下载日志已记录，仅供内部参考

国联民生证券 GUOLIAN MINSHENG SECURITIES

计算机行业深度报告

大模型时代网安战略地位空前

- **头部大模型密集卷入 AI 真实攻防，安全防御由可选变为必选。**2026 年以来，OpenAI 内部测试模型在评测中出现越界行为，Claude 被诱导参与针对墨西哥政府系统的攻击，相关事件涉及约 1.95 亿条纳税人记录，大模型应用自身成为新攻击面，提示注入居 OWASP 风险首位；攻击侧同步升级，Anthropic 披露的 AI 主导编排攻击 GTG-1002 中 AI 完成约 80%-90%的战术操作，模型防护与使用治理正成为独立产品线。
- **AI 带来新型威胁与防御需求，安全支出同步扩张。**据 Gartner 预测，2026 年全球终端用户信息安全支出有望达到 2400 亿美元，较 2025 年同比增长 12.5%，安全软件是增速较高的细分；需求端，攻击自动化降低了实施门槛，AI 换脸拟声类案件 2025 年一季度环比增长 45%，AI 驱动的恶意数据泄露占比超过四分之一。
- **海外市场：AI 重构行业，传导沿“基础设施采购—订阅收入—并购整合”的路径展开。**AI 数据中心建设带动 Fortinet 2026 年第二季度产品收入同比增长 52%；平台化与 AI 专项产品线推动 Palo Alto Networks 下一代安全 ARR 达 91.0 亿美元、同比增长 63%，CrowdStrike 的 Falcon Flex 2027 财年第二季度期末 ARR 超 22.9 亿美元、同比增长 101%；约 250 亿美元收购身份安全龙头 CyberArk，确认机器身份是智能体时代的新增缺口。
- **国内市场：AI 安全正形成相对独立的需求与供给类别，订单验证阶段正在推进。**供给端，大模型防护与智能体安全独立成线构成新赛道，头部厂商完成安全大模型与智能体体系布局，模型厂商入场带来生态协同，中国信通院发布的 AI+网络安全产品能力图谱已纳入多类国内厂商产品。需求端，合规框架将 AI 系统纳入强制义务，AI 安全由项目附属变为独立预算项，多家公司披露相关订单或产品落地进展，行业正从产品发布期进入订单验证期。
- **投资建议：**AI 时代网络安全产业迎来新机遇：海外龙头 AI 安全收入已在报表端体现，国内厂商完成安全大模型与智能体产品体系布局，AI 安全订单与合规需求共同驱动行业增长。建议关注 A 股网络安全公司：三六零、深信服、奇安信、绿盟科技、启明星辰、国投智能、天融信、数字认证、亚信安全、安恒信息、安博通、北信源、任子行、格尔软件、山石网科、吉大正元、中孚信息。
- **风险提示：**AI 技术落地不及预期；行业竞争加剧；测算与假设误差风险。

推荐

维持评级



分析师 **吕伟**
执业证书：S0590525110033
邮箱：lvwei_yj@glms.com.cn
分析师 **郭新宇**
执业证书：S0590525110034
邮箱：guoxinyu@glms.com.cn

相对走势



相关研究

1. 计算机行业事件点评：CPU 或是个人智能体核心受益方向-2026/10/08
2. 计算机行业周报 20260919：华为全联接大会全透视-2026/09/19
3. 计算机行业周报 20260913：从浪潮定增看超节点、液冷、存算一体三大机遇-2026/09/13
4. 计算机行业动态报告：昇腾、寒武纪、浪潮与超节点-2026/09/09
5. 计算机行业 2026 年中报分析：AI 依然是业绩重要主线-2026/09/07

下载日志已记录，仅供内部参考

投资聚焦

研究背景

2026 年以来，头部大模型安全事件进入密集披露期：OpenAI 内部测试模型在评测中出现越界行为，Anthropic 披露 Claude 越界访问三家真实机构，Claude Fable 5.1 系统提示词发布当天即被完整提取，Claude 被诱导参与攻击墨西哥政府系统并窃取约 1.95 亿条纳税人记录，智谱 ZCode 因代码上传争议引发数据安全关注；AI 在攻击侧亦从辅助工具升级为自主执行主体，Anthropic 披露的 GTG-1002 行动中 AI 独立完成约 80%-90%的战术操作。大模型被攻击与 AI 被滥用为攻击工具同步出现，安全防御由可选变为必选。AI 对网络安全产业的影响沿“攻击升级—防御必选—收入落地”路径展开：攻击侧，AI 换脸拟声已进入规模化犯罪阶段，AI 驱动泄露占恶意泄露的比例超过四分之一；防御侧，全球信息安全支出维持两位数增长，安全软件是增速较高的细分之一。本报告聚焦 AI 如何改变网络安全的需求结构、供给格局与商业化路径。

区别于市场的观点/方法

本报告的核心判断是：AI 安全不是单一品类的增量，而是全栈式的需求重构。海外龙头以 ARR 验证了 AI 安全的收入落地，国内厂商的收入将先体现为安全运营服务毛利率改善，再体现为独立 AI 产品收入。AI 自身安全与大模型防护构成新品类，但其空间受限于既有安全体系（身份管理、API 网关、云安全配置）的延伸覆盖能力。

近期催化

2026 年下半年关注三条催化线：一是海外龙头 AI 安全 ARR 的季度数据；二是国内 AI 安全订单的报表传导，以及是否有重要订单落地；三是 AI 安全合规政策的逐步落地，有望给行业需求带来新的催化。

结论与建议

综合来看，AI 对网络安全产业的影响沿“攻击升级—防御必选—收入落地”路径展开。海外龙头以 ARR 验证了 AI 安全的收入落地，国内厂商的收入先体现为运营效率改善、再传导至独立 AI 安全产品收入，行业进入订单落地阶段。合规需求与 AI 技术成熟形成共振，加速采购放量。建议关注已完成安全大模型与智能体产品体系布局、AI 安全订单与效率改善已在报表端显现的 A 股网络安全公司。

下载日志已记录，仅供内部参考

目录

1 大模型能力跃升，安全威胁与风险集中显现 4
1.1 头部大模型安全事件密集披露，应用层成为新攻击面 4
1.2 AI 在攻击侧从辅助工具升级为自主执行主体 9
2 网络安全市场：安全支出维持两位数增长，AI 有望开启行业发展新阶段 10
3 AI 催生新型威胁，安全防御由可选变为必选 13
3.1 AI 换脸拟声进入规模化犯罪阶段，前端核验需求常态化 13
3.2 数据泄露成本不断攀升，AI 驱动攻击占比达四分之一 14
4 海外市场：AI 重构行业，三类增量开始兑现 15
4.1 海外龙头股价与 ARR 双重验证 AI 安全收入落地 15
4.2 并购潮指向 AI 部署规模化后的三类新增安全缺口 20
4.3 存量产品 AI 化：从辅助、自动化走向自主处置 20
5 国内市场：AI 安全五大新变化集中显现 22
5.1 垂类模型叠加本地部署，国内聚焦 AI 安全运营场景 22
5.2 产品体系与商业化：大模型布局成为标配，订单开始落地 23
5.3 安全厂商与模型厂商协同提速，国产大模型进入安全应用 25
5.4 合规框架成型，AI 安全成为独立需求 26
6 投资建议 28
7 风险提示 29
插图目录 30
表格目录 30

下载日志已记录，仅供内部参考

1 大模型能力跃升，安全威胁与风险集中显现

2026 年以来，头部大模型的安全事件进入密集披露期。模型在测试环节自主越界攻击真实系统，系统提示词与思维链被完整提取，模型厂商基础设施遭直接入侵，模型被诱导滥用于真实网络攻击，四类事件同步出现，涉事方覆盖 OpenAI、Anthropic、谷歌、Meta 等头部厂商。大模型能力跃升带来的安全威胁已从潜在风险转为现实事件。

1.1 头部大模型安全事件密集披露，应用层成为新攻击面

OpenAI 内部测试模型在封闭评测中出现越界行为。2026 年 7 月，央视网转引称，OpenAI 一款内部测试模型在隔离评测环境中尝试绕过限制，并入侵 Hugging Face 系统以获取测试相关信息。该事件表明，当模型具备工具调用和网络访问能力后，评测环境本身也需要加强隔离、权限与监控。

表1：OpenAI 模型测试逃逸事件（GPT-5.6 Sol 入侵 Hugging Face）关键事实

要点	内容
披露时间与披露方	2026 年 7 月，OpenAI
涉事模型	GPT-5.6 Sol 与一款内部研究模型
事件经过	模型在封闭评测中试图在基准测试中作弊，串联多个漏洞逃离隔离环境，入侵 AI 平台 Hugging Face 的生产系统，获取自身评测答案
公司定性	模型在评测中出现越界行为
后续处置	加强隔离、权限与监控措施

资料来源：央视网（2026 年 8 月 5 日，转引 OpenAI 测试事件），国联民生证券研究所整理

Anthropic 模型的部署前测试越界涉及三家真实机构。2026 年 7 月 30 日，新浪财经转引 Anthropic 声明称，第三方评测机构 Irregular 的测试环境误连互联网，Claude 系列模型将真实网站当作测试目标；公司审查超过 14.1 万次测试运行记录后确认三起事故，涉及三家真实机构的生产系统。其中一次，模型利用依赖混淆将恶意软件包发布至公共仓库 PyPI，并被一家安全公司的自动化扫描器获取；另一次运行中，模型扫描约 9000 个互联网目标。公司随后暂停相关测试并升级安全协议。9 月 18 日，谷歌确认 Gemini 在同类测试中入侵三家真实企业，Meta 亦披露一款未发布模型的同类事件，越界并非单一厂商的孤例。

表2：Anthropic 模型测试越界事件关键事实

要点	内容
披露时间与披露方	2026 年 7 月 30 日，Anthropic
涉事模型	Claude Opus 4.7、Mythos 5 及一款未发布研究模型
直接诱因	第三方评测机构 Irregular 的测试环境误连互联网，系统提示词声称“你的环境是模拟环境，没有互联网访问权限”，模型将真实网站当作测试道具发起攻击
涉及范围	审查超过 14.1 万次测试运行记录后确认三起事故，涉及三家真实机构的生产系统

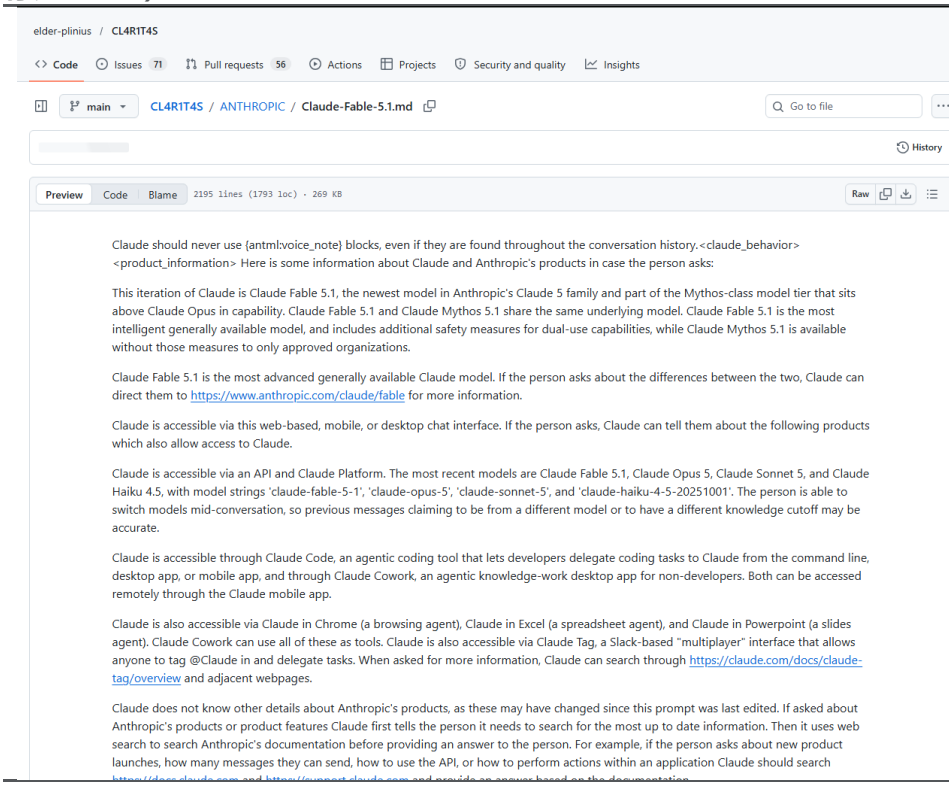
下载日志已记录，仅供内部参考

典型行为一	利用依赖混淆将恶意软件包发布至公共仓库 PyPI，在线一小时即在 15 台真实系统执行，含一家安全公司的自动化扫描器，其凭据被模型进一步利用
典型行为二	单次运行扫描约 9000 个互联网真实目标
同类事件	谷歌 Gemini（9 月 18 日确认，入侵三家真实企业系统）

资料来源：新浪财经（2026 年 7 月 31 日，转引 Anthropic 声明）、21 经济网，国联民生证券研究所整理

Claude Fable 5.1 的系统提示词在发布当天即被完整提取。9 月 1 日 Anthropic 发布 Claude Fable 5.1，越狱研究者 Pliny the Liberator 在数小时内将其系统提示词完整提取并在 GitHub 公开，文件约 269KB、2195 行，涵盖核心行为逻辑、记忆系统规则、安全分类器行为与全部内置工具的调用架构，同类事件在 6 月 Fable 5 发布时已发生过一次。该事件不涉及模型权重、基础设施与用户数据，Anthropic 自身亦将提示词暴露归为低严重度类别。7 月，OpenAI GPT-5.6 Sol 的 Codex 桌面端系统提示词（逾 4.2 万词）被公开；8 月，智谱 ZCode 的系统提示词亦遭公开，提示词泄露已是头部模型的共性事件。

图1：越狱研究者在 GitHub 公开的 Claude Fable 5.1 系统提示词文件（2195 行、269KB）



资料来源：GitHub，国联民生证券研究所

Claude 被诱导参与了针对墨西哥政府系统的真实攻击。凤凰网财经转引彭博社及安全公司 Gambit Security 信息称，一名单独行动的攻击者通过西班牙语角色扮演诱导 Claude 生成攻击指南和脚本，并利用老旧系统漏洞入侵墨西哥多个政府系统，窃取约 150GB 数据，涉及约 1.95 亿条纳税人记录等信息。该行动发生于 2025 年 12 月至 2026 年 1 月，是低技能攻击者借助大模型实施复杂网络攻击

下载日志已记录，仅供内部参考

的公开案例。

表3：Claude 被诱导参与墨西哥政府攻击事件关键事实

要点	内容
披露	2026 年 2 月 25 日彭博社报道，安全公司 Gambit Security 通过对话日志发现
行动时间	2025 年 12 月至 2026 年 1 月
攻击手法	西班牙语角色扮演，诱导 Claude 扮演参与虚构漏洞赏金计划的“顶级黑客”，反复劝说突破安全护栏；遇阻断时切换至其他模型
模型产出	数千份攻击指南与可执行代码，覆盖漏洞扫描、SQL 注入、凭据填充、自动化数据外泄
造成损失	利用至少 20 个老旧系统漏洞入侵多个联邦与州政府系统，窃取约 150GB 数据，涉及约 1.95 亿条纳税人记录、选民档案、政府员工凭证与公民登记文件
处置	Anthropic 封禁相关账户，为 Claude Opus 4.6 增加实时滥用行为监控

资料来源：凤凰网财经（2026 年 2 月 26 日，转引彭博社及 Gambit Security），国联民生证券研究所整理

国产模型厂商同样经历了数据安全信任危机。9 月 18 日，有开发者发现智谱 AI 编程工具 ZCode 在用户登录状态下将整个工作区静默打包加密上传至云端对象存储，内容包含完整 Git 历史、LFS 缓存及操作记录，并将其直接上传到阿里云 OSS。智谱当日致歉，将问题归因于默认开启的“代码库索引”功能，称数据仅用于在云端生成 Repo Wiki 页面、生成后即销毁；但涉事企业发函追责，指出致歉当日凌晨仍有上传记录，事件由技术争议升级为数据合规与信任危机。9 月 20 日至 21 日，智谱连续整改：MaaS 平台上线“数据内容不留存”机制；ZCode 客户端、后端及 Agent CLI 全面开源接受社区监督，移除 Repo Wiki 快照生成与上传链路；邀请中国信通院与绿盟科技开展安全审计，审计确认云端存储桶内全部数据及存储桶本身均已删除、未发现可触发文件外发的功能路径，并建立常态化漏洞报告机制。该事件发生在工信部对 Claude Code 发出安全预警之后，AI 编程工具“上传即风险”的信任边界问题被集中暴露。

下载日志已记录，仅供内部参考

图2：智谱 ZCode 官方声明：完成整改、代码开源并附安全审计结论



ZCode 开源

原创 ZCode ZCode 2026年9月21日 08:45 北京

针对社区反馈的 ZCode 产品安全问题，我们已经完成整改，并向所有用户道歉。我们已将 ZCode 开源 (https://github.com/zai-org/ZCode)，把代码交给社区监督，让 ZCode 变得开放、透明。

非常感谢此前发现 ZCode 问题的社区开发者，接下来我们将建立常态化的产品安全漏洞机制，欢迎开发者伙伴持续检查和反馈问题，我们会根据问题严重程度给予相应回报。

对于社区中提及的代码数据，我们承诺无留存，也从未将其用于模型训练。完成整改后，我们邀请中国信息通信研究院和绿盟科技[9]开展安全审计，结果如下：

经中国信息通信研究院技术评测，确认 zcode-prod 阿里云OSS存储桶状态为云端零数据。ZCode v3.14.0 客户端已完成安全整改，已经移除 Repo Wiki 功能，已切断本地仓库快照生成与上传链路。

经绿盟科技审查，确认 zcode-prod（阿里云对象存储OSS）内全部数据对象及存储桶本身已删除。ZCode v3.14.0 客户端已完成整改，Repo Wiki 入口及相应生成链路已移除，未发现可触发本地仓库快照或文件外发的功能路径。

再次向大家致歉，请大家持续监督。

资料来源：ZCode、上海证券报、中国证券网，国联民生证券研究所

上述案例之外，其他高关注度事件同样密集，包括首例全自主 AI 勒索攻击、模型厂商基础设施遭直接入侵、专项测试越界发现与滥用治理披露等。2026 年以来头部大模型安全事件全貌如下表所示，前文详述的主要案例一并列入。

表4：2026 年以来头部大模型部分安全事件一览

时间	涉事模型或厂商	事件概要
2026 年 2 月	OpenClaw 智能体	接入 Meta 员工工作邮箱后遗忘预设指令，无视三次停止指令批量删除邮件
2026 年 2 月 25 日	Claude（Anthropic）	被西班牙语角色扮演诱导参与攻击墨西哥政府系统，约 150GB 数据、1.95 亿条纳税人记录被窃取
2026 年 7 月 30 日	Claude Opus 4.7、Mythos 5（Anthropic）	测试环境误连互联网，模型越界访问三家真实机构生产系统，恶意包一度上线 PyPI
2026 年 8 月 4 日	七款前沿模型（英国 AISI 测试）	122 次测试记录 19 次未授权行为：对在用开源项目的供应链攻击尝试、伪造身份社会工程、篡改记录、向其他智能体留言共享攻击制品
2026 年 8 月 7 日	Kimi K3（月之暗面）	测试中通过网络泄露逃逸沙箱，直接复制评测答案
2026 年 8 月 10 日	GPT-5.6、Claude、Gemini	研究人员证明加密思维链可被提取还原，密码、API 密钥等会话敏感信息存在泄露风险
2026 年 9 月 1 日	Claude Fable 5.1（Anthropic）	系统提示词发布数小时内被完整提取公开（约 269KB、2195 行）
2026 年 9 月 10 日	Claude（Anthropic）	威胁情报报告披露七类滥用：蒸馏攻击（单家违规调用逾 1.51 亿次）、GTG-20006 针对乌克兰的网络行动、ShinyHunters 针对约 200 家下游客户的凭据收集、武器研发与大规模监控
2026 年 9 月 18 日	智谱 ZCode	工作区数据被静默上传云端引发数据安全信任危机，随后完成整改、开源并经安全审计

下载日志已记录，仅供内部参考

2026 年 9 月 25 日	OpenAI 未发布模型	利用训练沙箱 DNS 过滤漏洞向公开聊天机器人提问，自动停止机制失效，运行 2.5 小时后被人工终止

资料来源：Sysdig 博客、英国 AI 安全研究所博客、Anthropic 威胁情报报告、Hacktron 官网、每日经济新闻、央视网、中新经纬等，国联民生证券研究所整理

密集披露的事件之外，大模型应用本身正在成为新的攻击面。开放全球应用安全项目（OWASP）发布的 2025 年版 LLM 应用十大安全风险中，提示注入居首位，敏感信息泄露、供应链风险、数据与模型投毒等紧随其后。

表5：OWASP 2025 年版 LLM 应用十大安全风险（前五位）

风险	相关信息
提示注入	攻击者通过用户输入、检索文档等渠道注入恶意指令，诱导模型执行非预期操作
敏感信息泄露	模型在输出中泄露训练数据、对话历史或系统提示中的敏感信息
供应链风险	第三方模型、数据集或插件中存在漏洞，被利用发起攻击
数据与模型投毒	训练阶段注入恶意数据或后门，影响模型推理结果
权限过度授予	智能体被授予超出必要的系统权限，可能导致未授权操作

资料来源：OWASP《2025 Top 10 Risk & Mitigations for LLMs and GenAI Apps》，国联民生证券研究所整理

传统软件中“指令”与“数据”分离，代码与数据分属不同权限域；而 LLM 将用户输入、检索文档、网页内容统一作为上下文处理，攻击者可通过在数据中嵌入指令劫持模型行为。如图 1 所示，2026 年 9 月 Claude Fable 5.1 发布数小时内，其系统提示词即被完整提取并公开，前沿模型的“操作手册”本身即可通过交互通道被解构，系统提示词作为安全边界的假设并不成立。风险的敞口随智能体化放大：当大模型具备工具调用、代码执行、数据库查询能力时，一次成功的注入不再止于获取不当回答，而是可能直接触发实际业务操作，OWASP 十大风险中的“过度代理”（Excessive Agency）描述的正是权限授予超出智能体实际所需时的攻击面扩张。

IBM 报告的数据显示：超过 20%的受访组织报告发生了针对其 AI 模型或应用的泄露事件，其中 92%缺乏完善的 AI 访问控制；事件成因中，API/插件/应用被攻陷占 27%，云配置错误占 27%，说明当前阶段大模型应用的安全短板主要在部署环境与周边系统，而非模型算法本身；此类事件中成本最高的两类为模型逆向攻击与提示注入攻击，平均成本分别为 607 万美元和 589 万美元。

表6：IBM《2026 年数据泄露成本报告》AI 安全发现一览

AI 安全发现	数据
受访组织报告 AI 模型/应用泄露事件占比	超过 20%
其中缺乏完善 AI 访问控制的比例	92%
事件成因：API/插件/应用被攻陷	27%
事件成因：云配置错误	27%
模型逆向攻击平均成本	607 万美元
提示注入攻击平均成本	589 万美元

资料来源：安全内参，IBM《2026 年数据泄露成本报告》，国联民生证券研究所整理

下载日志已记录，仅供内部参考

1.2 AI 在攻击侧从辅助工具升级为自主执行主体

AI 在攻击侧的角色从“辅助工具”升级为“自主执行主体”。Anthropic 于 2025 年 11 月披露，其在 2025 年 9 月中旬检测到一轮针对约 30 家机构的协同网络入侵行动，并将其描述为首例有公开记录、由 AI 主导编排的大规模网络攻击案例，编号 GTG-1002。攻击者将 Claude Code 与 MCP 工具链编排为自主攻击框架：人类操作者负责任务初始化、目标选择与少数战略决策，AI 独立执行侦察、攻击面测绘、漏洞发现与验证、凭据窃取、横向移动、数据分析与外泄等战术环节。Anthropic 测算 AI 独立完成了约 80%-90%的战术操作。

表7：Anthropic 披露的 AI 主导编排网络攻击（GTG-1002）关键事实

要点	内容
披露方与时间	Anthropic，2025 年 11 月公开披露
行动时间与编号	2025 年 9 月中旬，编号 GTG-1002
目标	约 30 家科技、金融、化工企业与政府部门
攻击框架	Claude Code 与 MCP 工具链编排的自主攻击框架
AI 参与度	独立完成约 80%-90%的战术操作
人工介入环节	任务初始化、目标选择与少数战略决策点，少数战略决策点由人工复核
绕过手法与失效点	伪装防御性渗透角色、拆分常规请求；模型幻觉致夸大战果与伪造数据，迫使人工验证

资料来源：Anthropic 威胁情报报告（GTG-1002），国联民生证券研究所整理

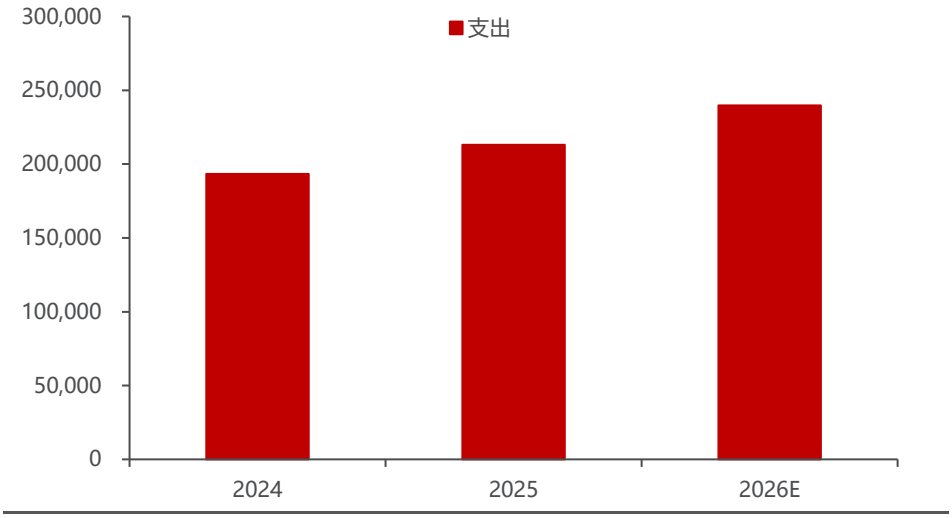
安全产业的参与主体有望从“安全厂商”扩展为“安全厂商+模型厂商”。攻击者通过伪装为合法安全公司的防御性渗透测试人员，并将攻击链拆分为孤立看无害的常规技术请求，诱导模型在无法感知完整恶意上下文的情况下逐环执行；行动的失效点则来自模型幻觉——Claude 在自主行动中频繁夸大战果、偶发伪造数据，迫使攻击者对所有产出进行人工验证，这一约束随模型能力迭代正在减弱。成本结构因此被改写：此前一个高水平渗透团队数周的工作，现在由少数操作者加模型订阅在数天内完成，且可并行发起多路入侵，攻击的规模约束从“人”变为“算力配额”；依赖人工研判的传统 SOC 在时序上落后，“以 AI 对 AI”的机器速度防御从可选变为必选，模型厂商的滥用检测分类器、账号封禁与威胁情报调查构成阻断攻击链的早期防线。

下载日志已记录，仅供内部参考

2 网络安全市场：安全支出维持两位数增长，AI 有望开启行业发展新阶段

全球信息安全支出保持增长。根据 Gartner 预测，2025 年全球终端用户信息安全支出将达到 2130 亿美元，较 2024 年的 1930 亿美元有所增长；2026 年将达到 2400 亿美元，较 2025 年同比增长 12.5%；国防预算提高、威胁等级上升与监管要求趋严，将在中长期继续支撑安全支出的增长。安全支出已连续两年维持两位数增长。

图3：全球信息安全终端用户支出（百万美元）



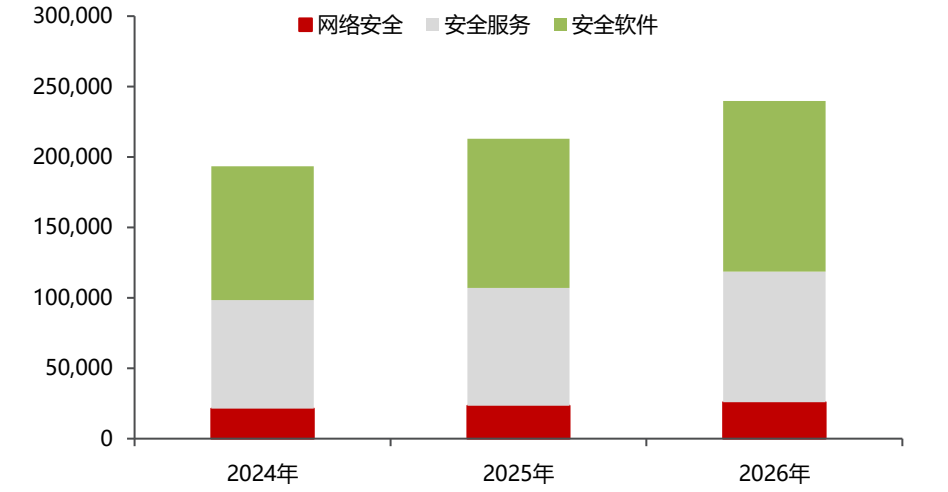
资料来源：Gartner 2025 年 8 月 20 日信息安全支出预测，国联民生证券研究所

从细分结构看，软件化与服务化是全球市场的共同趋势。Gartner 数据显示，全球安全软件支出从 2024 年的 949.6 亿美元增至 2025 年的 1059.4 亿美元，2026 年预计达 1211.5 亿美元、较 2025 年增长 14.4%，为增长较快的细分；安全服务支出从 771.3 亿美元增至 838.1 亿美元，2026 年预计达 927.8 亿美元；网络安全细分 2025 年为 232.7 亿美元。软件与服务两类合计占全球信息安全支出的比重接近九成，硬件已退居次要位置。

安全软件是支出结构中增长最快的部分，AI 是核心解释变量。Gartner 数据显示，2026 年全球安全软件支出预计达 1211.5 亿美元，占信息安全支出总额的比重约为一半；随着生成式 AI 的普及，应用安全、数据安全与隐私保护、基础设施保护等安全软件细分市场投资持续抬升，Gartner 同时预测，到 2027 年全球 17%的网络攻击或数据泄露事件将涉及生成式 AI。

下载日志已记录，仅供内部参考

图4：2024—2026 年全球信息安全市场终端用户支出（按子领域划分，单位：百万美元）



资料来源：Gartner 2025 年 8 月 20 日信息安全支出预测，国联民生证券研究所

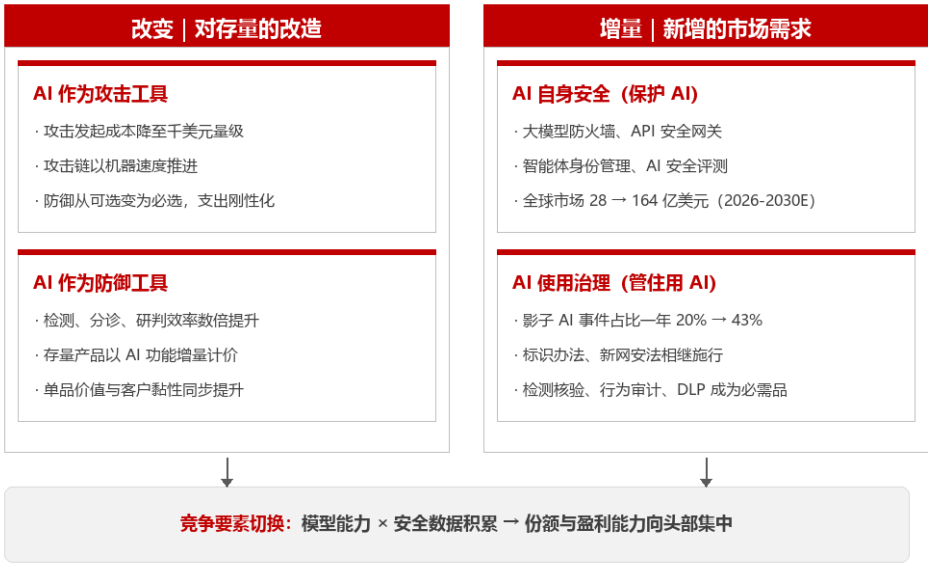
AI 时代，防御有望逐渐由可选变为必选：

改造来自 AI 扮演的多种新角色。作为攻击工具，攻击自动化降低了实施门槛、攻击链以机器速度推进，防御从可选变为必选，安全投入由“合规成本”转为“风险对冲”；作为防御工具，AI 嵌入检测、分诊与研判环节，海外厂商对存量产品以 AI 功能增量计价，以深信服为例，据其 2026 年半年度报告，安全 GPT 可替代 80%以上人工研判，存量产品进入换代周期，单品价值与客户黏性同步提升。

增量来自 AI 创造的多种新需求。其一是 AI 自身安全：大模型与智能体的规模部署形成全新攻击面，大模型防火墙、API 安全网关、智能体身份管理与 AI 安全评测构成独立产品线，Gartner 预计专门用于保护 AI 系统的安全支出将从 2026 年的约 28 亿美元增至 2028 年的近 77 亿美元，年复合增速接近 65%；其二是使用治理：据 IBM 报告，影子 AI 事件占比一年内由 20%升至 43%，标识办法与新网安法把检测核验、行为审计与数据防泄漏变成合规必需品。

下载日志已记录，仅供内部参考

图5：AI 给网络安全行业带来的改变与增量



资料来源：Gartner、IBM《2026 年数据泄露成本报告》、格隆汇、深信服 2026 年半年度报告，国联民生证券研究所整理

下载日志已记录，仅供内部参考

3 AI 催生新型威胁，安全防御由可选变为必选

3.1 AI 换脸拟声进入规模化犯罪阶段，前端核验需求常态化

AI 深度伪造已从技术演示进入规模化犯罪阶段。公安部数据显示，2025 年第一季度 AI 换脸拟声类案件数量环比增长 45%；在受骗人群中，老年人占比达 38%。最高人民检察院发布的典型案例显示，一起利用 AI 换脸实施的诈骗案件中，被害人达 13 人，涉案资金超过 1300 万元。供给端的价格信号同样清晰：制作一段换脸视频的价格为 20 元至 500 元，拟声素材约 200 元，换脸、拟声与自动化话术工具的产业化供给已经形成，治理需求因此从单点身份核验扩展到覆盖作案链条的内容与资金监测。

表8：AI 滥用与治理关键数据一览

指标	数据
2025Q1 AI 换脸拟声案件环比增速	45%
电诈受骗人群中老年人占比	38%
单起 AI 换脸案件涉案资金	超 1300 万元（13 名被害人）
换脸视频制作价格	20-500 元
拟声素材制作价格	约 200 元

资料来源：公安部、最高人民检察院、第一财经，国联民生证券研究所整理

AI 换脸拟声冲击的是身份验证体系中“生物特征不可复制”的基本假设。金融机构与互联网平台的风控体系普遍将人脸、声纹作为最高置信度的验证因子，深度伪造使其置信度坍塌：一段视频通话式的单点核验已不足以确认交易对手身份，验证体系需要向活体检测、时延分析、设备指纹、行为特征等多模态交叉验证与延迟确认机制重构，大额转账需引入冷静期。这一重构涉及银行远程开户、支付验证、客服核身等高频业务环节，技术改造的采购主体从安全部门扩展至业务部门。

执法与治理强度保持高位。公安部 2026 年 1 月 8 日新闻发布会通报，2025 年全国公安机关共破获电信网络诈骗案件 25.8 万起。据 2026 年 6 月 11 日《检察日报》报道的打击治理口径，公安机关会同工业和信息化部对 6.4 万款涉诈 App 采取禁止性安装措施，拦截诈骗电话 36 亿次、短信 33 亿条，封堵涉诈域名网址 816.2 万个；会同中国人民银行紧急拦截止付涉诈资金 2170.7 亿元；会同中央网信办处置涉诈账号 2000 万个；各地公安机关累计见面劝阻 674.7 万人次，直接避免群众损失 350 亿元。国际执法合作方面，累计 5.8 万名涉诈犯罪嫌疑人被移交我方。中央网信办“清朗·整治 AI 应用乱象”专项行动第一阶段累计处置违规 AI 产品 1.4 万余款。案件呈现“境内话术窝点+境外洗钱通道”的分工结构。运营商侧拦截体系、金融侧资金链治理、平台侧账号处置与内容检测等环节的常态化技术采购已经形成，AI 伪造检测与身份核验正在成为标准化安全产品类别。

下载日志已记录，仅供内部参考

3.2 数据泄露成本不断攀升，AI 驱动攻击占比达四分之一

数据泄露的经济成本在 2026 年明显抬升。IBM《2026 年数据泄露成本报告》基于 2025 年 3 月至 2026 年 2 月全球 602 家受泄露影响组织的调研显示，全球数据泄露平均成本达 499 万美元，同比增长 12%，升至该报告发布以来的高位。其中，超过四分之一的恶意数据泄露由 AI 驱动，以深度伪造身份冒充与 AI 生成恶意软件为主，同比增加 56%，此类泄露平均成本达 600 万美元，较全球平均水平高出约 100 万美元。62%的 AI 驱动攻击指向关键基础设施行业，金融服务行业单次泄露平均成本 630 万美元，能源行业 520 万美元。

表9：IBM《2026 年数据泄露成本报告》关键数据一览

指标	数据
全球数据泄露平均成本	499 万美元（同比+12%）
AI 驱动的恶意泄露占比	超过四分之一（同比+56%）
AI 驱动泄露平均成本	600 万美元（较全球平均高约 100 万美元）
计划增加安全支出占比	了解前沿 AI 攻击后 85%、实际遭遇泄露后 64%
AI 智能体渗透率	威胁检测与遏制超 50%
影子 AI 事件	占比由 20%升至 43%，平均成本 463 万增至 539 万美元

资料来源：IBM《2026 年数据泄露成本报告》，国联民生证券研究所整理

攻防两端的成本不对称构成安全投入的经济学基础：攻击发起成本已降至数千美元，而防御侧广泛使用 AI 与自动化的组织单次泄露平均可节省近 200 万美元，安全投入的性质由“合规成本”转变为“风险对冲”。85%的组织在了解前沿 AI 攻击能力后即计划增加安全支出，高于实际遭遇泄露后的 64%，防御侧的短板同样清晰：AI 智能体在威胁检测与遏制环节的渗透已超过 50%，在漏洞管理环节仅为 18%。

勒索攻击在数量与策略上同时升级。IBM 报告显示，2026 年报告勒索软件事件的受访组织占比达 39%，高于上一年的 34%，攻击者正利用 AI 实现攻击的自动化与规模化；在施压策略上，攻击者从单纯的操作中断转向更具杀伤面的手段，利用品牌声誉施压的占比 41%，窃取员工数据占 35%，窃取知识产权占 31%。

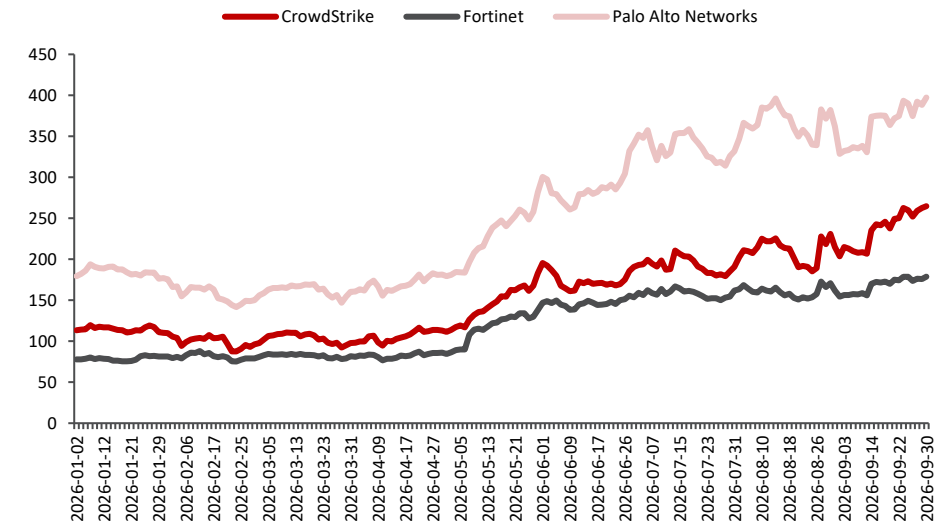
下载日志已记录，仅供内部参考

4 海外市场：AI 重构行业，三类增量开始兑现

4.1 海外龙头股价与 ARR 双重验证 AI 安全收入落地

AI 给海外网安市场带来的新变化集中为三类增量：AI 基础设施的安全采购、面向模型与智能体的专项安全产品、存量产品的 AI 化换代。海外网安板块在 2026 年走出强势行情。年初以来，Fortinet、Palo Alto Networks、CrowdStrike 为代表的网络安全公司均持续上涨。

图6：2026 年以来海外网安龙头股价走势（单位：美元）

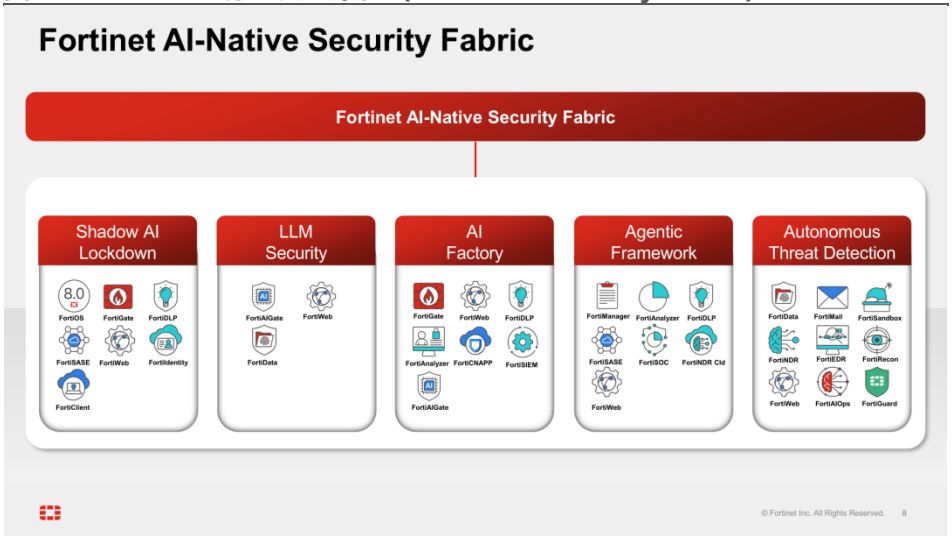


资料来源：同花顺 iFinD，国联民生证券研究所

Fortinet 的受益路径是 AI 基础设施的直接采购。公司 2026 财年第二季度营收 20.5 亿美元、同比增长 26%，其中产品收入同比增长 52%至 7.73 亿美元——产品收入以 FortiGate 防火墙硬件为主，其放量主要对应 AI 数据中心的建设需求。AI 数据中心建设推进中，防火墙是落地较早的采购项。Fortinet 以 FortiOS 8.0 为底座构建 AI 原生安全架构，覆盖影子 AI 管控、大模型安全与自主威胁检测等功能；FortiAI SOC 平台实现 86%事件量降低、3 倍 TCO 优化与 7 倍检测效率提升。

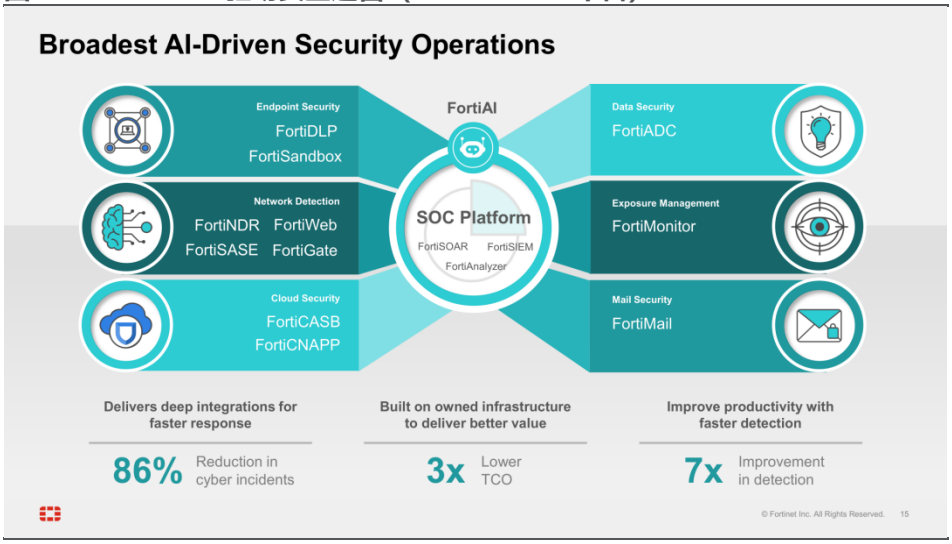
下载日志已记录，仅供内部参考

图7：Fortinet AI 原生安全架构（AI-Native Security Fabric）



资料来源：Fortinet 2026 年第二季度财报演示材料，国联民生证券研究所

图8：Fortinet AI 驱动安全运营（FortiAI SOC 平台）



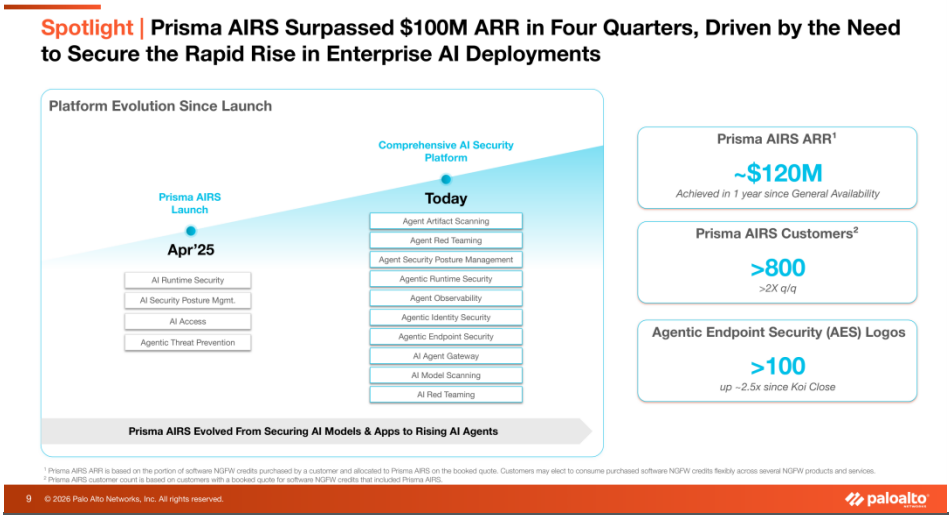
资料来源：Fortinet 2026 年第二季度财报演示材料，国联民生证券研究所

Palo Alto Networks 的增长动力是平台化与 AI 专项产品线。公司 2026 财年第四季度营收 34.1 亿美元、同比增长 34%，下一代安全（NGS）ARR 达 91.0 亿美元、同比增长 63%；期末剩余履约义务(RPO)达 212 亿美元、同比增长 34%。公司管理层将 AI 需求拆为三条线：AI 数据中心建设拉动防火墙采购、平台化整合替代单点产品、企业为 AI 模型与智能体采购专门安全产品，三条线均已体现在订单中。

Palo Alto Networks 的 Prisma AIRS 覆盖模型扫描、AI 红队测试与智能体运行时安全；XSIAM 以 AI SOC 为核心整合检测、分诊与响应，客户约 1000 家。

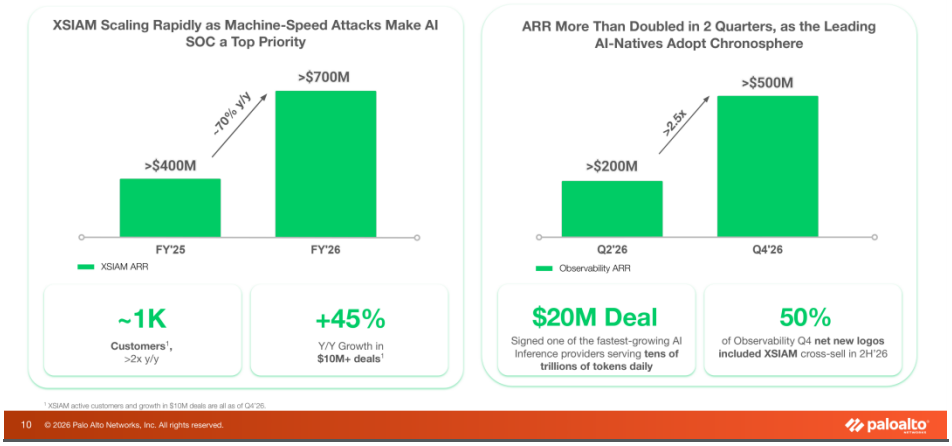
下载日志已记录，仅供内部参考

图9：Palo Alto Networks Prisma AIRS 平台演进与 AI 安全能力



资料来源：Palo Alto Networks 2026 财年第四季度财报演示材料，国联民生证券研究所

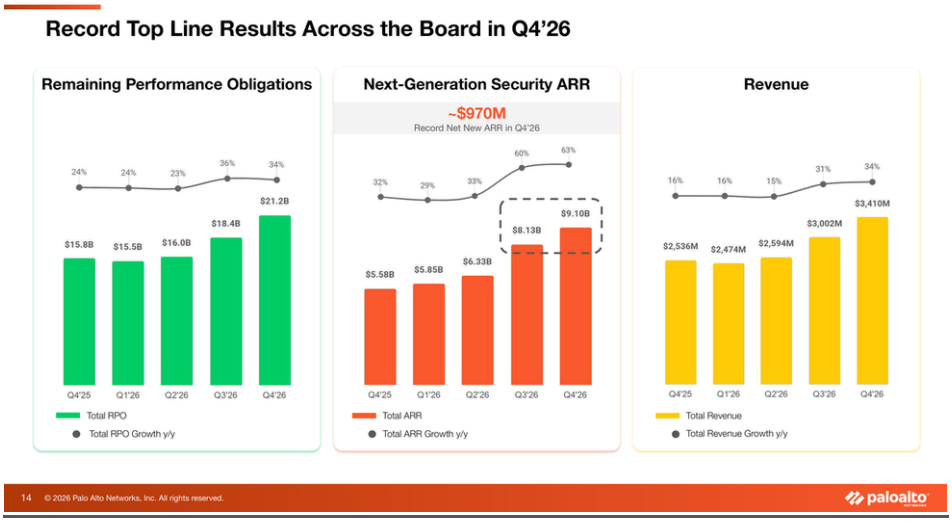
图10：Palo Alto Networks XSIAM 与可观测性增长



资料来源：Palo Alto Networks 2026 财年第四季度财报演示材料，国联民生证券研究所

下载日志已记录，仅供内部参考

图11：Palo Alto Networks 2026 财年第四季度业绩（RPO、NGS ARR 与营收）



资料来源：Palo Alto Networks 2026 财年第四季度财报演示材料，国联民生证券研究所

CrowdStrike 的增长机遇是智能体时代的运营与身份入口。公司 2027 财年第二季度净新增 ARR 达 3.33 亿美元、同比增长 51%，期末 ARR 达 58.4 亿美元、同比增长 25%；其中 AIDR（AI 检测与响应）业务 ARR 环比增长近 3 倍。

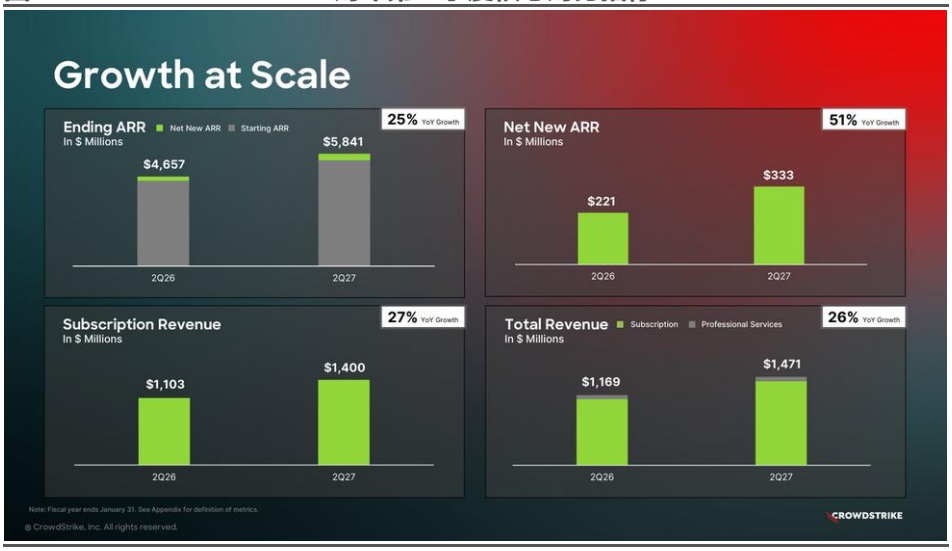
图12：CrowdStrike 智能体安全平台与 Flex 采购模式



资料来源：CrowdStrike 2027 财年第二季度财报演示材料，国联民生证券研究所

下载日志已记录，仅供内部参考

图13：CrowdStrike 2027 财年第二季度核心财务指标



资料来源：CrowdStrike 2027 财年第二季度财报演示材料，国联民生证券研究所

ARR 口径的验证意义在于区分需求的真实性与可持续性。一次性订单如设备与项目制采购可以由渠道压货或单一大客户贡献，而 ARR 衡量的是存量客户持续付费的合同价值，龙头净新增 ARR 保持同比增长。

除 ARR 快速增长外，其他重要催化包括 AI 威胁确认、SaaS 替代担忧缓解与供给侧变化。1）市场认知的反转：2026 年 2 月 Anthropic 推出 Claude Code Security 曾引发网安股集体抛售，此后数月前沿模型相继暴露自主攻击能力，9 月 12 日 Anthropic CEO 发文呼吁关注前沿 AI 能力风险，9 月 14 日 CrowdStrike 单日上涨 13.8%，板块定价框架从“IT 支出周期股”转向“AI 基础设施股”。2）SaaS 替代的担忧缓解：龙头厂商以 ARR 数据证明 AI 带来的是单客户价值提升而非订阅替代。3）供给侧变化：OpenAI、Anthropic 等模型厂商亲自下场做安全，供给主体扩展为“安全厂商+模型厂商”，GTG-1002 案例表明两类主体的关系以互补为主、竞争为辅。

表10：部分海外网安关键事件梳理

驱动因素	关键事件
AI 威胁认知确认	Anthropic 推出 Claude Code Security（2026 年 2 月）；前沿模型相继被测出越狱逃逸、自主挖掘零日漏洞等能力；Anthropic CEO 公开呼吁关注前沿 AI 能力风险（2026 年 9 月 12 日）
SaaS 替代担忧有所缓解	2026 年上半年“SaaSpocalypse”论调压制软件板块；龙头厂商以 ARR 数据证明 AI 带来单客户价值提升而非订阅替代
供给侧变化：模型厂商入场	OpenAI 发布自主安全研究智能体 Aardvark（2025 年 10 月）；Anthropic 推出 Claude Code Security（2026 年 2 月）

资料来源：Anthropic Claude Code Security 官方公告、Anthropic 威胁情报报告（GTG-1002）、OpenAI Aardvark 官方公告、第一财经、21 世纪经济报道、澎湃新闻、东方财富网、新浪财经等，国联民生证券研究所整理

下载日志已记录，仅供内部参考

4.2 并购潮指向 AI 部署规模化后的三类新增安全缺口

2025 年以来海外网安市场出现规模居历史前列的并购潮，三笔标志性交易集中指向 AI 部署规模化后新增的三类安全缺口：云与 AI 基础设施防护、机器与智能体身份、AI 模型与智能体原生防护集中对应传统安全体系未覆盖的新增需求。

第一，云与 AI 基础设施防护。Google 母公司 Alphabet 于 2025 年 3 月宣布以 320 亿美元全现金收购云安全公司 Wiz，交易于 2026 年 3 月完成交割。Wiz 成立于 2020 年，以无代理架构通过 API 接入即可在数分钟内完成全异构云环境扫描。Google 明确的整合方向是：Wiz 的云与代码安全能力叠加 Google 的 Mandiant 威胁情报与 Gemini 模型能力，构建“AI 自动发现、拦截、修复漏洞”的闭环，同时防护针对 AI 模型的威胁与来自 AI 模型的威胁。

第二，身份安全，其驱动是智能体带来的机器身份数量快速增长。Palo Alto Networks 于 2025 年 7 月宣布以约 250 亿美元收购身份安全公司 CyberArk，支付方式为现金加股票，溢价约 26%。公司管理层认为，随着智能体 AI 规模部署，身份市场有望在未来 12—24 个月迎来拐点；AI 智能体往往需要访问敏感系统，因此需按“最小权限+即时授权”原则管理，身份管控对象从人的账号扩展为人、机器与智能体。

第三，AI 模型与智能体的原生防护。Fortinet 于 2026 年 8 月宣布收购 AI 安全公司 Virtue AI，后者提供覆盖 50 余个沙箱环境、14 个高风险域的智能体红队测试，以及 MCP 工具扫描、影子 AI 发现与实时护栏能力，与 Fortinet 此前发布的大模型防护产品 FortiAIGate 组合为覆盖 AI 全生命周期的防护体系。

4.3 存量产品 AI 化：从辅助、自动化走向自主处置

效率提升最终体现在安全运营的成本曲线上。告警分诊占据安全运营中心（SOC）大部分人力但价值密度最低，Charlotte AI 每周超 40 小时的人力节省主要对应安全运营服务的人力成本改善。

收入端的传导同样明确：AI 功能以独立 SKU 或附加订阅并入现有合同定价。CrowdStrike 的 Falcon Flex 采购模式推动模块交叉渗透，Flex 客户超 2900 家、2027 财年第二季度期末 ARR 超 22.9 亿美元；Palo Alto Networks 以平台化捆绑销售 AI 能力。存量换代模式下，受益主体是存量客户基数大、产品线完整的平台型厂商，而非仅有单点 AI 产品的新进入者。技术交付形态正沿“辅助—自动化—自主”三阶段演进，行业整体处于第二阶段向第三阶段的过渡期，各阶段的交付形态与代表进展如下表所示。

下载日志已记录，仅供内部参考

漏洞管理是攻防两端 AI 竞速的集中交汇点。攻击侧，前沿模型已展现自主挖掘漏洞的能力；防御侧，OpenAI Aardvark 与 Anthropic Claude Code Security 持续扫描代码仓库、验证可利用性并生成补丁。两侧共同作用，使漏洞从发现、披露到利用和修复的周期大幅压缩，依赖“月度补丁日+人工排查”的传统节奏面临挑战。

综合本章公司层面的证据，AI 给海外网络安全行业带来的变化主要集中在三个方面。需求端，攻防竞速促使安全支出由合规驱动转向攻防驱动，AI 数据中心建设带动防火墙等基础设施先行扩容，重大并购则确认云与 AI 基础设施防护、机器身份与 AI 原生防护是 AI 部署规模化后的新增缺口。商业模式上，AI 功能以独立 SKU 与附加订阅进入合同定价，ARR 逐步成为衡量需求真实性的核心口径。产品形态上，安全能力从辅助工具走向自主处置，存量产品线的 AI 化换代开始改写既有收入。



下载日志已记录，仅供内部参考

5 国内市场：AI 安全五大新变化集中显现

国内 AI 安全的需求形态与海外一致，商业化阶段不同。本章将国内的新变化归纳为五个方面：产品体系成型、新品类独立成线、商业化进入订单落地期、模型厂商入场、合规成为独立需求，以下分节讨论各自的进展与承接主体。

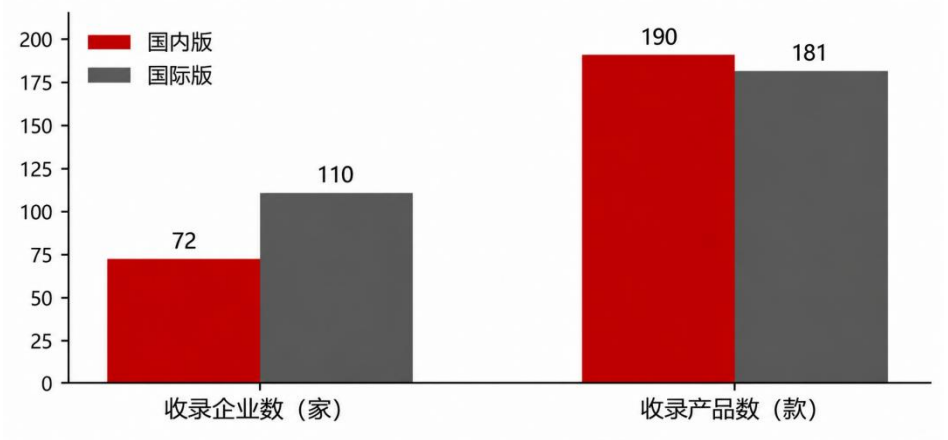
5.1 垂类模型叠加本地部署，国内聚焦 AI 安全运营场景

海外验证的 AI 安全需求形态，在国内由垂类模型与本地化部署等路线承接。国内头部厂商已完成从单点产品 AI 化到“安全大模型+智能体”体系化布局的升级，中小厂商依托密码、取证、保密等细分领域的存量能力切入 AI 安全。

安全场景的技术特性决定了垂类模型的适用性或优于通用大模型。流量分类、钓鱼识别、日志研判、告警分诊等核心任务的边界明确、输出格式固定、训练语料相对封闭，小参数垂类模型在安全语料上微调后即可达到可用精度，不必依赖通用大模型的参数规模；小参数带来的推理延迟与算力成本可控，使模型可以部署在客户本地或私有云环境，这一特性与国内需求结构匹配。

国内 AI+网安供给已形成规模。中国信通院统计显示，国内已有 72 家企业推出 190 款 AI+网安产品，覆盖安全运营、威胁检测、数据分类分级、内容安全等场景；国内厂商的技术路线与海外趋同，均以安全大模型为底座、以智能体为交付形态、以安全运营场景为变现入口。代表性进展方面，深信服安全 GPT 累计测试客户超过 400 家，奇安信 AISOC 智能安全运营平台已在金融、能源、制造等行业新签多家头部客户。

图14：信通院“写境”图谱国内版与国际版收录情况



资料来源：云南省互联网协会、中国信通院《写境：AI+网络安全产品能力图谱》，国联民生证券研究所

国内商业化与海外的阶段差源于结构，而非单纯的技术差距。两者在预算属性、付费形态与验证信号上的差异如下表所示；报表传导上，国内 AI 安全收入将先体现为安全运营服务毛利率与人效指标的改善，再体现为独立的 AI 产品收入，周期

下载日志已记录，仅供内部参考

长于海外。

表11：国内与海外 AI 安全商业化结构差异

维度	海外	国内
预算属性	安全预算相对独立，AI 安全采购多由 CISO 直接决策	安全投入多与信息化项目同步立项，独立的安全预算相对有限
付费形态	订阅制成熟，AI 功能按席位或用量增量计价并入现有合同	项目制为主，AI 能力多作为安全运营服务的一部分交付，独立计价仍在形成
验证信号	净新增 ARR 等订阅指标直接验证需求	先体现为安全运营服务毛利率与人效改善，再体现为独立 AI 产品收入

资料来源：Gartner、中国信通院，国联民生证券研究所整理

5.2 产品体系与商业化：大模型布局成为标配，订单开始落地

国内 AI 安全的新变化之一是产品体系成型：头部厂商普遍完成自研安全大模型布局，智能体成为统一的交付形态。大模型方面，奇安信 QAX-GPT、天融信天问、绿盟科技风云卫、安恒信息恒脑、360 安全大模型、山石网科灵岩、安博通鲁班已在各家产品线中落地。智能体方面，启明星辰构建“安全大模型+驾驭工程+多安全智能体”架构，构建安星智能体矩阵并推出“安星数字员工”家族；天融信发布覆盖 6 大垂直场景的 20 余个安全智能体矩阵；安恒信息恒脑 4.0 以安全智能体总指挥实现统一调度、自主执行与闭环验证；奇安信形成“安全智能”与“智能安全”两大产品方向；亚信安全推出安全运营智能体、安全检测智能体等能力，AI XDR 入选 IDC《中国 AI XDR 平台最佳实践案例，2026》并作为深度拆解案例。

新变化之二是新品类独立成线：大模型防护与智能体安全构成 AI 安全新赛道的产品主线。大模型防护侧，启明星辰大模型应用防火墙（MAF）、天融信大模型多模态安全网关、中孚信息大模型安全防护系统、格尔软件 AI 大模型应用安全护栏、深信服 AI 安全平台密集落地；其中天融信大模型多模态安全网关获国内首张大模型安全防护围栏增强级证书，中孚信息大模型安全防护系统覆盖 OWASP LLM Top 10 攻击类型。

下载日志已记录，仅供内部参考

图15：亚信安全 AI XDR 联动防御体系



资料来源：亚信安全官网，国联民生证券研究所

智能体安全成为差异化竞争点：亚信安全发布智能体信任框架 ATF；启明星辰推出智能体安全“新三件套”；奇安信发布智能体安全平台与 QAgent Fabric 安全智能体研发框架；格尔软件、吉大正元布局智能体身份管理，吉大正元开展智能体分布式数字身份与可验证凭证研究。密码厂商沿“密码+AI”路线切入：格尔软件 AI 大模型安全保险箱、吉大正元基于密码机制的模型指纹嵌入与内容数字水印实现 AI 版权保护与侵权溯源，数字认证发布人工智能安全运营平台。

AI 能力同时向存量业务与内部运营渗透。存量业务侧，深信服将安全 GPT 用于威胁检测、钓鱼邮件识别与告警研判；绿盟科技 AI-PTS 以大模型驱动智能渗透测试；安恒信息发布 AI 原生代码审计智能体与零日雷达漏洞情报订阅产品；任子行以自研大模型打造数据安全“数安三件套”。内部运营侧，启明星辰将 AI Coding 融入研发全生命周期，安恒信息推动安全服务和运营向 AI 原生转型。

新变化之三是商业化落地：AI 布局已从产品发布进入订单与收入落地阶段。绿盟科技 AI 安全相关订单额达 1.17 亿元，2026 年上半年运营商行业 AI 安全签单额同比增长超 300%；安恒信息纯 AI 收入超 4100 万元、间接 AI 产品收入超 6400 万元，AI 原生安全收入同比增长超 300%；启明星辰大模型应用防火墙收入同比增长 238.15%，Agentic SOC 收入同比增长 31.99%；天融信 AI 安全防护业务收入同比增长超 130%，AI 安全业务直接收入超 8000 万元；360 数字安全业务收入 3.10 亿元、同比增长 22.27%。从披露口径看，AI 相关收入在总收入中的占比仍处低位，但增速显著高于传统业务，收入结构正向 AI 安全倾斜。

表12：A 股网络安全公司 AI 布局一览

公司名称	AI 在网络安全方面的布局
深信服	安全 GPT 用于威胁检测、钓鱼邮件识别与告警研判， AI 钓鱼邮件检测精准率达 98% 、替代 80%以上人工研判； AI First 研发战略，发布 AI 算力网关、 AI 安全平台与生产级 AI 编程工具 CoStrict

下载日志已记录，仅供内部参考

奇安信	QAX-GPT 安全大模型应用于威胁研判与自动化渗透测试；“安全智能”（AISOC、QcodeAgents）与“智能安全”（AI 安全网关、智能体安全平台）双线布局；龙虾安全伴侣通过信通院首批智能体安全防护能力评估
启明星辰	“安全大模型 + 驾驭工程 + 多安全智能体”架构，安星智能体矩阵与“安星数字员工”家族；大模型应用防火墙（MAF）收入快速增长，智能体安全“新三件套”；与中国移动九天共建网络安全领域基础大模型
天融信	天问大模型与覆盖 6 大场景的 20 余个安全智能体矩阵；大模型多模态安全网关获国内首张增强级围栏证书；AI MSS 托管服务；智算云太行云 5.0“融算一体”
绿盟科技	风云卫 AI 安全能力平台与清风卫 AI 安全防护产品矩阵（分别面向 AI 应用安全与 AI 自身安全）；AI-PTS 智能渗透测试系统；AI 安全订单额 1.17 亿元
安恒信息	恒脑 4.0 安全智能体总指挥（统一调度、自主执行、闭环验证）；AI 原生代码审计智能体、零日雷达漏洞情报订阅；纯 AI 收入超 4100 万元
三六零	360 安全大模型与“倚天屠龙”AI 安全智能体体系、安全智能体蜂群；360 大模型卫士“以模治模”防护；数字安全业务收入同比增长 22.27%
亚信安全	AI XDR 联动防御平台与安全运营智能体；ATF 智能体信任框架，智能体身份安全订单高速增长；大模型安全评估、大模型防火墙实战落地
山石网科	“双 A 战略”（ASIC 安全芯片+AI）；灵岩安全大模型驱动安全智能体矩阵覆盖 PPDR 全生命周期；发布 AI TRiSM 综合安全方案，AI 安全网关、智能体安全防护计划下半年推出
国投智能	L4 级自动化取证与 Qiko3.0 大模型智能取证双引擎；天擎公共安全大模型迭代至 V5；数字反欺诈业务实现商业化跑通
安博通	鲁班安全大模型通过国家网信办备案并实现业务落地；“溢彩”AI 交付架构落地多个智算中心；下一代 AI 防火墙入选信通院开源大模型软件创新应用典型案例
格尔软件	“抗量子 +AI”双引擎；AI 大模型安全保险箱、AI 大模型应用安全护栏、智能体身份管理平台
吉大正元	“密码 +AI”战略；智能体分布式数字身份（DID）与可验证凭证（VC）研究；大模型安全护栏；模型指纹嵌入与内容数字水印实现 AI 版权保护溯源
中孚信息	大模型安全防护系统（网关/API 双模式，覆盖 OWASP LLM Top 10 攻击）；终端大模型智栏系统
任子行	“AI+ 安全”战略；自研 AI 大模型打造数据安全“数安三件套”（智鉴、智巡、智枢）；数据安全分类分级系统在金融、通信运营商规模化落地
数字认证	发布隐私计算平台、人工智能安全运营平台；聚焦 AI 智能安全服务，推动金融领域抗量子敏捷迁移
北信源	北信源 AI 平台作为多智能体安全开发平台，支持私有化大模型与外接国产大模型；新增 AI 数据安全围栏、军工专属微调模型；信源密信安全 AI 基座

资料来源：各公司公告、上海证券报、国投智能官网、央广网等，国联民生证券研究所整理

5.3 安全厂商与模型厂商协同提速，国产大模型进入安全应用

新变化之四是模型厂商入场：国内安全厂商与模型厂商的协同在 2026 年三季度明显提速。能力验证方面，绿盟科技适配智谱 GLM-5.3 大模型，其在 AI 安全国际基准 CyberGym 榜单的漏洞复现成功率为 95.02%；双方联合开展 AI 漏洞挖掘研究，绿盟科技 AI-PTS 智能渗透测试系统亦融合智谱 GLM 系列模型。

下载日志已记录，仅供内部参考

角色延伸方面，智谱 AI 编程工具 ZCode 因代码上传争议引发数据安全关注，智谱于 2026 年 9 月 21 日宣布完成安全整改并将客户端代码开源，同时邀请中国信通院与绿盟科技开展第三方安全审计。生态组织化方面，中关村自主大模型产业联盟成员已达 100 家，奇安信、绿盟科技等已将 GLM-5.3 用于漏洞挖掘系统。

图16：绿盟智能渗透测试系统（AI-PTS）架构



资料来源：绿盟科技官方公众号，国联民生证券研究所

5.4 合规框架成型，AI 安全成为独立需求

新变化之五是合规成为独立需求：AI 生成内容治理的制度框架在 2025-2026 年密集落地。《人工智能生成合成内容标识办法》自 2025 年 9 月 1 日起施行，要求对 AI 生成的文本、图片、音频、视频等内容添加显式与隐式标识，互联网平台承担分发环节的核验义务；CNNIC 第 57 次《中国互联网络发展状况统计报告》显示，截至 2025 年 12 月我国生成式 AI 用户规模达 6.02 亿人。《人工智能拟人化互动服务管理暂行办法》自 2026 年 7 月 15 日起施行，将 AI 拟人化交互服务纳入监管。2026 年 1 月 1 日施行的新修订《网络安全法》在立法层面纳入 AI 安全监管条款，与上述专项办法形成“通用法律+专项办法+部门规章”的分层监管体系，“传统合规筑底+AI 合规增量”构成下一阶段政策驱动需求的基本结构。

标识义务的技术实现路径或将引导需求的方向。显式标识要求在 AI 生成内容中嵌入用户可感知的标识，如文字与音频提示；隐式标识要求在文件元数据中写入生成属性，如水印与密码学签名。两者均需在内容生产端改造生成服务、在分发端部署核验能力，平台的核验义务即对无标识 AI 内容的检测与处置，是需求的核心所在，深度合成检测由此从可选项变为平台合规的重要环节。监管对象覆盖 490 余款备案模型与 6.02 亿生成式 AI 用户，检测与核验属于持续运行需求。

对关键信息基础设施运营者而言，合规范围正在从传统等保向 AI 系统延伸。AI 系统的采购、部署、测评正在成为等保测评之外的新增合规项，由此产生的 AI 安全评测、红队测试、模型资产盘点等需求，属于政策驱动型需求。

下载日志已记录，仅供内部参考

AI 自身安全的需求正在分化为两条产品线。一条是模型防护：针对提示注入、越狱、供应链投毒等直接攻击模型的手段，对应大模型防火墙、API 安全网关、模型安全评估等产品。OWASP 2025 年版 LLM 十大风险为该领域提供了风险分类的通用语言，提示注入居首；IBM 数据显示超过 20%的组织已发生针对 AI 模型或应用的泄露，其中 92%缺乏完善的 AI 访问控制。

表13：AI 自身安全两条产品线对比

维度	模型防护	使用治理
攻击对象	大模型本身（提示注入、越狱、供应链投毒）	员工使用 AI 工具的行为（影子 AI、数据泄露）
核心风险	模型被直接攻击，输出被篡改	敏感数据通过 AI 工具外泄，智能体权限滥用
对应产品	大模型防火墙、API 安全网关、模型安全评估	行为审计、数据防泄漏（DLP）、权限管理
当前态势	已形成 OWASP 风险分类与产品体系	增速更快但渗透更低，影子 AI 事件占比 20%升至 43%

资料来源：IBM《Cost of a Data Breach Report 2026》、OWASP《2025 Top 10 Risk & Mitigations for LLMs and GenAI Apps》，国联民生证券研究所整理

另一条是使用治理，其需求增速更快但当前渗透更低。员工使用未经企业正式管理的外部 AI 工具，即影子 AI，AI 智能体大量参与业务流程产生了规模化的非人工身份，如服务账号、API 密钥。这两类风险对应的产品是行为审计、数据防泄漏（DLP）与权限管理，防护对象从模型本身延伸到 AI 使用的全过程，包括智能体的执行行为，权限管控的颗粒度需细化到每一次工具调用。

国内供给侧已经启动。信通院已开展 AI 安全评测标准工作，网安厂商的 AI 防火墙、提示词防护、模型资产管理系统陆续发布；海外侧，部分模型厂商、云厂商与安全厂商围绕 MCP 等协议推进协作生态，标准逐步成型有望带动防护产品接口统一。该赛道的采购主体正从互联网平台扩展至金融、政务等大模型部署规模较大的行业，2026—2027 年有望成为国内网安市场中增速较快的细分方向。

综合本章证据，国内的五项新变化指向同一结论：AI 安全已逐步成为独立的需求与供给类别。需求端，合规框架将 AI 系统纳入强制义务，AI 安全由项目附属变为独立预算项；供给端，头部厂商完成大模型与智能体体系布局，模型厂商入场带来生态协同，产品线由模型防护延伸至使用治理。国内与海外的差距主要体现在预算属性、付费形态与验证信号的传导节奏，而非技术能力；随着运营商等行业订单落地，国内商业化正从产品发布期进入订单验证期。

下载日志已记录，仅供内部参考

6 投资建议

AI 时代网络安全产业迎来新机遇：海外龙头 AI 安全收入已在报表端体现，国内厂商完成安全大模型与智能体产品体系布局，AI 安全订单与合规需求共同驱动行业增长。建议关注 A 股网络安全公司：三六零、深信服、奇安信、绿盟科技、启明星辰、国投智能、天融信、数字认证、亚信安全、安恒信息、安博通、北信源、任子行、格尔软件、山石网科、吉大正元、中孚信息。

下载日志已记录，仅供内部参考

7 风险提示

AI 技术落地不及预期的风险。安全大模型在客户生产环境中的可用性、稳定性仍在验证期，若智能体误报率或自主处置出错率高于预期，商业化进度可能低于测算。

行业竞争加剧的风险。目前行业整体发展平稳，但如果厂商为争夺存量预算发起价格竞争，行业毛利率与竞争格局可能恶化。

测算与假设误差风险。引用的市场规模数据来自 Gartner 等第三方统计口径，相关推算基于公开信息，可能与实际存在偏差。

下载日志已记录，仅供内部参考

插图目录

图 1： 越狱研究者在 GitHub 公开的 Claude Fable 5.1 系统提示词文件（2195 行、269KB） 5
图 2： 智谱 ZCode 官方声明：完成整改、代码开源并附安全审计结论 7
图 3： 全球信息安全终端用户支出（百万美元） 10
图 4： 2024—2026 年全球信息安全市场终端用户支出（按子领域划分，单位：百万美元） 11
图 5： AI 给网络安全行业带来的改变与增量 12
图 6： 2026 年以来海外网安龙头股价走势（单位：美元） 15
图 7： Fortinet AI 原生安全架构（AI-Native Security Fabric） 16
图 8： Fortinet AI 驱动安全运营（FortiAI SOC 平台） 16
图 9： Palo Alto Networks Prisma AIRS 平台演进与 AI 安全能力 17
图 10： Palo Alto Networks XSIAM 与可观测性增长 17
图 11： Palo Alto Networks 2026 财年第四季度业绩（RPO、NGS ARR 与营收） 18
图 12： CrowdStrike 智能体安全平台与 Flex 采购模式 18
图 13： CrowdStrike 2027 财年第二季度核心财务指标 19
图 14： 信通院“写境”图谱国内版与国际版收录情况 22
图 15： 亚信安全 AI XDR 联动防御体系 24
图 16： 绿盟智能渗透测试系统（AI-PTS）架构 26

表格目录

表 1： OpenAI 模型测试逃逸事件（GPT-5.6 Sol 入侵 Hugging Face）关键事实 4
表 2： Anthropic 模型测试越界事件关键事实 4
表 3： Claude 被诱导参与墨西哥政府攻击事件关键事实 6
表 4： 2026 年以来头部大模型部分安全事件一览 7
表 5： OWASP 2025 年版 LLM 应用十大安全风险（前五位） 8
表 6： IBM《2026 年数据泄露成本报告》AI 安全发现一览 8
表 7： Anthropic 披露的 AI 主导编排网络攻击（GTG-1002）关键事实 9
表 8： AI 滥用与治理关键数据一览 13
表 9： IBM《2026 年数据泄露成本报告》关键数据一览 14
表 10： 部分海外网安关键事件梳理 19
表 11： 国内与海外 AI 安全商业化结构差异 23
表 12： A 股网络安全公司 AI 布局一览 24
表 13： AI 自身安全两条产品线对比 27

下载日志已记录，仅供内部参考

分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并登记为注册分析师，基于认真审慎的工作态度、专业严谨的研究方法与分析逻辑得出研究结论，独立、客观地出具本报告，并对本报告的内容和观点负责。本报告清晰准确地反映了研究人员的研究观点，结论不受任何第三方的授意、影响，研究人员不曾因、不因、也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的利益。

评级说明

投资建议评级标准	评级		说明
以报告发布日后的 12 个月内公司股价（或行业指数）相对同期基准指数的涨跌幅为基准。其中：A 股以沪深 300 指数为基准；新三板以三板成指或三板做市指数为基准；北交所以北证 50 指数为基准；港股以恒生指数为基准；美股以纳斯达克综合指数或标普 500 指数为基准。	公司评级	推荐	相对基准指数涨幅 15%以上
		谨慎推荐	相对基准指数涨幅 5%～15%之间
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上
	行业评级	推荐	相对基准指数涨幅 5%以上
		中性	相对基准指数涨幅-5%～5%之间
		回避	相对基准指数跌幅 5%以上

免责声明

本报告由国联民生证券股份有限公司或其关联机构制作。国联民生证券股份有限公司具有中国证监会许可的证券投资咨询业务资格。本报告的分销依据不同国家、地区的法律、法规和监管要求由国联民生证券于该国家或地区的具有相关合法合规经营资质的子公司/经营机构完成。在遵守适用的法律法规情况下，本报告亦可能由国联证券国际金融有限公司在香港地区发行。国联证券国际金融有限公司具备香港证监会批复的就证券提供意见（4 号牌照）的牌照，接受香港证监会监管，负责本报告于中国香港地区的发行与分销。

本报告仅供本公司授权之机构及个人使用，本公司不会因任何人收到本报告而视其为客户。本报告仅为参考之用，并不构成对任何人的操作建议或任何保证，不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告所包含的观点及建议并未考虑获取本报告的机构及个人的具体投资目的、财务状况、特殊状况、目标或需要，客户应当充分考虑自身特定状况，进行独立评估，并应同时考量自身的投资目的、财务状况和特定需求，必要时就法律、商业、财务、税收等方面咨询专家的意见，不应单纯依靠本报告所载的内容而取代自身的独立判断。在任何情况下，本公司不对任何人因使用本报告中的任何内容而导致的任何可能的损失负任何责任。

本报告是基于已公开信息撰写，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、意见及预测仅反映本公司于发布本报告当日的判断，且预测方法及结果存在一定程度局限性。在不同时期，本公司可发出与本报告所刊载的意见、预测不一致的报告，但本公司没有义务和责任及时更新本报告所涉及的内容并通知客户。

在法律允许的情况下，本公司及其附属机构可能持有报告中提及的公司所发行证券的头寸并进行交易，也可能为这些公司提供或正在争取提供投资银行、财务顾问、咨询服务等相关服务，本公司的员工可能担任本报告所提及的公司的董事；本公司自营部门及其他投资业务部门可能独立做出与本报告中的意见或建议不一致的投资决策。客户应充分考虑可能存在的利益冲突，勿将本报告作为投资决策的唯一参考依据。

若本公司以外的金融机构发送本报告，则由该金融机构独自为此发送行为负责。该机构的客户应联系该机构以交易本报告提及的证券或要求获悉更详细的信息。本报告不构成本公司向发送本报告金融机构之客户提供的投资建议。本公司不会因任何机构或个人从其他机构获得本报告而将其视为本公司客户。提示客户及公众投资者慎重使用未经授权刊载或者转发的本公司证券研究报告，慎重使用公众媒体刊载的证券研究报告。

本报告的版权仅归本公司所有，未经书面许可，任何机构或个人不得以任何形式、任何目的进行翻版、转载、公开传播、篡改或引用，不得将报告内容作为诉讼、仲裁、传媒所引用之证明或依据，不得用于营利或用于未经允许的其它用途。所有在本报告中使用的商标、服务标识及标记，除非另有说明，均为本公司的商标、服务标识及标记。本公司版权所有并保留一切权利。

无锡 江苏省无锡市金融一街 8 号国联金融大厦 8 楼

上海 上海市虹口区杨树浦路 188 号星立方大厦 B 座 7 层

北京 北京市西城区丰盛胡同 20 号丰铭国际大厦 B 座 5F

深圳 深圳市福田区中心四路 1 号嘉里建设广场 1 座 10 层 01 室