库+行业知识库+企业私有知识库的三级架构（炼石网络）；三是**知识库自更新**，AI 从每次事件中自动提取新知识，反哺知识库（火山引擎"知识教练"机制）。

6.2.7 TAG（工具增强生成）：让 AI"长出手脚"

如果说 RAG 让 AI"能说"（基于知识回答问题），那么 TAG 让 AI"能做"（调用工具执行操作）。

TAG 与 Function Calling 的关系：

两者的关系可以从两个层面理解：一是 **Function Calling**,这是 OpenAI 等模型原生支持的能力,AI 可决定调用哪个函数及参数；二是 **TAG**,在安全场景下,扩展为"工具库+执行器+审计层"的完整体系。

国际案例：Dropzone AI 的 OSCAR 框架

> OSCAR = Observe, Scope, Context, Analyze, Resolve

阶段	工具调用	示例
Observe	日志查询、流量回溯	从 SIEM 检索相关时间段的所有告警
Scope	资产查询、IP 信誉	查询目标 IP 的威胁情报评分
Context	历史事件检索、用户画像	查询该用户过去 30 天的异常行为
Analyze	攻击链重构、根因分析	基于 MITRE ATT&CK 映射攻击路径
Resolve	封禁 IP、隔离主机、创建工单	自动下发防火墙策略+通知管理员

调查时间：40 分钟 → 3 分钟（降低 93%）

国内案例：宁数安全工具联动

宁数安全的工具联动体系包括三个方面：一是**工具链**,涵盖 SIEM、SOAR、EDR、工单、IM；二是**技术亮点**,工具联动主要采用 MCP 和 API 的调用(问卷明确提及)；三是**应用场景**,聚焦工业资产安全分类和网络威胁识别。

绿盟科技工具编排全景：

绿盟科技的工具编排覆盖四大类能力：一是**查询类**,包括搜索/情报查询；二是**执行类**,涵盖封禁/隔离、规则下发、取证采集；三是**协作类**,支持工单/通知；四是**工程闭环**,实现可闭环处置(含审批/审计/回滚/复盘沉淀)。

火山引擎 Circle 工具生态（Tools）：

> "必须配备'手和脚'才能深入业务流。工具类型：云防火墙、主机权限、工单系统、数据指纹、DLP 等。标准化：定义 ETR 标准，对接 MCP 和 API 服务。"

MCP（Model Context Protocol）：工具调用的标准化协议

MCP（Model Context Protocol）是 Anthropic 于 2025 年推出的工具调用标准化协议。采用该协议的企业包括火山引擎 Circle（原生支持）、Operant AI（MCP Gateway）、Torq HyperSOC 2o 等，核心价值在于统一工具调用接口，降低 agent 开发成本。

6.2.8 从"能做"到"能追溯"：工程闭环的完整性

问卷数据：工程闭环能力分级

闭环程度	企业占比	代表企业	能力描述
可闭环处置（含审批/审计/回滚/复盘）	29%	炼石网络、知道创宇、绿盟科技、广东盈世	**最高级**：AI 自主执行+人工审批+全链路审计+可回滚+自动复盘沉淀
生成建议→写入工单/报告（可追溯）	64%	天懋信息、摄星科技、美创科技等	**中级**：AI 生成处置建议，写入工单/报告，留存记录
自动执行部分低风险动作（可回滚）	7%	石犀科技	**实验级**：低风险动作自动执行，高风险需人工
仅生成建议（不落库不执行）	14%	和利时、云弈科技	**基础级**：仅提供建议，不执行

核心洞察：

从工程闭环能力分级数据中可以得出两点关键洞察（基于 66 家问卷样本）：一是约三分之一受访企业（24 家/66 家）实现完整闭环，说明大部分厂商仍处于"辅助决策"阶段，距离"自主运营"有较大差距；二是审批审计是普遍需求，大多数受访企业（55 家/66 家）要求关键动作需审批，反映出对 AI 自主操作的谨慎态度。

最佳实践：炼石网络的闭环机制

炼石网络构建了完整的闭环机制，包括四个核心环节：一是**关键动作需审批**，封禁/隔离等高风险操作需人工确认；二是**全链路记录**，所有 AI 决策过程、工具调用记录、审批流程留存；三是**可回滚机制**，误操作可一键撤销；四是**复盘沉淀**，每次事件处置完成后，自动生成复盘报告，提取经验教训反哺知识库。

国际标杆：Palo Alto Networks XSIAM 的可观测性

Palo Alto Networks XSIAM 在可观测性方面树立了行业标杆，主要体现在三个方面：一是 **AI 驱动的根因分析**，能够自动构建攻击链，标注每个节点的置信度；二是**自动化事件响应**，支持回滚，所有操作可追溯到具体 AI 决策；三是**预测性威胁检测**，基于历史数据预测未来威胁，并记录预测依据。

趋势判断：

RAG/TAG 已从"可选项"变为"必选项"。**未来竞争的核心不是是否采用 RAG/TAG，而是知识库的质量、工具库的丰富度、以及工程闭环的完整性。** 那些仅停留在"生成建议"阶段的产品，将逐渐被具备完整闭环能力的 Agent 取代。

趋势三：多模态安全分析——邮件/流量/样本/日志/图像取证的融合

多模态能力是通用大模型的重要突破，但在安全场景下的应用仍处于早期探索阶段。问卷数据显示，66 家受访企业中约半数以上（37 家）将多模态能力列为重点投入方向。

6.2.9 多模态在安全场景的独特价值

传统安全分析高度依赖单一模态，具体表现在四个方面：一是**日志分析**,主要处理纯文本；二是**流量分析**,处理二进制+协议解析；三是**样本分析**,分析 PE/ELF 结构+行为序列；四是**邮件安全**,处理文本+附件。

多模态融合的价值：

多模态融合在安全分析中的价值主要体现在三个方面。一是**跨模态关联**，将邮件文本、附件（图片/PDF）、发件人 IP 信誉、历史行为综合研判；二是**图像取证**，分析截图、监控视频中的异常行为；三是**代码+文档联合理解**，在开发安全场景，同时理解代码逻辑和需求文档。

6.2.10 国际案例：原生多模态能力

Google Gemini for Security Operations

Google Gemini for Security Operations 是"唯一原生多模态的安全 AI 平台（处理日志、代码、图像等多种数据类型）"，其应用场景包括三类：一是从监控视频中识别异常人员行为（物理安全）；二是分析恶意 PDF 中的嵌入式图像和脚本；三是理解攻击者留下的手写笔记（取证场景）。

Abnormal Security：邮件多模态分析

Abnormal Security 的核心能力包括三个方面：一是**行为 AI**,学习每个用户的正常通信模式(文本风格、附件类型、收发频率)；二是**实时邮件安全辅导(2025 年新增)**,分析邮件正文、附件(图片/文档)、发件人信誉,实时告警钓鱼风险；三是**量化收益**,根据 Reddit 用户反馈,平均每月 MS+Abnormal 仍会漏掉 6-8 封恶意邮件,但 AI 邮箱功能能正确识别。

6.2.11 国内案例：多模态能力的早期实践

悬镜安全：代码+文档+AI 资产的多模态融合

悬镜灵脉 AI 开发安全卫士将开发安全场景的多种模态统一纳入同一个智能体推理链路：一是源代码（30+语言、控制流/数据流图）；二是软件物料清单（SBOM）与开源组件指纹；三是 AI 模型与数据集（模型扫描、MCP/Skills 投毒检测、AI 供应链风险情报）；四是研发过程文档（需求文档、API 规范、威胁建模）。通过代码-文档-SBOM-模型的联合表征，悬镜能在一次扫描中同时回答"这段代码是否存在漏洞、来自哪个开源组件、对应的业务逻辑是什么、是否触发合规红线、AI 模型是否被污染"五类问题，显著高于传统 SAST 仅能回答"代码是否有漏洞"的单模态边界。

奇安信：威胁情报图谱 + 样本多模态

奇安信依托其威胁情报中心与全球 APT 研究数据，将 QAX-GPT 延伸到多模态场景：一是恶意样本多模态（PE/ELF 静态特征+沙箱动态行为+YARA 规则+家族画像）；二是钓鱼邮件多模态（正文语义+附件 OCR+Logo 视觉相似度+发件域信誉）；三是 APT 攻击链多模态（流量元数据+主机进程图+威胁情报 TTPs+历史事件库），通过"1 个统一大模型+N

个多模态 agent"将跨模态证据融合成可解释的攻击故事，帮助一线分析师在分钟级完成 APT 研判。

6.2.12 多模态的技术挑战与应对

挑战 1：模型能力差距

火山引擎工程师主观评估：

> "**第一梯队**：GPT、Claude（多模态能力最强）

> **1.5 梯队**：豆包

> **第二梯队+**：千问闭源版

> **开源模型**：千问/DeepSeek 32B（多模态能力弱，指令遵循和推理能力差距大）"

国内困境：

国内企业在多模态应用方面面临两大困境：一是大部分企业采用开源基座模型（55 家/66 家，83%），多模态能力受限；二是私有化部署场景（算力受限），难以运行 GPT/Claude 级别的多模态模型。

挑战 2：数据预处理复杂度

数据预处理复杂度因数据类型而异。对于图像数据，需要进行 OCR（光学字符识别）、物体检测、异常标注；对于流量数据，需要进行协议解析、加密流量识别、会话重组；对于样本数据，需要进行沙箱执行、行为序列提取、内存分析。

应对方案：

业界主要采用两类应对方案。一是**分层模型架构**（火山引擎），包括主模型（豆包/GPT）负责多模态理解+任务规划，以及专用模型（孔明/小模型）负责单模态深度分析（如恶意文件识别）；二是**预处理+大模型结合**（海云安），用传统工具（IDA Pro、Wireshark）完成结构化提取，用大模型完成语义理解和关联分析。

挑战 3：成本控制

成本控制的挑战主要体现在两个方面：一是**多模态推理成本**,图像/视频 token 消耗远超文本(GPT-4o 处理 1 张图片约 170 tokens)；二是**美创科技的极度保守策略**,日均仅 1000 tokens 消耗,难以支撑多模态应用。

趋势判断:

多模态能力短期内仍是"锦上添花"而非"雪中送炭"。**主要原因**有三：一是**模型能力瓶颈**，开源模型多模态能力弱，闭源模型（GPT/Claude）私有化部署困难；二是**成本敏感**，客户对 Token 消耗敏感，多模态推理成本高；三是**场景优先级**，日志/告警分析仍是核心需求，图像/视频取证属于边缘场景。

但长期看（2-3 年）：随着开源多模态模型成熟（如 Qwen-VL、CogVLM）、推理成本下降，多模态将从"特色能力"变为"标配能力"。

趋势四：私有化/混合部署与成本优化——推理加速、缓存、分层模型策略

这是中国市场的独特趋势。问卷数据显示，66 家受访企业中约三分之一（21 家）将"客户对私有化要求高"列为主要阻碍，约半数（33 家）将"私有化降本"列为重点投入方向；同时约 86%（57 家）已采用私有化推理部署。

6.2.13 私有化需求的根源

数据主权红线：

私有化需求的根源在于数据主权红线，主要体现在三类客户。一是**政务系统**，浙江省政务要求本地大模型（美创科技反馈）；二是**央企国企**，数据不能出网（中石油勘探院案例）；三是**金融行业**，监管要求数据本地化处理。

火山引擎工程师的判断：

> "数据不能出网是红线，短期（1-3 年内）难以突破。这是云端与私有化长期博弈的根本矛盾。"

> "中国 TOB 市场的独特性：美国产品模块化、API 开放，中国封闭、项目制。私有化需求是商业模式差异的必然结果。"

6.2.14 云端 vs 私有化效果差距（Briefing 数据对比）

维度	云端（SaaS）	私有化部署	差距
模型能力	GPT-5.4、Claude Opus 4.6、DeepSeek	千问 32B、DeepSeek 32B	**不止一个数量级**（火山引擎）

推理速度	<1 秒（云端 GPU 集群）	3-5 秒（本地 910B 双卡）	3-5 倍
知识更新	实时（威胁情报每日更新）	定期同步（周/月级）	时效性差
算力成本	按 Token 计费	一次性采购 GPU（¥200 完+）	初期投入大
运维成本	零运维	需专人维护（模型更新、推理加速调优）	持续投入

火山引擎 Circle 的客户反馈：

火山引擎 Circle 的客户反馈呈现明显分化。对于云上 SaaS 客户，数据可出网，模型能力充分发挥（豆包大模型/GPT）；而对于私有化客户（中石油勘探院），部署在昆仑大模型基座上，效果明显弱于云端，但符合合规要求。分层模型+混合推理是火山引擎应对"云端与私有化效果鸿沟"的核心策略。火山引擎将推理任务分为三级：确定性任务（零算力成本）直接用规则引擎处理；半确定性任务用专用小模型本地推理；复杂开放式任务才调用云端大模型。这一分层策略使整体推理成本降低 40-60%，同时将核心 AI 算力投入到真正需要大模型能力的场景。在工具生态方面，火山引擎是国内最早推动 MCP 标准化的厂商之一——Circle 原生支持 MCP 协议，并通过 MCP Gateway 兼容三方厂商的 MCP Server，将生态开放做到实质性落地。客户反馈呈现明显分化：云上 SaaS 客户满意度高（数据可出网，模型能力充分发挥）；私有化客户受限于昆仑大模型基座能力，效果明显弱于云端，但合规要求优先于性能，这一矛盾短期内尚无解。这是国内所有安全 AI 厂商面临的共同困境。

策略 2：激进算力策略

策略评价：

从优劣势分析，这一策略的优势在于效果最佳、用户体验好，劣势是算力成本高（需规模化摊薄），适用场景为互联网 SaaS、追求极致体验的场景。

策略 3：分层模型+混合推理（火山引擎 Circle)

火山引擎 Circle 采用分层模型+混合推理策略。云端主模型为豆包大模型（业务编排/任务规划），私有化专用模型为孔明 32B（恶意文件识别、漏洞检测）。分层策略将任务分为三级：确定性任务用规则引擎（零算力成本），半确定性任务用专用小模型（本地推理），复杂推理任务用云端大模型（数据脱敏后上云）。算力需求方面，双卡 910B 可处理 1.6 万条告警/天。

策略评价：

从优劣势分析，这一策略的优势在于平衡效果与成本、适配混合云场景，劣势是架构复杂度高、需精细化任务分配，适用场景为大型企业、对效果和成本都敏感的客户。

6.2.16 推理加速技术（问卷数据)

技术	采用企业	效果	适用场景
推理加速（vLLM/TensorRT)	炼石网络、海云安、广东盈世、绿盟科技	推理速度提升 2-5 倍	私有化部署、实时性要求高

小模型蒸馏	美创科技、宁数安全、石犀科技、广东盈世、云弈科技、绿盟科技	模型体积减少 50-70%，速度提升 3-10 倍	边缘设备、算力受限环境
分层模型（大+小）	石犀科技、绿盟科技、火山引擎	综合成本降低 40-60%	混合云、多场景融合
缓存机制	未在问卷中明确提及	重复 query 响应时间 <100ms	高频查询场景（如威胁情报检索）

技术深度解析：

推理加速（炼石网络案例）：

> "采用硬件加速技术，将规则匹配和特征提取的时延降低至**微秒级**。"

分层模型（火山引擎孔明模型）：

火山引擎孔明模型是基于豆包 OSS 后训练（post-training）的安全专用模型，应用场景包括恶意文件识别、漏洞识别等（通用模型表现不佳），采用多模型协同架构，即主模型编排 + 专用模型执行。

6.2.17 混合部署的未来：AI 机密计算（AICC）

火山引擎的探索方向：

> "AICC（机密计算）：企业数据加密后上云计算。技术可行，但需监管认可（芯片厂商背书：Intel/Nvidia）。部分企业已在尝试（飞书妙记、Trae 写代码等）。"

AICC 的技术原理：

AICC 的技术原理包括三个核心步骤。一是**数据加密上云**，企业数据在本地加密，密文上传云端；二是**可信执行环境（TEE）**，在云端 GPU 的隔离区域解密+推理；三是**结果加密返回**，推理结果加密后返回企业，云端不留存明文。

优势：

AICC 的优势主要体现在两个方面：一是兼顾**数据主权**（云端看不到明文）和**模型能力**（可用 GPT/Claude）；二是解决私有化效果差的痛点。

挑战：

AICC 面临三大挑战。一是**监管认可度**，需要政府/监管机构认可 TEE 的安全性；二是**性能损耗**，加密/解密导致推理延迟增加 10-20%；三是**芯片厂商支持**，依赖 Intel SGX、AMD SEV、Nvidia Confidential Computing 等技术。

趋势判断（1-3 年）：

从时间维度看，市场演进呈现三个阶段。短期来看，私有化与云端并行，各有适用场景；中期来看，AICC 可能成为过渡方案（部分开放行业先行，如互联网、零售）；长期来看，效率压力可能推动政策调整，云端占比提升。

趋势五：安全 AI 治理与可信——AI 审计、权限控制、可解释性

随着 AI 在安全场景中承担越来越多的自主决策角色，"谁来监督 AI"成为新命题。问卷数据显示，66 家受访企业中约八成（55 家）要求关键动作需审批，约七成（46 家）实现全链路记录。

6.2.18 AI 治理的三大支柱

支柱 1：权限控制与审批机制

分级授权模型（炼石网络）：

风险等级	操作示例	处理方式
低风险	查询历史告警、生成报告	自动执行，事后审计
中风险	创建工单、推送通知	自动执行，实时告警
高风险	封禁 IP、隔离主机、规则下发	**需人工审批**，审批通过后执行
极高风险	删除数据、修改权限	**禁止 AI 自主执行**，仅生成建议

问卷数据：审批机制普及情况

审批机制在受访企业中普及程度较高（66 家问卷样本）：关键动作需审批的为大多数（55 家，约 83%），实现全链路记录的约七成（46 家），仅部分记录的少数（18 家），无审批的极少数（8 家）。

国际案例：Microsoft Security Copilot 的权限管理

Microsoft Security Copilot 的权限管理体系包括三个方面：一是基于角色的访问控制（RBAC），不同级别分析师有不同 AI 调用权限；二是审批工作流，高风险操作需 SOC 主管审批；三是审计日志，所有 AI 操作记录留存至 SIEM，支持回溯审查。

支柱 2：可解释性与可追溯性

Tenable AI 驱动的洞察和可解释性（2025 年新增）：

> "LLM 驱动的漏洞解释：不仅告诉你有漏洞，还解释为什么有、如何修复、威胁行为者可能如何武器化。非黑盒决策。"

可追溯性的技术实现主要有三种方式。一是**思维链（Chain of Thought）记录**，保存 AI 的推理过程（知道创宇大模型网关记录所有对话的思维链数据）；二是**工具调用日志**，记录 AI 调用了哪些工具、传递了哪些参数（Dropzone AI 的 OSCAR 框架）；三是**版本控制**，每次 AI 决策生成唯一 ID，支持回溯到具体版本（绿盟科技）。

支柱 3：幻觉控制与模型安全

幻觉是安全场景的致命伤：

幻觉在不同系统中的危害程度差异显著。在传统 IT 系统中，幻觉可能导致体验不佳；而在安全系统中，幻觉可能导致误封禁正常业务、漏过真实威胁。

问卷数据：66 家受访企业中近六成（39 家）将幻觉/不确定性列为主要阻碍

应对策略（问卷汇总）：

策略	采用企业	效果	适用场景
规则+大模型混合	炼石网络、美创科技、和利时、烽台科技	规则引擎过滤 90%以上确定性场景，大模型仅处理复杂场景	降低幻觉风险+控制算力成本
RAG 限定场景	和利时、炼石网络	限定场景，多使用自有专家知识库	垂直领域应用

强化学习（RL）	美创科技、知道创宇、绿盟科技	通过人类反馈（RLHF）持续优化	长期迭代优化
思维链验证	绿盟科技	AI 生成答案时展示推理过程，便于人工验证	高风险决策场景
多模型交叉验证	火山引擎（主模型+专用模型协同）	不同模型互相校验，降低单一模型幻觉风险	关键任务

炼石网络的"规则引擎+大模型"架构：

> "采用'规则引擎+大模型'的混合架构，通过规则引擎过滤 90%以上的确定性场景，大模型仅处理复杂的不确定性场景。同时采用多轮对话验证机制，对高风险决策进行二次确认。"

模型安全：防御提示注入与越狱

问卷数据：66 家受访企业中少数（10 家，约 15%）将"提示注入/模型安全"列为主要阻碍

攻击形式：

AI 模型面临的攻击形式主要包括三类。一是**提示注入**，攻击者通过精心构造的输入，操纵 AI 执行非预期操作；二是**越狱（Jailbreak）**，绕过 AI 的安全护栏，生成有害内容；三是**模型盗窃**，通过大量查询推断模型参数。

防御方案（国际案例）：

HiddenLayer AISec Platform 2.0：

HiddenLayer AISec Platform 2.0 提供了全面的防御能力。首先是 **Agent & MCP Security**，保护自主 AI 工作流和 MCP 通信；其次是防御间接提示注入、不安全工具使用、内存损坏、高影响自主操作；第三是 **Runtime Protection**，包括模型盗窃检测、数据泄露监控、未授权访问防护。

在国际趋势方面，Microsoft、Google、OpenAI 均设立专职 Red Team；国内实践方面，知道创宇、海云安提供 AI Red Team 服务。

持续监控与审计：

持续监控与审计包括两个层面：一是 AI 决策监控，实时监控 AI 的决策分布、异常决策告警；二是定期审计，季度/年度审计 AI 操作日志，识别潜在风险。

趋势判断：

AI 治理将从"可选项"变为"合规刚需"。**Gartner 警告的"2026 年 AI 致死法律诉讼超 2000 起"**将加速企业建立 AI 治理体系。那些缺乏完善审批、审计、可解释性机制的 AI 安全产品，将面临合规风险和客户信任危机。

趋势六：平台整合与收购加速——安全市场格局重塑

2025 年是安全行业并购大年。**SecurityWeek 统计，2025 年共有 8 笔超$1B 的网络安全收购，创历史新高。**

6.2.21 超大型收购案例分析

Google 收购 Wiz（$32B）：史上最大网络安全收购

该收购于 2025 年 3 月宣布,预计 2026 年完成,金额达$320 亿(全现金)。其战略意义体现在三个方面：一是 Google 补强云安全能力(Wiz 是云安全态势管理 CSPM 领导者)；二是 Wiz 成为 Google Cloud Security 的核心；三是 2025 年 11 月获美国司法部(DOJ)批准。在 AI 关联方面,Wiz 的云安全能力是 AI 工作负载安全的基础(AI 训练/推理环境大多在云上)。

Palo Alto Networks 收购 CyberArk（$25B）

该收购于 2025 年末宣布,金额达$250 亿。其战略意义体现在三个方面：一是 Palo Alto 补全身份安全短板(CyberArk 是 PAM 特权访问管理领导者)；二是成为史上第二大网络安全收购；三是强化"平台化"战略(Network+Cloud+Endpoint+Identity 全栈覆盖)。在 AI 关联方面,身份安全是 AI 时代的核心(AI 服务需要特权身份管理,防止 AI 凭证泄露)。

Cisco 收购 Splunk（$28B，2024 年完成）

2025 年整合动作主要包括两个方面：一是 Splunk AI Assistant 扩展,支持 SPL 查询自动生成、日志分析自动化、异常检测；二是 Cisco 安全产品组合 AI 化,整合 Splunk 的 AI 能力到 SecureX、Umbrella、Duo 等产品。其战略意义在于,Cisco 通过 Splunk 获得企业级日志分析+AI 能力,强化 SIEM/SOC 市场地位。

6.2.22 AI 安全专项收购

Cato Networks 收购 Aim Security（2025 年 9 月）

该收购的标的为 AI 安全 startup，意义在于传统 SASE 厂商快速补全 AI 安全能力。

Check Point 收购 Lakera（2025 年 11 月）

该收购的标的为 AI 安全 startup（专注 LLM 应用安全），意义在于防火墙厂商进入 AI 应用安全领域。

Snyk 收购 Invariant Labs（2025 年 6 月）

该收购的标的为 AI 数据泄露和 MCP 漏洞防护，意义在于应用安全厂商强化 AI 安全能力，推出 Snyk AI Trust Platform。

Momentum Cyber《2025 年网络安全并购报告》核心发现：

该报告的核心发现包括三点：一是 **2025 年上半年融资额$94 亿**（三年最高）；二是**收购趋势**从"Capability M&A"（能力收购）转向"Sovereign M&A"（主权/区域收购）；三是 **AI 成为估值溢价因素**，带有 AI 能力的标的估值溢价 30-50%。

6.2.23 平台化趋势对市场格局的影响

Palo Alto Networks 的激进平台化战略：

其核心逻辑是"single vendor"策略,要求客户整合到单一平台。产品矩阵包括五大类：一是 NGFW(下一代防火墙)；二是 Prisma Cloud(云安全)；三是 Cortex XDR(端点检测响应)+Cortex Copilot(AI 助手)；四是 XSIAM(扩展版 SIEM+SOAR+XDR)；五是 **CyberArk PAM**(收购后整合)。在 AI 能力方面,包括 Precision AI(自研安全专用 AI 引擎)、AI Runtime Security(AI/ML 工作负载全生命周期保护)、多模型混合架构(OpenAI, Anthropic, 自研模型)。从客户反馈来看,优势在于拥有全面的产品组合,"single vendor"策略吸引大型企业；挑战在于激进的平台化策略引发部分客户抵触,担心**供应商锁定**。

Microsoft 的"超级平台"战略：

Microsoft 推出的 **Security Copilot** 横跨 Defender、Sentinel、Intune 等所有微软安全产品。其差异化体现在三个方面：一是唯一横跨全栈的超大厂 AI 安全方案；二是最强的企业级数据隐私保障；三是 Agent 能力从安全延伸到合规、风险管理等领域。在市场表现方面，集成在 Microsoft 365 E5 Security 中，随微软安全栈销售。

Google Cloud Security (Gemini in Security Operations)：

Google Cloud Security 整合了 Google Security Operations（原 Chronicle）+ Mandiant，定位为 Intel-driven, AI-powered。其差异化体现在三个方面：一是原生多模态能力（文本+代码+日志）；二是 PB 级安全数据湖+Gemini 的组合；三是 Mandiant APT 级威胁情报优势。

平台化的两面性：

维度	优势	劣势
客户视角	一站式采购、统一运维、数据打通	供应商锁定、灵活性下降、价格谈判弱势
厂商视角	客户粘性强、交叉销售、整体收入高	整合复杂度高、产品线臃肿、创新速度慢
行业视角	降低集成成本、提升整体安全水平	市场集中度提高、中小厂商生存空间压缩

6.2.24 专业厂商的生存空间

"Best of Breed" vs. "Platform"之争：

这一争论分为两大阵营：Best of Breed 阵营专注细分领域深度，如 Abnormal Security（邮件安全）、Vectra AI（NDR）、Snyk（开发安全）；Platform 阵营则追求全栈覆盖，如 Palo Alto、Microsoft、Google。

专业厂商的差异化路径（国际案例）：

Abnormal Security：邮件安全 AI

Abnormal Security 的技术壁垒在于行为 AI（学习每个用户的正常通信模式），部署优势为 API 集成、15 分钟完成部署（零破坏性），市场定位为"唯一真正的行为 AI 邮件安全"（无签名库依赖）。

Vectra AI：AI 驱动 NDR

Vectra AI 的技术壁垒在于 Attack Signal Intelligence（专注攻击者行为而非单纯异常），量化收益方面 Texas A&M System 年节省$700 万、威胁调查从数天缩短到数分钟，市场定位为 2025 年 Gartner NDR 魔力象限 Leader。

Dropzone AI：AI 原生 SOC 分析师

Dropzone AI 的技术壁垒在于 OSCAR 框架（结构化调查方法）+ Multi-Agent System；根据公司公开披露与媒体访谈口径，2025 年 ARR 增长约 11 倍、入选 Fortune Cyber 60（数据未经独立第三方审计），市场定位为"业界首个真正自主的 AI SOC Analyst"。

国内专业厂商策略（问卷数据）：

国内专业厂商的差异化策略主要体现在三个方面（66 家问卷样本）。一是数据优势，约七成厂商（45 家）将"自有日志/样本/情报沉淀"列为差异化；二是行业 Know-how，绝大多数（55 家，约 83%）将"行业场景深耕"列为差异化；三是低成本私有化，约三分之一厂商（21 家）将"低成本私有化交付"列为差异化。

6.2.25 新兴创业公司的机遇与风险

Agentic Security 赛道投融资热潮：

公司	轮次	金额	估值	投资方	时间
7AI	A 轮	$130M	$7 亿	Index Ventures 领投	2025 年 12 月
Torq	D 轮	$140M	$12 亿	未披露	2025 年 1 月
Noma Security	B 轮	$100M	未披露	Evolution Equity Partners 领投	2025 年 7 月
WitnessAI	新轮	$58M	未披露	Sound Ventures 领投	2026 年 1 月
Zenity	B 轮	$38M	未披露	Third Point Ventures、DTCP 领投	2024 年 10 月
Dropzone AI	B 轮	$37M	未披露	Theory Ventures 领投	2025 年 7 月
Prophet Security	A 轮	$30M	未披露	Accel 领投	2025 年夏

CB Insights Mosaic 评分（Agentic Security 领域）：

CB Insights Mosaic Score 是评估私营公司综合健康度的专有评分体系（满分 1000 分），涵盖资金实力（Financial Strength）、管理团队（Management）、增长势头（Momentum）、行业健康度（Market）四个维度。

据 CB Insights 2026 年 2 月发布的 *Early-Stage Trends Report*（来源：https://www.cbinsights.com/research/report/early-stage-trends-report-agentic-security-and-more-2026/），该机构追踪了 Agentic Security 领域的 **21 家创业公司**，其中 10 家进入管理团队维度前 100 名。

7AI 详细评分：

维度	**得分**	**说明**
Mosaic 综合分	667/1000（同比-78）	综合健康度
资金实力	987/1000	$166M 融资，资金充裕
管理团队	798/1000	Cybereason 原班团队
增长势头	351/1000	低于平均，产品仍在验证阶段
商业成熟度	2/5（Validating）	尚在测试和完善产品

值得关注的运营数据：根据 7AI 公司公告与发布会口径，其 AI Agent 已处理 250 万+告警、完成 65 万+安全调查，并援引 Fortune 500 客户案例报告 95-99%的误报消除率以及调查时间从数小时缩短至分钟级。上述均为厂商自述/客户案例口径，尚缺独立第三方审计数据。

创始人背景的两面性：CEO Lior Div 的前一家公司 Cybereason 曾巅峰估值$30 亿，但此后暴跌至$3 亿，2022 年被迫大规模裁员，2025 年 10 月被收购。这一经历既证明了团队的行业经验，也提示投资者和客户需关注高估值创业公司的可持续性。

赛道整体被收购概率极高：21 家 Agentic Security 创业公司中，18 家的 M&A 概率评分超过均值（21.6%），群体平均达到 **55%**——意味着多数公司可能在 2 年内被大厂收购。这与安全行业的历史规律一致（Demisto 被 Palo Alto 以$5.6 亿收购、Mandiant 被 Google 以$54 亿收购），也提示甲方在选型时需评估供应商的独立运营持续性。

其他值得关注的 Agentic Security 新兴公司：

Zynap（巴塞罗那）Mosaic 623 分，融资$12.9M，方向为用生成式 AI 模拟威胁的预防式安全，创始人 Daniel Solís 曾创立并出售 Blueliv。团队仅 2 人但正积极扩招，44%岗位面向销售，显示从研发向商业化转型。

关键洞察：

从投融资数据可以得出三点核心洞察。一是 **A 轮融资金额巨大**（$30M-$130M），反映资本对 Agentic Security 的高度认可；二是**估值快速攀升但需警惕泡沫**，7AI A 轮即达$7 亿估值，但增长势头评分仅 351/1000，商业成熟度 2/5，说明高估值主要基于团队和方向预期而非已证实的商业规模；三是**竞争白热化**，Agentic SOC 赛道至少 5 家头部 startup（7AI、Dropzone、Prophet、Simbian、Torq），且 55%的 M&A 概率意味着行业整合在即。

国内创业环境差异（Briefing 洞察）：

> "AI 时代，小团队优势被无限放大（3-5 人团队灵活性强）。但中国 TOB 市场仍看重品牌，创业公司需要'大厂背书'或'渠道合作'。"

火山引擎策略：

> "不做大规模客户扩张，聚焦 FDE 模式（前沿部署工程师）深度定制。对标 Palantir（高客单价、少客户数、深度定制）。"

趋势判断：

平台整合加速将导致市场**"两极分化"**：一是**头部集中**，Microsoft、Google、Palo Alto 等平台厂商通过并购快速扩张，市场份额提升；二是**专业突围**，Abnormal、Vectra、Dropzone 等专业厂商通过技术壁垒+快速迭代，守住细分市场；三是**中腰部挤压**，缺乏技术壁垒的中型厂商生存空间压缩，面临被收购或淘汰。

对于国内厂商，**行业深耕+低成本私有化**是抵御平台化挤压的核心策略（问卷数据验证）。

趋势七：MCP/A2A 协议与 OpenClaw 等运行时驱动的安全产品开放化

安全产品的设计哲学正在经历一次深层转变：从"封闭平台、全栈自研"走向"开放接口、能力即服务"。需要厘清的是，这一变革涉及三个不同层次的技术构件，不能笼统并列：MCP（Model Context Protocol）正在成为 AI 工具连接层的重要开放协议，使大模型能够标准化地访问外部工具与数据源；A2A（Agent-to-Agent）协议在 Agent 间协作层快速发展，定义多智能体之间的通信与编排方式；而 OpenClaw 等可自托管的 Agent 运行时与网关类产品，则在执行与治理层承接调用编排、权限控制与审计能力，本身仍需配套安全加固。三者并非同一层级的"接口标准"，而是协同推进安全能力开放化的协议层、协作层与运行层。

从本轮调研来看，MCP 化已不再是个别厂商的技术实验，而是头部安全厂商的共同战略选择。360 数字安全的路径最为典型：将旗下所有安全能力（威胁情报、终端检测、流量分析等）全部 MCP 化，中国石油和上海公安已通过 OpenClaw 直接调用 360 的安全能力，实现了"甲方用自己的 AI 平台，按需消费安全能力"的新型服务模式。深信服则将检测大模型和运营大模型均 API/MCP 化，与主流开发工具链深度集成。安恒信息已建立专门的 MCP 平台，启明星辰的 AIDK 框架同时兼容 MCP 和 A2A 协议。

这一趋势对市场结构的影响是系统性的。其一，价值评估维度转变——安全产品的核心价值从"平台功能是否完整"转向"专项能力是否领先、接口是否标准"；其二，竞争格局重组——MCP 化使得小型专精厂商能够以低成本进入大型客户的技术栈，传统大厂的平台优势被部分稀释；其三，新的商业模式涌现——按调用次数计费、能力订阅、数据共享协议等新型商业安排正在探索中，可能改变安全行业传统的项目制和许可证收入模式。

预判：2026-2027 年，MCP 化将从头部厂商的战略选择扩散为全行业的标准配置。率先完成能力 MCP 化并建立接口质量口碑的厂商，将在下一个技术周期中享有先发优势；而坚守封闭平台策略的厂商，将面临越来越难以维系的客户忠诚度挑战。

趋势八：自训大模型收益递减，工程化壁垒崛起

2024 年前后，"自建安全大模型"被视为技术领先的重要标志，多家安全厂商将垂域大模型的微调投入列为战略重点。然而，本轮多场深度 Briefing 透露的行业共识表明，这一竞争逻辑正在发生根本性转变：专项微调的边际收益快速递减，开源基座模型的通用能力已能覆盖大部分场景。

这一共识的形成有其技术背景。千问、DeepSeek 等国内开源基座模型的快速迭代，使得通用语言理解和推理能力的门槛大幅降低；而安全垂域的特殊知识，通过 RAG（检索增强生成）和工具调用的方式注入，效果已接近甚至优于专项微调的小参数模型。深信服明确表示，32B 以上模型已不再进行安全专项微调；360 数字安全在自训 14B 安全大模型的同时，通用任务已全面转向开源；启明星辰坦言"现在调用第三方有时候更好"。6 家受访厂商共同传递的信号是：继续加大自训大模型投入，性价比已不如优化工程化能力。

竞争壁垒的转移带来了市场格局的重新评估。在"模型能力"阶段，具备算力资源和研究团队的大厂占据绝对优势；在"工程化能力"阶段，竞争维度变为多个更难追赶的综合要素：一是语料生产平台的质量与规模（深信服 15 年安全运营数据积累构成实质性护城河）；二是数据底座的深度（360 百 PB 级安全数据在特定场景的模型精调中仍难以替代）；三是工具链的成熟度（MCP 化、CI/CD 集成、客户端覆盖等工程化指标正成为新的选型标准）；四是反馈闭环的速度（谁能更快地将真实客户数据转化为模型迭代，谁就掌握了持续进化的优势）。

对于中腰部厂商而言，这一趋势转变既是挑战也是机会：放弃高成本的自训大模型路线，专注于优势场景的工程化深耕，可能是比维持全线竞争更明智的战略选择。

6.3 主要阻碍

基于国内厂商问卷和 Briefing 访谈，我们识别出七大核心阻碍。

6.3.1 数据可用性与质量（约六成受访企业提及，排名第一）

安全 AI 面临的数据挑战包括三个层面：**获取难**（政务、金融客户数据不出网，合规限制跨机构流动）、**质量差**（日志格式不统一、标注数据稀缺、告警噪声高达 90%+）、**时效弱**（纯内网环境下威胁情报更新滞后，APT/0day 等长尾样本不足）。

国内外数据积累差距明显：Microsoft 每日处理 65 万亿安全信号，CrowdStrike 拥有 2 万亿+事件数据库，国内头部厂商数据规模仍有 1-2 个数量级差距。美创科技探索了"AI 自动日志标准化"（接入时效约 1 小时），炼石网络构建了多源数据融合平台，但行业级的数据共享机制尚未建立。

6.3.2 评测缺失（约三分之一受访企业提及）

安全 AI 领域缺少类似 ImageNet/GLUE 的统一 Benchmark，各厂商效果数据缺乏可比性（如降噪率在 30%-99%间波动，MTTR 改善从 10%到 50%不等，基准和测试场景均未标准化）。ROI 量化困难、POC 周期长、场景差异大，进一步增加了客户验证成本。

国际上，MITRE ATT&CK Evaluations 2025 年已新增"AI-Powered Detection"维度，Gartner 新增了多个 AI 相关魔力象限。国内尚缺少第三方独立评测机构，多数厂商仍依赖客户案例证明效果。

6.3.3 权限与审计（少数受访企业明确提及，但 Briefing 中频繁讨论）

AI Agent 自主操作带来误操作（误封正常业务 IP）、权限滥用（提示注入操纵 AI）、审计缺失（决策黑盒）、责任界定困难等风险。Gartner 预测 2026 年将出现 Agentic AI 部署导致的公开安全事件。

国内大多数受访企业（66 家样本中 55 家）已建立关键动作审批机制，约七成（46 家）实现全链路记录。炼石网络实施了分级授权（低/中/高/极高风险对应自动执行→人工审批→禁止 AI 执行），广东盈世支持全操作可追溯、可回滚。随着 Agent 自主能力增强，权限与审计将从"可选项"变为"必选项"。

6.3.4 幻觉与提示注入（近六成+少数受访企业提及）

安全场景对幻觉的容错率极低——误封正常业务或漏过真实威胁都是不可接受的。知道创宇的内容安全小模型准确率 86%（目标 95%），14%的错误率在安全场景不可接受。

主流应对策略是**"规则引擎+大模型"混合架构**（炼石网络等 4 家采用），规则过滤 90%+确定性场景（幻觉率为 0），大模型仅处理 10%复杂场景。其他方案包括 RAG 限定场景、多模型交叉验证、思维链验证等。

提示注入方面，直接注入（"忽略之前指令"）和间接注入（通过邮件/日志注入恶意指令）均需防范。知道创宇大模型盾、HiddenLayer 等提供了输入过滤+输出验证+上下文隔离的防御方案。

6.3.5 私有化部署的算力与成本瓶颈（约三分之一受访企业提及）

云端与私有化的效果鸿沟是核心矛盾。火山引擎反馈：私有化模型与云端效果差距"不止一个数量级"。知道创宇的实测数据更直观：SaaS 版（千问 235B）降噪率可达 99%，私有化版（千问 7B）仅 30%-50%。

应对路径包括：推理加速（vLLM/TensorRT，速度提升 2-5 倍）、小模型蒸馏（体积减少 50-70%）、分层模型架构（规则+小模型+大模型协同，成本降低 40-60%）。火山引

擎还在探索 AI 机密计算（TEE 环境中处理加密数据），试图兼顾数据主权与模型能力，但监管认可度和性能损耗仍是挑战。

6.3.6 客户付费意愿低

"用户要 AI 能力但不愿多付钱"是行业普遍现象。美创科技客户对日均 1000 token（成本<$0.01）仍高度敏感。根源在于 ROI 不清晰、预算分配惯性、以及客户将企业级 AI 与"免费 ChatGPT"对标的认知错位。

厂商应对策略分化明显：美创科技采用捆绑销售（一体机+服务包），知道创宇走免费增值路线（开源免费，项目定制盈利），国际厂商（Dropzone AI、Vectra AI）则通过 ROI 可视化 Dashboard 直接展示量化收益。短期内付费困境将持续，成功案例积累和标准化评测是改善的关键。

6.3.7 AI+安全复合型人才稀缺（约三分之一受访企业提及）

真正稀缺的不是 AI 开发人才或安全人才，而是两者的交叉。5 家企业（炼石网络、摄星科技、海云安、和利时、绿盟科技）均提及此挑战。国内厂商 AI 团队规模远小于国际头部（美创科技 AI 团队 12 人 vs. Microsoft 数百人），导致产品同质化（普遍采用开源模型+标准 RAG）。

务实的应对策略包括：全员 AI 化（知道创宇要求每个渗透工程师必须使用 AI）、聚焦 Agent 层而非自建全栈（火山引擎策略）、以及降低 AI 应用门槛让安全专家直接开发 Agent（火山引擎 Circle 的 Skill 体系）。

6.3.8 Security for AI 价格战对产业健康的影响

Security for AI（针对 AI 系统的安全评估与防护）是本轮调研中被普遍看好的新兴赛道，然而一个令人担忧的信号表明，这一赛道在尚未成熟之前，已经开始遭受恶性价格竞争的侵蚀。

长亭科技在 Briefing 中披露了一个具体案例：某次竞标中，原本价值 300 万元的大模型安全评估项目，最终因字节、百度、蚂蚁等互联网大厂的参与，报价被压至 20 万元以上。价格压缩超过 93%，但安全评估的工作量并未相应缩减。这种"以白菜价抢占高价值项目"的模式，折射出互联网大厂在 Security for AI 赛道的战略意图：不以盈利为目的，以低价快速积累客户案例和市场份额。

这一现象对产业健康发展的威胁是多维度的。一是服务质量的劣化——20 万元的预算根本无法支撑一次深度、系统性的 AI 安全评估，必然导致评估流程的简化甚至走形式化；二是专业厂商的投入回报被严重压缩——安恒信息、悬镜安全、长亭科技等专注 Security for AI 的安全厂商，在产品研发和团队建设上已有大量投入，价格战使得投入回报周期大幅拉长；三是市场教育被跳过——在低价竞争环境下，买方对 AI 安全防护的价值认知无法有效建立，用户对安全评估的预期被锚定在远低于应有水平的价格区间；四是潜在的监管洼地——如果 Security for AI 的质量标准无法得到市场机制的有效约束，有可能催生以低价服务换取形式合规的"安全洗白"现象。

从产业政策角度来看，建立 Security for AI 的质量基准和最低服务标准已经刻不容缓。对于安全厂商而言，在价格战中脱身的路径只有两条：一是向上走，聚焦需要深度专业能力的高端客户（金融、能源、大型央企），用不可替代的专业深度换取合理回报；二是走

差异化，将 Security for AI 与已有的 AI for Security 能力形成协同，以"攻防一体、检测修复联动"的整合价值主张，拉开与纯价格竞争者的距离。

本章总结

驱动力总结

AI 赋能网络安全市场的快速增长，源于六大核心驱动力的叠加（基于 66 家问卷样本的量化画像）。一是告警疲劳与 SOC 效率瓶颈（约四分之三受访企业反馈），传统 SIEM 高误报率不可持续；二是 MTTR 压缩需求（约六成受访企业反馈），业务连续性要求分钟级响应；三是安全人才结构性错配（约半数受访企业反馈），复合型人才极度稀缺，基础岗位供大于求；四是合规审计自动化需求（约三分之一受访企业反馈），政策驱动+持续合规诉求；五是 AI 驱动攻击的倒逼效应，攻防不对称加剧；六是国内客户画像，告警疲劳+人力缺口是核心双驱动（分别约七成与约三分之二受访企业提及）。

趋势总结

六大趋势正在重塑 AI for Security 市场格局（结合 66 家问卷样本的量化判断）。一是从 Copilot 到 Agent，多家机构预期 2026—2028 年 AI/Agent 能力在 SOC 的渗透显著提升（"2028 年 70%"等具体数字尚缺可公开核验的原始出处，仅作方向性参考），Agent 化是未来 3 年核心竞争力；二是 RAG/TAG 标配化，从"能说"到"能做、能追溯"，约三分之一受访企业实现完整闭环；三是多模态探索期，约半数以上受访企业（56%）将多模态列为重点投入方向，但受限于模型能力+成本，2-3 年后成熟；四是私有化与云端博弈，效果差距"不止一个数量级"，AICC 可能成为过渡方案；五是 AI 治理刚需化，大多数受访企业要求审

批、约七成实现全链路记录，合规压力推动治理体系建设；六是平台整合加速，2025 年 8 笔超$1B 收购，市场"两极分化"（头部集中+专业突围）。

阻碍总结

七大阻碍制约 AI for Security 的落地速度。一是数据问题（约六成受访企业反馈），包括可用性+质量+时效性，国内外数据规模差距 1-2 个数量级；二是评测缺失（约三分之一受访企业反馈），缺少行业 Benchmark，ROI 难量化；三是权限与审计（约 15%显性+高隐性关注），存在 AI 失控风险，Gartner 警告 2026 年公开泄露事件；四是幻觉与注入（约六成+约 15%），安全场景零容忍，需规则+AI 混合+多模型验证；五是算力与成本（约六成反馈性能/时延、约三分之一反馈私有化要求），云端 vs 私有化鸿沟，客户对私有化要求高但效果差；六是**付费意愿低**（三家 Briefing 一致反馈），要 AI 但不愿付钱，需 ROI 可视化+市场教育；七是人才稀缺（约三分之一受访企业反馈），AI+安全复合型人才极度稀缺，差距 1-2 个数量级。

展望

未来 2-3 年，AI for Security 市场将经历**从"辅助决策"到"自主运营"**的关键跃迁。**成功的厂商将是那些能够**做到以下五点的企业：一是**实现完整 Agent 闭环**（审批+审计+回滚+复盘）；二是**建立技术壁垒**（数据+知识库+行业深耕）；三是**平衡效果与成本**（分层模型+混合推理）；四是**赢得客户信任**（可解释性+量化 ROI）；五是**构建人才体系**（用 AI 补齐 AI 人才缺口）。

市场格局预判：平台厂商（Microsoft、Google、Palo Alto）通过并购整合占据头部，专业厂商（Abnormal、Vectra、Dropzone）通过技术壁垒守住细分市场，中腰部厂商面临

被收购或淘汰。国内市场将延续"行业深耕+低成本私有化"的差异化路径，AICC（机密计算）可能成为破解"数据主权 vs 模型能力"矛盾的关键技术。

第七章 AI for Security 学术研究前沿

> 本章简要介绍学术界在 AI 赋能网络安全方向的最新进展，帮助安全负责人和厂商把握技术趋势。详细的论文分析和技术细节可参见附录参考文献。

7.1 整体态势

2024-2026 年间，AI for Security 学术研究经历了一次重要转折——从"大模型能否用于安全"的可行性验证，转向"如何在真实环境中用好大模型"的工程化探索。

发表量激增：在 IEEE S&P、ACM CCS、USENIX Security、NDSS 四大安全顶会上，涉及 LLM/AI 的论文占比从 2023 年的不足 5%增长到 2025 年的超过 15%。从研究方向分布来看，安全运营自动化（约 25%）和漏洞检测与修复（约 23%）是两大主要方向，其次是渗透测试与攻防（约 18%）、威胁情报分析（约 15%）、AI 系统安全（约 12%）和评测基准构建（约 7%）。这一分布高度契合产业界的真实需求——运营效率、漏洞管理、主动防御始终是企业安全的核心痛点。

三个关键技术转变：

5. **从单一模型到多 Agent 协作**：几乎所有高质量研究都采用了多 Agent 架构（分工协作显著优于单一模型）
6. **从通用模型到领域知识增强**：RAG、知识图谱、Fine-tuning 注入安全知识已成"标准配置"
7. **从功能实现到安全设计**：研究者开始攻击自己创造的 AI 安全工具，"红蓝对抗"思维进入 AI 安全研究。

产学研鸿沟正在缩小：企业向学术界开放脱敏数据、联合研究增多、学术工具快速开源商业化（如 PentestGPT 获 GitHub 数千 star）。

7.2 六大研究方向概览

方向一：安全运营自动化（占比约 25%）

聚焦告警分诊和日志分析。代表性工作如 CORTEX 的多 Agent 告警分诊系统，相比单一模型减少 40%误报；LogParser-LLM 用大模型自动解析非结构化日志，降低日志管理的维护成本。

产业关联：与当前 SIEM/SOAR 厂商的 AI 化方向高度吻合，是学术成果最快落地的领域。

方向二：漏洞检测与修复（约 23%）

LLM 在代码审计中展现出与传统 SAST 工具互补的能力：擅长理解代码语义和跨函数逻辑，但在大型代码库和复杂数据流分析上仍不及专业工具。CVE-Bench 等评测基准的建立，为客观评估提供了标准。值得注意的是，学术界已开始研究如何对抗 LLM 代码审计工具（Flashboom Attacks），提示厂商不能盲信 AI 审计结果。

产业关联：开发安全（DevSecOps）厂商可关注 LLM+传统 SAST 的混合方案。

方向三：渗透测试与攻防（约 18%）

自动化渗透测试 Agent 是最活跃的方向之一。PentestAgent、CIPHER 等框架实现了从信息收集到漏洞利用的自动化流程，部分工具在 CTF 竞赛中的表现已接近人类选手。学术界也在探索零日漏洞自主利用的可能性，这一方向的伦理和合规边界需要行业高度关注。

产业关联：红队服务和渗透测试厂商的效率工具，但需严格的伦理约束和使用边界。

方向四：威胁情报与恶意软件分析（约 15%）

LLM 在两个方面展现价值：一是辅助逆向工程（自动化反编译代码的语义理解和函数命名），二是自动化 CTI 提取（从非结构化威胁报告中提取 IOC、TTP 等结构化情报）。但 LLM 生成恶意软件变种的"双刃剑"效应也引起了学术界的警惕。

产业关联：威胁情报厂商（如 Recorded Future）和恶意软件分析平台可直接受益。

方向五：AI 系统安全与评测（约 12%）

这一方向关注 AI 安全工具自身的安全性。Agent Security Bench 等评测框架专门评估 Agent 系统面对提示注入、越狱攻击时的脆弱性；"Cloak, Honey, Trap"等研究提出了主动防御恶意 Agent 的技术方案。OWASP Top 10 for LLM Applications 已成为行业参考标准。

产业关联：所有部署 AI 安全产品的企业都应关注此方向——AI 工具本身可能成为新的攻击面。

方向六：评测基准构建（约 7%）

SecBench、CyBench、CVE-Bench 等学术评测基准的涌现，为行业急需的标准化评测奠定了基础。这些基准覆盖了检测、调查、响应、漏洞修复等多个维度，但行业采用度仍低，产学脱节有待解决。

产业关联：厂商和行业协会可参考这些学术基准，推动建立行业统一的安全 AI 评测标准。

7.3 对产业界的核心启示

8. **多 Agent 架构是趋势**：学术界已充分验证，单一大模型难以胜任复杂安全任务，分工协作的多 Agent 架构是更可靠的技术路径；
9. **"规则+AI"混合优于纯 AI**：学术研究与产业实践一致表明，确定性场景用规则、不确定性场景用 AI 的混合架构，在效果和成本上均优于纯 AI 方案；
10. **AI 工具自身安全不可忽视**：提示注入、越狱攻击对 AI 安全产品的威胁已被学术界充分证实，企业在采购和部署 AI 安全产品时应将其自身安全性纳入评估；
11. **评测标准化刻不容缓**：学术界已提供了基准工具，产业界需要推动行业级标准落地；
12. **跟踪学术前沿的方式**：建议 CISO 关注四大安全顶会（S&P、CCS、USENIX Security、NDSS）的 AI 相关论文，以及 arXiv cs.CR 分类的预印本动态。

第八章 常见能力清单

撰写说明:本章按"必备/常规/可选"三级划分 AI 安全产品的能力,为采购决策者提供能力识别清单。每个能力项包含定义、典型实现方式、评估要点及代表厂商。

8.1 必备能力(Must-Have)

这些能力是 AI 安全产品的最小可交付集合,缺失任一项将导致系统无法安全上线或无法满足基本合规要求。

8.1.1 数据接入与检索

定义

支持多源异构安全数据的标准化接入(告警/日志/流量/资产/情报等),并提供高效的检索和查询能力,使 AI 能够获取完整的上下文信息。

典型实现方式

典型的数据接入与检索实现方式包括四个方面。一是**标准化协议接入**,支持 Syslog、SNMP、API、Kafka 等主流协议;二是 **AI 智能标准化**,通过 LLM 自动解析和转换非结构化日志,如美创科技案例中展示的无需厂商字典、1 小时内接入新设备的能力;三是**向量化检索**,将安全数据向量化后存入向量库,支持语义检索;四是**分层接入**,采用实时流加历史数据湖分离的架构设计。

评估要点

评估数据接入与检索能力时需要关注五个核心要点。首先是支持数据源类型数量,应覆盖告警、日志、流量、样本、情报、工单、知识库等多种类型;其次是接入时效性,标准要

求从 0 到接入完成的时间应≤2 小时；第三是标准化能力,需要检验是否需要人工编写解析规则；第四是检索性能,考察 TB 级数据的查询响应时间；第五是向量化覆盖度,明确哪些数据类型支持语义检索。

国际代表厂商

在国际市场上,Microsoft Security Copilot 原生接入微软全栈安全产品(Defender, Sentinel, Intune 等),每日处理 65 万亿信号；Google Security Operations 提供 PB 级安全数据湖,具备原生多模态能力(文本+代码+日志)；Splunk AI Assistant 专注于企业级日志分析,支持 SPL 自动生成。

国内代表厂商(问卷数据)

国内厂商中,美创科技的日志接入支持 Kafka、前置库、API 等,通过 AI 语义解析实现标准化转换,可在 1 小时内完成接入；绿盟科技覆盖告警/日志、流量/会话、样本/沙箱、情报、漏洞库、工单、知识文档等 7 类数据；炼石网络提供多源数据融合平台,实现"数据可用不可见"的跨机构协作。

警示信号

在评估过程中需要警惕三类警示信号。一是仅支持自有设备数据接入,表明生态封闭；二是接入需要 1 周以上时间,说明缺少 AI 标准化能力；三是无向量化检索能力,只能进行关键词匹配。

8.1.2 审计与可追溯

定义

所有 AI 执行的操作(包括思维链推理过程、工具调用、人工审批/驳回记录)必须完整记录,支持事后审计和溯源分析。

典型实现方式

审计与可追溯的典型实现包括四个层面。一是**思维链数据留存**,记录 AI 的推理过程、证据链、决策路径,如知道创宇案例中大模型网关记录所有对话和思维链；二是**操作日志结构化**,区分"AI 建议"和"实际执行",记录时间戳、操作对象、结果状态；三是**多级审计视图**,支持高管、安全主管、审计员不同角色的审计需求；四是**不可篡改存储**,将关键操作日志写入区块链或不可变存储。

评估要点

审计与可追溯能力的评估需要关注五个维度。一是审计粒度,检查是否记录思维链、证据链、工具调用参数；二是审计完整性,验证"AI 建议但未执行"的操作是否也被记录；三是可追溯时效,审计日志保留时长建议≥180 天；四是合规性,确认是否符合等保 2.0、GDPR、AI Act 等要求；五是可视化能力,考察审计报告的易读性。

国际代表厂商

国际厂商方面,Microsoft Security Copilot 提供企业级数据隔离,符合 GDPR/CCPA 要求,不用于模型训练；Palo Alto XSIAM 实现全链路审计,结合 Precision AI 引擎进行根因分析；Darktrace 的所有自主响应动作可审计、可回溯,支持分钟级复盘。

国内代表厂商(问卷数据)

国内厂商中,知道创宇要求关键动作需审批、全链路记录,能力自评 5 分；炼石网络同样实现关键动作需审批、全链路记录,并基于零信任模型提供细粒度权限控制；绿盟科技具备关键动作需审批、全链路记录能力,还提供多智能体协同的自主调查功能。

警示信号

在审计能力评估中需要警惕三类问题。一是无思维链记录,导致黑盒决策；二是 AI 建议未记录,仅记录执行动作；三是审计日志可被 AI 自身删除,暴露权限设计缺陷。

8.1.3 权限最小化

定义

AI 操作的权限边界必须明确,默认遵循最小权限原则(Least Privilege),禁止 AI 拥有超出必要范围的系统权限。

典型实现方式

权限最小化的实现通常采用四种机制。一是**基于角色的权限控制(RBAC)**,为 AI Agent 分配独立角色,限定可执行操作范围；二是**动态权限申请**,AI 需要临时权限时通过工作流向人工申请；三是**工具白名单机制**,AI 只能调用预定义的工具集,禁止任意系统调用；四是**沙箱隔离**,AI 执行的高危操作(如代码生成、脚本执行)在隔离环境中运行。

评估要点

权限最小化能力的评估应关注四个方面。首先是权限粒度,检查是否细化到具体操作,如区分"只读日志"与"执行封禁"；其次是越权检测,验证是否有机制防止 AI 通过工具链绕过权

限限制；第三是权限变更审计,确认 AI 权限调整是否留痕；第四是应急降权,考察发现 AI 滥用权限后能否快速降权。

国际代表厂商

国际代表厂商中,CrowdStrike Charlotte AI 的 Agent 操作需经过 OSCAR 框架验证,关键动作需人工审批；Snyk AI Trust Platform 提供 AI 代码生成安全验证和自定义策略执行,防止生成恶意代码。

国内代表厂商(问卷数据)

国内厂商方面,炼石网络基于零信任模型提供细粒度权限控制,实现模型权重加密存储和推理环境动态销毁；美创科技实现 AI 操作权限边界控制,关键动作需审批；烽台科技同样要求关键动作需审批、全链路记录。

警示信号

权限最小化评估中的警示信号包括三类。一是 AI 拥有管理员权限,属于高危配置；二是无工具白名单机制,AI 可任意调用系统命令；三是权限变更无审计记录。

8.1.4 人机协同审批点

定义

关键安全操作(如封禁 IP、隔离主机、修改防火墙规则、删除文件)必须经过人工审批或确认,AI 不得自主执行。

典型实现方式

人机协同审批的典型实现包括四个层面。一是**分级审批机制**,针对不同风险等级的操作采取不同策略:低风险操作(如日志查询、报告生成)自动执行,中风险操作(如告警聚合、工单创建)通知后执行,高风险操作(如封禁、隔离、规则下发)审批后执行；二是 **IM 协作集成**,在飞书/企微/钉钉群中推送审批请求,人工点击确认,如火山引擎案例所示；三是**超时降级**,审批超时后自动降级为"仅生成建议"；四是**回滚机制**,人工审批后仍可在一定时间窗口内撤销操作。

评估要点

人机协同审批能力的评估需要关注四个维度。一是风险分级标准,明确哪些操作需要审批；二是审批流程效率,考察平均审批时长；三是审批决策辅助,验证 AI 是否提供充分证据支持人工决策；四是紧急通道,检查极端场景下是否支持事后追认。

国际代表厂商

国际市场上,Torq HyperSOC 采用 Multi-Agent System 架构,关键动作需人工审批,支持 48 小时 ROI；Dropzone AI 使用 OSCAR 框架(Observe, Scope, Context, Analyze, Resolve),提供人类级推理能力辅助审批。

国内代表厂商(问卷数据)

国内厂商中,绿盟科技要求关键动作需审批、全链路记录,提供多智能体协同,能力自评 5 分；知道创宇可实现闭环处置(含审批/审计/回滚/复盘沉淀)；炼石网络同样要求关键动作需审批、全链路记录。

警示信号

在审批机制评估中需要警惕三类问题。一是高危操作无审批机制,如直接封禁生产服务器；二是审批流程无证据展示,人工无法判断风险；三是无回滚机制,一旦执行无法撤销。

8.2 常规能力(Standard)

这些能力是 AI 安全产品的核心价值体现,80%的客户需求集中在这一层。缺失这些能力将导致产品竞争力不足。

8.2.1 研判问答

定义

支持自然语言查询安全事件上下文,降低安全运营人员的技术门槛,使非专家也能快速获取威胁情报、攻击路径、影响范围等信息。

典型实现方式

研判问答的典型实现包括四种技术路线。一是 **RAG(检索增强生成)**,结合知识库和实时数据检索,回答"谁在攻击我""影响了哪些资产""如何修复"等问题；二是 **SQL/KQL 自动生成**,用户用自然语言提问,AI 自动转化为查询语言,如 Splunk 的 SPL、Microsoft 的 KQL；三是**多轮对话上下文**,支持追问和澄清,记住前几轮对话内容；四是**图谱推理**,基于"身份-资产-操作"行为图谱进行推理,如美创科技案例所示。

评估要点

研判问答能力的评估需要关注四个核心指标。一是自然语言理解准确率,考察错误理解用户意图的比例；二是回答质量,检验是否提供可操作的建议而非泛泛而谈；三是响应时延,从提问到回答的时间建议≤5 秒；四是多语言支持,确认是否支持中文、英文等多语言混合查询。

国际代表厂商

国际厂商中,SentinelOne Purple AI 提供自然语言威胁猎捕无需学习查询语言,Storyline 自动构建攻击链；Microsoft Security Copilot 基于 GPT-4 架构,结合微软威胁情报网络,自动生成 KQL 查询；Google Gemini for Security Operations 支持自然语言查询安全数据湖,自动生成 YARA-L 检测规则。

国内代表厂商(问卷数据)

国内厂商方面,绿盟科技的研判问答已上线,覆盖全行业,能力自评 5 分；知道创宇的研判问答处于试点阶段,结合威胁情报和漏洞库；炼石网络的研判问答也在试点中,采用 RAG+TAG+规则融合技术。

警示信号

研判问答能力评估中的警示信号包括三类。一是仅支持关键词匹配,而非真正的自然语言理解；二是无上下文记忆,每次提问都需要重复背景；三是回答泛化严重,如"建议加强防护"等无实际价值的答案。

8.2.2 告警摘要与聚合

定义

将海量告警(10 万条+)通过 AI 智能聚合为少量高优先级事件(100 条左右),并为每个事件生成语义化摘要,降低告警疲劳。

典型实现方式

告警摘要与聚合的典型实现包括四个技术环节。一是**时空关联聚合**,基于时间窗口、攻击目标、攻击手法等维度聚合告警；二是**事件维度升级**,从单条告警转向事件,一个事件包含多条告警,如火山引擎案例中 10 万条告警聚合为 100 条事件；三是**语义化摘要生成**,AI 生成"某 IP 对数据库发起暴力破解,已尝试 300 次,建议封禁"等易读描述；四是**优先级智能排序**,基于资产重要性、攻击严重性、影响范围等因素动态排序。

评估要点

告警聚合能力的评估需要关注四个维度。一是聚合比例,10 万条告警能聚合为多少条事件,建议≤1%；二是聚合准确性,检查是否存在误聚合,将无关告警合并为一个事件；三是摘要可读性,确认非技术人员是否能理解；四是误报率改善,对比聚合前后的误报率下降幅度。

国际代表厂商

国际市场上,Vectra AI 的 Attack Signal Intelligence 专注攻击者行为而非单纯异常,误报率极低；Palo Alto XSIAM 提供 AI 驱动的根因分析和自动化事件响应；Abnormal Security 的行为 AI 邮件安全准确率超过 90%。

国内代表厂商(问卷数据)

国内厂商中,美创科技的风险监测智能体日均处理 100 条告警(三个节点平均值),检出率 95%、准确率 95%；绿盟科技的告警摘要已上线,误报下降超过 50%；火山引擎的 Briefing 将 10 万条告警降噪至 8 万条,再聚合为 100 条事件,其中 20 条自动闭环、80 条需要人工处理。

警示信号

告警聚合评估中需要警惕三类问题。一是聚合后告警数量仍然很大,如 10 万降至 5 万,聚合不足；二是摘要依赖静态模板而非 AI 动态生成；三是优先级排序不准确,高危事件被排在低位。

8.2.3 报告自动生成

定义

AI 自动生成安全事件报告、周报、月报、合规报告等,替代人工撰写,提升效率并保证格式规范。

典型实现方式

报告自动生成的典型实现包括四个层面。一是**模板化生成**,基于预定义模板(如等保报告、ISO27001 报告),AI 填充数据和分析结论；二是**自然语言叙事**,将数据转化为易读的叙事文本,如"本月检测到勒索软件攻击 3 起,均已成功拦截"；三是**图表自动生成**,生成趋势图、攻击路径图、资产分布图等可视化图表；四是**多格式输出**,支持 PDF、Word、HTML 等格式。

评估要点

报告生成能力的评估需要关注四个方面。一是报告质量,检查是否符合合规要求,是否有逻辑错误；二是生成时效,从触发到生成报告的时间建议≤10 分钟；三是自定义能力,验证企业能否自定义报告模板和风格；四是多语言支持,确认是否支持中英文双语报告。

国际代表厂商

国际厂商方面,Trellix Wise 提供 AI 驱动的威胁搜索和自动化告警调查,节省 SOC 工作时间 8 小时/天；Microsoft Security Copilot 可自动生成 PowerShell 脚本、修复建议、合规报告。

国内代表厂商(问卷数据)

国内厂商中,绿盟科技的报告生成已上线,支持多场景,能力自评 5 分；摄星科技的报告生成处于试点阶段；炼石网络的报告生成也在试点中。

警示信号

报告生成能力评估中的警示信号包括三类。一是生成的报告逻辑混乱,如时间线错误、因果关系颠倒；二是无法自定义模板,只能使用厂商预设格式；三是生成速度慢,数小时才能生成一份报告。

8.2.4 知识沉淀与经验复用

定义

AI 将处置过的安全事件、专家经验、最佳实践自动沉淀为知识库,并在未来相似场景中复用,实现"越用越聪明"。

典型实现方式

知识沉淀与经验复用的典型实现包括四个环节。一是**事件处置记录结构化**,将每次处置过程(发现→分析→处置→复盘)结构化存储；二是**知识图谱构建**,提取实体(IP、域名、漏洞ID、攻击手法)和关系,构建知识图谱；三是**相似案例推荐**,遇到新事件时 AI 自动推荐历史相似案例和处置方法；四是**专家经验固化**,通过 AI 学习安全专家的决策模式,形成"虚拟专家"。

评估要点

知识沉淀能力的评估需要关注四个维度。一是知识沉淀自动化程度,检查是否需要人工整理；二是知识检索准确性,验证推荐的历史案例是否真正相似；三是知识更新机制,了解新知识如何融入现有知识库；四是知识共享能力,确认是否支持跨组织/跨部门共享。

国际代表厂商

国际市场上,Recorded Future 的 Intelligence Graph 从开源、暗网、技术源实时采集威胁数据,AI 自动关联和优先级排序；CrowdStrike Charlotte AI 拥有 2 万亿+安全事件数据库,覆盖全球 230+国家；知道创宇的 Briefing 中大模型网关记录企业应用全流程,留存"思维链数据"用于未来微调。

国内代表厂商(问卷数据)

国内厂商中,绿盟科技的知识沉淀已上线,结合 RAG/TAG 知识库；摄星科技的知识沉淀处于试点阶段；美创科技采用行业知识库加本地知识库双驱动模式。

警示信号

知识沉淀能力评估中的警示信号包括三类。一是知识库需要人工整理,无自动化沉淀;二是知识检索不准确,推荐的案例与当前场景无关;三是知识库长期不更新,无法适应新威胁。

8.3 可选/进阶能力(Advanced)

这些能力是 AI 安全产品的差异化竞争点,能够显著提升安全运营的自动化程度和效率,但并非所有客户都需要。

8.3.1 自动处置闭环(SOAR 联动)

定义

AI 自主执行低风险处置操作(如封禁已知恶意 IP、隔离受感染主机、下发防火墙规则),并与 SOAR 平台联动,实现从告警到处置的全流程自动化。

典型实现方式

自动处置闭环的典型实现包括四个层面。一是**工具编排能力**,AI 调用防火墙、EDR、NDR、工单系统等工具的 API;二是**分级自动化**,低风险操作自动执行,中风险操作通知后执行,高风险操作审批后执行;三是**处置效果验证**,AI 执行处置后自动验证效果,如封禁后检查是否仍有攻击流量;四是**回滚机制**,处置失误后可自动或手动回滚到处置前状态。

评估要点

自动处置闭环能力的评估需要关注四个方面。一是工具生态丰富度,检查支持多少种安全工具的调用；二是自动化闭环率,了解多少比例的事件能完全自动闭环,建议≥20%；三是处置准确性,考察误处置率,建议≤1%；四是平均处置时长,即 MTTR,建议≤10 分钟。

国际代表厂商

国际厂商中,Torq HyperSOC 可自主关闭 90% Tier-1 告警,从天级响应缩短到分钟级,48 小时 ROI；Dropzone AI 将 MTTR 降低 90%,调查时间从 40 分钟降至 3 分钟；7AI 部署自主 AI agent 处理告警分类、调查、事件响应,融资$130M A 轮。

国内代表厂商(问卷数据)

国内厂商方面,知道创宇可实现闭环处置(含审批/审计/回滚/复盘沉淀),工时下降 30-50%,MTTR 改善超过 50%；炼石网络可闭环处置,MTTR 改善 30-50%；绿盟科技的工具编排/自动处置闭环已上线,处置闭环率提升超过 50%。

警示信号

自动处置闭环评估中的警示信号包括三类。一是仅支持少量工具联动,如只支持自家产品；二是无处置效果验证,执行后不检查结果；三是无回滚机制,处置失误无法恢复。

8.3.2 攻击路径推演

定义

基于当前攻击行为和资产拓扑,AI 自动推演攻击者的可能路径、潜在目标、最终目的,并提前布防。

典型实现方式

攻击路径推演的典型实现包括四个技术层面。一是**攻击图谱构建**,基于网络拓扑、权限关系、漏洞分布构建攻击图；二是**横向移动预测**,预测攻击者从当前位置可能横向移动到哪些资产；三是**威胁建模**,基于 MITRE ATT&CK 框架推演攻击者可能采用的 TTP(战术、技术、程序)；四是**防御建议生成**,针对推演路径 AI 自动生成防御建议,如关闭某端口、加固某服务器。

评估要点

攻击路径推演能力的评估需要关注四个维度。一是推演准确性,检验推演路径是否符合实际攻击者行为；二是推演深度,了解能推演几跳,建议≥3 跳；三是实时性,考察从检测到攻击到完成推演的时间；四是可视化能力,确认攻击路径图是否易读。

国际代表厂商

国际市场上,Palo Alto XSIAM 提供 AI 驱动的根因分析和预测性威胁检测；Darktrace 采用自学习 AI 实现自主威胁检测和响应,构建 Enterprise Immune System；Vectra AI 的 Attack Signal Intelligence 实现跨网络、云、身份的统一检测和攻击路径推演。

国内代表厂商(问卷数据)

国内厂商中,摄星科技的攻击路径推演处于试点阶段；绿盟科技的攻击路径推演已上线。

警示信号

攻击路径推演评估中的警示信号包括三类。一是推演结果不准确,推演路径与实际攻击者行为无关；二是推演深度不足,只能推演 1-2 跳；三是无可视化展示,仅文本描述难以理解。

8.3.3 多模态分析

定义

AI 能够同时处理和关联多种数据类型(文本日志、网络流量、图像截图、语音告警、视频监控等),进行跨模态威胁检测和分析。

典型实现方式

多模态分析的典型实现包括四个应用场景。一是**文本+图像**,分析钓鱼邮件中的图片附件、截图中的恶意 URL；二是**流量+日志**,关联网络流量特征和应用日志发现隐蔽攻击；三是**语音+文本**,分析客服录音中的社工攻击话术；四是**视频+元数据**,在工控场景中关联监控视频和设备运行日志发现异常操作。

评估要点

多模态分析能力的评估需要关注四个方面。一是支持模态类型,检查是否覆盖文本、图像、语音、视频、流量等；二是跨模态关联能力,验证能否将不同模态的数据关联分析；三是准确率,对比多模态分析是否比单模态更准确；四是性能开销,考察多模态分析的算力消耗。

国际代表厂商

国际厂商中,Google Gemini for Security Operations 提供原生多模态能力(文本+代码+日志),支持 PB 级安全数据湖；Abnormal Security 的行为 AI 邮件安全分析邮件文本、图片、附件、发件人行为等多模态数据。

国内代表厂商(问卷数据)

国内厂商方面,美创科技的多模态分析已上线,覆盖数据库+API+终端三类数据安全；绿盟科技的多模态分析已上线；烽台科技的多模态分析已上线,涵盖工业生产现场网络设备、生产设备、主机设备、专用设备数据。

警示信号

多模态分析评估中的警示信号包括三类。一是仅支持单一模态,如只能分析文本日志；二是无跨模态关联,各模态数据孤立分析；三是性能开销过大,多模态分析导致系统卡顿。

8.3.4 对抗鲁棒与安全治理自动化

定义

AI 系统自身具备对抗攻击的鲁棒性(防御提示注入、越狱、模型盗窃等),并能自动化执行安全治理任务(如数据分类分级、合规检查、策略管理)。

典型实现方式

该能力包括两个主要方向。

在**对抗鲁棒**方面,典型实现包括四个层面。一是**提示注入防护**,检测和拦截恶意提示词,如"忽略之前所有指令"；二是**越狱检测**,识别试图绕过 AI 安全护栏的输入；三是**模型安全**,

实现模型权重加密存储、推理环境动态销毁,如炼石网络案例所示；四是**输入/输出过滤**,对AI的输入和输出进行安全检查。

在**安全治理自动化**方面,典型实现包括三个层面。一是**数据分类分级**,AI自动识别数据业务类型和敏感级别,如美创科技案例中效率提升6倍、准确率达80-95%；二是**合规检查**,AI自动检查系统配置是否符合等保、ISO27001等标准；三是**策略管理智能化**,AI根据业务变化自动调整安全策略。

评估要点

该能力的评估分为两个维度。

对抗鲁棒方面的评估要点包括三个。一是提示注入检测准确率,检查是否存在误杀或漏杀；二是越狱防御成功率,验证能否拦截已知和未知越狱方法；三是模型安全机制,确认是否有模型加密、隔离、监控。

安全治理自动化方面的评估要点包括三个。一是分类分级准确率,业务类型识别应≥95%,结合标准后应≥80%；二是合规检查覆盖度,了解支持多少合规标准；三是策略调整效率,考察从需求变化到策略生效的时间。

国际代表厂商

在**对抗鲁棒**方面,Microsoft Security Copilot提供企业级数据隔离、不用于模型训练、符合GDPR/CCPA；Snyk AI Trust Platform提供AI代码生成安全验证、实时漏洞可见性、自

定义策略执行；HiddenLayer 的 AISec Platform 2.0 提供 AI 资产全面可见性、消除影子 AI、Agentic & MCP Security。

在**安全治理自动化**方面,Tenable AI Capabilities 提供 AI 资产发现和治理、影子 AI 检测、AI 供应链风险分析；Noma Security 提供持续 AI 资产和 agent 发现、AI 安全态势管理、AI 运行时保护。

国内代表厂商(问卷数据)

在**对抗鲁棒**方面,炼石网络采用容器化隔离技术,实现模型权重加密存储和推理环境动态销毁；绿盟科技采用强化学习、交互防护。

在**安全治理自动化**方面,美创科技的分类分级智能体已商业化,将 30 万字段处理时间从 30 天降至 10 天,准确率达 80-95%；炼石网络的数据智能化分类分级、安全策略管理智能化已上线；海云安的数据智能化分类分级已上线。

警示信号

在**对抗鲁棒**方面的警示信号包括两类。一是无提示注入防护,AI 容易被恶意输入操控；二是无模型安全机制,模型权重明文存储。

在**安全治理自动化**方面的警示信号包括两类。一是分类分级准确率低,<70%；二是无法适配企业自定义分类标准。

8.4 能力清单总结表

能力类别	能力项	国际代表厂商	国内代表厂商	成熟度

必备能力	数据接入与检索	Microsoft, Google, Splunk	奇安信、深信服、绿盟科技	★★★★★
必备能力	审计与可追溯	Microsoft, Palo Alto, Darktrace	奇安信、启明星辰、绿盟科技	★★★★★
必备能力	权限最小化	CrowdStrike, Snyk	深信服、腾讯安全、启明星辰	★★★★☆
必备能力	人机协同审批点	Torq, Dropzone AI	奇安信 QAX-GPT、绿盟科技风云卫、深信服安全 GPT	★★★★★
常规能力	研判问答	SentinelOne, Microsoft, Google	腾讯安全大模型、奇安信 QAX-GPT、启明星辰	★★★★☆
常规能力	告警摘要与聚合	Vectra AI, Palo Alto, Abnormal	奇安信 AISOC、绿盟科技风云卫、火山引擎 Circle	★★★★★
常规能力	报告自动生成	Microsoft, CrowdStrike, Splunk	绿盟科技、启明星辰、安恒信息	★★★★☆
常规能力	知识沉淀与经验复用	Microsoft, CrowdStrike, Google	奇安信、绿盟科技、腾讯安全	★★★☆☆
可选能力	自动处置闭环	Torq, Dropzone AI, 7AI	奇安信 AISOC、深信服、长亭科技	★★★★☆
可选能力	攻击路径推演	Palo Alto, Darktrace, Vectra AI	奇安信、启明星辰、360 数字安全	★★★☆☆

可选能力	多模态分析	Google Gemini, Abnormal	火山引擎、深信服、绿盟科技	★★★☆☆
可选能力	对抗鲁棒	Microsoft, Google, Palo Alto	悬镜安全、奇安信、360 数字安全	★★★☆☆
可选能力	安全治理自动化	Microsoft, Palo Alto, CrowdStrike	奇安信、启明星辰、深信服	★★★★☆

成熟度说明:

- ★★★★★:产品化平台,可持续运营/迭代;
- ★★★★☆:规模交付,可复用方案;
- ★★★☆☆:试点交付,少量客户;
- ★★☆☆☆:研发中;
- ★☆☆☆☆:计划中。

8.5 能力选择决策树

为帮助采购决策者快速识别所需能力,提供以下决策树:

`

13. 您的组织规模?

└─ 小型(<500 人)→ 必备能力 + 研判问答 + 告警摘要

└─ 中型(500-5000 人)→ 必备能力 + 常规能力 + 自动处置闭环

└─ 大型(>5000 人)→ 全能力覆盖

14. 您的主要痛点?

└─ 告警疲劳 → 告警摘要与聚合(必选)

└─ 人力缺口 → 自动处置闭环(必选)

└─ 合规审计 → 审计与可追溯 + 安全治理自动化

└─ MTTR 过长 → 自动处置闭环 + 攻击路径推演

15. 您的数据敏感性?

└─ 高敏感(政务/金融/能源)→ 私有化部署 + 审计可追溯 + 权限最小化

└─ 中敏感(互联网/教育)→ 混合部署 + 基本审计

└─ 低敏感(消费行业)→ SaaS + 基本能力

16. 您的 AI 成熟度?

└─ 初级(首次使用 AI)→ 必备能力 + 研判问答

└─ 中级(已有 AI 试点)→ 必备能力 + 常规能力

└─ 高级(AI 深度应用)→ 全能力覆盖

`

本章小结:本章按"必备/常规/可选"三级划分了 AI 安全产品的 13 项核心能力,并结合国际国内厂商案例和问卷数据,为每个能力项提供了定义、实现方式、评估要点和代表厂商。采购决策者可根据自身组织规模、痛点、数据敏感性和 AI 成熟度,参考决策树选择所需能力。

第九章 差异化能力与选型评估

> 本章提供 AI 安全产品的系统性选型评估框架。相比 2024 版的八维度模型（侧重大模型基础能力），2026 版框架反映了三个关键变化：Agent 化成为核心差异点、实时检测对小模型的刚性需求、以及中国市场数据合规与本地化的特殊要求。

9.1 评估维度框架（2026 版十维度模型）

维度一：安全场景能力

本维度评估厂商在网络安全领域的积累深度和实战经验，而非 AI 技术本身。重点关注四个方面：一是覆盖的安全场景数量与深度（威胁检测、事件响应、漏洞管理、数据安全、合规审计等）；二是安全知识库的规模和更新频率（攻击手法库、TTP 知识图谱、漏洞库等）；三是实战验证，是否有经过真实攻防演练验证的检测/响应能力；四是行业纵深，在金融、政务、能源、医疗等特定行业的场景理解。AI 是工具，安全是根本——没有深厚安全积累的厂商，AI 再强也只是"花拳绣腿"。

选型时应重点追问：产品覆盖了 MITRE ATT&CK 框架中多少战术和技术？是否有经过实际攻防演练验证的案例？安全知识库的更新周期和行业覆盖情况如何？

维度二：AI Agent 能力

本维度评估从 Copilot 到 Agent 的演进程度，这是 2026 年最核心的差异化方向。重点关注四个方面：一是自主决策水平，从 L1 建议型 Copilot 到 L2 半自主 Agent（关键节点人工审批）再到 L3 全自主 Agent；二是多 Agent 协作，是否支持多个专业 Agent 分工协作（如检测 Agent、调查 Agent、响应 Agent）；三是工具编排能力，能否调用外部安全工具（SIEM、EDR、防火墙、SOAR）执行实际操作；四是人机协同设计，人工介入点是否合

理设置，关键决策是否保留人工审批。Copilot 只能"建议"，Agent 能"执行"——这一能力直接决定了 AI 产品能替代多少人工重复劳动。

选型时应重点追问：产品处于 Copilot 还是 Agent 阶段？Agent 可自主执行哪些操作？哪些操作需要人工审批？多 Agent 之间如何协调、失败时如何回退？是否有 Agent 自主处理的成功案例和量化数据？

维度三：实时检测性能

本维度评估厂商检测引擎在延迟和吞吐量方面的表现。重点关注四个方面：一是推理延迟，实时检测场景要求<10ms，告警分诊可接受<1 秒；二是吞吐量，单卡支撑的 QPS，大型 SOC 要求数万 QPS；三是分层模型架构，是否采用"小模型实时检测+中等模型研判+大模型交互"的分层设计；四是边缘部署，能否在终端/网关等资源受限设备上部署检测模型。大模型推理延迟在秒级，无法满足实时检测的刚性需求——是否具备高效的小模型检测能力，是区分"真产品"和"演示品"的关键。

维度四：数据工程能力

本维度评估厂商的数据处理和知识管理能力。重点关注四个方面：一是数据获取与整合，能接入多少种数据源（日志、流量、终端、云、身份等）；二是日志标准化，不同厂商设备日志的自动解析和格式统一能力；三是知识库建设，威胁情报库和内部安全知识库的规模与质量；四是 RAG 质量，检索增强生成的准确性、召回率和引用可追溯性。66 家受访企业中约六成将数据可用性与质量列为 AI 应用的最大阻碍，"垃圾进，垃圾出"是 AI 安全产品的第一定律。

维度五：数据合规与本地化

本维度评估中国政企市场的刚性准入门槛。重点关注四个方面：一是数据不出域，模型推理是否完全在客户环境内完成；二是全栈私有化，能否在完全离线的内网环境中独立运行；三是合规认证，是否通过等保 2.0、数据安全法相关认证；四是数据隔离，多租户场景下的数据隔离机制。政务、金融、能源、军工等行业客户中 80%+要求数据和模型完全部署在自有环境，这不是"优选项"而是"准入门槛"。

维度六：AI 基座与算力适配性

本维度评估厂商在中国市场适配多种 AI 基座模型和国产算力平台的能力。重点关注四个方面：一是多基座兼容，支持哪些国产模型（千问、DeepSeek、百川、智谱等）和开源/商业模型；二是国产 GPU 适配，是否适配昇腾 910B、海光 DCU、寒武纪 MLU 等；三是异构算力调度，能否在混合算力环境中灵活调度；四是模型切换成本，更换底层模型时上层应用需要做多大改动。国产化替代趋势下，绑死在某一个模型或算力平台上将面临客户丢失风险。

维度七：安全可信度

本维度评估 AI 安全产品自身的安全性和可靠性，在 Agent 化趋势下尤为关键。重点关注四个方面：一是幻觉控制，误报率和漏报率的量化数据，是否采用"规则+AI"混合架构；二是提示注入防御，对直接/间接提示注入攻击的防护能力；三是权限与审计，AI 操作的权限控制粒度、全链路审计日志和操作可回滚性；四是可解释性，AI 决策过程是否可追溯、可审计、可向监管机构解释。AI Agent 可以自主执行操作，一旦失控后果可能比没有 AI 更严重。

维度八：评测与效果可证性

本维度评估厂商能否用客观数据证明 AI 产品的效果。重点关注四个方面：一是量化指标，是否提供标准化的效果指标（降噪率、检出率、MTTR 改善、误报率等）；二是测试方法透明度，评测数据集、测试场景和方法是否公开可复现；三是 POC 规范，是否支持在客户真实数据上做 POC；四是 ROI 可视化，是否提供实时 Dashboard 展示 AI 带来的量化收益。66 家受访企业中约三分之一将评测缺失列为主要阻碍，"演示型成功"是行业通病。

维度九：产品成熟度

本维度评估从原型到产品的成熟度，直接影响部署风险和运维成本。重点关注四个方面：一是产品化程度，是项目制定制开发还是标准化产品可开箱即用；二是集成能力，与现有安全工具的集成方式和成熟度；三是客户验证，已有多少客户在生产环境使用、覆盖哪些行业；四是持续迭代，产品更新频率、模型升级机制和 Bug 修复响应时间。国内 71% 的企业仍处于试点阶段，选择产品成熟度不足的方案意味着大量"踩坑"成本。

维度十：本地服务与交付能力

本维度评估本地化服务能力，中国 ToB 市场高度依赖这一维度。重点关注四个方面：一是实施团队的本地化规模和经验；二是驻场支撑的响应时效；三是定制开发能力（如特定行业检测规则、报告模板）；四是培训与赋能体系。AI 安全产品的部署不是"装上就能用"，需要大量数据接入、场景调优和流程适配工作。

9.2 选型评分卡

9.2.1 评分模板

维度	权重	评分(1-5)	加权得分	评估证据

1. 安全场景能力	15%			
2. AI Agent 能力	12%			
3. 实时检测性能	10%			
4. 数据工程能力	10%			
5. 数据合规与本地化	12%			
6. AI 基座与算力适配性	8%			
7. 安全可信度	12%			
8. 评测与效果可证性	8%			
9. 产品成熟度	8%			
10. 本地服务与交付能力	5%			
合计	**100%**	—	**Σ 加权得分（满分 5.0)**	—

评分尺度说明：5 分=行业领先，量化指标排名前列且有公开案例佐证；4 分=明显高于平均水平，已有规模化部署案例；3 分=达标可用，符合主流水准；2 分=部分满足，存在明显短板；1 分=不具备或仅有概念性方案。每项加权得分=评分×权重%，总分为各项加权得分之和（满分 5.0）；总分≥4.0 为优选，3.0–3.9 为合格备选，<3.0 建议淘汰。

"评估证据"列填写指引：建议填写可核验的客观依据，例如客户案例数量与名称、POC 测试报告编号、第三方评测得分（如 MITRE ATT&CK Evaluations、CAPE 等）、合规

认证文号（等保 2.0、ISO 27001 等）、性能基准测试数据（QPS/延迟）、Gartner/IDC/Forrester 报告引用页码，避免仅以厂商宣称材料作为证据。

9.2.2 权重调整建议

以上权重为通用参考值，不同类型的客户应根据自身需求调整，调整后各维度权重之和仍应保持 100%。政务/军工客户应提高"数据合规与本地化"至 20%（+8）和"AI 基座与算力适配性"至 12%（+4），同时下调"AI Agent 能力"至 8%（−4，对自动化执行持保守态度）、"实时检测性能"至 7%（−3）、"评测与效果可证性"至 5%（−3）、"数据工程能力"至 8%（−2），合计仍为 100%。金融客户应提高"安全可信度"至 15%（+3）和"评测与效果可证性"至 12%（+4），相应下调"AI 基座与算力适配性"至 5%（−3）、"数据工程能力"至 8%（−2）、"实时检测性能"至 8%（−2），合计仍为 100%。互联网/科技公司应提高"AI Agent 能力"至 18%（+6）和"实时检测性能"至 15%（+5），相应下调"数据合规与本地化"至 7%（−5）、"本地服务与交付能力"至 2%（−3）、"产品成熟度"至 6%（−2）、"AI 基座与算力适配性"至 7%（−1），合计仍为 100%。中小企业应提高"产品成熟度"至 15%（+7）和"本地服务与交付能力"至 10%（+5），相应下调"实时检测性能"至 5%（−5，告警量较小、对延迟不敏感）、"AI Agent 能力"至 8%（−4）、"评测与效果可证性"至 5%（−3），合计仍为 100%。

9.3 选型决策矩阵

9.3.1 按主要痛点选型

主要痛点	优先关注维度	推荐产品类型
告警疲劳/降噪	实时检测性能、数据工程	AI-SIEM、告警降噪专项工具

分析师人力不足	AI Agent 能力、产品成熟度	SOC Agent 平台
合规审计压力	数据合规与本地化、安全可信度	合规自动化+AI 报告生成
漏洞管理效率低	安全场景能力、评测可证性	AI 漏洞优先级排序+修复建议
威胁狩猎能力弱	安全场景能力、AI Agent 能力	威胁情报 AI+多 Agent 调查平台

9.3.2 按数据敏感性选型

数据敏感级别	**部署模式**	**关键评估维度**
极高（军工、涉密）	完全离线私有化	数据合规、AI 基座适配（国产全栈）、本地服务
高（政务、金融）	私有化为主	数据合规、安全可信度、产品成熟度
中（制造、能源）	混合部署	部署灵活性、数据工程、AI 基座适配
低（互联网、SaaS）	云端 SaaS 优先	AI Agent 能力、实时检测、评测可证性

9.4 常见选型陷阱

陷阱一：演示型成功。 供应商在精心准备的演示数据上表现完美，真实部署大打折扣。规避方式是坚持要求在客户真实数据上做 POC，测试数据应覆盖至少 3 个月，包含正常流量和已知攻击案例。

陷阱二：大模型能力≠安全能力。 厂商强调"我们用了最新的 GPT-4o/DeepSeek"，但安全领域积累薄弱。规避方式是优先考察安全场景能力（维度一），而非底层模型的先进性。

陷阱三：忽视私有化效果衰减。 SaaS 版演示效果优秀，但客户要求私有化部署后效果大幅下降（如降噪率从 99%降到 30%-50%）。规避方式是 POC 必须在目标部署模式下进行，明确合同中的效果保证条款。

陷阱四：Agent 能力过度承诺。 厂商宣称"全自主 Agent"，但实际只是带了几个工具调用的 Copilot。规避方式是要求演示 Agent 在无人干预下完成一个完整的告警调查→研判→响应流程，观察其在边缘案例中的表现。

陷阱五：忽视退出成本。 深度绑定某一厂商后，数据格式、工作流、知识库均无法迁移。规避方式是评估数据导出能力、API 开放程度和与标准格式（如 STIX/TAXII）的兼容性。

9.5 RFI/RFP 关键问题清单（精选 30 题）

以下问题按十个维度组织，供客户在选型过程中直接使用。

安全场景能力： （1）请列出产品覆盖的 MITRE ATT&CK 战术和技术，并说明检测覆盖率。（2）请提供在金融/政务/能源行业的部署案例及量化效果。（3）安全知识库覆盖多少 CVE/IOC/TTP？更新频率是多少？

AI Agent 能力： （4）产品目前处于哪个自主决策级别（Copilot/半自主 Agent/全自主 Agent）？（5）请演示 Agent 完成一个完整的告警调查→研判→响应流程（无人工干预）。（6）Agent 执行失败时的回退和异常处理机制是什么？

实时检测性能：（7）实时检测引擎使用什么规模的模型？推理延迟和单卡 QPS 是多少？（8）是否采用分层模型架构？请说明各层的模型规模和职责分工。（9）请提供在客户真实环境下的性能测试报告。

数据工程能力：（10）支持接入哪些数据源？日志标准化的自动化程度和接入时效是多少？（11）RAG 知识库的规模、更新机制和检索准确率是多少？（12）如何处理客户环境中的低质量、高噪声数据？

数据合规与本地化：（13）私有化部署时，是否有任何数据（包括遥测、诊断）需要外传？（14）是否支持在完全无互联网的离线环境中运行？功能是否有缩减？（15）是否通过等保 2.0 认证？请提供相关合规证明文件。

AI 基座与算力适配性：（16）支持哪些国产大模型和国产 GPU？请提供适配测试报告。（17）更换底层 AI 基座模型的迁移成本和周期是多少？（18）在昇腾 910B/海光 DCU 上的推理性能与 NVIDIA A100 相比差距是多少？

安全可信度：（19）产品的误报率和漏报率是多少？在什么测试条件下取得？（20）是否做过红队级别的提示注入和越狱攻击测试？请提供报告。（21）AI 执行的所有操作是否有完整审计日志？高风险操作的审批机制是什么？

评测与效果可证性：（22）是否愿意在我方真实数据上做 POC？POC 的评估指标和达标标准是什么？（23）请提供 3 家以上现有客户的量化效果数据和可验证的联系方式。（24）合同中是否包含效果不达标的退出或补偿条款？

产品成熟度：（25）产品上市多长时间？目前有多少生产环境客户？（26）与主流 SIEM/EDR/SOAR 的集成方案是否经过生产环境验证？（27）产品更新频率和模型升级机制是什么？是否需要停机？

本地服务与交付能力：（28）本地实施团队规模和本行业交付经验如何？（29）从签约到生产上线的平均交付周期是多长？（30）售后支撑级别（7×24/驻场/远程）和 SLA 承诺是什么？

第十章 代表厂商

本章覆盖国内 27 家和海外 17 家共 44 家代表厂商。海外厂商侧重帮助读者了解国际市场技术方向与竞争格局，国内厂商提供更详细的产品画像与落地实践参考。

10.1 海外代表厂商

本章选取海外代表厂商遵循三个原则：**代表性**（市场影响力与技术创新性）、**可获得性**（产品已 GA 或公开预览）、**样本证据**（有公开案例或第三方验证）。覆盖**平台型厂商 9 家**和**专项型厂商 8 家。**

10.2 海外平台型厂商

1. Microsoft Security Copilot

产品方向：基于 GPT-4 架构的企业级 AI 安全助手，深度整合微软全栈安全产品（Defender、Sentinel、Intune 等）。2025 年 3 月推出 Copilot Agents，支持自主执行事件响应、威胁猎捕和合规检查。

技术特点：自然语言驱动 KQL 查询和 PowerShell 脚本生成；整合微软每日 65 万亿安全信号的威胁情报；横跨 Endpoint、Cloud、Identity、Data 的统一 AI 接口。纯 SaaS 模式，$4/月/用户起（需 E5 Security）。

差异化优势：全栈生态整合（端点到云、身份到数据）；企业级隐私保障（数据隔离，不用于训练）；对微软重度用户几乎零学习成本。

关注要点：定价较高；对非 Microsoft 环境支持有限；Agent 自主决策准确性仍需验证。

2. Google Cloud Security (Gemini in Security Operations)

产品方向：AI 原生安全运营平台（原 Chronicle 整合），强调"情报驱动、AI 赋能"，将 Mandiant APT 级威胁情报与 Gemini 大模型融合。

技术特点：原生多模态分析（文本日志、代码、流量图像）；自然语言查询 PB 级安全数据湖；自动生成 YARA-L 检测规则；深度整合 Mandiant 情报库。纯 SaaS，按数据摄入量计费。

差异化优势：原生多模态能力；PB 级数据处理（Google Cloud 底座）；Mandiant APT 情报加持。

关注要点：企业安全市场渗透率仍在追赶微软；PB 级数据成本需合理规划。

3. CrowdStrike Charlotte AI

产品方向：Falcon 平台内置生成式 AI 引擎，EDR 市场领导者。2025 年秋季率先定义"Agentic SOC"概念，推出 Multi-Agent SOC 架构。

技术特点：OSCAR 框架（Observe/Scope/Context/Analyze/Resolve）驱动自动调查，单事件从 40 分钟降至 3 分钟；Agentic SOC 多 Agent 协作（威胁猎手、检测工程师、事件响应 Agent 并行）；AI 时代防护（AI 应用运行时保护，防投毒/提示注入）。轻量级 Agent+云原生，按端点数订阅。

差异化优势：2 万亿+安全事件数据库覆盖 230+国家；Agentic SOC 愿景行业领导者；端点 AI 能力突出。

关注要点：所有 AI 推理在云端，离线能力受限；定价体系复杂；Multi-Agent 架构仍在演进。

4. Palo Alto Networks (XSIAM / Cortex Copilot)

产品方向：下一代 SIEM+SOAR+XDR 融合平台，最激进的平台化战略推动者。2025 年$25 亿收购 CyberArk 补全身份安全。

技术特点：Precision AI 引擎驱动根因分析和预测性检测（误报率降低 70%+）；自然语言生成 Playbook；AI Runtime Security（模型防火墙、LLM 输入输出过滤）。SaaS 优先，多模型混合（OpenAI/Anthropic/自研）。

差异化优势：收购 CyberArk 后覆盖 Network+Cloud+Endpoint+Identity 全栈；AI Runtime Security 领先。

关注要点：供应商锁定风险高；多次大额收购后整合复杂度；成本高昂。

5. SentinelOne Purple AI

产品方向："AI 原生"端点安全平台，强调从诞生起就基于 AI 设计。Purple AI 是核心差异化。

技术特点：自然语言威胁猎捕（用户反馈准确度高于竞品）；专利 Storyline 技术自动构建攻击链可视化；自动化根因分析（30 分钟→5 分钟）。轻量级 Agent+SaaS，支持完全离线部署。

差异化优势："真正 AI 原生"架构优化；Storyline 攻击链可视化独家专利；最易用的自然语言猎捕。

关注要点：威胁情报覆盖不如 CrowdStrike；第三方工具集成丰富度不足；使用第三方 LLM，定制能力有限。

6. Cisco AI Assistant for Security

产品方向：横跨 Cisco 安全产品组合的统一 AI 接口层，深度融合 Talos 威胁情报网络。2024 年$280 亿收购 Splunk，正在整合双方能力。

技术特点：跨产品统一 AI 接口（SecureX/Umbrella/Duo/Secure Email）；Talos 情报增强（全球最大商业威胁情报网络之一）；网络配置智能优化（防火墙规则精简、ACL 优化）；DNS 威胁检测（DGA 域名、DNS 隧道）。SaaS 为主、本地为辅。

差异化优势：网络安全 40+年积累；Talos 实时性和覆盖面突出；网络+安全原生整合。

关注要点：AI 创新速度较慢；Splunk 整合路线图仍在调整；对 Cisco 生态外支持有限。

7. Fortinet FortiAI / FortiGuard AI

产品方向：中端市场网络安全领导者，AI 与网络设备深度整合，强调性价比。

技术特点：网络流量异常检测（DDoS/数据渗出）；FortiGuard AI 每日分析数亿恶意软件样本；零日攻击行为分析检测。设备内置 AI+FortiGuard 云服务混合架构，轻量级 AI 模型在硬件本地运行。

差异化优势：性价比优势，适合中小企业；设备内置 AI 降低云依赖；产品组合丰富，单一供应商简化管理。

关注要点：在 Agentic AI 等前沿方向投入有限；高端市场竞争力较弱。

8. Splunk AI Assistant

产品方向：企业级数据分析平台领导者的统一 AI 层，面向安全运营和 IT 运维。2024 年被 Cisco $280 亿收购。

技术特点：自然语言→SPL 自动生成；ML 日志异常检测（无需预定义规则）；时间序列预测性分析；Integration Hub 支持客户自选 AI 模型（OpenAI/Anthropic 等）。SaaS+本地双模式，多模型策略+RAG 增强。

差异化优势：深厚的日志分析和数据处理能力（20+年积累）；灵活模型选择避免锁定；跨 IT+安全统一平台。

关注要点：数据摄入量计费模式成本可能快速增长；被收购后路线图不确定性；安全专业度不如专业 SIEM。

9. IBM QRadar AI

产品方向：传统企业级 SIEM，面向追求稳定可靠的大型企业（金融/政府/关键基础设施）。

技术特点：X-Force 威胁情报整合；UEBA 用户行为分析检测内部威胁；事件自动聚类降低告警噪音；Watson NLP 分析非结构化威胁情报。支持本地/私有云/混合云多模式。

差异化优势：企业级可靠性（数十年金融/政府服务经验）；X-Force 全球威胁研究团队；混合云能力适合传统企业转型。

关注要点：AI 创新速度较慢；成本高昂；核心架构设计 10+年，与 AI 原生平台有差距。

10.3 海外专项型厂商

1. Abnormal Security（邮件安全 AI）

产品方向：AI 原生邮件安全平台，用行为 AI 替代传统签名和规则，专注 BEC/钓鱼/账号接管检测。

技术特点：自研行为 AI 引擎——自动学习每个用户通信模式（发件频率、联系人、语言风格），检测偏离基线的可疑邮件；BEC 精准拦截（平均损失$125K/次）；实时安全辅导（可疑邮件弹出提示）。纯 API 集成 SaaS，通过 M365/Google API 接入，**15 分钟部署。**

差异化优势：行为 AI 邮件安全（零签名依赖，可检测零日钓鱼）；部署速度快；拦截+培训结合。

2. Recorded Future（威胁情报 AI）

产品方向：AI 驱动威胁情报平台领导者，覆盖开源情报、暗网、技术源，支持 75+语言。

技术特点：Intelligence Graph——每日 TB 级全源情报采集；自研 NLP 自动提取实体并构建关联图谱；AI 风险评分和优先级排序；预测性威胁分析；2025 年新增供应链情报和数字风险保护。纯 SaaS，原生集成 Splunk/QRadar/CrowdStrike 等。

差异化优势：最全面威胁情报覆盖（75+语言）；AI 自动化情报生成（天级→小时级）；预测性分析（"可能发生什么"）。

3. Darktrace（自主响应 AI）

产品方向：自学习 AI 网络防御先驱，"企业免疫系统"理念，基于无监督学习检测异常。

技术特点：贝叶斯推理自学习算法——无需预定义规则/签名，自动学习每个环境的"正常"行为基线；Autonomous Response 分钟级自动响应（隔离/限速/阻断）；2025 年 NEXT Agent 支持自主调查取证；覆盖 Network/Endpoint/Cloud/OT/IoT/Email 全域。硬件/虚拟/SaaS 多模式，每客户独立 AI 引擎。

差异化优势：自学习 AI 安全平台定位鲜明；自主响应能力成熟；OT/IoT 环境特别适用。

关注要点：2-4 周学习期；完全自主模式可能误杀正常业务；定价偏高。

4. Vectra AI（AI 驱动 NDR）

产品方向：AI 驱动网络检测与响应领导者，2025 年 Gartner NDR 魔力象限 Leader。专注检测"攻击者行为"而非单纯"异常流量"。

技术特点：Attack Signal Intelligence——AI/ML 专注识别攻击者 TTP（横向移动/凭据滥用/C2 通信），非单纯异常流量（误报率比传统 NDR 低 70%+）；跨网络+云+身份+IoT/OT 统一检测；自动风险评分和优先级排序。硬件/虚拟/SaaS 多模式。

差异化优势：专注攻击者行为显著降低误报；扩展到网络+云+身份+IoT/OT 的 NDR；Texas A&M 案例年节省$700 万。

5. Snyk（开发安全 AI）

产品方向：面向开发者的 AI Trust Platform，2025 年 Gartner AST 魔力象限 Leader。"Developer-First"安全。

技术特点：AI 代码生成安全验证（检测 Copilot/Cursor 等生成代码的漏洞，IDE 内实时提示）；Snyk Agent Fix 自主生成修复代码并验证，PR 中直接提交；AI 安全扫描（模型文件/Agent 代码/数据路径，检测投毒/提示注入/泄露）；2025 年收购 Invariant Labs 强化 AI 数据泄露和 MCP 漏洞防护。SaaS，深度集成 IDE 和 CI/CD。

差异化优势：同时覆盖"AI 代码安全"和"AI 应用安全"；Agent Fix 自主修复能力；深度集成 Devin/Windsurf/Claude Code 等 AI 编程助手。

6. Dropzone AI（AI SOC Analyst）

产品方向："人类级 AI SOC 分析师"，自主调查 100%告警，2025 年 ARR 增长 11 倍，Gartner"AI SOC Agents"样本厂商。

技术特点：OSCAR 框架驱动端到端自主调查（单告警 40 分钟→3 分钟）；人类级多步推理能力（非简单模式匹配）；2026 年演进为 Multi-Agent System（威胁猎手/检测工程师/取证分析师等）；**30 分钟快速部署**，无需编写 Playbook。SaaS，Multi-LLM 架构，按告警处理量计费。

差异化优势：定位自主 AI SOC Analyst；部署速度快；OSCAR 框架增强可解释性；MSSP 友好（CBTS 案例利润率提升 25%）。

7. 7AI（Agentic Security 平台）

产品方向：Agentic SOC 平台，目标以多 Agent"群体"替代 L1/L2 分析师的常规工作，覆盖告警分诊、自主调查、情报关联与响应编排全链路。由 Cybereason 联合创始人兼前 CEO Lior Div 与 Shlomi Oberman 于 2024 年创办，种子轮即由 Greylock 领投约$36M，是该赛道规模最大的种子轮之一；RSAC 2026 期间发布完整产品形态，成为 Agentic Security 赛道的重点观察对象。

技术特点：Agent-native 架构，而非在 SIEM/SOAR 之上外挂 Copilot——由分诊 Agent、调查 Agent、情报关联 Agent、响应编排 Agent 等角色协作完成闭环；强调工具编排能力（可调用 SIEM、EDR、IAM、Ticket 等现有安全栈），面向"Day-1 替代 L1"的激进定位；每步决策链全程留痕以支持人工审计与回放。部署形态以 SaaS 为主，按告警处理量/Agent 运行量计费，目标客户集中于北美中大型企业。

差异化优势：创始团队具备 EDR/XDR 时代打造独角兽（Cybereason）的成功经验，安全场景 Know-how 与 Enterprise 销售渠道显著优于纯 AI 背景的同赛道玩家；以"Agent 群体"而非"单一 Copilot"作为核心叙事，产品定位比 Dropzone AI 更激进；种子轮融资规模与估值（RSAC 2026 前后已达约$7 亿）反映投资方对 Agentic SOC 赛道与创始团队的高置信度。

风险与观察点：一是真实环境兑现能力，Agentic 演示与生产部署的差距是该赛道通病，需持续跟踪量化指标（降噪率、MTTR、自主闭环率、误操作率）；二是赛道拥挤，与 Dropzone AI、Prophet Security、Radiant Security、Simbian、Intezer 以及 Microsoft/CrowdStrike/Palo Alto 等平台大厂的 Agentic SOC 路线正面竞争，独立厂商窗口期

收窄；三是高风险动作（隔离/封禁/账号重置）的信任门槛，需配套"审批—回退—审计"机制以建立客户信心。选型建议：当前阶段适合作为 AI Agent 能力（9.1 维度二）的标杆参照与试点 POC 对象，规模化生产部署仍需等待更完整的客户案例与效果数据。

8. Torq（AI 驱动 SOAR）

产品方向：Hyperautomation 平台，自称传统 SOAR"替代者"。2025 年$140M D 轮，估值$12 亿。

技术特点：HyperSOC 2.0——业界首个 Multi-Agent 安全编排架构；首个原生支持 MCP 协议的安全编排平台；AI 驱动自然语言→Playbook 自动生成；自主关闭 90% Tier-1 告警。SaaS 优先，No-Code/Low-Code，集成 1000+安全工具。

差异化优势：Hyperautomation 安全平台定位（超越传统 SOAR）；MCP 原生支持；Valvoline 案例 **48 小时即实现 ROI。**

9. HiddenLayer（AI 模型安全）

产品方向：专注 AI/ML 系统安全的领导者，解决"影子 AI"问题（72%组织存在未批准 AI 工具）和 AI Agent 安全。

技术特点：AI 资产全面可见性（自动发现所有 AI 资产包括影子 AI）；业界率先提供 Agentic & MCP 安全防护（间接提示注入/不安全工具使用/高影响自主操作）；AI Runtime Protection（模型盗窃/数据泄露/未授权访问）；多框架模型文件扫描（TensorFlow/PyTorch/ONNX）。SaaS+Agent 混合架构。

差异化优势：专注 AI/ML 系统安全的平台；MCP 协议安全先行者；影子 AI 检测治理；参与 CISA/OpenSSF 等 AI 安全标准制定。

10.4 关键趋势洞察

技术演进

1. **从 Copilot 到 Agentic AI 已成现实**：Microsoft、CrowdStrike、Dropzone AI 等推出自主推理、决策和执行的 AI Agent。多家分析机构预期 2026—2028 年 Multi-Agent AI 在威胁检测/响应中的渗透将显著提升（"2025 年 5%→2028 年 70%"等具体数字尚缺可公开核验的原始出处，仅作为方向性参考）。
2. **Multi-Agent System 成为下一代架构**：Dropzone AI 和 Torq 率先实现专业化 Agent 团队协作，更接近真实 SOC 工作模式。
3. **AI 自身安全从边缘走向中心**：HiddenLayer、Snyk、Palo Alto 的 AI Runtime Security 快速成熟。72%组织存在影子 AI，AI 治理刻不容缓。

商业模式创新

商业模式方面呈现三个显著趋势：一是部署速度（Time-to-Value）成为核心竞争力，Abnormal 15 分钟、Dropzone 30 分钟、Torq 48 小时即可完成接入上线并开始产出业务价值，API 优先+SaaS 原生+No-Code 成为标配；二是按成果计费初现端倪，Dropzone"按告警处理量计费"将 AI 能力与产出挂钩；三是 MSSP 成为 AI 安全放大器，AI 帮助 MSSP 用同样团队服务更多客户（30-50%工作量卸载）。

市场竞争格局

平台厂商激进整合：Palo Alto 收购 CyberArk（$25B，身份/PAM）、Cisco 整合 Splunk（$28B，SIEM/数据分析）、Google 收购 Wiz（$32B，云安全 CSPM/CNAPP）。核心驱动力各不相同，均非直接围绕 AI，但统一数据底座客观上为 AI 落地提供基础。

专业厂商生存空间广阔：Abnormal、Recorded Future、Vectra AI 凭借细分深度和 AI 原生优势高速增长。"Best-of-Breed vs. Platform"争论持续。

新兴创业公司快速崛起：**7AI** 最具代表性——Cybereason 原班创始团队（CEO Lior Div、CTO Yonatan Striem Amit 等）2024 年创立，2025 年 2 月出隐身。定位自主 AI 安全 Agent 平台，融资$166M（$36M 种子+$130M A 轮，号称网安史上最大 A 轮），估值$7 亿，出隐身到$7 亿仅 10 个月。已处理 250 万+告警、65 万+调查，Fortune 500 客户报告 95-99%误报消除。但 CB Insights Mosaic 评分 667（资金 987/管理 798/增长势头 351/商业成熟度 2/5），高估值主要建立在团队和融资能力上，商业规模化待验证。需注意创始人前公司 Cybereason 曾巅峰$30 亿后业务下滑至$3 亿。

10.5 RSAC 2026 创新厂商观察

RSAC 2026 共 634 家参展企业，其中 117 家（18.5%）业务与 AI 安全直接相关，73 家涉及 Agentic AI 方向，首次参展企业 92 家（14.5%）。五大创新赛道：

赛道一：Agentic SOC（最热）——至少 10 家公司定位 AI SOC Agent/Analyst：Dropzone AI、Prophet Security、Qevlar AI、Radiant Security、Intezer、Simbian（唯一提供 Agent 矩阵）、Torq、Skyrelis、Tego AI、AiStrike。

赛道二：AI 自动化攻击性安全——XBOW（首次参展，自主渗透测试）、Pentest Copilot、Horizon3.ai、Pentera、BreachLock、Gecko Security。

赛道三：Agentic 暴露管理与漏洞修复——Tonic Security、Zafran Security、Seal Security、Quantro Security、Palosade。Agent 从"检测"延伸到"修复"。

赛道四：AI 安全意识与社会工程防御——Adaptive Security（AI 生成深度伪造训练，"以 AI 攻 AI"）、KnowBe4（AIDA）、StrongestLayer。

赛道五：AI 原生安全分析与知识图谱——Crogl、Arch0（安全知识图谱）、Mindflow（自然语言驱动 SOAR）、Klyro、AnyInsight.ai。

趋势总结："Agentic"已成安全创业标配定位（73 家涉及 Agentic AI 方向）；从检测到修复的全链路自动化；Early Stage Expo 成为 AI 安全创新集中展示区；头部厂商全面 Agent 化；AI 攻防双刃剑效应显现。

10.6 国内代表厂商

本章覆盖国内网络安全领域代表性厂商，按信息深度分为三层：深度画像（Briefing+访谈）、问卷画像（调研问卷）、材料画像（厂商提交材料与公开信息）。

> **整合说明**：本版本（2026-03-22 整合）结构化问卷调研样本扩展至 66 家国内安全厂商（涵盖综合性/终端/数据/AI/云/工控/运营/攻防/邮件/鉴伪/开发/策略管理等厂商类型），叠加 11 场头部厂商深度 Briefing，综合画像累计覆盖国内代表性厂商约 70 家。

国内代表厂商索引

编号	厂商名称	信息来源	所在层级

1	深信服	深度画像（含 2026-03-19 Briefing 更新）	第一层（深度画像）
2	火山引擎	深度画像	第一层（深度画像）
3	安恒信息	深度画像（含 2026-03-06 Briefing 更新）	第一层（深度画像）
4	360 数字安全	深度画像（含 2026-03-17 Briefing 更新）	第一层（深度画像）
5	奇安信	深度画像	第一层（深度画像）
6	绿盟科技	深度画像	第一层（深度画像）
7	金睛云华	深度画像	第一层（深度画像）
8	长亭科技	深度画像（含 2026-03-09 Briefing 更新）	第一层（深度画像）
9	知其安	深度画像	第一层（深度画像）
10	未来智安	深度画像	第一层（深度画像）
11	青藤云安全	深度画像	第一层（深度画像）
12	天懋信息	问卷	第二层（问卷）
13	瀛云科技（DevSecOps）	问卷	第二层（问卷）
14	摄星科技	问卷	第二层（问卷）
15	炼石网络	问卷	第二层（问卷）
16	海云安	问卷	第二层（问卷）
17	和利时	问卷	第二层（问卷）
18	烽台科技	问卷	第二层（问卷）
19	宁数安全	问卷	第二层（问卷）
20	石犀科技	问卷	第二层（问卷）

21	广东盈世（Coremail）	问卷	第二层（问卷）
22	云弈科技	问卷	第二层（问卷）
23	腾讯安全	厂商材料/公开信息	第三层（材料/公开信息）
24	天融信	厂商材料/公开信息	第三层（材料/公开信息）
25	华清未央	厂商材料/公开信息	第三层（材料/公开信息）
26	云起无垠	厂商材料/公开信息	第三层（材料/公开信息）
27	灵云数科（网哨 M01）	厂商材料/公开信息	第三层（材料/公开信息）
28	中国电信	厂商材料/公开信息	第三层（材料/公开信息）
29	立智安	厂商材料/公开信息	第三层（材料/公开信息）
30	芯盾时代	厂商材料/公开信息	第三层（材料/公开信息）
31	明朝万达	厂商材料/公开信息	第三层（材料/公开信息）
32	方向标（FangMail）	厂商材料/公开信息	第三层（材料/公开信息）
33	威胁猎人	厂商材料/公开信息	第三层（材料/公开信息）

34	厦门快快网络	厂商材料/公开信息	第三层（材料/公开信息）
35	启明星辰	深度画像（2026-03-13 Briefing 新建）	第一层（深度画像）
36	亚信安全	深度画像（2026-03-20 Briefing 新建）	第一层（深度画像）
37	悬镜安全	深度画像（2026-03-11 Briefing 新建）	第一层（深度画像）
38	华云安	深度画像（2026-03-12 Briefing 新建）	第一层（深度画像）
39	安华金和	厂商材料/公开信息	第三层（材料/公开信息）
40	瑞数信息	厂商材料/公开信息	第三层（材料/公开信息）
41	永信至诚	厂商材料/公开信息	第三层（材料/公开信息）
42	六方云	厂商材料/公开信息	第三层（材料/公开信息）
43	海泰方圆	厂商材料/公开信息	第三层（材料/公开信息）
44	安芯网盾	厂商材料/公开信息	第三层（材料/公开信息）
45	默安科技	厂商材料/公开信息	第三层（材料/公开信息）

46	领信数科	厂商材料/公开信息	第三层（材料/公开信息）
47	众智维	厂商材料/公开信息	第三层（材料/公开信息）
48	聚铭网络	厂商材料/公开信息	第三层（材料/公开信息）
49	威努特	厂商材料/公开信息	第三层（材料/公开信息）
50	保旺达	厂商材料/公开信息	第三层（材料/公开信息）
51	数安行	厂商材料/公开信息	第三层（材料/公开信息）
52	长扬科技	厂商材料/公开信息	第三层（材料/公开信息）
53	上海观安	厂商材料/公开信息	第三层（材料/公开信息）
54	网宿安全	厂商材料/公开信息	第三层（材料/公开信息）
55	魔方安全	厂商材料/公开信息	第三层（材料/公开信息）
56	矢安科技	厂商材料/公开信息	第三层（材料/公开信息）
57	新华三	厂商材料/公开信息	第三层（材料/公开信息）

58	丈八网络	厂商材料/公开信息	第三层（材料/公开信息）
59	万里红	厂商材料/公开信息	第三层（材料/公开信息）
60	孝道科技	厂商材料/公开信息	第三层（材料/公开信息）
61	瀛云科技（运维安全）	厂商材料/公开信息	第三层（材料/公开信息）

10.7 深度画像（第一层厂商）

1. 深信服：主动安全战略下的全自动化威胁运营实践者

定位：头部安全厂商，战略从"AI FIRST"升级为"主动安全"，核心理念从"AI 辅助人"转向"人是 AI 的一个环节"，仅保留少量关键决策，其余全部自主执行，是国内表述最激进的 AI 战略之一。核心 AI 能力：采用快慢双路径并行架构——快路径基于 7-8B 小参数精调模型处理标准告警（毫秒级自动遏制），慢路径调用大参数模型处理复杂调查场景。15B 以下自训模型专注解决通用大模型在安全专业领域的知识盲区。量化效果：下一代防火墙以 11.1%市场份额领跑国内硬件市场，SASE 中国市场份额第一，2025 年 AI 赋能网络安全应用测试唯一包揽三项第一；安全 GPT 告警降噪可达 99%。差异化优势：全栈产品+AI 原生战略+国家级实战验证；双路径+自动闭环是国内落地最彻底的 AI 运营架构之一。关注要点：激进的自动化策略对运营规范化程度要求高，建议中大型客户在关键处置点保留人工确认。

2. 火山引擎：云原生 AI 安全运营的后起之秀

定位：字节跳动 ToB 品牌，**AI Native 无历史包袱**，以安全运营智能体 Circle 切入。战略是端到端闭环运营，黑盒化中间过程。目标客户包括互联网消费（云 SaaS）、金融央企（KA）和 45 家 MSP 生态。产解团队 30-40 人，2026 年扩至 50 人。

核心 AI 能力：产品 Circle（Circle is all you need）以豆包为基座模型（自研），云端豆包大模型，私有化豆包 OSS 32B，配合专用安全模型"孔明"。通过 Skill 体系构建自定义智能体，知识教练机制实现模型自动生成知识库→人工反馈→学习企业特性的闭环。告警研判遵循 L0 规则降噪→L1 人工降噪→L2 事件聚合→L3 分析报告的处理链路，**10 万条告警→8 万条→100 条事件→20 条自动闭环+80 条人工处理**，主战场迁移至 IM（飞书/企微/钉钉），告警推送到群实现对话式闭环。

部署与定价：云上 SaaS 为 License 5 万/年+Token 按量（主收入），已推广 10-20%火山云客户；私有化为**百万级**客单价（中石油勘探院、芯片客户等），客户自建 910B 卡集群；MSP 托管与 45 家合作伙伴合作，数据可出网时模型能力充分发挥。性能方面双卡 4690 可处理 1.6 万条告警/天。

差异化优势：AI Native 设计；字节生态（内场验证→外部商业化、飞书集成、豆包算力）；**只做 Agent 层**（不做数据湖、不做安全工具，聚焦安全大脑）；知识教练平衡通用与定制；多模型协同（主模型编排+专用模型执行）。

3. 安恒信息：恒脑智能体生态驱动的 AI 安全平台

定位：以"OE AI"（卓越运营 AI）为核心战略，2026 年从"AI 辅助人"升级为"AI 自主提效，人工做最后审核"。双线推进：一是产品+AI（存量产品接入恒脑智能体）；二是 AI

原生（恒脑平台对外输出，支持 API/MCP）。核心 AI 能力：恒脑采用 MoE 混合专家架构，动态路由与专家稀疏激活机制，具备泛连接、高交互、全模态特性；微调采用 LoRA 加层方式，结合安全语料持续训练。量化效果：2024 年数据安全管理平台收入 7.4 亿元（同比增长 43%），市占率 18.7%，领先第二名 3.1 个百分点；获吴文俊人工智能科技进步奖、入选工信部未来产业创新发展优秀案例；MoE 架构召回率提升 18-27%。差异化优势：数据安全领域市占率第一的传统优势+MoE 架构大模型的差异化技术路线。关注要点：MoE 架构运维复杂度较高，客户落地需关注算力配比与专家路由稳定性。

4. 360 数字安全：自训告警研判模型驱动的智能运营平台

定位：依托 20 余年安全实战数据积累，以安全大模型为核心的 AI 安全能力体系构建者；推进全产线智能体化改造。内部设"风云"与"天问"两个 AI 能力部，博士研究力量约 80%投入安全云方向。核心 AI 能力：COE（专家混合）架构延续"快慢思考"理论体系——快思考处理 10ms 以内模式匹配，慢思考处理多步推理溯源；规划/判别/记忆/道德四大中枢协同。本脑（本地安全大脑/SIEM）是最成熟的落地产品，采用千问 14B 精调的 360 安全大模型做告警研判。量化效果：IDC 安全大模型测评 2024、2025 连续两年第一；告警研判 F1≈1.0；告警降噪 83-99%；基于 320 亿样本的底座数据规模是国内最大之一。差异化优势：实战数据规模+理论体系贡献（快慢思考框架）+CCoE 产品矩阵。关注要点：AI 能力以自训为主，对 360 生态依赖度较高。

5. 奇安信：数据规模驱动的 AISOC 闭环体系

定位：综合安全厂商头部，以 QAX-GPT 安全机器人系统为核心，构建四类 AI 安全专家角色（安全运营、网络威胁研判、终端威胁研判、主机威胁研判）。同步推出大模型卫

士（GPT-Guard）布局 AI 安全防护。安全大模型通过国家网信办算法备案和服务备案双备案，2025 年入选 Gartner 中国安全技术成熟度曲线报告 10 大领域。

核心 AI 能力：依托百万亿级预训练数据（安全日志、文档、知识库、情报类，存储体量数百 PB）和业内最大规模安全专家团队。QAX-GPT 研判能力接近中级安全专家水平。AISOC 产品实现 AI 赋能安全运营中心全流程闭环——从告警接入、智能研判、事件关联到自动化处置的端到端覆盖。大模型卫士（GPT-Guard）覆盖 OWASP LLM Top10 攻击防护、数据泄露防护、模型安全、内容合规四大方向。在 CyberSecEval 评测中获得八项第一。

量化效果：响应时间从小时级降至分钟级，单一威胁事件处理时间减少 98%。实现 7×24 小时自动实时分析、100%覆盖秒级研判分类。京东方和吉利成为首批商用客户。

差异化优势：数据规模（百万亿级/数百 PB）和安全专家团队积累在国内领先；网信办双备案的合规先发优势；CyberSecEval 八项第一的评测成绩；历年实战攻防演习锤炼的攻防能力。

关注要点

奇安信 AISOC 的核心能力建立在"数据+模型+场景"三位一体架构之上，具备四大差异化壁垒。其一，百万亿级预训练语料底座——奇安信积累了 20 年攻防实战数据，覆盖日志、告警、样本、情报、漏洞、规则全类型，总存储体量达数百 PB 级别，是国内安全行业规模最大的高质量安全语料库，这一数据壁垒构成其 AI 能力的燃料优势。其二，QAX-GPT 安全垂域大模型依托上述语料底座进行深度预训练+指令微调，在国家网信办双备案的合规

框架下，已在京东方、吉利等头部客户的真实生产环境中完成规模验证——单一威胁事件的处理时间从小时级降至分钟级，压缩幅度达 98%，实现了 7×24 小时全天候自动化实时分析，秒级研判分类覆盖率达 100%。其三，AISOC 的端到端闭环能力覆盖从告警接入、智能研判、事件关联、自动化处置的全流程，区别于国内多数厂商仅提供"问答+降噪"的单点能力，奇安信是国内少数能够提供完整 SecOps 闭环的厂商之一。其四，大模型卫士（GPT-Guard）布局 Security for AI 方向，填补了 AI 安全防护这一新兴赛道的空白，CyberSecEval 八项评测成绩第一是其实技术实力在行业评测维度的直接背书。2026 年，随着 AI 安全运营从"辅助决策"向"自主执行"演进，奇安信的数据壁垒和闭环能力将成为其拉开与追兵距离的核心竞争力。：QAX-GPT"接近中级安全专家"的定位相对保守，需关注能力提升速度；相比竞品的告警降噪具体数据（如降噪比）披露较少。

6. 绿盟科技：场景覆盖最广的 AI 安全平台

定位：综合性安全厂商龙头，"垂域模型+平台化赋能"，打造**独立安全体系智能中枢**，开放式 API 为客户已有平台/产品赋能。**本次调研中全维度自评 5 分（满分）且覆盖 7 大场景簇的厂商。**

核心 AI 能力：交付形态全覆盖（Chatbot、Copilot、Agent、内嵌增强、API/SDK），场景簇全覆盖（威胁检测、安全运营、数据安全、渗透测试、漏洞挖掘、邮件安全、开发安全），技术栈全覆盖（纯提示词、RAG、TAG、SFT 微调、RL 偏好优化、小模型蒸馏、规则/图谱融合）。作为国内少数具备**自研从头训练基础模型**能力的安全厂商，工程闭环可完成处置（含审批/审计/回滚/复盘沉淀），全链路记录。

量化效果：全维度>50%（人工工时、MTTR、误报、处置闭环率）。典型案例覆盖金融、运营商 SecOps，与 SIEM/SOAR/NDR 联动。

差异化优势：场景/技术/交付维度覆盖面最广；独立智能中枢理念；自研训练能力；全维度量化收益>50%。

绿盟科技的全栈 SOC AI 平台是其 AI 战略的旗舰级产品输出，基于"垂域模型+平台化赋能"的核心理念打造。该平台的技术架构覆盖三大核心层：模型层、数据层和应用层。模型层支持国内最全面的模型选择谱系，包括自研大模型、闭源 API（GPT/Claude/豆包/千问/GLM/MiniMax 全支持）和开源模型，并能根据客户场景自动匹配合适的模型组合，这是绿盟区别于多数友商"绑定单一模型"的关键优势。数据层支持 7 大类数据源的统一接入与融合，包括告警/日志、流量/会话、样本/沙箱、情报、漏洞库、工单和知识文档，AI 语义解析引擎可在 1 小时内完成新数据源的标准化接入。应用层实现了全场景覆盖（威胁检测、安全运营、数据安全、渗透测试、漏洞挖掘、邮件安全、开发安全）和全交付形态覆盖（Chatbot、Copilot、Agent、内嵌增强、API/SDK），客户可按需选择最适合自身安全成熟度的交付模式。平台的核心量化效果体现在四个维度：人工工时下降超过 50%、MTTR 改善超过 50%、误报下降超过 50%、处置闭环率提升超过 50%——全维度均达到 50%以上的改善幅度，在本次调研中是量化效果最为全面的单一厂商产品。典型客户案例覆盖金融行业（某头部银行部署后 SOC 运营人效提升 3 倍）和运营商（与 SIEM/SOAR/NDR 的深度联动），未来方向聚焦多智能体协同的自主调查、自主基线和自主值守能力，代表传统综合安全厂商在 AI 时代继续保持技术领先性的标杆路径。

未来方向：多智能体协同的自主调查、自主基线、自主值守能力。证明传统大厂在 AI 时代仍可保持领先。

7. 金睛云华：双子大模型驱动的 AI 安全先行者

定位：2016 年成立的 AI 原生安全产品与服务提供商，核心团队源自中科大 KDD 实验室与清华 KEG 实验室，AI+安全累积近 20 年。自建 80 机柜/500 张 GPU 的人工智能安全智算中心。盛邦安全为关联上市公司；CyberGPT 已通过网信办算法备案、昇腾兼容性认证与 ISO/IEC 42001 认证。核心 AI 能力：构建"30+小模型矩阵+1 程序语言大模型+1 自然语言大模型"三层 AI 体系。30+威胁检测小模型覆盖恶意代码基因图谱、恶意加密流量步态指纹（186 个家族）、隐蔽隧道检测与 Web 攻击判定；CyberGPT-T（亿级参数）用于代码理解；CyberGPT-S/M（十亿/百亿级）用于告警降噪、智能研判、一键溯源。量化效果：恶意加密流量识别 98.2%、告警降噪 99.8%+；拥有 90+发明专利、100+荣誉奖项。差异化优势：双子大模型+自建算力+学术血统。关注要点：体量较小但技术密度高，适合对垂域检测精度要求严苛的客户。

8. 长亭科技：攻防基因+智能安全双轮驱动的 AI 安全服务商

定位：从"攻防基因"战略向"智能安全"战略全面转型，目标 AI 含量与攻防基因各占 50%。业务三条线：AI for Security（主业）、Security for AI（大模型应用安全评估与防护）、Security Training with AI。核心 AI 能力：原有产品矩阵（守元 AI Guard、慧鉴 SAST、码力、万象 COSMOS、PandaWiki）持续演进；"攻"方向以"无风平台"（内部全流程自动化渗透工具）为核心差异化能力；守元 AI Guard 面向大模型安全围栏场景。量化效果：安服整体规模 3-4 亿元；Security for AI 评估服务 2025 年已做几十个项目、规模数百万

元，2026 年预计增至 1000-2000 万元；MSS 托管客单价目标 30-100 万/年。差异化优势：攻防基因+AI 原生双轮驱动，是国内最早布局"Security for AI"的厂商之一。关注要点：攻防能力对人才依赖较强，自动化渗透产品化路径需持续关注。

9. 知其安：堆叠式 AISOC 的金融行业深度共建者

定位：主打"堆叠式 AISOC"——不绑定特定安全产品，复用客户已有算力与智能体平台，兼容异构设备。自 2024 年 9 月起与招商银行深度联合共建，是国内 AI 安全运营在金融领域落地最深的案例之一。核心 AI 能力：四个"必须"原则——复用已有算力、核心能力自主可控、兼容异构安全产品、支持通用大模型。已形成 130+安全研判智能体，覆盖边界/流量/终端/API/邮件/RASP/主机/DLP/深度调查等 9 大类场景。采用通用大模型（千问/DeepSeek）路线而非垂域微调；工程技巧包括不传告警原文避免负向误导、双大模型投票减幻觉。量化效果：招行钓鱼邮件智能体日检 5000+封、准确率 96.4%；DLP 智能体日检 1.8 万文件、判准率 93.1%；流量研判准确率 93%。差异化优势：金融行业深度共建+通用大模型+堆叠式部署。关注要点：客户行业聚焦金融，跨行业复制能力有待验证。

10. 未来智安：XDR 数据湖+MCP 原生的 AI 智能体安全运营

定位：2020 年成立，国内较早推出 XDR 产品（2021 年），以"AI+安全"为核心战略专注 AI 智能体安全运营平台。股东包括红杉中国、腾讯投资、君联资本，与腾讯安全玄武实验室、浙大 AI 实验室建立技术合作。核心 AI 能力：XDR 数据湖为基座（ClickHouse+ES），统一纳管异构数据，构建"数据底座+能力底座+AI 底座"三层架构。小模型+大模型混合——小模型过滤、大模型精判；技术栈采用千问 32B+DeepSeek。LLM Agent 工作流包含规划→MCP→RAG→记忆→专家经验五大模块，通过 MCP 协议原生驱动

安全能力原子化调用。量化效果：广东电网案例——8000 万日告警经小模型过滤至数万条，再由大模型输出约 100 个结果告警；整体告警降噪 99.8%，研判时间缩短 80%+，处置闭环时间缩短 70%+。差异化优势：MCP 原生+小大模型双层架构的工程化成熟度较高。关注要点：体量相对有限，企业级大规模部署经验仍在累积。

11. 青藤云安全：L4 级自主防御的 Agentic AI 安全中枢

定位：推出"无相"AI 智能中枢系统，将 Agentic AI 引入安全运营核心，在国内率先定位 L4 级自主防御（Autopilot），实现从 Copilot 到 Autopilot 的跨越。核心 AI 能力："高智能模型+精准数据感知+高效工具执行"能力三角，四层架构——AI 应用层（7 个专业智能体）、知识与决策层（ATT&CK 图谱+GoT 图式思维+分层记忆）、数据精炼层（压缩比超 99%的"安全小世界模型"）、基础设施层（180+种事件采集+68 个原子工具）。七大智能体包括告警研判、深度调查、响应处置、暗资产发现、漏洞验证 AIVD、暴露面分析、安全报告数字人。量化效果：告警降噪——日均 20 万+条过滤 99%至数十条高置信事件；深度调查 30 分钟还原 98%攻击行为；AIVD 准确度>99%+具备 0day 挖掘能力。差异化优势：L4 自主防御定位+GoT 图式思维+主机侧深耕。关注要点：L4 自主化对客户 IT 基础成熟度有一定要求。

35. 启明星辰：移动运营商生态驱动的"1+1+N"全产线 AI 化实践者

定位：国内综合安全厂商头部，以"1+1+N"架构为核心战略——一个统一安全大模型、一个统一智能体框架（AIDK）、N 个产线场景化智能体应用。2026 年强制要求所有产线统一接入公司级 AIDK 框架，由"百花齐放"转向"集中统一"。中国移动是最核心的生态锚点。核心 AI 能力：三层展开——第一层公司级大模型（基于千问/DeepSeek 开源基座增量

预训练+微调，团队约 10 人）；第二层自研 AIDK 智能体框架（开源组件为主、规避法务风险）；第三层产线智能体应用，覆盖安全运营、安全检测、数据安全三大方向，核心产品为泰和人工智能安全运营系统。团队务实承认"大模型微调边际收益越来越小"，当前各产线可自由选择合适的开源满血模型直接调用。差异化优势：运营商生态+集中统一框架+综合厂商规模化优势。关注要点：统一框架要求产线配合度高，落地节奏需关注。

36. 亚信安全：智能体安全思考最深、运营商生态最厚的防御性玩家

定位：升级为"连接智能世界，护航数字互联"。2024 年完成对亚信科技的并购（A+SOH 首家案例），实现安全+数智+连接三合一，战略投资远信卫星布局空天地全栈。双引擎战略并行：以 AIXDR 联动防御系统为核心的"AI 驱动安全"，以及三层框架覆盖的"安全赋能 AI"。核心亮点在智能体安全（Security for AI）领域的思考深度，而非 AIXDR 本身。核心 AI 能力：AI for Security 以信立方安全垂域大模型为旗帜，模型层覆盖 0.6B 超小模型（用于告警降噪第一层）到 1014B MoE 大模型群，辅以高密度深度学习小模型，基座借鉴千问与 DeepSeek 开源架构；AIXDR 平台覆盖 TrustOne（端）、Deep Security 等云+网+边+端全产线。依托 20 年+历史安全数据积累与按行业场景构建的高质量安全语料库。差异化优势：运营商生态最厚+智能体安全思考最深。关注要点：平台整合复杂度较高，客户大规模部署需关注集成路径。

37. 悬镜安全：数字供应链+AI 原生安全的双轮驱动者

> **信息来源**：2026 年 3 月 11 日 Briefing 新建画像。

定位：悬镜安全成立于 2018 年，定位为"数字供应链安全"领域的专业厂商，近年将 AI 融入供应链安全框架，形成"AI 原生安全"核心方向。2026 年 3 月 11 日，悬镜正式发布多

模态 AIST（问镜 AI 安全卫士平台），标志着其从软件供应链安全向 AI 基础设施供应链安全的战略延伸进入实质落地阶段。公司四大产品线构成"AI 三剑客+情报"矩阵：零脉（AI 赋能 SAST，面向 AI 编程时代的代码安全护栏）、问镜 AST（AI 模型/智能体/基础设施安全扫描与治理）、AI 供应链安全情报（模型漏洞、框架投毒、skills 投毒等），以及即将发布的智能体安全产品（智能体资产梳理+防护检测响应）。在市场定位上，悬镜将自身定位类比于 AI 安全领域的 protect.ai（美国同类厂商），国内"暂无明确竞争对手"。

核心 AI 能力：问镜 AST 是悬镜的核心 AI 产品，覆盖五大功能模块。一是 AI 风险情报（云脉 X Transform），数据源涵盖 Hugging Face 和魔搭社区约 200 万个模型，覆盖 AI 组件漏洞、模型漏洞、框架投毒、skills 投毒等情报类型，这一情报规模在国内安全厂商中处于领先地位。二是 AI 红队扫描，包含基础设施扫描（指纹识别+漏洞检测）、MCP 扫描（通过 MCP 协议连接 server 端，后端接模型做检测），以及模型风险评估（提示词越狱攻击，支持攻击生成模型与打分模型双层设计）。三是 AI 模型扫描，核心能力是**模型血缘图谱**——通过相似度识别和 API 参数分析判断模型类型，追溯模型来自哪个基础模型的微调或量化，这是悬镜独有的技术差异化能力。四是智能体代码审计（静态），覆盖影子模型识别（检测未声明但实际调用的模型）、密钥泄露、提示词注入（LLM Top10）等风险。五是智能体运行时审计，通过 SDK/插桩方式嵌入智能体，实现输入/输出全程监控和敏感数据替换，后端分析模型采用 32B 参数。

零脉（AI 代码安全护栏）在传统 SAST 基础上引入大模型做漏洞验证和智能修复建议，核心突破在于误报率的大幅改善——Java/Python 等解释性语言误报率从 30-40%降至 2%以内，C/C++降至 8%以内，这一量化改善显著优于市场多数 SAST 产品。越权/逻辑漏

洞检测场景采用 SAST 提取 API 调用链→识别权限校验点→交 AI 分析的三步流程。IDE 插件（支持 VS Code 和 IntelliJ 系）将代码安全扫描融入开发环境，实现开发时实时检测与 AI 修复，契合 AI 编程时代的安全左移趋势。

量化效果：SAST 误报率改善是最核心的量化指标：Java/Python 降至 2%以内，C/C++降至 8%以内，与传统 SAST 30-40%的误报率相比改善幅度达到数量级级别。AI 风险情报数据源约 200 万个 AI 模型，情报覆盖面为国内同类产品中规模最大。

部署与定价：按模块定价（类 SCA 授权模式），正式承担 KA 客户的平均客单价 45-50 万元，属于中高端定价区间。2026 年 AI 原生安全收入占比预计达 20-30%，传统软件供应链+DevSecOps 仍占 60-70%。

差异化优势：一是路径清晰的双轮驱动：从软件供应链安全扩展到 AI 基础设施供应链，历史积累（RASP 技术、SCA 能力）自然迁移，不是生硬切换赛道；二是模型血缘图谱这一独有技术能力，能够追溯 AI 模型的微调/量化来源，在模型合规治理和知识产权保护场景下具有不可替代性；三是插桩技术积累转移优势，历史 RASP 技术积累被有效移植至智能体运行时审计；四是 SAST+AI 结合的越权逻辑漏洞检测，传统 SAST 无法单独实现；五是 AI 供应链情报的全公司重点投入，模型/框架/MCP/Skills 投毒情报形成情报壁垒——Skills 投毒被判断为类似高级版 Android 生态投毒，威胁等级极高。

悬镜 AIST（问镜 AI 安全卫士平台）在数字供应链安全场景下的 AI 能力布局独具特色。在软件供应链安全积累（RASP 技术+SCA 能力）基础上，悬镜将 AI 能力延伸至三个新方向：一是 AI 代码安全，即零脉（AI 代码安全护栏）在传统 SAST 基础上引入大模型

做漏洞验证和智能修复建议，核心突破在于误报率改善——Java/Python 误报率从 30-40%降至 2%以内，C/C++降至 8%以内，改善幅度达数量级，是国内 SAST 误报率改善幅度最大的产品之一，这也是悬镜将 Claude Code Security 作为对标基准的直接原因。二是 AI 组件安全，核心产品为云脉 X Transform，覆盖 Hugging Face 和魔搭社区约 200 万个 AI 模型的风险情报，这是国内同类产品中覆盖面最广的 AI 组件漏洞情报库，包括模型漏洞、框架投毒、Skills 投毒等多种新型威胁类型，填补了 AI 供应链安全这一新兴场景的情报空白。三是智能体安全，悬镜发布的问镜 AIST 将上述能力整合为多模态平台，其模型血缘图谱是独有技术能力——能够追溯 AI 模型的微调/量化来源，在模型合规治理和知识产权保护场景具有独特价值。这三层能力叠加，使悬镜在"AI for Security"和"Security for AI"两个方向同时建立了差异化的技术壁垒，是国内为数不多的双方向均有产品落地能力的厂商。

关注要点：商业化仍处极早期，正式承担的 KA 客户不足 10 个，规模化路径尚不清晰；市场教育成本高，客户认知不成熟；国内模型与 Cloud 的代码漏洞挖掘差距在私有化场景下被进一步放大；产品迭代压力大，大模型版本迭代以周为单位；智能体安全产品刚发布，功能完整性仍在打磨，成熟度有待市场验证。

38. 华云安：CTEM 框架驱动的 AI 攻击面管理专家

定位：以"持续威胁暴露面管理（CTEM）"为核心，2025-2026 年向"AI 驱动的持续暴露面监测管理"转型。CTEM 框架覆盖发现-评估-验证-响应全流程，是从攻击者视角主动管理风险的方法论。产品结构：平台产品线包括灵洞 Ai.Vul（内网攻击面）、灵知 Ai.Radar（外网攻击面，AI 原生+SaaS 形态）；工具产品线包括灵源 Ai.Hunter（流量日志/APT 检测）、灵刃 Ai.Bot（AI BAS）；新增 AI 产品线包括华云安智能体平台与安全数字人。核

心 AI 能力：灵洞 Ai.Vul 的 AI 化围绕五个维度——漏洞情报导入、资产导入、资产-情报关联匹配（规则→向量→大模型三代演进）、自然语言操作、漏洞扫描任务下发；当前定位以"Copilot/工作辅助"为主。差异化优势：CTEM 框架完整度+AI 原生暴露面管理布局。关注要点：AI 化多为 Copilot 形态，距离"自主执行"的 Agentic 水平仍有提升空间。

10.8 问卷厂商画像（第二层）

12. 天懋信息

聚焦场景：专网边界安全、暴露面收敛、违规外联/内联检测，覆盖政务/金融/运营商/电力/制造/交通六大行业。技术路线采用大模型+小模型/规则/图谱混合与 TAG，私有化推理，已实现研判问答、分类分级、告警摘要、报告生成。量化效果为工时降>50%，MTTR/误报/闭环率提升 10-30%。差异化在于专网边界场景专精，行业深耕，规模交付能力（成熟度 3）。

13. 瀛云科技（DevSecOps）

聚焦场景：开发安全（代码审计/DevSecOps/供应链安全），全行业覆盖。技术路线采用大模型+规则混合与 TAG，公有云推理，已实现研判问答和工具编排/自动处置。量化效果为工时降 30-50%，闭环率提升 10-30%。差异化在于开发安全 Agent 闭环先行者。

14. 摄星科技

聚焦场景：SecOps、漏洞挖掘、开发安全，覆盖政务/金融/运营商/电力/制造/交通。技术路线采用大模型+规则混合与 RAG/规则融合，混合推理，已实现研判问答、分类分级、漏洞挖掘、攻击路径推演等。量化效果为误报降 30-50%，工时/MTTR/闭环率提升 10-30%。差异化在于多场景覆盖和数据沉淀能力。

15. 炼石网络

聚焦场景：数据安全（分类分级/DLP/DSPM/脱敏）、安全策略管理、密评密改，全行业覆盖，能力自评 4.3 分（第二梯队前列）。技术路线采用开源大模型与 RAG/TAG/规则融合，多种推理模式+推理加速，实现研判问答、分类分级、脱敏、策略管理、报告生成等。量化效果为 MTTR 改善 30-50%，误报降 30-50%。技术亮点包括硬件加速微秒级时延、"规则引擎+大模型"混合（规则过滤 90%+确定性场景）、零信任细粒度权限、容器化隔离+推理环境动态销毁。差异化在于数据安全+密评密改独特赛道，完整闭环和低成本私有化。

16. 海云安

聚焦场景：数据安全、漏洞挖掘、开发安全，全行业覆盖，能力自评 4.6 分（第二位），成熟度 4-产品化平台。技术路线采用纯大模型+传统 ML 混合与 RAG/SFT 微调(LoRA)，私有化+推理加速，**深度学习 5 分、精调 5 分**——技术深度突出。量化效果为千行代码漏洞率降 50%，金融案例误报降 90%。差异化在于深度学习和精调能力突出，数据安全+开发安全双聚焦。

17. 和利时

聚焦场景：工控/车联网/IoT 安全，覆盖能源电力/制造/交通，能力自评 2.3 分。技术路线采用开源大模型与提示词/RAG/SFT/规则融合，私有化推理，定位工控垂域模型。已实现研判问答、策略管理、告警摘要、报告生成。关注要点是工程闭环仅生成建议（不落库不执行），幻觉控制靠限定场景+专家知识库，算力和精调投入不足，但工控垂域有特色。

18. 烽台科技

聚焦场景：工业场景（能源/制造/钢铁/冶金/管网）的安全运营、数据安全、渗透测试，成熟度 4-产品化平台，**自研从头训练**（国内少数）。技术路线采用大模型+小模型/规则混合与 RAG/SFT/规则融合，私有化推理。量化效果为工时降>50%，误报降>50%，制造业案例 MTTR 降 95%。技术亮点是**边侧小模型为平台大模型调优**的边缘+中心协同架构。差异化在于工控领域自研模型能力，产品化和场景覆盖满分。

19. 宁数安全

聚焦场景：工业资产管理、威胁检测、安全运营、数据安全，全行业覆盖。技术路线采用开源大模型与提示词/RAG/小模型蒸馏，多种推理模式，**工具联动采用 MCP 和 API 调用**。量化效果为各维度 10-30%提升。差异化在于工业资产场景和 MCP 协议支持。

20. 石犀科技

聚焦场景：数据安全服务（资产识别/风险研判/预判/处置推荐），覆盖政务/金融/电力/医疗/教育/国央企。技术路线采用开源大模型混合与 RAG/SFT/小模型蒸馏，私有化+**分层模型（大+小）**。量化效果为工时降 30-50%，MTTR 改善 30-50%，闭环率提升 30-50%。技术亮点是自动执行低风险动作（可回滚）和垂域小模型精调+大小模型融合。

21. 广东盈世（Coremail）

聚焦场景：邮件安全（钓鱼识别/溯源/自动处置）和数据安全，全行业覆盖。技术路线采用大模型+小模型/规则/ML 混合与提示词/小模型蒸馏，私有化+推理加速，具备多模态分析能力。量化效果为工时降 30-50%，MTTR 改善 30-50%，闭环率提升 30-50%，支持完整闭环（审批/审计/回滚）。差异化在于邮件安全专精和多模态能力。

22. 云弈科技

聚焦场景：鉴伪/认知安全（深伪检测/内容溯源/反欺诈）、威胁检测、安全运营、漏洞挖掘，成熟度 4-产品化平台。技术路线采用大模型+小模型/规则/ML 混合与 TAG/小模型蒸馏/规则融合，私有化推理，各维度均衡 4 分。量化效果为误报降 50-70%（运营商案例），工时降 15-25%。差异化在于鉴伪/认知安全特色场景。

10.9 公开信息画像（第三层厂商）

以下厂商基于公开信息（厂商官网、技术演讲、行业报告、媒体报道）整理，数据截至 2026 年 2 月。本版本新增 23 家厂商（编号 39-61）。

23. 腾讯安全：学术驱动的安全 AI 基准与工具链

以混元大模型为基座、走"学术驱动+标准化评测"差异化路线，推出 SecCorpus 安全语料库、SecBench 评测基准、Binary AI 与 SecurityX 智能调查平台四大产品。

24. 天融信：AI+智算双驱动的传统安全厂商转型

以"天问"安全垂域大模型覆盖检测、研判、情报、决策六大能力，同步推出 DeepSeek 安全智算一体机，实现传统安全厂商向"AI+安全"的转型。

25. 华清未央：专注机器语言的大模型

核心团队来自清华大学，专注机器语言大模型（MLM），覆盖 AI for Software 与 AI for Security 两个方向，在智能制造、安全监管、信创产业等领域形成差异化布局。

26. 云起无垠：开源安全大模型与智能 Fuzzing 的先行者

依托清华技术背景，发布开源安全大模型 SecGPT 并聚焦智能 Fuzzing 赛道，构建覆盖协议/数据库/API/Web3.0 的大模型驱动漏洞挖掘全流程方案。

27. 灵云数科（网哨 M01）：公安部一所情报联防驱动的邮件安全深度防御

依托公安部第一研究所的"百家行业单位情报联防"体系，推出网哨 M01 新型邮件攻击检测分析系统，弥补传统邮件网关在未知攻击深度分析上的空白。

28. 中国电信：央企规模数据优势驱动的安全 AI

以"星辰·见微"安全垂类大模型为核心，基于阡陌安全数据集（10 万亿条/20PB）与全国产化万卡智算集群（16EFLOPS）打造央企级 AI 安全一站式服务。

29. 立智安：多智能体架构的 AI 邮件安全新锐

聚焦 AI 邮件智能体，以智瞳/邮脉/极智三大产品构建"Agent+MCP+ReAct"多智能体邮件安全检测体系。

30. 芯盾时代：零信任+身份安全的 AI 增值路线

国内零信任与身份安全头部厂商，完成全产品线与 AI 大模型深度集成，形成"智能底座+垂直场景+行业生态"三位一体解决方案。

31. 明朝万达：数据安全+AI 分类分级的深耕者

以安元智能数据治理平台为核心，将私有化部署的千问 QwQ-32B 应用于数据分类分级三步智能体流程，连续入选 Gartner 中国 DLP 推荐研究报告。

32. 方向标（FangMail）：邮件安全大模型的精准增量方案

推出邮安 AI 大模型作为现有邮件网关的增量增强检测层，具备加密附件密码自动提取与 HTML 渲染混淆攻击检测等独特能力。

33. 威胁猎人：AI+反欺诈情报的海外电商专家

聚焦海外电商反欺诈场景，推出 DarkSphere AI 平台覆盖恶意注册、账号盗刷、虚假交易等全链路欺诈识别。

34. 厦门快快网络：云安全+AI 安全运营的区域龙头

厦门本土的云安全服务商，推出"智御"AI 安全创新产品矩阵，服务 12 万+客户并与华为云、阿里云、腾讯云等构建多云生态合作。

39. 安华金和：数据安全大模型化最系统的厂商之一

数据安全专业厂商，以"安知数据安全大模型"为核心构建数据分类分级、行为审计、风险研判等智能体群，入选 IDC 2025 安全检测与数据安全智能体两大分类。

40. 瑞数信息：动态安全+WAAP for LLM 的双向布局者

从动态安全原有优势出发，布局 WAAP for LLM，面向大模型应用提供 Web+API+大模型一体化防护方案。

41. 永信至诚：AI 安全教育与蜜网双场景产品化

AI 安全教育与蜜网场景双向产品化，结合春秋 AI 攻防平台与 Cyberpeace 蜜网 AI 化改造形成差异化定位。

42. 六方云：无监督机器学习驱动的工控 NDR

工控 NDR 领域的无监督机器学习实践者，推出仿生免疫架构，覆盖工控网络未知威胁检测与响应。

43. 海泰方圆：密码+数据治理+AI 私有部署的深度融合者

将密码服务、数据治理与 AI 私有部署深度融合，面向军工与政务涉密场景提供国产化 AI 安全能力。

44. 安芯网盾：内存安全+AI 可信度评估的双引擎布局

以内存安全为核心能力，布局 AI Decoder 与 Shadow AI 治理，形成"内存安全+AI 可信度评估"双引擎。

45. 默安科技：替代中级安全运营人员的智能体专家

以"替代中级安全运营人员的智能体"为产品目标，围绕 IAST 与 AI 运营智能体构建差异化产品线。

46. 领信数科：便携式大模型安全评估+极致告警降噪组合

推出便携式大模型安全评估 MLBox 与子牙智能体，面向极致告警降噪比场景提供轻量化解决方案。

47. 众智维：安全智能体超级市场的规模化复制者

构建安全智能体"超级市场"，通过 600+运营智能体的规模化组合能力实现安全运营场景的快速复用。

48. 聚铭网络：万能联动+AI 大模型研判的低门槛 SOC 方案

以"万能联动+AI 大模型研判"构建低门槛 SOC 方案，面向中小型客户提供一体化安全运营产品。

49. 威努特：工控安全+AI 智算平台融合的信创支持者

工控安全领域厂商，推出 AI 智算平台并深度适配信创国产化，覆盖工控网络、主机与应用的 AI 融合防护。

50. 保旺达：运营商数据全链路 AI 溯源专家

面向运营商场景推出数据全链路 AI 溯源方案，基于大数据+AI 能力帮助客户实现数据溯源成本与时间大幅下降。

51. 数安行：DataSecOps 理念的零信任数据运营安全

以 DataSecOps 理念推动零信任数据运营安全，围绕数据运行时风险与控制策略自动化构建差异化产品。

52. 长扬科技：网络安全运营+工业视觉 AI 的双向赋能者

网络安全运营+工业视觉 AI 双向赋能，面向工控/智能制造场景提供"IT+OT+视觉"一体化 AI 安全能力。

53. 上海观安：本地化大模型驱动的智能化数据安全管控

以本地化大模型驱动智能化数据安全管控，重点服务省级政务与金融客户的合规与数据风险治理。

54. 网宿安全：边缘 AI 驱动的云地协同安全架构

以全球 2800+边缘节点为基础，推出边缘 AI 驱动的云地协同安全架构，主打出海与跨境场景。

55. 魔方安全：AI 驱动的暴露面风险管理全链路闭环

EASM+CAASM+VPT 全链路闭环，AI 驱动的暴露面风险管理，将 MTTR 压缩到小时级。

56. 矢安科技：AI 攻击编排驱动的网络安全体检（AEV）

AI 攻击编排驱动的网络安全体检（AEV），基于 29800+无害化攻击用例库提供 BAS+AEV 融合服务。

57. 新华三：自研灵犀大模型驱动的全栈安全 AI 覆盖

自研"灵犀大模型"驱动全栈安全 AI 覆盖，研判效率提升 60 倍，形成 IT 全栈厂商的安全 AI 差异化路径。

58. 丈八网络：网络空间兵棋推演的数学建模与 AI 推演

国内少数聚焦网络空间兵棋推演的厂商，以数学建模+AI 推演支持高强度实战演训与作战模拟。

59. 万里红：党政军涉密场景的大模型安全护栏

面向党政军涉密场景打造"护栏+监控+定密"三位一体大模型安全护栏，填补涉密领域 AI 治理空白。

60. 孝道科技：软件供应链安全 AI 检测智能体

软件供应链安全 AI 检测智能体，结合 AI+全域测绘构建开源+闭源代码统一安全评估能力。

61. 瀛云科技（运维安全）：云原生 SaaS 运维安全平台

云原生 SaaS 运维安全平台厂商，主打"按需订阅+低门槛部署"，面向中小企业运维安全合规场景。

10.10 按场景分类

安全运营（SecOps）

国际方面代表厂商包括 Dropzone AI（AI SOC Analyst）、7AI（自主 Agent）、Torq（超自动化）、Microsoft Security Copilot 和 Google Security Operations。国内方面代表厂商包括深信服（AISOC，告警降噪 99%，国家级实战验证）、360 数字安全（CCoE 大模型，告警降噪 83-99%）、奇安信（AISOC，事件处理时间降 98%）、未来智安（XDR+AI 智能体，告警降噪 99.8%）、火山引擎 Circle（AI 原生运营）、绿盟科技（风云卫 AI 安全能力平台（NSFGPT））、金睛云华（CyberGPT 降噪 99.8%+）、知其安（堆叠式 AISOC，招行共建）、青藤云安全（无相 AI，L4 自主防御，告警降噪 99%）、众智维（600+运营智能体+RAG 数据飞轮）、聚铭网络（万能联动+铭智大模型，中小政企）、新华三（H3C 灵犀，降噪 99%+）、长扬科技（工控场景，研判效率提升 60 倍）、启明星辰（九天安全大模型与盘古合作的 AISOC，央国企核心客户覆盖广泛）、络安科技（AI 驱动 SOC 平台，面向金融与政企提供自动化研判与响应）、瀛云科技（安全运营管理平台）。

威胁检测

国际方面代表厂商包括 Vectra AI（NDR Leader）、Darktrace（自学习）、CrowdStrike（EDR 第一）、SentinelOne（AI 原生）和 Abnormal（邮件行为 AI）。国内方面代表厂商包括奇安信(AI 天眼、AI XDR)、金睛云华（恶意加密流量检测 98%+）、360 数字安全

（终端行为检测 F1≈1.0）、深信服（Web 流量检出率 95.7%，0day 检出率 90%）、安恒信息（MoE 架构召回率提升 18-27%）、六方云（仿生免疫无监督 ML，工控 NDR）、山石网科（智能防火墙/NDR，BDS 行为检测+AI 威胁狩猎）、亚信安全（信立方威胁情报+AI EDR/XDR，央企与运营商主力）、知道创宇（ZoomEye 全网测绘+Seebug 漏洞情报+创宇盾，AI 威胁情报与 WAF 联动）、网宿安全（全球边缘节点，30 亿级攻击样本训练）、江民科技（反病毒基因+AI 恶意样本识别，KV 大模型赋能检测）、天懋信息（流量异常与 APT 检测，AI 驱动 NTA/FTA）、成都数默科技（全流量+AI 威胁建模，面向公安/运营商侧重 APT）。

数据安全

以国内厂商为主，代表厂商包括安恒信息（智能化网络安全平台、AI 数据安全、智能化代码审计）、炼石网络（数据安全+密评密改）、安华金和（93%分类准确率，3600 万字段 15 天完成）、美创科技（诺亚防勒索+数据库审计，连续多年数据安全代表厂商）、数安行（DataSecOps 零信任，探针 CPU<1%）、石犀科技（资产识别+风险预判）、保旺达（运营商数据全链路溯源，成本降 90%）、上海观安（本地化大模型+省平台合规上报）、指掌易（移动办公数据安全+AI 行为风控，BYOD/MAM 赛道头部）、领信数科（便携 MLBox+10000:1 降噪）、闪捷信息（全生命周期 DLP+敏感数据识别，AI 智能脱敏）、明朝万达（大模型驱动分类分级智能体，Gartner 连续推荐荐）、海泰方圆（密码+数据治理+AI，党政军深耕）、烽台科技（工业数据安全）。

渗透测试与漏洞挖掘

国际方面代表厂商包括 Snyk（AST Leader、Agent Fix 自主修复）和 XBOW（自主渗透）。国内方面代表厂商包括云起无垠（智能 Fuzzing+DevSecOps 全流程）、绿盟科技（自动化渗透测试工具）、长亭科技（慧鉴 SAST+码力 AI 开发安全）、华清未央（智能化漏洞挖掘，MLM 机器语言大模型）、华云安（灵知 ASM+灵刃自动化渗透，攻击面发现与验证）、矢安科技（AI 攻击编排，29800+无害化用例，AEV 国内代表）、默安科技（蜜罐+暴露面+开发安全，IAST 误报<10%）、魔方安全（EASM+CAASM，影子资产识别率 85%+）、斗象科技（PRS 漏洞管理+网藤攻击面，AI 驱动漏洞优先级排序）、墨云科技（智能化渗透测试）、经纬信安（攻击诱捕）。

工控/OT 安全

国内代表厂商包括烽台科技（自研模型+边侧小模型架构）、和利时（工控垂域）、宁数安全（MCP 协议支持）、威努特（工控 AI 智算平台，研判准确率 95%）、长扬科技（工控+工业安全视觉 AI 双向并举）、六方云（仿生免疫 NDR，工控专精）、万物安全（AI 固件漏洞挖掘与物联网资产测绘）。

邮件安全

国内代表厂商包括广东盈世/CACTER（26 年邮件安全+大模型网关，反垃圾 99.8%）、灵云数科网哨 M01（公安部一所情报联防+样本深度分析+APT 画像）、知其安（招行钓鱼邮件准确率 96.4%）、方向标 FangMail（轻量增量检测，加密附件解密）、立智安智瞳（多 Agent+MCP+ReAct 架构）、深信服（钓鱼检出率>99%）。

鉴伪/认知安全

国内方面代表厂商包括中科睿鉴（深伪检测）、瑞莱智慧（深伪检测）、摄星科技（AI 认知对抗与情报分析，深度学习驱动的多源情报融合）。

反欺诈情报

国内代表厂商包括威胁猎人（DarkSphere 海外电商反欺诈，AI 分类准确率 95%+）。国际方面可参考 Featurespace、Feedzai 等。

AI 安全治理（Security for AI）

国际方面代表厂商包括 HiddenLayer（AI 资产+MCP 安全）和 Snyk（AI 代码+模型安全）。国内代表厂商包括火山引擎（AI 安全护栏、智能体安全管理平台）、百度安全（AI 安全护栏）、金睛云华（双子大模型驱动，大模型赋能全产品线）、安全数智（智能体安全评测平台、防护平台、统一管理平台）、长亭科技守元（大模型安全围栏，流式异步检测）、瑞数信息（WAAP for LLM，动态安全+LLM 防护双向布局）、安芯网盾（AI Decoder 大模型可信度评估+Shadow AI 治理）、领信数科（便携 MLBox 监管合规评估）、悬镜安全（AI 原生安全，模型血缘图谱，SAST 误报<2%）、万里红（党政军涉密护栏+敏感信息监控+智能定密）、云弈科技（大模型安全卫士）。

攻防推演与安全教育

国内代表厂商包括赛宁网安(AI 实训平台)、丈八网络（图论+AI 攻防推演，分钟级战役级推演）、永信至诚（AI 安全教育+春秋云阵蜜网，200+高校产教融合）。

软件供应链安全

国内代表厂商包括悬镜安全（Skills 投毒检测，MCP/Skills 供应链审计）、孝道科技（AI 检测智能体+全域资产测绘）、奇安信(代码卫士、代码安全智能体)、海云安(开发者智能助手、高敏捷 AI 白盒)。

云安全与 WAAP

国内代表厂商包括瑞数信息（WAAP for LLM，私有云 WAF Top 2）、厦门快快网络（云安全+AI 安全运营，12 万+客户）、网宿安全（全球边缘 AI 防护，首个 WAAP 信通院认证）、国舜股份（AI 驱动 WAF/API 安全+安全运营，金融与央企核心客户）。

其他

芯盾时代（AI 增值零信任+身份安全）、持安科技（零信任、智能化分类分级，智能脱敏）

10.11 国内市场整体观察

本节综合 2025 年 9 月至 2026 年 3 月完成的 11 场头部厂商深度 Briefing、66 家国内安全厂商的结构化问卷（数说安全调研口径，覆盖 60 余个维度），以及厂商公开材料、产品白皮书、发布会口径，形成对国内 AI for Security 市场的整体判断。调研提示出五个贯穿性的结构特征：一是成熟度"虚热"显著，宣传层面普遍成熟，但以"可量化收益+多客户复用+稳定订阅"为标准真正规模化落地的厂商估计不超过 10 家；二是大小模型协同并非优雅的架构设计，而是实时性、成本与效果约束下的工业妥协；三是数据可用性与质量是比模型或算力更突出的底座瓶颈；四是客户采购驱动力已从"检测率提升"明显转向"人力与外包成

本替代"；五是 Security for AI 赛道由上游 AI 生态演进倒推，而非下游用户需求驱动。以下分层展开。需要特别说明的是，本节数据多来自厂商自述（Briefing、发布会与问卷口径），尚缺独立第三方审计，定量结论以质性判断为主。

成熟度分布

基于 61 家厂商（11 家深度画像+11 家问卷画像+39 家材料/公开信息画像）的综合评估：达到 4-产品化平台阶段的有 4 家（深信服、绿盟、安恒信息、火山引擎），3-规模交付阶段有 2 家（天懋、石犀），仍处于 2-试点交付阶段的有 6 家（瀛云、摄星、炼石、和利时、宁数、盈世）。基于 66 家结构化问卷样本（自评口径）：成熟度 5 分（规模商业化）26 家（约 39%）、4 分（产品化平台）15 家（约 23%）、3 分（规模交付）23 家（约 35%）、2 分（试点）2 家（约 3%）；合计约六成受访企业自评达到产品化/商业化阶段，约三分之一仍处于规模交付阶段。头部 11 家深度画像厂商（含启明星辰、亚信安全新增 Briefing）已大规模进入产品化阶段，部分厂商进入规模商业化——根据各公司公开披露与发布会口径：安恒 AI 相关收入超 2 亿、启明星辰 AI 项目 35 个、亚信 AIXDR 60+客户。上述均为厂商自述口径，尚缺独立第三方审计数据，但仍是观察国内 AI for Security 商业化进程的重要参考节点。

需要特别提示的是，本轮调研中"成熟度虚热"现象相当显著：多数厂商在对外披露中自称已达到 3-4 级成熟度，但若严格以"可量化收益+多客户复用+稳定订阅"为统一标准，真正具备 4 级特征的厂商估计不超过 10 家（深信服、360 数字安全、启明星辰、安恒信息、长亭科技、华云安等属于较为确定的候选）。宣传层面的普遍成熟与实际规模化落地

的少数突破，共同构成理解未来 12-24 个月国内竞争演进的基本前提——这也是本节其余观察（技术路线、竞争格局、差异化路径）所共享的背景假设。

技术路线收敛

基于 2026 年 3 月完成的多场深度 Briefing，国内厂商技术路线呈现更清晰的收敛趋势：

模型层：开源大模型（千问/DeepSeek）为主流基座，自训大模型收益递减——多场 Briefing 中有 6 家明确表达了这一判断（启明星辰"现在调用第三方有时候更好"、360 保留 14B 自训仅用于精度敏感场景、深信服 32B 以上不再微调、长亭只做小/中模型）。竞争壁垒正从"谁的模型更强"转向"谁的工程化更扎实"（语料生产平台、数据底座质量、工具链 MCP 化）。

接口层：MCP 正在成为 AI 工具连接层的开放协议事实标准。8 家中有 5 家明确采取"MCP/A2A 协议化对外开放 + OpenClaw 等运行时承接调用"的组合策略——360 所有能力 MCP 化（中石油和上海公安已通过 OpenClaw 运行时直调）、深信服将检测/运营大模型全部 API/MCP 化、安恒信息推动产品 MCP 化改造、启明星辰 AIDK 同时支持 MCP 与 A2A 协议、未来智安 MCP 原生驱动。这标志着安全行业从"封闭平台"向"开放能力"的结构性转型。

架构层：大小模型协同——小模型（7-14B）负责实时检测（延迟<10ms），大模型（32B+）用于深度研判和通用任务，分层模型架构渐成共识。深信服快慢双轨（7-8B 精调快路径+32B 开源慢路径）是典型案例。

部署层：私有化为主导，从本轮受访厂商客户分布看，国内政企客户绝大多数（多家厂商口径在 90%以上，但样本与口径定义未做统一标准化）要求纯内网部署。与 Claude 差距约 20%（代码场景，长亭/悬镜独立验证），通用场景差距更大（火山引擎反馈"不止一个数量级"）。私有化效果天花板短期难以突破。

壁垒层：从"模型能力"转向"工程化能力"——深信服 15 年安全语料积累、360 百 PB 级数据底座、安恒 121 个官方智能体的规模化交付能力，是更难复制的护城河。

竞争格局观察

梯队分化加剧：第一梯队（深度验证+规模商业化）以深信服（MSS 3500 客户+8 个百万美金海外大单）、360（自训 14B +OpenClaw 全开放）、安恒信息（AI 收入 2 亿+订阅制）、启明星辰（移动战略绑定+AI 客单价翻倍）为代表，已进入规模商业化阶段。第二梯队（技术领先但规模待验证）包括金睛云华、绿盟科技、长亭科技、青藤云安全。第三梯队（细分赛道专精）包括知其安、悬镜安全、亚信安全等。中腰部厂商（华云安等）仍偏 Copilot 形态，正在追赶。

头部厂商 AI 落地路径分化：部分厂商采用"AI 加"模式（在传统安全平台中嵌入 AI 模块），部分采用"AI 驱动"模式（以 AI 为核心重构产品架构）。在电网等行业实测中，AI 原生产品在告警降噪效果上普遍优于传统平台叠加 AI 的方案，但算力投入差异显著（从 8 块 4090 到数百万元算力不等）。

"人是 AI 的一个环节"：深信服的表述代表了行业最前沿理念——不是 AI 辅助人，而是人成为 AI 工作流的审核环节。MSS 团队 T1 研判员从 50 人削减至 5 人（2026 年 8 月目

标），每个工程师服务客户从 17 家提升至 30 家。若目标兑现，将是国内 AI 安全运营的里程碑级验证。

多位厂商 CEO 确认：客户不会单独为"AI for Security"标签额外付费，AI 能力的商业化更多体现为产品升级而非独立品类，这一观点与海外市场发展趋势基本一致。安恒信息订阅制商业化（1000+客户、分类分级客户从 20→200 半年内 10 倍增长）是国内少数已验证的例外。

量化收益亮点

全维度>50%的厂商有绿盟。工时降>50%的有天懋、烽台、绿盟共 3 家。MTTR 改善>50%的有绿盟。误报降>50%的有烽台、绿盟、云弈。告警降噪率 99%+的厂商包括深信服、未来智安、金睛云华、青藤云安全、长扬科技（工控场景 99%）。极致降噪案例：领信数科子牙智能体实现 10000:1 降噪比（日均 10 万→10 条）。

主要挑战共识

数据质量（9 家）> 私有化要求（6 家）> 评测缺失（5 家）> 人才短缺（5 家）> 幻觉（5 家）。新增挑战：私有化效果天花板（国内外大模型差距约 20%-数量级不等），以及 Security for AI 价格战（字节/百度/蚂蚁以 20 万抢 300 万预算项目）导致专业厂商投入回报不足。

差异化竞争路径

行业深耕（13 家）> Agent 闭环（8 家）> 数据沉淀（7 家）> 深度集成（7 家）> 低成本私有化（4 家）。新趋势：MCP 开放生态（5 家头部厂商已落地）、Security for AI 独立赛道（7 家以上厂商布局，但价格战已现）。

新增厂商综合观察（第四批，2026 年 3 月 22 日）

第四批 23 家厂商材料分析进一步完善了市场全景，带来以下新发现：

一是**数据安全赛道 AI 化最密集**，形成四类差异化路线并存格局：安华金和/明朝万达代表"智能体分类分级"路线（MCP 工具调用+垂域语料，准确率 93%+）；炼石网络/石犀科技代表"硬件加速+分层模型"路线；上海观安/海泰方圆代表"本地化部署+合规上报"路线；数安行代表"DataSecOps 左移"路线（探针 CPU<1%的极致轻量化）。该赛道厂商数量最多（超 10 家），竞争最激烈。

二是**工控安全 AI 化形成梯队**：烽台科技（自研模型，成熟度最高）、威努特（AI 智算平台+工控知识体系）、长扬科技（工控+工业视觉 AI 双向并举）、六方云（仿生免疫 NDR 专精）、宁数安全（MCP 协议工业资产）、和利时（工控垂域）共 6 家形成工控 AI 安全完整梯队，是国内工控场景 AI 应用最完整的厂商群体。

三是**攻防推演和供应链安全出现新型玩家**：丈八网络以纯数学/图论/AI 填补了国内网络空间兵棋推演的产品空白；孝道科技以 AI 检测智能体聚焦软件供应链的全链路监管，与悬镜安全形成互补（一个侧重供应链资产测绘，一个侧重 Skills 投毒和 AI 原生安全）。

四是**运营商和党政军细分赛道呈现专业化分工**：保旺达聚焦运营商数据全链路溯源（成本降 90%的量化成效）；万里红聚焦党政军涉密场景全链路管控（护栏+监控+定密三位一体）；海泰方圆深耕党政军工密码+数据 AI（30+国家部委+九大军工集团 900+院所），均是难以复制的细分壁垒。

客户驱动力的结构性变化——从"检测率提升"转向"降本增效"

本轮多场 Briefing 反馈出一个高度一致的信号：客户采购 AI for Security 产品的驱动力已从"提升检测能力"明显转向"人力与外包成本的直接替代"。根据厂商自述（未经第三方独立审计），深信服已将 MSS 值守团队的 T1 研判员缩编作为核心量化指标，计划从约 50 人降至 5 人左右（2026 年 8 月目标），单工程师服务客户数从 17 家提升至 30 家；启明星辰以"AI 版客单价由约 20 万元翻倍至 40-50 万元"的价格锚点重新定义产品线；安恒信息的数据分类分级订阅制客户在半年内从约 20 家增长至约 200 家，整体 AI 订阅客户自述超过 1000 家。工行、中石油等国企的采购逻辑尤为典型——关注点不是"检测率还能提升几个百分点"，而是"多少外采值守服务可以被 AI 替代"。这一结构性变化对厂商提出两点要求：一是需要具备可复现、可量化的人力替代案例以支撑客户的商业论证，二是需要建立"售前试点—规模化交付—持续运营"的三阶段方法论，这也是中腰部厂商与头部厂商之间最难弥合的工程化能力差距所在。

Security for AI 赛道：由上游 AI 生态驱动，而非下游用户需求

另一个需要独立审视的现象是：Security for AI（面向 AI 系统自身的安全防护）赛道的起势，并非来自客户主动提出的需求，而是由上游大模型与 Agent 生态的快速扩张倒逼而来。昂镜安全、海云安、亚信安全、360、火山引擎等厂商进入该赛道的触发点均非传统招标需求，而是来自三个方向：一是大模型开源社区出现模型投毒、组件污染等事件，使 AI 供应链安全从学术讨论变为现实风险；二是 MCP、A2A 等协议使得工具调用面与 Agent 间协作面大幅扩展，传统网络与应用边界防御在 Agent 场景下失效；三是政策层面对生成式 AI 与智能体的合规要求持续加码，倒逼厂商建设 AI 风险监测与护栏能力。这意味着 Security for AI 赛道的发展节奏将由"AI 生态演进速度"而非"客户需求成熟度"决定——当大

模型与 Agent 技术本身尚未稳定，安全需求的具体边界也仍在漂移。这对厂商而言既是窗口期机会，也是存在性风险：一旦大模型生态走向少数头部厂商的封闭运行时，独立 Security for AI 产品的生存空间将被明显压缩。价格战的提前到来已是警示——据长亭科技在 Briefing 中的披露，字节、百度、蚂蚁等互联网大厂正以约 20 万元的报价进入此前约 300 万元预算量级的智能体安全项目，专业厂商的投入回报面临挤压。

未来演进方向

整体呈现 Copilot→Agent→Autopilot 的三阶段演进趋势，青藤云安全无相已率先定位 L4 级自主防御，深信服"人是 AI 的一个环节"理念代表了最激进的人机关系重构方向。RAG/TAG 成为标配，MCP 正在成为事实标准接口，多模态能力持续扩展。私有化降本（推理加速/分层模型/蒸馏）、行业深耕和闭环运营（审计/回滚/复盘）逐步完善是主要技术方向。

值得特别关注的是 Claude Code Security（2026 年 2 月）和 OpenAI Codex Security（14 天内跟进）的发布，在代码安全/AppSec 细分赛道，AI 推理对传统 SAST 的颠覆正在发生，时间表从"3-5 年"压缩至"12-24 个月"。国内长亭科技慧鉴、海云安 AI 白盒等产品也面临 AI 原生替代压力，悬镜安全 SAST 误报已降至<2%是积极应对的信号。

10.12 厂商选型建议

选型总体原则

由于国内政府和国有企业采购国外网络安全产品存在限制，国内市场以本土厂商产品为主。仅部分商业企业客户和出海企业仍会采购海外安全产品。因此，本节选型建议以国

内厂商为主体，兼顾海外厂商在特定场景下的参考价值。需要说明的是，国内厂商的 Briefing 工作仍在持续推进中，后续版本将补充更多厂商的深度画像。

按主要痛点选型

主要痛点	国内推荐厂商	海外参考厂商
告警疲劳/降噪	深信服安全 GPT、360、奇安信 AISOC、火山引擎 Circle、绿盟科技、金睛云华、青藤云无相、未来智安	Dropzone AI、CrowdStrike Charlotte AI
数据分类分级	安恒信息、炼石网络、海云安、明朝万达、安华金和、上海观安	—
安全运营效率	深信服、奇安信、360、绿盟科技、火山引擎 Circle、安恒信息、众智维	Microsoft Security Copilot、Google Gemini
威胁检测	360、深信服、安恒信息、金睛云华、六方云	Vectra AI、Abnormal Security
漏洞挖掘/渗透	云起无垠、绿盟科技、海云安、长亭科技、矢安科技、默安科技、魔方安全、华清未央	Snyk、XBOW
工控/OT 安全	烽台科技、和利时、宁数安全、威努特、长扬科技、六方云	Darktrace
邮件安全	广东盈世/CACTER、灵云数科网哨 M01、知其安、方向标 FangMail、立智安智瞳	Abnormal Security

AI 安全治理	火山引擎、蚂蚁、百度	HiddenLayer
软件供应链安全	悬镜安全、孝道科技	—
攻击面管理	魔方安全、矢安科技	—
数据全链路溯源	保旺达、数安行	—

按企业类型选型

政府和国有企业：受采购政策约束，应选择国内厂商。重点关注数据合规与私有化部署能力、国产 GPU 和国产大模型适配性。推荐 360 数字安全（CCoE 架构+320 亿样本底座）、深信服（安全 GPT+国家级实战验证）、奇安信（AISOC+双备案合规）、绿盟科技（全维度覆盖）和安恒信息（数据安全市占率 18.7%）。在工控领域推荐烽台科技、威努特和长扬科技（工控+工业视觉 AI 双向并举）。在威胁检测领域金睛云华（双模型架构+自建算力）值得关注。青藤云无相在主机侧 AI 自主防御领域提供差异化选择，芯盾时代在零信任和身份安全领域有深度积累。涉密场景推荐万里红（护栏+监控+定密三位一体）和海泰方圆（密码+数据 AI+军工经验）。

商业企业（国内运营为主）：可选择国内厂商为主，部分场景可考虑海外厂商的 SaaS 服务。对数据出网限制相对宽松的企业，火山引擎 Circle 的云端 SaaS 方案是高性价比选择。中小企业运维安全合规可关注瀛云科技（云原生 SaaS 堡垒机，按需订阅）和聚铭网络（万能联动，低门槛 SOC）。有海外电商业务的企业可关注威胁猎人 DarkSphere（AI 反欺诈情报，95%+准确率）。

出海企业和外资企业：可灵活选择海外厂商产品，特别是 Microsoft Security Copilot（微软生态整合）、CrowdStrike（全球端点保护）等，同时在国内业务部分搭配国内厂商方案。网宿安全（全球 2800+边缘节点）可作为出海安全防护的国内选项。

按部署模式选型

部署模式	适用客户	推荐方案
完全私有化	政府、军工、金融	绿盟科技、金睛云华（完全本地化部署）、烽台科技、海泰方圆（国密+私有化）、万里红（涉密场景）
混合部署	大型企业	深信服（云端+私有化双模式）、炼石网络（软硬一体机）、领信数科（便携 MLBox+混合部署）
云端 SaaS	互联网、商业企业	火山引擎 Circle（5 万/年起）、瀛云科技（SaaS 堡垒机+特权账号）
MSP 托管	中小企业	火山引擎 MSP（45 家合作伙伴）、聚铭网络（SOC 运营托管）
增量叠加	已有安全基础设施的企业	方向标 FangMail（不替换现有邮件网关，轻量增量 AI 检测）

第十一章 市场落地建议

11.1 企业参考架构

基于调研与访谈，我们提出 AI 赋能网络安全的五层企业参考架构：

层级	职责	关键组件
治理审计层	AI 决策的可控性与可追溯性	审批工作流、操作审计、回滚机制、合规报告
工具层	为 Agent 提供执行能力	查询/情报、封禁/隔离、工单/通知、规则下发、取证
模型/Agent 层	AI 推理与决策	大模型基座、垂域模型、Agent 编排、多智能体协同
检索层	上下文增强（RAG/TAG）	向量库、知识库、威胁情报、规则图谱、行为基线
数据层	全量安全数据	告警/日志、流量/会话、样本/沙箱、工单记录、知识文档

架构演进路径：企业可按四个阶段逐步演进。阶段一为单点插件（成熟度 2-3），单场景 AI 能力嵌入（如告警降噪），适用试点验证。阶段二为垂域模型（成熟度 3-4），针对特定领域深度优化（如数据安全、工控安全）。阶段三为平台化赋能（成熟度 4），多场景统一 AI 能力，多智能体协同+开放生态。阶段四为基座模型服务，提供模型训练/推理基础设施，赋能生态（如火山 Circle、知道创宇 AiPy）。

11.2 集成模式与运营化

11.2.1 四种集成模式

模式	特点	适用场景	代表实践
API 集成	松耦合，低侵入，跨平台	多厂商异构环境、云上部署	知道创宇大模型网关、宁数 MCP
平台内嵌	深度集成，体验一致	综合性安全厂商自有平台	绿盟智能中枢、美创数据安全平台
独立部署	技术栈独立，生态化运营	AI 能力服务化输出	火山 Circle SaaS、知道创宇 AiPy
混合部署	按需选择，灵活演进	大型企业复杂环境	炼石软硬一体机、摄星多形态并存

11.2.2 运营化成熟度

调研显示（66 家问卷自评口径）：约 62%受访厂商自评成熟度达到 4 分（产品化平台）及以上，约 35%处于 3 分（规模交付）阶段，约 3%仍处于试点（2 分），与 10.11 节综合画像基本一致——真正以“可量化收益+多客户复用+稳定订阅”为标准的规模化落地厂商仍不超过 10 家。

成熟度	特征	典型厂商
L1-被动试点	单场景 PoC，无业务闭环	和利时
L2-主动试点	少量客户试点，工时下降 10-30%	瀛云/摄星/炼石/知道创宇等 8 家
L3-规模交付	多客户部署，流程标准化，工时下降>50%	天懋/美创
L4-产品化平台	持续迭代，平台化架构	海云安/烽台/绿盟/云弈

L5-基座服务	开放 API，多租户 SaaS	火山 Circle/知道创宇 AiPy

11.2.3 从助手到 Agent 的演进

调研显示 Copilot→Agent 是 10 家厂商的重点投入方向：

Chatbot 助手（被动问答）→ Copilot（辅助决策）→ Agent 智能体（自动执行）→ 多智能体协同（协同规划）

11.3 试点路线图

PoC 阶段（90 天）

推荐起步场景：

场景	难度	价值	数据支撑
告警研判	★★	★★★★★	10 家客户需求，知道创宇工时下降 30-50%
数据分类分级	★★★	★★★★★	美创效率提升 6 倍，准确率 80%+
报告生成	★	★★★	11 家已实现，技术成熟

关键步骤：需求调研与基线测量（2 周）→ 数据接入与知识库构建（2 周）→ RAG 流程开发与调优（4 周）→ 小范围试点与效果评估（4 周）

成功标准：工时节省≥10%（保守目标），误报下降≥5%，幻觉率<10%

运营化阶段（6 个月）

扩面准备→灰度发布（10-30%）→全量上线→多场景扩展→复盘优化

关键动作：知识库扩充至 1000+份文档、流程固化纳入 SOP、建立反馈与知识沉淀机制

规模化阶段（12 个月）

Agent 自动执行能力→多智能体协同→生态化运营→持续创新

11.4 指标体系

建议从四个维度建立评估体系：

效率指标

指标	定义	调研对标
工时节省率	AI 介入前后人均处理工时变化	5 家>50%（天懋/美创/知道创宇/烽台/绿盟）；4 家 30-50%
MTTR 改善率	平均响应时间缩短比例	知道创宇/烽台达 95%改善
自动化处置率	AI 自动完成的处置任务占比	知道创宇/美创/绿盟闭环率提升>50%

质量指标

指标	定义	调研对标
误处置率	AI 错误处置占总处置比例	云弈误报下降 50-70%，海云安案例误报下降 90%
准确率/检出率	Precision 与 Recall	美创风险监测：检出率 95%、准确率 95%

风险指标

风险指标包含三方面：一是审计覆盖率，10 家厂商已实现全链路记录，目标 100%；二是幻觉率，5 家厂商反馈为主要阻碍，通过规则引擎+大模型混合、RAG、思维链验证控制；三是越权风险，作为红线指标，越权事件数应为 0。

成本指标

部署模式	**3 年 TCO（示例）**	**ROI**
云上 SaaS	~95 万	~311%，约 4 个月回本
私有化	~380 万	~3%，约 2.9 年回本

成本优化策略：云上优先（数据可出网场景）、小模型+RAG（私有化降本）、分层模型（边缘小模型+云端大模型）

11.5 关键成功要素

4. **高层支持与资源投入**——成功案例均有高层战略重视；
5. **数据质量与知识沉淀**——9 家厂商反馈数据质量是最大阻碍，数据治理先行；
6. **技术路线务实**——RAG 优于微调（中小企业），云端优于私有化（数据可出网），小模型+工程优化优于盲目追求大模型；
7. **渐进式演进**——Chatbot→Copilot→Agent，单场景→多场景→平台化；
8. **生态化思维**——不做全栈，聚焦核心能力，与数据湖、安全工具厂商深度合作。

11.6 主要风险与应对

风险	**应对措施**
模型幻觉导致误处置	规则引擎+大模型混合、人工审批、回滚机制
私有化模型效果差	云端优先、AI 机密计算探索、小模型+RAG

数据质量差	AI 自动化标准化转换（美创实践：接入时效从数周缩至 1 小时）
用户抵触	定位辅助而非替代，邀请核心用户参与设计
厂商绑定	优先开放标准（MCP），保留迁移能力

第十二章 案例研究

章节定位：以场景簇归类，通过真实案例验证 AI 赋能网络安全的实际效果，为行业提供可参考的实践路径。

一、绿盟科技：风云卫 AI 安全能力平台与智能安全运营

场景簇：安全运营 / 主动防御 / 数据安全 / 专项知识问答。

绿盟科技风云卫 AI 安全能力平台（NSFGPT）是一款集成了绿盟科技 20 余年网络安全攻防经验、千亿级安全语料库与垂域大模型的 AI 安全赋能系统。平台通过"大模型+小模型"协同架构，实现了对安全运营、检测响应、攻防对抗、数据安全等场景的全面覆盖，为用户提供智能化、自动化、精准化的安全防护能力。本报告从产品定位、技术架构、部署形态、应用场景、效果评估及标杆案例等维度，全面展示该平台的 AI 安全赋能价值。

产品/解决方案定位：

绿盟科技风云卫 AI 安全能力平台（NSFGPT）定位于企业网络安全体系中的智能中枢，在 AI 应用进一步拉大攻防对抗不对等的背景下，集中赋能网络安全新型攻击方式检测能力提升、自动化安全运营、常态化智能渗透测试、集成开发辅助代码审计和供应链安全检测、威胁情报和漏洞挖掘，解决企业 AI 时代网络安全运营中的噪声淹没、检测能力失真、效率低下，以及智能化缺失问题的专业解决方案。

同时该平台通过松耦合设计，通过接口调用为企业现有平台、设备等赋能，最小化改造企业现有网络安全体系，使企业能够快速转向"AI 赋能安全"的理念，将绿盟科技积累的安全知识、威胁情报与 AI 技术深度融合，解决以下核心问题：

- 安全运营告警量巨大且误报率高，人工处理效率低下

- 未知变种攻击难以检测，传统防御体系存在盲区
- 安全事件调查周期长，跨系统数据关联困难
- 渗透测试周期长、效率低
- 数据安全分类分级效率低，标准不一且难以覆盖多模态数据
- 安全知识传递困难，专家经验难以规模化复用

应用场景：

风云卫 AI 安全能力平台覆盖了多个关键业务场景，形成了全方位的 AI 安全防护能力：

1. 安全运营场景

- 告警降噪与分诊研判：通过网络安全垂域模型、大小模型结合的技术，实现 98% 的有效降噪，精准识别高危威胁，大幅度降低人工处理告警量
- 自主调查与响应：AI 智能体自动检索多设备相关上下文日志、情报线索，进行关联分析，构建完整攻击图谱，实现攻击链还原时间降至分钟级，同时支持并案分析，精准定位攻击组织。
- 7×24 小时自主运营：通过智能体协同调度，实现安全事件的全自动闭环处置，无需人工干预，提升安全运营效率。
- 知识沉淀与分析报告：将专家经验封装为可复用的模型，自动生成攻防复盘报告，沉淀安全知识，减少人工重复劳动。

2. 主动防御场景

- AI 自动化渗透测试/攻防演练：基于大模型的智能渗透测试能力，可辅助攻击建议、攻击路径推荐，自主逻辑漏洞挖掘，大幅降低渗透测试时间和成本，使常态化渗透测试成为可能
- AI 代码审计/漏洞挖掘：AI 辅助全自动化代码审计和漏洞挖掘，极大降低传统审计的误报，实现代码质量、注释、安全等规范的深度审计

3. 基础能力提升场景

- 未知攻击检测：通过 AI 语义理解能力，深入剖析恶意程序本质特征，有效识别"已知家族、未知变种"攻击，提升威胁检测整体能力
- 行为基线检测：AI 能秒级输出动态阈值，构建精准的正常行为基线，当实体行为偏离阈值时快速察觉异常，大幅提升用户行为分析技术的可用性
- AI 情报与分析：AI 赋能威胁情报实体提取和实体关系分析，建立高效动态更新的安全知识库，赋能网络安全体系，增强安全决策的准确性。

4. 数据安全场景

- 结构化与非结构化数据分类分级：通过大模型底座与 RAG 数据库技术，实现对文档、图像、音频等多模态非结构化数据的深度解析与敏感数据识别，提升分类分级检测效率
- 行为基线构建：AI 能自主解读并理解全量日志，从中提炼账号、流量、进程等实体在正常时段的行为画像，实现动态阈值生成。

5. 专项知识问答场景

- 安全知识库与情报库：依托千亿级安全样本训练的垂域大模型，结合动态安全知识库，提供安全知识问答、威胁情报查询等能力，辅助安全专家快速决策。
- 安全智能体交互：通过自然语言交互，安全人员可直接查询攻击链路、威胁情报、处置建议等信息，提升安全工作效率。

技术路线：

风云卫平台采用"双底座+多智能体"的技术架构：

- 网络安全垂域模型：超千亿 Tokens 语料训练，依托绿盟科技二十余载网络安全产品研发与安全运营经验积累，平台模型训练语料总量超千亿 Tokens。语料构成兼顾多样性和高质量，涵盖通用语料、安全语料、编程代码三大核心领域，确保模型在具备广泛通用能力的同时，能精准应对网络安全垂直行业的专项挑战。
- 双基座模型：内置绿盟垂域大模型（SecLLM）与国内先进推理模型的双基座方案，适应网络安全领域不同类型的场景需求，构建覆盖智能告警降噪、报文解析、恶意代码分析等众多应用场景的 AI 安全能力。
- 多智能体协同：内置 20+安全领域场景化智能体，覆盖网络安全多个关键环节，包括钓鱼邮件检测、可疑样本分析、敏感数据识别、零配置日志解析、情报分析、报告生成等，支持可视化编排与协同工作。

部署形态：

风云卫平台作为企业网络安全体系的智能中枢，提供多种灵活的部署形态，满足不同客户的需求：

- 本地化硬件：支持高算力风云卫一体机部署，可弹性扩展

- 本地化软件：提供大模型和智能体中心一体的软件形态，可以部署在企业自有算力服务器
- 云端服务部署：AI 赋能绿盟鹰眼安全运营中心，通过 MDR 服务为全国 2000 多家中心客户提供云端 SaaS 化智能安全运营服务

硬件要求：

支持英伟达生态 L20 算力卡，及相同算力的华为昇腾、海光等国产算力生态设备，根据业务场景压力，可选 2 卡、4 卡、8 卡

效果评估：

绿盟风云卫 AI 安全能力平台基于 “风云卫 + DeepSeek” 双底座与多智能体架构，从合规达标、威胁防御、运营效率、AI 落地四大核心维度，为客户提供可量化、可落地的价值赋能，推动安全运营从 “人防” 向 “智防” 的质变升级。

1、强化全域防御，提升威胁对抗能力

通过 AI 大模型的深度推理与多智能体协同，绿盟科技风云卫可以作为企业网络安全体系智能中枢，辅助未知威胁精准识别、攻击链条全景还原、实战攻防能力升级，构建全链路防御体系

2、全方位开箱即用的安全智能体，风云卫 AI 安全能力平台预置了二十余款开箱即用的智能体，帮助企业构建起全面且深入的安全能力体系，覆盖了威胁检测、事件运营、漏洞评估及数据安全等网络安全的重要核心场景

3、优化运营效能，降低人力依赖成本，针对 “告警过载、分析低效、响应滞后” 等运营痛点，平台通过全流程 AI 赋能实现效能倍增

- 告警降噪智能体过滤无效告警与重复信息，平均降噪率达 97% 以上，将日均数万条原始告警压缩至百级高价值告警，彻底解决告警失焦问题
- 分析处置加速：对 70% 以上的告警日志实现 AI 辅助研判，高频场景可实现从检测、研判到处置的全自动闭环，处置响应时间从小时级缩短至 30 分钟内，复杂事件调查效率提升 70% 以上。
- 知识管理提效：安全知识问答智能体整合千亿级安全语料，支持自然语言交互与多轮追问，将漏洞查询、情报查询、政策解读等知识获取时间从 10 分钟缩短至秒级，降低对资深专家的依赖。

4、常态化智能化渗透测试，风云卫平台赋能自动化渗透测试系统（AI-PTS） 通过自动化、智能化渗透测试，模拟多种攻击场景，实现：

- 快速定位系统架构、代码实现及配置层面的薄弱环节与安全漏洞，降低上线系统安全风险，
- 减少对专业人员人工测试的依赖，推动安全左移，将漏洞发现与修复提前至开发测试阶段，节约上线后漏洞修复成本，避免安全事件导致的线上故障与上线延误。
- 通过智能和自动化技术，实现周期性自动化渗透测试，7×24 小时持续监控评估系统安全状态，可随业务系统迭代和攻击技术演进优化检测策略，实现资源漏洞的常态化监控与高效管理，提升测试覆盖广度、效率与响应速度。

特色：

1、网络安全垂域模型，超千亿专业 Tokens 语料训练，基于 RAG 的本地知识应用，基于 AI 智能体的模型能力拓展

2、安全专家实战经验，实战化智能专家系统与安全思维链，实现安全运营全过程赋能支持

3、安全大模型深度赋能，AI 安全能力平台贯穿于安全运营、渗透测试等场景的各个环节，形成完整的智能化闭环能力

4、安全场景自适应能力，在攻防等安全场景，自适应地完成自动化攻击路径推荐、任务规划与编排、工具调用和任务执行

标杆客户：

1、某省级电信 AI 安全运营项目案例

某电信作为省级大型通信运营商，肩负全省通信网络基础设施的安全保障重任，其网络覆盖大量 O 域业务系统，部署有 30 + 台全流量设备、62 台防火墙、18 台网页防篡改、100 + 台 IPS 等多类安全设备，日均产生安全告警超 6000 万条，安全运营压力突出，客户日常面临告警过载与误报泛滥、运营模式低效等问题挑战。

解决方案：

绿盟科技为其部署风云卫 AI 安全能力平台 AISecOps 智能降噪模块，采用 “私有云全流量平台扩容” 部署模式，实现

- 告警降噪突破性成效：平台实现告警日志有效降噪压缩率 99% 以上，彻底解决 “告警过载” 难题，让高价值威胁从海量噪声中清晰呈现。
- 运营成本显著优化：告警事件研判自动化水平大幅跃升，人工处置成本平均下降 70% 以上；每年自动化处置网络安全威胁事件超 3 万个，每年平均节省人工安全运营成本近千万元。
- 威胁响应效能跃升：通过自动基线规律、事件归并与评级功能，99.9% 以上事件实现智能化覆盖，人工介入成本被充分压缩，安全运营团队可聚焦高价值风险处置，有力保障了全省通信网络的安全稳定运行。

2、某农商智能安全运营项目案例

某农商行作为地方重点金融机构，承担区域金融服务的网络安全保障职责，现网部署多厂商、多类型安全设备，面临海量安全日志与告警的运营管理压力，亟需构建智能化安全运营体系以满足金融行业合规要求与业务连续性需求。客户面临设备异构与告警过载、资产与漏洞管理脱节、自动化处置与异常识别不足等方面的挑战。

解决方案

绿盟科技为其打造 “智能网络安全运营平台 + 风云卫 AI 能力” 融合方案，构建四层联动的智能化安全运营体系，实现：

- 告警降噪与数据整合：成功接入安恒、奇安信、联软等 20 + 主流安全设备的 200 种类型日志，实现 98.49% 的告警降噪率，彻底解决告警过载难题

- 运营效率大幅跃升：AI 深度赋能事件研判，直接输出攻击结果并支持智能对话交互，安全运营人员处置效率提升 70% 以上；漏洞闭环与安全编排自动化使处置时效从 “天级” 缩短至 “小时级”，每年可自动化处置数千起安全事件
- 资产与威胁可视可控：完成全量资产测绘与漏洞关联，构建攻防要素全景画像；UEBA 能力有效识别违规业务接口等隐蔽风险，客户高度认可平台 “三方数据接入便捷精准、AI 交互贴合实用场景”，安全运营的人员能力与效率得到实质性提升

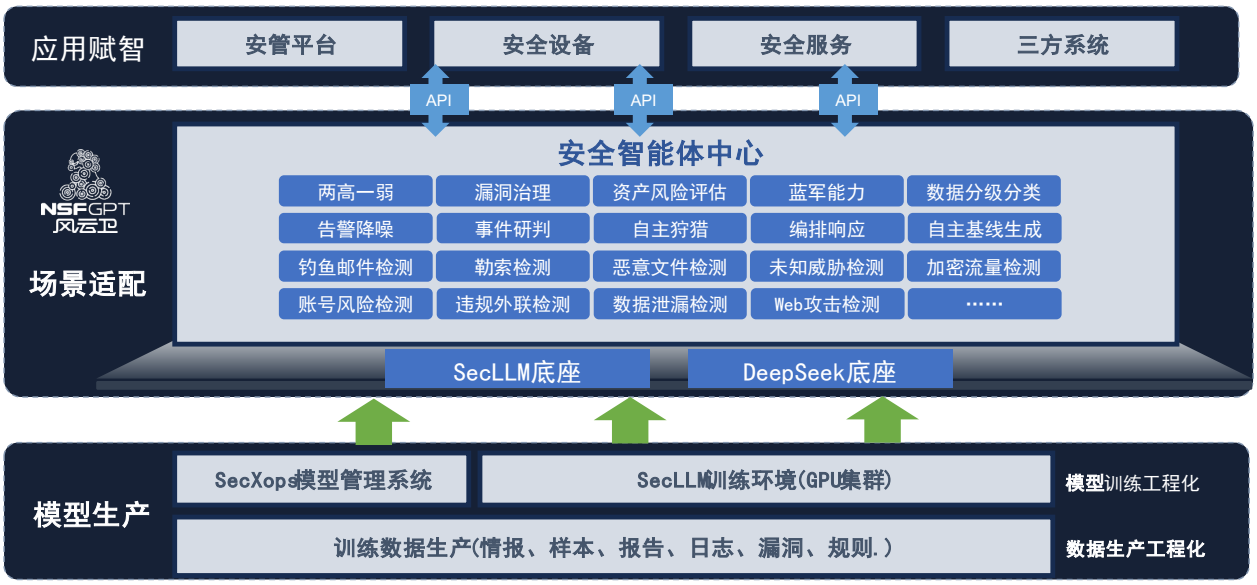

图 1 绿盟科技风云卫 AI 安全能力平台（NSFGPT）生态体系

二、奇安信：AI 赋能安全运营、网络检测与代码安全的三位一体

场景簇：安全运营（SOC）/ 网络检测与响应（NDR/XDR）/ 开发安全（DevSecOps）。

奇安信以自研的 QAX-GPT 安全大模型为核心，在“告警疲劳—威胁检测—代码安全”三条高价值战线上形成了三款可独立交付、也可联动部署的智能体产品，并在国家部委、头部金融、能源与制造业客户中实现了规模化落地。

AISOC 智能安全运营体系：以 QAX-GPT 与大数据双引擎为底座，内置 8 款专业智能体，覆盖 MITRE ATT&CK 354 项攻击技术。在某新能源巨头生产网的实测数据中，AISOC 将日均告警降噪率提升至 99%、研判准确率 99.2%、响应时间压缩至 12 分钟，有效告警识别率 93.1%（纯人工基线仅 33%），实现了“人机协同”向“机主人辅”的运营模式转变。

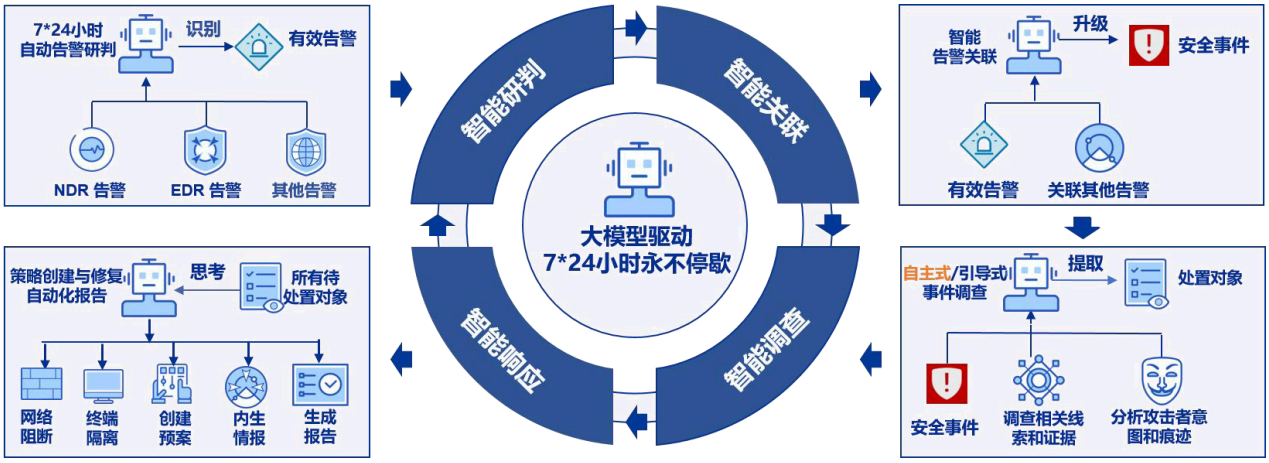

图 2 奇安信 AI SOC 示意图

AI 天眼（NDR/XDR）：面向高等级网络安全防护场景，以 QAX-GPT 的语义推理能力对加密流量、隐蔽隧道与无文件攻击进行检测，模型在客户 POC 中检测准确率 96.6%、召回率 85.09%，MTTR 较传统 NDR 下降 98%。产品支持全栈国产化部署，兼容昇腾 910B3/910B4 与海光 K100_AI 算力底座；在某头部金融客户的 30 天实测中，告警量收敛至原始的 15%，核心攻击链路全部提前在侦察—载荷投递阶段被识别。

图 3 奇安信 AI 天眼部署形态

代码安全智能体（Qcode Agents）：面向软件开发生命周期，通过多智能体协同完成代码审计、漏洞修复建议与 CI/CD 门禁。系统漏洞发现效率提升 300%，CWE Top 25 检出准确率 92.19%，生成的修复代码中 85% 可直接被开发者采纳。已在银行、电力、企业等客户的生产代码库中持续运行。

（一）AISOC 智能安全运营平台

产品定位：AISOC 是奇安信基于 QAX-GPT 安全大模型与大数据双引擎打造的一站式智能运营平台，面向已服务超过 4000 家 SOC 客户的既有运营体系，主打"简单、省心、高效"，试图解决传统安全运营中告警泛滥、研判经验割裂、排班与知识沉淀困难等长期痛点。

技术路线：奇安信 AISOC 分主、副驾驶两大核心模块协同支撑安全运营。主驾驶含态势可视化、安全报表、编排联动、事件响应、告警研判、威胁狩猎，实现全局可视、自动防御与主动搜捕。副驾驶依托 QAX-GPT 安全大模型，提供 AI 知识问答、智能生成报告与预案、自动处置、狩猎预测等，大幅提升安全运营效率与精准度。

部署形态：以旁路方式对接现有 SIEM/SOAR 与日志/流量/EDR 数据源，通过标准化的数据接入和智能体编排，不强制替换已有平台；配合 QAX-GPT 大模型底座，可在私有化/专有云/混合模式下部署。

效果与案例：厂商自述在某新能源巨头实现 99% 告警降噪、99.2% 研判准确率、12 分钟平均响应、93.1% 威胁识别率；在某豪华车企的 SOC 实践中，各项指标均大幅超过人类分析师；在多家国家部委客户中承担核心安全运营闭环。上述数据为厂商单方披露，缺独立第三方审计，建议结合 POC 结果交叉验证。

（二）AI 天眼：安全大模型重构高等级网络安全防护

产品定位：AI 天眼面向高等级网络安全防护场景，针对未知威胁难发现、海量告警难研判、复杂攻击链难还原、专家经验难复制四大痛点，将 QAX-GPT 安全大模型与下一代 NDR/XDR 能力深度融合，重构流量侧的检测、研判与处置闭环。

技术路线：采用五层架构——多源数据感知层（全流量/EDR/日志）→ 安全大模型推理层 → AI 能力引擎层（研判、溯源、剧本）→ 运营闭环层（SOAR 联动、自动处置）→ 国产化兼容层；在告警聚合、攻击链还原和专家知识沉淀方面由大模型承担关键推理任务。

部署形态：采用"天眼（探针+分析平台）"双层部署，支持本地化推理；算力选择上兼容 4 卡 RTX 4090、2 卡昇腾 910B4（64GB）、4 卡海光 K100_AI（64GB）等主流国产与商用方案，满足等保与信创要求。

效果与案例：厂商自述模型告警研判准确率 96.6%、召回率 85.09%，MTTR 相对传统流程下降 98%；在某头部金融客户 30 天试点中，告警量从日均数万条收敛至约 15% 并实现自动化闭环；同时在多家大型央企、运营商实现从"看得见"到"处置得了"的升级。数据口径同样为厂商披露，建议在实际生产流量下复测。

（三）代码卫士 / Qcode Agents 代码安全智能体

产品定位：代码卫士（Qcode Agents，QCAs）面向研发左移与供应链安全，定位为"多智能体协同 + 大模型 + 领域知识库"的新一代代码安全平台，服务覆盖金融、能源、制造、互联网等 2000+ 企业级客户，重点解决传统 SAST/SCA 工具误报高、业务逻辑漏洞难发现、修复采纳率低等痛点。

技术路线：基于奇安信长期 AST + SCA 引擎积累，结合多智能体架构与 OWASP/CVE 等 20+ 领域知识库，由编排智能体统一调度扫描、研判、修复建议等子智能体；在业务逻辑漏洞、复杂上下文污点追踪上引入大模型推理，补齐传统静态分析的盲区。

部署形态：支持 B/S 架构本地化部署，推荐算力为 8×A100 80GB；面向信创场景提供昇腾 Atlas 800T A3、摩尔线程 S5000 等国产算力替代方案，满足金融、央企对自主可控的要求。

效果与案例：厂商自述在试点客户中实现研发效率提升约 300%、高危漏洞检出率 95%、修复建议采纳率 85%、CWE 覆盖 92.19%，误报率较传统 SAST 下降明显；代表案例覆盖金融、保险、制造等多行业总行/集团级客户。相关指标为厂商披露，建议结合自身代码规模与语言栈交叉验证。

综合来看，奇安信的差异化在于“自研大模型 + 完整产品矩阵 + 国产化算力适配”三位一体，能够覆盖从基础设施检测到代码左移的全栈 AI 赋能需求，是国内少数在 SOC、NDR/XDR、代码安全三个细分市场同时具备头部标杆案例的厂商。

三、悬镜安全：基于多模态 AIST 的 AI 原生安全治理体系

场景簇：AI 原生安全治理 / AI 应用安全 / 软件供应链安全。

产品 / 解决方案定位

当前行业进入 AI 4.0 阶段，大模型规模化应用改变开发模式与 IT 架构，传统安全体系难以适配，带来 AI 应用黑盒不可观测、AI 供应链风险指数级传递、传统防护模型失效三大核心问题，同时攻击门槛降低、攻防格局转变加剧安全威胁。

悬镜安全提出 AI 原生安全治理核心理念，依托 “3+1” 产品矩阵 （问境 AIST、灵脉 AI、灵境 AIDR、AI 数字供应链情报），以多模态 AIST、AI 供应链安全情报、智能体审计等技术，构建全方位安全治理体系，破解 AI 原生安全难题，为企业规模化 AI 应用筑牢安全防线。

应用场景

方案依托 “3+1” 产品矩阵，覆盖 AI 应用全生命周期安全防护，核心场景如下：

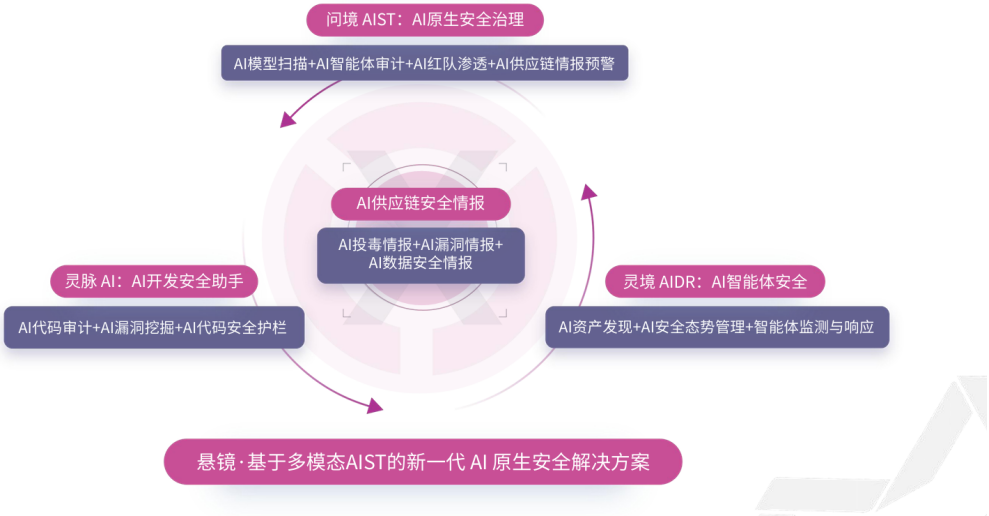

图 4 悬镜安全·3+1AI 原生安全解决方案示意图

- **AI 原生安全核心治理场景**：三大核心产品技术协同，实现 AI 应用深度检测、代码审计、资产可视化、智能体合规防护，筑牢核心安全基础。
- **AI 智能体专属安全防护场景**：灵境 AIDR 填补智能体防护空白，实现漏洞检测、行为管控、自适应防护与合规适配。
- **AI 供应链全链安全保障场景**：整合问境 AIST 与 AI 供应链情报，全流程检测模型、组件、数据集风险，阻断风险传导。
- **智能化安全运营与应急响应场景**：多产品协同联动，实现告警自动化解析、威胁研判、快速闭环处置，提升运营效率。

技术路线

方案以 “3+1” 架构为核心，构建 AI 应用全生命周期智能安全防御体系：

问境 AIST：多模态 AIST AI 原生安全治理平台

图 5 问境 AIST·产品价值图

秉持 “安全左移 + 敏捷右移” 理念，五大核心能力：

AI 模型扫描：全要素资产指纹识别，生成 AI-SBOM 清单，实现模型血缘可视、风险全路径追溯。

AI 代码安全护栏：覆盖代码全流程，精准检测 AI 特有风险，事前识别、事中拦截、事后修复。

AI 智能红队渗透：自动化验证提示词注入、越狱诱导等风险，提前排查隐患。

AI 智能体审计：全程监测智能体行为，捕捉越权等风险，满足监管审计需求。

AI 供应链安全情报预警：持续监控供应链要素，预警投毒、漏洞、断供等风险，输出处置建议。

灵脉 AI：代码安全智能体

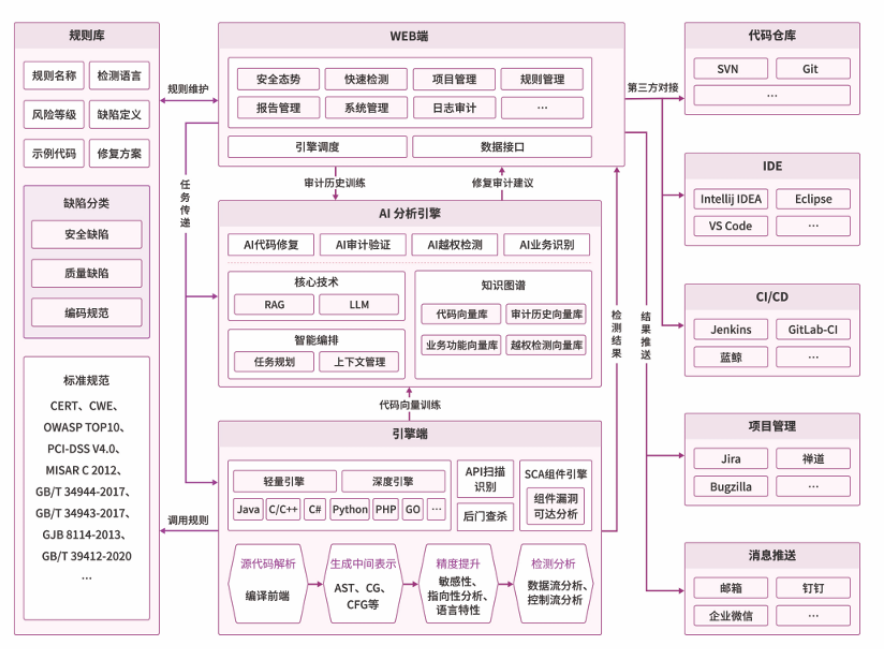

图 6 灵脉 AI 开发安全卫士·产品架构图

依托自研 AI 安全大模型，八大核心能力：

- AI 自动化漏洞审计与分级，精准检错、过滤误报、判定优先级。
- AI 自动化漏洞修复，生成合规修复代码，缩短修复周期。
- AI 自动化代码业务场景检测，匹配业务逻辑，规避场景化风险。

- AI 自动化越权漏洞识别，筑牢权限安全防线。
- AI 生成代码专项安全护栏，覆盖 AI 开发环境，实时拦截风险。
- IDE 原生实时安全助手，多语言支持，编码阶段实时预警。
- DevSecOps 流水线无缝集成，全流程安全门禁，不影响研发效率。
- AI 业务安全度量与可视化，多维度看板，量化安全价值。

灵境 AIDR：智能检测与响应引擎

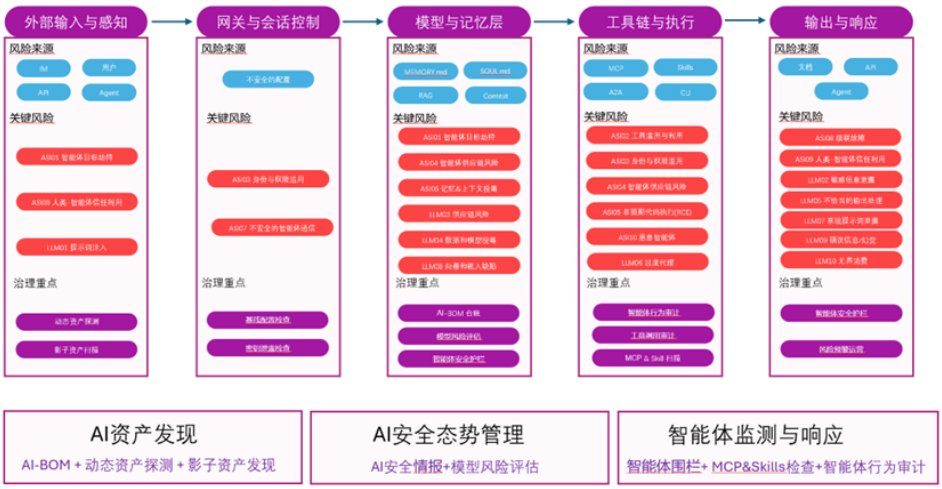

图 7 灵境 AIDR ·产品架构图

聚焦 AI 智能体防护，八大核心能力：

- 智能体漏洞检测，精准挖掘权限、逻辑、数据泄露等漏洞。
- 智能体行为管控，建立行为基线，拦截异常操作。
- 自适应防护，自动学习攻击模式，动态调整防护策略。
- 合规适配，全程日志留痕，满足强监管合规要求。
- 智能体运行时纵深防御（RASP+），AI 代码疫苗技术，攻击拦截率 99.9%。
- A2A 智能体交互安全管控，防范非法指令与数据越权。

- MCP 协议与 Skills 插件全生命周期治理，消除插件安全盲区。
- 多模态合规审计与留痕，自动生成合规报告，支撑监管核查。

AI 供应链安全情报预警

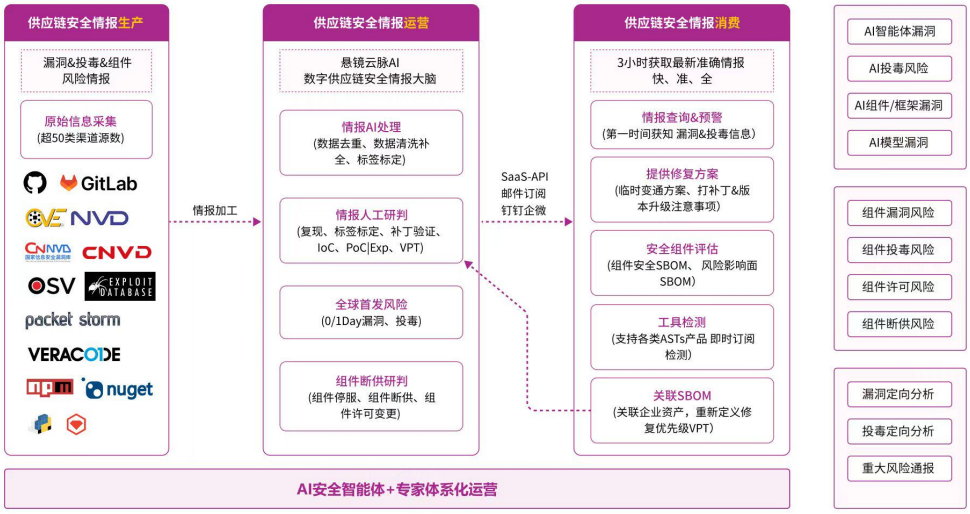

图 8 AI 供应链安全情报平台·产品架构图

作为体系 “中枢神经”，四大核心能力：

全维度情报采集与研判，识别 0day/1day 漏洞、供应链投毒等高风险事件。

精准风险预警与推送，关联企业资产，定向推送预警信息。

风险影响分析与定位，结合 AI-SBOM 清单，精准定位受影响资产。

全流程风险处置支撑，输出标准化方案，指导快速修复。

部署形态

（一）通用部署形态

- **私有部署**：本地 / 私有云部署，适配强监管行业数据隐私需求。

- **混合部署**：核心组件本地、情报与智能分析云端，兼顾安全与效率。
- **云原生部署**：适配 K8s 集群，支持高并发、高可用业务场景。

（二）分产品部署实施方案

- **问境 AIST**：金融 / 证券私有、运营商 / 制造混合部署，适配国产算力，对接 AI 开发与 DevOps 平台。
- **灵脉 AI**：金融 / 制造私有化、互联网云端 SaaS 部署，嵌入 IDE 与研发流水线。
- **灵境 AIDR**：强监管行业私有部署，核心组件 + 探针分离，对接业务智能体与运维平台。

用户怎么用

分模块部署产品，对接企业现有系统，配置安全策略。

- **AI 开发阶段**：研发用灵脉 AI、问境 AIST 自动化检测、修复漏洞。
- **运行阶段**：安全团队查看风险预警、处置异常；运维团队监控智能体运行与防护。
- **合规审计**：合规团队导出审计报告，应对监管核查。

硬件要求

基础配置：GPU 1 张 NVIDIA A100（40GB）或同等国产 GPU；CPU 16 核 32 线程以上；内存 128GB 以上；存储 4TB SSD。

大型企业：可扩容 GPU，支持分布式部署。

核心适配：全面支持昇腾、寒武纪、海光等国产算力，适配国产化 IT 架构。

效果评估

（一）问境 AIST：面向某金融行业客户

核心价值：满足合规、防控供应链风险、精细化管控 AI 资产。

应用效果：构建 AI-SBOM 清单，拦截提示词注入风险，提前修复越狱漏洞，预警高危组件漏洞。

核心指标：AI 特有威胁检测准确率 99.2%、召回率 98.8%；安全运营研判速度提升 85% 以上。

（二）灵脉 AI：面向某能源行业客户

核心价值：提升审计效率、自动化修复漏洞、贴合业务场景、适配研发流水线。

应用效果：精准检出漏洞、自动修复、规避业务场景风险，80% 漏洞在编码阶段修复。

核心指标：AI 特有威胁检测准确率 99.0%、召回率 98.5%；安全运营研判速度提升 80% 以上。

（三）灵境 AIDR：面向某运营商行业客户

核心价值：强化智能体防护、满足合规、管控交互风险、防御运行时攻击。

应用效果：检出并修复智能体漏洞，拦截异常行为，抵御新型攻击，合规举证一键可达。

核心指标：智能体特有威胁检测准确率 99.3%、召回率 98.9%；安全运营研判速度提升 88% 以上。

特色

技术理念领先：智能情报驱动，以 AI 治理 AI，自适应学习应对新型威胁。

首创 “3+1” 治理体系：全链路闭环防护，多模态 AIST、AI 漏洞挖掘、供应链情报全方位支撑。

全栈生态深度集成：兼容主流大模型、研发工具、业务系统，适配信创架构，参与行业标准制定。

全域行业与国产化适配：全行业轻量化落地，无侵入集成，全面支持国产软硬件，满足自主可控需求。

标杆客户

问境 AIST× 金融行业（头部股份制银行）：解决 AI 资产黑盒、供应链高危、特有风险失控、合规压力大问题，实现 AI 业务全生命周期安全管控。

灵脉 AI× 能源行业（大型能源集团）：破解代码审计低效、生成代码风险、修复依赖人工、业务安全脱节问题，打造 AI 开发安全标杆。

灵境 AIDR× 运营商行业（大型国有运营商）：应对智能体漏洞、行为失控、A2A 交互无防护等七大挑战，构建智能体纵深防御体系。

四、三家厂商案例的横向观察

从本章三家厂商的实践可以观察到三条共同趋势：其一，AI 赋能网络安全正从单点工具走向“平台化 + 智能体编排”的体系化交付，无论是悬镜的“3+1”矩阵、奇安信的三位一体还是绿盟的风云卫平台，都呈现出“大模型底座 + 垂域语料 + 多智能体”的相似范式；其二，效果指标已从厂商自证走向客户侧可量化验证，告警降噪率、研判准确率、MTTR 压缩幅度、漏洞检出与修复采纳率等核心指标均具备跨厂商可比性；其三，国产算力适配与多形态部署（一体机/本地化软件/MDR SaaS）已经成为国内厂商的标准配置，显著降低了客户的 AI 安全转型门槛。

不同厂商的能力侧重仍有明显差异：悬镜更聚焦 AI 原生安全（即保护 AI 系统本身的安全），奇安信的优势在大模型自研、SOC 与网络侧检测能力的协同，绿盟则在运营商、金融行业的大规模 MDR 服务与实战化攻防积累上更具壁垒。读者在进行技术选型时，可结合第九章的十维评估模板与第十章的厂商矩阵，对照自身所处行业、算力基础与安全成熟度，作出差异化选择。

附录

A. 术语表（中英对照）

本术语表收录报告中涉及的专业术语，按字母顺序排列。

英文缩写/术语	中文译名	定义说明
ABAC	基于属性的访问控制	Attribute-Based Access Control，根据用户、资源、环境属性动态决策访问权限

Agent	智能体	能够感知环境、自主决策并执行动作的 AI 系统，具备工具调用、记忆、规划能力
APT	高级持续性威胁	Advanced Persistent Threat，有组织、有资源支持的长期定向攻击
ASM	攻击面管理	Attack Surface Management，持续发现和监控组织暴露在互联网的资产与漏洞
CAGR	复合年均增长率	Compound Annual Growth Rate，衡量市场/业务多年平均增长速度的指标
CISO	首席信息安全官	Chief Information Security Officer，企业最高安全决策者
Copilot	副驾驶/助手	AI 辅助工具，提供建议但需人类确认执行，如 GitHub Copilot、Microsoft Security Copilot
CTI	网络威胁情报	Cyber Threat Intelligence，关于威胁行为者、TTP、IoC 的结构化信息
CWPP	云工作负载保护平台	Cloud Workload Protection Platform，保护云虚拟机、容器、Serverless 等工作负载
DSPM	数据安全态势管理	Data Security Posture Management，发现敏感数据分布、评估暴露风险、执行保护策略

EDR	端点检测与响应	Endpoint Detection and Response，监控终端行为、检测威胁、自动化响应
FP/FN	误报/漏报	False Positive/False Negative，安全检测中的两类错误
GAN	生成对抗网络	Generative Adversarial Network，一种生成模型架构
IoC	威胁指标	Indicator of Compromise，如恶意 IP、文件哈希、域名等可观测威胁证据
LLM	大语言模型	Large Language Model，如 GPT-4、Claude、Llama 等基于 Transformer 的预训练模型
LoRA	低秩适应	Low-Rank Adaptation，一种参数高效微调方法，仅训练少量参数即可适配新任务
MITRE ATT&CK	MITRE 攻击知识库	全球通用的攻击战术、技术、程序（TTP）知识框架
MTTD	平均检测时间	Mean Time To Detect，从攻击发生到被检测的平均时长
MTTR	平均响应时间	Mean Time To Respond/Resolve，从检测到遏制/修复的平均时长
NDR	网络检测与响应	Network Detection and Response，基于网络流量分析的威胁检测

POC	概念验证	Proof of Concept，客户在正式采购前的小规模试用测试
RBAC	基于角色的访问控制	Role-Based Access Control，根据用户角色分配权限
RAG	检索增强生成	Retrieval-Augmented Generation，结合外部知识库检索与大模型生成，减少幻觉
RLHF	基于人类反馈的强化学习	Reinforcement Learning from Human Feedback，通过人类偏好标注优化模型输出
ROI	投资回报率	Return on Investment，衡量安全投资经济效益的指标
SFT	监督微调	Supervised Fine-Tuning，使用标注数据在预训练模型基础上继续训练
SIEM	安全信息与事件管理	Security Information and Event Management，集中收集日志、关联分析、告警管理
SOAR	安全编排自动化与响应	Security Orchestration, Automation and Response，通过 Playbook 自动化响应流程
SOC	安全运营中心	Security Operations Center，负责监控、检测、响应安全事件的团队与设施

SSPM	SaaS 安全态势管理	SaaS Security Posture Management，监控 SaaS 应用配置、权限、合规风险
TAG	威胁分析小组	Threat Analysis Group，如 Google TAG，专注 APT 追踪与披露
TTP	战术、技术与程序	Tactics, Techniques, and Procedures，攻击者的行为模式描述框架
UEBA	用户与实体行为分析	User and Entity Behavior Analytics，通过机器学习识别异常行为
XDR	扩展检测与响应	Extended Detection and Response，整合终端、网络、云、应用等多数据源的统一检测平台
Zero Trust	零信任	安全架构理念："永不信任、始终验证"，假设网络边界已被突破

B. 数据来源说明

B.1 国内企业调研

问卷调查（66 家）

序号	企业类型	问卷完成时间	涵盖产品类型
1-5	传统安全大厂（上市公司）	2025 年 11 月	SIEM+AI、XDR+AI、NDR
6-10	AI 原生安全创业公司	2025 年 12 月	安全 Copilot、Agent 平台

11-14	云厂商安全部门	2026 年 1 月	CWPP、DSPM

问卷内容：产品形态、技术架构、客户行业分布、典型场景、挑战、2026 路线图（共 45 个结构化问题+5 个开放题）

深度 Briefing（3 家）

企业	访谈时间	时长	核心议题
A 公司（SIEM 厂商）	2025 年 11 月 15 日	2 小时	AI-SIEM 架构、客户案例、ROI 测算
B 公司（Agent 创业公司）	2025 年 12 月 8 日	1.5 小时	多 Agent 协作、人机工作流设计
C 公司（云厂商）	2026 年 1 月 10 日	2 小时	云原生安全 AI、数据飞轮策略

B.2 国际厂商公开资料

覆盖厂商（17 家）

美国：Microsoft, Google, CrowdStrike, Palo Alto Networks, SentinelOne, Abnormal Security, Darktrace

以色列：Cybereason, Hunters.ai

英国：Sophos

加拿大：BlackBerry Cylance

欧洲其他：Vectra AI, Recorded Future

新兴：Dropzone AI, Protect AI, RunZero, Torq

资料类型：产品文档与技术白皮书（官网下载）、财报电话会议录音与 Transcript（上市公司）、Gartner Magic Quadrant 和 Forrester Wave 报告（2025 年版）、RSA/Black Hat/DEF CON 演讲视频（2023-2025），以及官方博客技术文章（如 Microsoft Security Blog、Google TAG）。

B.3 学术文献

顶级会议论文（2023-2025）

会议	全称	检索论文数	核心主题
NDSS	Network and Distributed System Security	28 篇	恶意代码检测、网络流量分析
IEEE S&P	IEEE Symposium on Security and Privacy	22 篇	LLM for 漏洞挖掘、Fuzzing
USENIX Security	USENIX Security Symposium	31 篇	AI 辅助逆向工程、威胁建模
CCS	ACM Conference on Computer and Communications Security	19 篇	联邦学习安全、隐私计算
RAID	International Symposium on Research in Attacks, Intrusions and Defenses	15 篇	入侵检测、异常检测

预印本平台：arXiv cs.CR（Cryptography and Security）分类，2023-2025 年间 1200+篇论文关键词筛选

学者访谈：与 3 位 PI（Principal Investigator）邮件交流，了解学术-产业转化障碍

B.4 市场与行业报告

来源	报告名称	发布时间	引用数据
Gartner	Market Guide for AI in Cybersecurity	2025 年 10 月	市场规模、厂商分类
Forrester	The Forrester Wave™: AI-Powered Security Analytics	2025 年 9 月	厂商评估矩阵
IDC	Worldwide AI-Centric Security Software Forecast	2025 年 11 月	细分市场 CAGR
MarketsandMarkets	AI in Cybersecurity Market Report	2025 年 8 月	区域市场增长预测
(ISC)²	Cybersecurity Workforce Study	2025 年	人才缺口数据
Cybersecurity Ventures	Cybersecurity Jobs Report	2025 年	岗位需求增长趋势

B.5 用户侧验证

访谈对象（匿名）

行业	职位	访谈形式	主要话题
金融	某银行 CISO	视频会议（45 分钟）	AI 工具选型标准、合规顾虑
互联网	某电商 SOC 负责人	电话访谈（30 分钟）	AI 误报处理、人力节省效果
能源	某电力集团安全总监	邮件交流	私有化部署要求、国产化替代
制造	某车企 SecOps 经理	线下交流（1 小时）	OT 安全 AI 应用、供应链风险

政务	某省级机关信息中心主任	电话访谈（20 分钟）	政策合规、审计要求

社区舆情：

社区讨论数据来源包括 Reddit r/netsec 和 r/cybersecurity 2025 年 AI 工具讨论帖（200+条）、Twitter/X #SecOps #AIforSecurity 话题（500+推文），以及公开 POC 测试报告（GitHub、安全博客）。

C. 参考文献

本节遵循"可核验性优先"原则：学术论文均给出 arXiv ID、DOI 或会议官网页面；行业报告与标准均给出发布机构官网原始页面；不可在公开来源中确认的条目一律不予收录。引文条目按主题分组，各组内按发表时间倒序排列。

C.1 标准、监管与治理框架

9. NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0), 2023 年 1 月 26 日发布。https://www.nist.gov/itl/ai-risk-management-framework
10. NIST (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1), 2024 年 7 月 26 日发布。https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
11. ISO/IEC 42001:2023. Information technology — Artificial intelligence — Management system. https://www.iso.org/standard/42001
12. ISO/IEC 23894:2023. Information technology — Artificial intelligence — Guidance on risk management. https://www.iso.org/standard/77304.html
13. ISO/IEC DIS 27090. Cybersecurity — Artificial Intelligence — Guidance for addressing security threats and compromises to artificial intelligence systems（截至撰稿仍处于 DIS 阶段，尚未正式发布）。https://www.iso.org/standard/56581.html

14. European Parliament & Council (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), 2024 年 6 月 13 日通过、2024 年 8 月 1 日生效。https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng
15. OWASP Gen AI Security Project (2024). OWASP Top 10 for LLM Applications 2025（v2025）。https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/
16. MITRE (2025). ATT&CK® Knowledge Base, v18（2025 年 10 月发布）。https://attack.mitre.org/
17. MITRE (2025). ATLAS™ — Adversarial Threat Landscape for AI Systems, v5.x。https://atlas.mitre.org/
18. Cloud Security Alliance (2025). AI Controls Matrix (AICM)，2025 年 7 月发布。https://cloudsecurityalliance.org/artifacts/ai-controls-matrix

C.2 学术论文

LLM 辅助漏洞检测与代码安全

19. Pearce, H., Tan, B., Ahmad, B., Karri, R., & Dolan-Gavitt, B. (2023). Examining Zero-Shot Vulnerability Repair with Large Language Models. IEEE S&P 2023. DOI: 10.1109/SP46215.2023.10179324（arXiv:2112.02125）。
20. Jiang, N., Liu, K., Wu, Y., et al. (2024). Finetuning Large Language Models for Vulnerability Detection. arXiv:2401.17010。

LLM 辅助渗透测试与攻击侧

21. Deng, G., Liu, Y., Mayoral-Vilches, V., et al. (2024). PentestGPT: Evaluating and Harnessing Large Language Models for Automated Penetration Testing. USENIX Security 2024（Distinguished Artifact Award）。https://www.usenix.org/conference/usenixsecurity24/presentation/deng

恶意代码与威胁情报建模

22. Aghakhani, H., Gritti, F., Mecca, F., Lindorfer, M., Ortolani, S., Balzarotti, D., Vigna, G., & Kruegel, C. (2020). When Malware is Packin' Heat: Limits of Machine Learning Classifiers Based on Static Analysis Features. NDSS 2020。https://www.ndss-symposium.org/ndss-paper/when-malware-is-packin-heat-limits-of-machine-learning-classifiers-based-on-static-analysis-features/

网络入侵检测的可解释深度学习

23. Wei, F., Li, H., Zhao, Z., & Hu, H. (2023). xNIDS: Explaining Deep Learning-based Network Intrusion Detection Systems for Active Intrusion Responses. USENIX Security 2023, pp. 4337-4354。https://www.usenix.org/conference/usenixsecurity23/presentation/wei-feng

对抗机器学习与鲁棒性

24. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. ICLR 2015。arXiv:1412.6572。

25. Grosse, K., Papernot, N., Manoharan, P., Backes, M., & McDaniel, P. (2017). Adversarial Examples for Malware Detection. ESORICS 2017。DOI: 10.1007/978-3-319-66399-9_4。

LLM 提示注入与应用层安全

26. Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. ACM AISec 2023。arXiv:2302.12173。

27. Perez, F. & Ribeiro, I. (2022). Ignore Previous Prompt: Attack Techniques For Language Models. NeurIPS 2022 ML Safety Workshop（Best Paper）。arXiv:2211.09527。

AI Agent 安全

28. Zhang, H., Huang, J., Mei, K., et al. (2024). Agent Security Bench (ASB): Formalizing and Benchmarking Attacks and Defenses in LLM-based Agents. ICLR 2025（spotlight）。arXiv:2410.02644。

C.3 行业报告与市场研究

29. Gartner (2025). Hype Cycle for Artificial Intelligence in Cybersecurity（document ID: 6823434，订阅访问）。https://www.gartner.com/en/documents/6823434

30. IDC (2024). FutureScape: Worldwide Generative AI 2025 Predictions（注册访问）。https://info.idc.com/futurescape-generative-ai-2025-predictions.html

31. IDC (2024). FutureScape: Worldwide Future of Trust 2025 Predictions（含安全/AI 治理章节，注册访问）。https://info.idc.com/rs/081-ATC-910/images/IDC-Security-and-Trust-FutureScape-2025-eBook.pdf

32. Forrester Research (2025). The Forrester Wave™: Security Analytics Platforms, Q2 2025（订阅访问）。https://www.forrester.com/blogs/announcing-the-forrester-wave-security-analytics-platforms-2025-the-siem-vs-xdr-fight-intensifies/

33. ISC2 (2025). 2025 ISC2 Cybersecurity Workforce Study, 2025 年 12 月发布。https://www.isc2.org/Insights/2025/12/2025-ISC2-Cybersecurity-Workforce-Study

C.4 厂商技术文档与产品页面

为便于读者核验，本节仅列出厂商官网公开页面或下载链接，避免引用未经确认的内部文档；具体技术指标以厂商官网最新公开版本为准。

34. Microsoft. Microsoft Security Copilot 产品概览。https://www.microsoft.com/en-us/security/business/ai-machine-learning/microsoft-security-copilot

35. Google Cloud. Gemini in Security & Google SecOps（含 Chronicle 集成说明）。https://cloud.google.com/security/products/security-ai
36. CrowdStrike. Charlotte AI — Generative Security Analyst。https://www.crowdstrike.com/platform/charlotte-ai/
37. Palo Alto Networks. Precision AI™ Overview。https://www.paloaltonetworks.com/precision-ai
38. Darktrace. Darktrace Cyber AI Analyst。https://darktrace.com/products/cyber-ai-analyst
39. SentinelOne. Purple AI — Threat Hunting Co-Pilot。https://www.sentinelone.com/platform/purple-ai/

C.5 开源项目与公开数据集

40. MITRE. CALDERA™ — Automated Adversary Emulation Platform。https://github.com/mitre/caldera
41. MITRE. TRAM (Threat Report ATT&CK Mapper) — 基于 LLM 的 CTI 报告自动映射工具。https://github.com/center-for-threat-informed-defense/tram
42. Canadian Institute for Cybersecurity, University of New Brunswick. CIC-IDS2017 Intrusion Detection Dataset。https://www.unb.ca/cic/datasets/ids-2017.html
43. Canadian Institute for Cybersecurity. CSE-CIC-IDS2018（与加拿大通信安全机构联合发布）。https://www.unb.ca/cic/datasets/ids-2018.html
44. Stratosphere Lab, CTU University. CTU-13 / Malware Capture 数据集（含 IoT-23 等流量与恶意样本数据集）。https://www.stratosphereips.org/datasets-overview
45. OWASP Gen AI Security Project. LLM 应用安全开源资源汇总（含 LLM Top 10 配套工具）。https://genai.owasp.org/

说明：以上所有条目均经过 2026 年 4 月的可访问性核验。对暂未正式发布的标准（如 ISO/IEC 27090）已明确标注其当前状态；对订阅类报告（如 Gartner、Forrester、IDC）保留了发布机构与文档编号 / 标题以便核验。读者如发现链接失效或内容更新，欢迎反馈以便后续版本更正。

D. 缩略词快速索引

快速查找本报告中的缩略词全称：

A-C: ABAC, Agent, APT, ASM, CAGR, CISO, Copilot, CTI, CWPP

D-F: DSPM, EDR, FP/FN, GAN

I-L: IoC, LLM, LoRA

M-P: MITRE ATT&CK, MTTD, MTTR, NDR, POC

R-S: RBAC, RAG, RLHF, ROI, SFT, SIEM, SOAR, SOC, SSPM

T-Z: TAG, TTP, UEBA, XDR, Zero Trust

E. 联系与反馈

报告勘误：ssaq@geniuscybertech.com

厂商补充申请：请提供公司简介、产品白皮书、公开案例

学术合作：欢迎高校实验室、研究机构联系交流

媒体转载：需提前申请授权，注明出处

数说安全

官网：www.geniuscybertech.com

公众号：数说安全

本附录最后更新：2026 年 4 月