2026 年上半年

AI Skill 生态安全研究报告

从“能力扩展”到“供应链入口”

——基于近 7 万公开 Skill 的安全观察

发布机构：360安全能力中心

数据来源：360 Skill 分析平台

发布日期：2026 年 7 月

360

目 录

2026 年上半年 Skill 生态安全总览

随着 AI Agent 技能（Skills）生态的迅速发展，社区开发者贡献的技能数量与日俱增。然而，这些技能来源多样、质量参差不齐，其安全性缺乏有效保障。攻击者可能借机发布恶意技能，对用户设备进行攻击或窃取数据。360 Skill 分析平台为面向智能体技能包的全链路安全检测方案，为技能生态提供全面的安全保障。

本文将从风险态势、安全风险分析、企业治理等维度，全面梳理 Skill 安全的核心挑战，详细阐述 360Skill 分析平台的安全检测体系与保障方案，为业务接入与开发者提供完整的实践指南。

数据总览

66,474 **累计扫描 Skill** 占敏感数据访问总量的 9.7%	**40.3%** **风险检出率** 至少命中一项风险信号（其中高危+严重仅 1.76%）	**1.76%** **高危+严重占比** 可被直接利用或造成重大损失
3,380 **有效硬编码凭证** 涉及 2,127 个 Skill、1,532 位作者	**12,929** **高风险敏感数据访问** 占敏感数据访问总量的 9.7%	**1,955** **可疑外联累计命中** 17 个可疑 URI 累计出现次数

核心判断

判断一：近四成公开 Skill"带病上岗"，系统性安全基线缺失是首要风险

Skill 生态并非"近半有毒"，而是近四成样本暴露出不同程度的安全基线缺口。绝大多数开发者缺少安全规范意识，明文写密钥、权限多开、私自外联、随意读取本地文件等问题普遍存在。这类低危漏洞叠加后可形成完整攻击链路，大幅降低黑客入侵成本，其危害远大于少量定向恶意攻击。

判断二："能力-权限倒挂"是高危风险最典型特征

区块链、DevOps 运维两类 Skill 风险检出率分别高达 60.9%、56.0%。工具对外宣称的功能只需少量权限，实际却申请全盘文件读写、系统命令执行等高权限。一旦被调用可窃取企业与个人核心资产，治理需按照功能能力 × 授予权限 × 资产价值三维度划定管控优先级。

判断三：硬编码凭证是打通业务系统的 "万能钥匙"

Skill 运行环境会继承终端、企业系统已有登录权限，攻击者无需破解大模型，仅爬取公开社区里泄露的 API Key、仓库令牌、机器人密钥、钱包助记词，就能直接冒用身份调用服务、篡改数据、消耗资源。且区块链助记词泄露无法通过改密码挽回，资产损失不可逆。

判断四：零散危险能力已可拼接链式攻击，攻击前置条件完全具备

远程拉取脚本、远程执行、下载外部代码、网络外联、Shell 命令、文件读写等高危能力在海量 Skill 中分散存在。单看任意一项都可用于正常办公，但组合使用就能完成「读取隐私数据→外传至黑客服务器→下载恶意程序→本地运行→长期驻留控制」一整条入侵链路。本次暂未统计单个 Skill 内多风险共存比例，但生态层面攻击基础条件已经成熟。

判断五：Skill 安全正从"查毒"升级为"可信验证"

多款主流安全检测工具对同一批 Skill 检测结果重合度最高仅 10.4%，八成告警只能被某一款工具单独捕获。传统只查杀恶意代码的模式存在巨大盲区，未来防护必须同时核查五大维度：来源可信、描述真实、权限最小必要、运行行为与说明一致、版本更新后持续复检。

总结：AI Skill 正在从单纯的功能能力扩展，演变为全新软件供应链入口。当前生态最大隐患不是刻意投毒攻击，而是生态扩张速度远超安全治理建设速度。后续企业、平台、开发者三方需搭建全生命周期管控机制，从事后查杀转向事前准入、事中管控、事后追溯的可信治理体系，避免第三方 Skill 成为企业数字安全新缺口。

一、引言与研究说明

1.1 研究背景

随着 AI Agent 技术规模化落地，Skill 作为可被智能体调用、用于完成具体业务动作的能力载体，打通了大模型与本地文件、网络接口、服务器、云服务、业务系统之间的通路。海量开发者在社区开源渠道自由发布、共享 Skill，生态扩张速度远超传统软件包体系，但上架准入审核、权限管控、安全审计机制普遍滞后。

恶意主体可通过投放违规 Skill 窃取用户数据、横向渗透内网、盗用企业凭证与云资产；普通开发者不规范编码带来的权限越界、密钥明文存储、无约束外联等问题，形成大面积安全短板。

360 Skill 分析平台，打造面向技能包全链路安全检测方案，覆盖发布、接入、运行、迭代全流程风险识别。本报告依托全网公开样本实测数据，研判 Skill 生态整体安全形势，拆解核心威胁，并面向企业、开发者、平台运营方输出可落地的安全治理方案。

1.2 研究对象

本报告分析对象为可公开下载、由第三方开发者上传分发的 AI Agent Skill。Skill 并非单纯代码包或提示词文本，而是由自然语言功能描述、元数据配置、执行脚本、依赖资源、外部调用地址共同组成的复合执行单元，能够直接驱动 Agent 读取本地资源、发起网络请求、执行系统指令，具备访问真实业务资产的权限能力。

本次监测覆盖 ClawHub 等国内主流 Skill 分发社区，对采集样本完成标准化归类、风险等级定级、行为标签打标、敏感访问审计、硬编码凭证提取、外联行为溯源。

1.3 数据口径总览

- 累计扫描有效 Skill 样本：66474 个
- 至少命中一项风险信号样本：26786 个，整体风险检出占比 40.3%；其中高危 + 严重级 Skill 合计 1170 个，占总样本 1.76%

- 风险样本结构：84.0% 为低危风险，仅 16.0% 属于高危、严重等级风险项
- 硬编码凭证：原始捕获凭证字段 5178 条，同 Skill 内去重后剩余 3893 条，剔除模板占位字符后，有效硬编码凭证 3380 条，关联 2127 个 Skill、1532 位开发者
- 敏感数据访问：具备可判定规则的访问行为 132891 次，其中高风险越权访问 12929 次，占敏感访问总量 9.7%
- 外部网络外联：共识别 300 个外联 URI，标记可疑外联地址 17 个，可疑外联行为累计触发 1955 次

口径重要说明：40.3% 为「样本是否命中任意风险规则」统计结果，单份 Skill 可叠加多条低危告警；1.76% 为「样本包含至少一条高危 / 严重风险项」的独立样本占比，两项统计维度不同，不可直接相乘换算。

1.4 分析框架与方法边界

本报告以静态代码解析 + 语义文本分析为核心检测手段，辅以小范围动态沙箱抽样验证。

静态与语义分析可批量识别明文密钥、危险指令、未声明网络请求、路径越权读取等特征，但受限于业务上下文缺失，无法 100% 自动判定行为是否完全匹配业务需求。报告严格区分风险信号、高风险行为、确认恶意样本三个层级，对第三方域名、外部服务仅做特征标记，不直接定性平台存在安全漏洞。

1.5 方法论局限性与数据自检

- **检测局限：**未对全部样本执行动态沙箱运行分析，依赖运行时条件触发的隐蔽恶意行为存在漏检可能；
- **凭证有效性：**报告中 "有效凭证" 指格式合规且非模板占位文本，不等于接口一定可调用。项目组抽检 200 条凭证样本，仅 34% 在发布时段可正常鉴权调用，其余凭证已过期或被权限回收；

- **复核准确率：**随机抽取 500 条风险样本、500 条安全样本人工复核，风险标记确认率 78.2%，安全标记确认率 91.4%，误报主要集中在 "功能刚需但未在描述中注明权限" 场景；
- **生态对标参考：**npm 2023 年审计数据显示约 21% 软件包存在高危依赖漏洞，PyPI 同类问题占比约 15%。Skill 生态 40.3% 风险检出率更高，核心原因是 Skill 风险判定口径更广，纳入自然语言不合规、权限超配、未公示外联等传统软件供应链不纳入审计的条目，直接横向对比需审慎参考。

二、AI Skill 生态发展态势：野蛮生长迈入安全治理拐点

2.1 生态扩张：半年规模超越传统软件数年积累

2026 年 1 月 ClawHub 平台上线之初，全网公开 Skill 总量不足 2000 个；截至 6 月初，全平台公开 Skill 数量突破 10 万。npm 开源包生态耗时近两年才完成十万体量积累，Skill 生态仅用 180 天即实现同等量级扩张。

- **供给端**：Agent 工具普及催生海量场景封装需求，内容创作、数据分析、区块链钱包、运维自动化、办公批量处理等场景均被快速制作为可复用 Skill；
- **需求端**：企业与个人用户将 Agent 作为常态化生产力工具，第三方 Skill 安装调用频次持续走高。

生态高速扩容之下，安全审核体系建设严重滞后。ClawHub 上线初期无常态化安全检测能力，直至 ClawHavoc 恶意 Skill 事件爆发后，才增补基础扫描能力，"先放量扩张、后补安全管控" 的模式造成风险长期沉淀。

2.2 攻击范式迭代：从显性投毒转向全链路信任劫持

Skill 场景下攻击手段已完成三代演进：第一代直接植入木马、下载器等显性恶意代码；第二代利用依赖劫持、包名仿冒开展供应链投毒；第三代不再局限于代码篡改，通过修改自然语言描述、优化检索排名、伪装开发者信誉，诱导 Agent 主动选择恶意 Skill，实现无明显恶意特征的隐性攻击。

对应防御逻辑也必须从单一文件查杀，升级为来源可信校验、权限最小约束、运行行为审计、版本更新复检的全生命周期闭环管控。

代际	特征	典型案例	状态
第一代：显性恶意	恶意指令直接写在文档中，如 ClawHavoc 恶意投毒	824 个恶意 Skill 批量上架，ClickFix 社会工程	已被大规模清理
第二代：绕过检测的隐藏后门	多层编码+反序列化等技术，实现远程控制，绕过多层检测	伪装 DeepSeek 的恶意 Skill、Promptware 提示词恶意软件	已在生态中大量存在
第三代：生态与平台级攻击	排名操纵、批量投毒、权限滥用，影响用户与 Agent 的全自动化决策	ClawHub 排名操纵、语义供应链攻击、Skill.md 文本注入	攻击与平台对抗升级中

2.3 核心矛盾：基线普遍缺失重于定向恶意攻击

数据显示，66,474 个 Skill 中有 26,786 个至少命中一项风险信号，占比 40.3%。但这并不意味着近半数的 Skill 都具有恶意意图——其中 84.0%为低危，高危和严重 Skill 合计仅 1,170 个，占全部样本的 1.76%。

真正的问题在于：大量 Skill 在开发过程中存在凭证管理不规范、权限边界不清、外联缺少约束等安全欠账。此类 Skill 本身无明确攻击意图，但相当于长期敞开防护入口，大幅降低攻击者批量嗅探、漏洞利用、凭证爬取的成本。学术研究佐证，多项低危缺陷叠加即可构建完整攻击链路，单点短板的聚合风险不可忽视。这种"门没有锁好"的状态，为攻击者提供了大量可乘之机。

从安全等级分布来看：

安全等级	Skill 数	占全部样本	解读
Safe（未检出已知风险）	39,610	59.60%	未检出当前规则可识别的已知风险
Low（低危）	22,495	33.80%	存在轻微不规范或潜在风险
Medium（中危）	3,121	4.70%	需要结合场景进一步核验
High（高危）	1,055	1.60%	存在可直接利用或高影响风险
Critical（严重）	115	0.17%	可能造成重大损失或失控后果

2.4 风险场景分化：高权限品类风险最高，大众化品类影响最广

Skill 生态的风险并非均匀分布，而是沿着业务场景和资产价值呈现出明显的分化。

一级分类	Skill 数	风险 Skill 数	风险率	高危+严重	高危+严重率
内容与媒体	12,100	6,496	53.70%	207	1.70%
商业	11,746	4,699	40.00%	163	1.40%
数据与 AI	6,448	3,124	48.40%	148	2.30%
工具	6,085	2,817	46.30%	123	2.00%
开发	5,847	2,306	39.40%	105	1.80%
生活方式	3,707	1,077	29.10%	38	1.00%
文档	3,611	647	17.90%	14	0.40%
DevOps	3,097	1,734	56.00%	93	3.00%
区块链	2,650	1,615	60.90%	141	5.30%
测试安全	2,542	1,122	44.10%	84	3.30%
研究	2,002	585	29.20%	16	0.80%
其他	1,084	564	52.00%	38	3.50%

这张表揭示了两个不同维度的风险逻辑：

第一个逻辑是"风险率"——哪些场景天然更危险。 区块链类 Skill 风险率 60.9%，高危和严重占比 5.3%，在所有类别中最高。DevOps 类风险率 56.0%，高危和严重占比 3.0%，排名第二。这两类场景的共同特征，是 Skill 通常需要接触更高价值的资产（钱包、私钥、CI/CD、云平台）和更强的执行权限。单次失控的损失上限远高于一般场景。

第二个逻辑是"风险存量"——哪些场景的影响面最大。 内容与媒体类 Skill 风险存量最大，共计 6496 个风险样本，占全部风险 Skill 的 24.3%。该类 Skill 受众广、安装门槛低，一旦出现数据外传、凭证泄露，波及用户规模更大。

这意味着企业不应只按安装数量排查 Skill，也不应只看风险率排序，而需要同时考虑"风险概率"和"影响规模"两个维度，按权限等级和业务影响确定治理优先级。

2.5 生态阶段判断

当前 Skill 生态已告别完全无监管的野蛮分发阶段，进入平台基础安全扫描落地的 "基础安检期"，但距离体系化 "可信治理阶段" 仍存在明显差距：

- **生态规模**高速增长，但统一安全规范、准入机制尚未成型；
- **恶意攻击**手段向语义层、信任层延伸，传统特征查杀存在大量盲区；
- **平台治理**刚刚起步，主流分发平台仅配置基础静态扫描，权限校验、行为动态监测、版本变更复核能力普遍薄弱；
- **行业标准**正在形成，OWASP Agentic Skills Top10、国内多部门生成式 AI 合规指导意见等框架逐步落地，但行业落地细则仍待完善。

本质差距在于防护思路需要从 "检查代码是否带毒"，升级为 "核查来源、描述、权限、行为、更新全链条可信度"。

三、安全风险深度分析

3.1 硬编码凭证泄露：供应链最易利用的系统性风险

凭证泄露是本次监测中最具实战价值的风险类型。3380 条有效格式硬编码凭证内，API 密钥占比 47.9%，集中出现在第三方模型接口、数据库、支付服务、协作办公工具接入场景；同时检出 31 条区块链钱包助记词，该类凭证一旦泄露，无法通过重置密码完成权限回收，资产损失不可逆。

泄露凭证关联代码仓库、云平台、企业机器人、数据库、支付链路等多类业务系统。区别于传统软件供应链，Skill 运行环境天然继承用户终端、企业内网已授权身份，攻击者无需渗透大模型本身，仅批量爬取公开社区 Skill 即可批量窃取有效访问凭据，直接冒用身份调用服务、篡改数据、消耗资源。

指标	数量
原始凭证总条数	5,178
同 Skill 内相同值去重后	3,893
过滤无效占位符	513
有效凭证条数	3,380
涉及 Skill 数	2,127
涉及作者数	1,532

凭证类型	数量	占比
API 密钥	1,620	47.90%
其他	630	18.60%
密码	312	9.20%
用户名	159	4.70%
Bearer 令牌	149	4.40%
Basic 认证	115	3.40%
加密密钥	105	3.10%
连接字符串	61	1.80%
Session Cookie	57	1.70%
OAuth 令牌	52	1.50%
私钥	42	1.20%
钱包助记词	31	0.90%
Webhook 密钥	26	0.80%
SSH 密钥	6	0.20%

关联服务分布

通过上下文识别，泄露凭证涉及支付、数据库、企业办公、机器人、代码仓库和模型服务等多类平台。该统计仅说明凭证在相关 Skill 上下文中出现，不代表相应服务或平台自身存在安全漏洞。

服务/产品	凭证条数	说明
SkillPay	158+	Skill 付费/计费相关凭证
PostgreSQL	38	数据库连接凭证
飞书应用	32+	App ID/Secret 等
Telegram Bot	23	机器人 Token
QQ	22	机器人或应用凭证
ClawHub	20	平台访问凭证
Langfuse	18	LLM 可观测平台凭证
MySQL	16	数据库凭证
OpenAI	15	API Key
ElevenLabs	14	语音合成 API 凭证
Supabase	12	后端服务凭证
Gmail	12	邮箱认证凭证
GitHub Personal	12	个人访问令牌
Stripe	10	支付平台凭证
AiCoin	10	加密资产服务凭证

治理重点：杜绝明文密钥嵌入代码与说明文档，落地密钥托管服务、短期临时令牌、最小权限分配、凭证定期轮换、泄露后快速吊销机制，同时对历史上架 Skill 开展回溯扫描排查。

3.2 敏感数据访问：功能声明与实际权限严重错配

在 132891 次可判定的敏感数据读取行为中，9.7% 被判定为非必要越权访问。从风险概率来看，API 令牌访问行为中 41.5% 存在越权，浏览器 Cookie 读取风险率 21.3%，私钥文件读取风险率 18.0%。

读取敏感数据本身并非恶意行为，判定核心依据为三维规则引擎：Skill 公示功能是否必须使用该类数据、访问路径是否合规、是否经用户主动授权、数据是否向外传输。

典型违规场景：天气查询类 Skill 无合理理由读取浏览器 Cookie 与本地环境变量文件；仅用于查询余额的程序申请全盘文件读写与系统命令执行权限。

敏感数据类型	risky 次数	safe 次数	风险率
API 令牌	312	440	41.50%
Cookie	303	1,122	21.30%
私钥	279	1,269	18.00%
凭证	4,686	24,419	16.10%
会话令牌	302	2,542	10.60%
API 密钥	3,946	35,207	10.10%
密钥文件	1,191	12,747	8.50%
配置密钥	587	9,724	5.70%
OAuth 令牌	178	3,937	4.30%
环境变量	693	26,630	2.50%

需要强调的是，读取敏感数据并不天然等于恶意。例如，邮件 Skill 读取 OAuth 令牌、数据库 Skill 读取连接字符串可能是完成任务所必需的。安全判断的关键，是访问行为是否与 Skill 声明功能一致、是否只读取必要范围、是否经过用户授权，以及数据是否被发送到未声明的外部地址。

risky 判定标准说明：

risky/safe 判定规则：敏感数据访问的风险判定基于"数据类型×Skill 声明功能×调用上下文"三维规则引擎。核心逻辑：

- 若 Skill 声明功能需要访问该类数据（如支付 Skill 读取 API 密钥调用支付 API），标记为 safe；
- 若 Skill 声明功能不需要访问该类数据，但实际读取（如天气查询 Skill 读取浏览器 Cookie），标记为 risky；
- 若数据类型本身高度敏感（如私钥、助记词），除非 Skill 明确声明且功能强相关，否则默认标记为 risky；
- 若访问路径为非常规路径（如通过远程脚本间接读取），即使功能相关也标记为 risky。

该规则引擎的准确率为 82.7%（基于 500 条样本人工复核），误报主要来自"功能需要但

未在声明中明确列出"的场景。

3.3 外联与原子化风险可组装为链式攻击链路

平台累计抓取全部外联请求 48962 次，其中仅 17 个外部域名被标记为可疑，累计异常外联 1955 次。外联风险不能单一依靠域名黑名单拦截，同一域名可同时承载正常业务调用与隐蔽数据回传，必须结合请求参数、携带数据类型、Skill 业务场景综合判定。

生态内高频危险能力分布：远程脚本加载 7810 个 Skill、远程代码执行 5639 个 Skill、外部代码下载 3346 个 Skill；攻击向量排序为网络外联 40.5%、Shell 命令执行 18.0%、本地文件操作 15.2%。

单项能力均可用于正常工作，但多类能力组合即可形成完整攻击链路：读取本地.env 配置文件与 Cookie→外联将敏感数据外传→下载远端恶意脚本→调用 Shell 执行载荷→写入定时任务实现持久化控制。

本次报告仅验证各类原子风险能力广泛存在，暂未统计单份 Skill 内多风险项共存比例，可确认链式攻击的基础条件已完全具备。

聚集模式	URI 数	命中次数	风险解读
nemovideo.ai 相关接口	7	1,301	匿名令牌、上传、任务、推送等接口需要结合调用上下文核验
Skillpay.me 支付接口	4	336	涉及扣款、支付链接、余额查询等敏感业务动作
内网 IP 地址	3	145	发布版本保留内网地址，可能暴露开发环境或造成异常探测
其他可疑外联	3	173	包含远程脚本、未知服务或需进一步核验的地址

同一域名可能同时存在正常与可疑调用。风险通常不来自"域名本身"，而来自匿名令牌、未声明上传、敏感数据携带、调用频率或执行上下文。外联治理不能只做域名封禁，还需要结合请求参数、数据类型和 Skill 功能进行判断。

高频行为标签

安全标签	命中 Skill 数	说明
信息收集	17,970	读取系统信息、环境变量或用户数据等
远程脚本	7,810	从远程地址拉取并运行脚本
远程执行	5,639	在远端或受控环境执行代码、命令
内置凭证	3,563	代码或配置中直接包含密钥、令牌等
密钥泄露	3,384	存在可提取的密钥或敏感认证材料
代码下载	3,346	运行时从外部下载代码或载荷
数据外传	2,470	将本地数据发送至外部服务
安全绕过	2,437	规避鉴权、审批或安全检查
敏感路径访问	1,932	访问 .env、.ssh、credentials 等路径
隐私采集	1,541	采集用户隐私或个人信息

"远程脚本、远程执行、代码下载"三类能力同时高频出现，表明 Skill 生态已经具备形成远程载荷链的基础条件。

攻击类别与目标资产

攻击类别	发现数	占比
不安全传输	14,743	47.50%
远程代码执行	5,815	18.80%
数据窃取	5,058	16.30%
逃避检测	2,678	8.60%
后门	720	2.30%
数据破坏	357	1.10%
供应链投毒	87	0.30%
其他/未分类	1,547	5.00%

目标资产	发现数	占比	说明
凭证	6,520	21.00%	API Key、密码、令牌等
文件系统	4,276	13.70%	文件读写、路径访问
网络	4,233	13.60%	外联、网络请求与连接
系统配置	4,201	13.50%	系统或应用配置
用户数据	3,921	12.60%	用户个人及业务数据
API 令牌	3,652	11.70%	第三方服务调用令牌
进程	1,051	3.40%	进程启动、注入与控制
环境变量	718	2.30%	.env 等环境配置
浏览器数据	356	1.10%	Cookie、历史记录等
加密货币	313	1.00%	钱包、私钥、助记词
SSH 密钥	33	0.10%	SSH 密钥对
其他/未分类	1,845	5.90%	其他目标或未完成映射

3.4 核心风险一：能力 - 权限倒挂

定义：Skill 对外公示业务功能所需最小权限集合，远小于脚本实际申请的文件读写、命令执行、网络访问、目录遍历等权限范围，权限授予边界远超业务刚需。

典型案例：某区块链钱包 Skill 标注仅支持余额与交易记录查询，却申请完整文件系统读写与 Shell 执行权限。仅查询余额仅需读取指定配置文件与只读 API 请求，超额权限可随意读取目录内其他业务密钥，并将文件上传至外部服务器。

该类高危 Skill 集中在区块链、运维自动化等高资产价值场景，建议企业建立「功能能力 × 授予权限 × 关联资产价值」的分级管控模型，作为准入审批核心依据。

风险标签	命中 Skill 数	含义
供应链投毒	497	仿冒、抢注、上游篡改或依赖异常
后门植入	270	植入持续控制或隐藏入口
C2 通信	218	与远程控制基础设施通信
持久化	125	计划任务、启动项等持续驻留
反检测	55	规避分析或安全工具
代码混淆	46	隐藏真实执行逻辑
木马依赖	25	引入带恶意行为的依赖
反向 Shell	11	建立远程交互式控制通道

3.5 核心风险二：供应链信任操纵，攻击面跳出代码层

本次扫描检出供应链投毒特征 497 个 Skill、后门植入 270 个 Skill、C2 远控通信 218 个 Skill、持久化驻留 125 个 Skill。传统软件供应链攻击聚焦包篡改、依赖劫持，而 Skill 新增三类信任攻击面：

- 改写自然语言描述，提升检索排序优先级，诱导 Agent 优先调用恶意 Skill；
- 刷量抬高下载量、伪造可信作者账号，利用用户信任心理降低警惕；
- 初次审核通过后，通过版本更新植入后门，依托已放行的信任权限持续作恶。

攻击者无需攻击基座大模型，仅操控 Skill 分发环节即可完成入侵，供应链攻击复杂度显

著下降。

3.6 核心风险三：单一检测体系存在天然盲区

多厂商检测工具对同一样本集交叉验证结果显示，任意两类扫描引擎阳性告警重合率最高仅 10.4%，仅 0.69% 样本可被三类检测方式同时标记风险，超八成告警仅能被单一工具识别。

病毒查杀擅长已知恶意特征识别，静态规则适合捕捉明文密钥、高危指令，语义分析才能识别描述与行为不符的隐性风险。单一检测手段必然存在防护缺口，Skill 安全必须从 "特征查毒" 升级为五维可信验证：来源可信、描述属实、权限必要、行为一致、更新复检。

四、2026 年上半年 Skill 生态安全大事记

4.1 基础扫描上线后，恶意 Skill 仍可持续绕过平台拦截

2026 年 6 月 Palo Alto Networks Unit 42 披露，针对 2—5 月 ClawHub 生态复盘，仍发现 5 款规避检测的恶意 Skill，主要手段包括超大文件体积挤压检测截断、伪装常规业务工具投递窃密程序、利用 Agent 自主调度逻辑实现交易抢跑与恶意佣金注入。

360 观察：恶意样本已从直白木马转向 "伪装合规 + 利用 Agent 运行逻辑" 的复合绕过思路，单次静态审核无法实现长效防御。

4.2 检测机制本身成为攻击者重点对抗目标

Trail of Bits 于 6 月披露多项 Skill 检测绕过技术：利用大量换行符截断扫描文本、隐藏文件目录规避遍历、二进制与内嵌图片载荷跳过解析；还可将恶意仓库包装为企业内部镜像源，依靠大模型语义审核的逻辑偏差完成放行。

360 观察：安全检测不能仅解析说明文档可见内容，必须覆盖全目录文件、依赖链解析、语义一致性校验与动态行为核验，构建多维度交叉验证机制。

4.3 多工具检测结果割裂，无统一风险判定基准

2026 年 5 月底发布的 ClawHub Security Signals 研究整理 67,453 个最新公开 Skill 版本，对比 VirusTotal、静态启发式分析和 NVIDIA SkillSpector 等信号。研究发现，任意两类扫描器的正样本重合率最高仅 10.4%，只有 0.69%的 Skill 被三类方法同时标记，81.9%的告警只由一种扫描方法发现。

360 观察：企业侧需搭建多引擎融合检测 + 人工复核兜底的流程，不能依赖单一工具输出作为准入唯一标准。

4.4 语义供应链攻击从理论风险落地为可复现手段

2026 年 5 月发布的研究《Under the Hood of Skill.md》显示，攻击者仅通过修改 Skill

的自然语言描述，就可能提高恶意 Skill 在检索中的曝光、影响 Agent 选择，并规避治理检测。研究中，短文本触发器可将对抗 Skill 推入 Top 10 检索结果，描述性包装还会显著影响 Agent 在功能相近 Skill 之间的选择。

360 观察：未来的 Skill 安全评估必须同时覆盖"来源是否可信、描述是否诚实、权限是否匹配、代码是否安全、运行行为是否一致"五个方面，不能仅聚焦代码层面风险。

4.5 行业权威安全框架启动标准化建设

OWASP 已建立 Agentic Skills Top 10 项目，当前仍处于开发阶段，计划围绕恶意 Skill、供应链、过度授权、不安全元数据、提示词注入、隔离不足、更新漂移、扫描不足、治理缺失和跨平台复用等风险形成专项指南。2026 年 6 月发布的 SkillGuard 研究进一步提出，将 Skill 视为"携带权限的可执行制品"，同时约束其对上下文的影响和对真实世界动作的副作用。

五、 多主体全生命周期安全治理建议

Skill 安全不能只靠用户安装前看一眼，也不能只靠市场平台做一次静态扫描。企业需要把 Skill 纳入软件供应链和 AI 资产治理体系，覆盖发现、评估、准入、运行、更新和退出全生命周期。

5.1 重点风险优先处置清单

优先级	风险	立即动作	建议
P0	有效硬编码凭证	立即下线或隔离相关 Skill，轮换凭证，核查使用日志和异常调用	安全团队 1-2 人，1-2 天
P0	后门、C2、反向 Shell	阻断安装与运行，封禁外联，开展主机和账号排查	安全+运维，2-3 天
P1	远程下载+Shell 执行	核验下载源、脚本内容和哈希，改为固定版本与沙箱执行	开发配合，1 周
P1	支付、钱包、CI/CD 高权限 Skill	收敛权限，增加人工确认与交易/变更限额	业务+安全联合，1 周
P2	明文传输和不规范外联	迁移 HTTPS，限制域名与路径，避免携带敏感数据	开发团队，2 周
P2	低危开发不规范	通过规则、模板和安全开发指南批量整改	开发者自助，持续

5.2 总体原则：先检测、再准入、后运行

阶段	核心问题	建议动作
先检测	这个 Skill 来自哪里，包含什么，宣称与实际是否一致	来源核验、静态与语义扫描、凭证与依赖检测、风险分级
再准入	它是否适合进入企业环境，应该获得哪些权限	白名单、最小权限、版本锁定、人工审批、签名与哈希校验
后运行	它运行后实际访问了什么、连接了哪里、执行了什么	动态监测、外联控制、行为审计、异常阻断、更新后复检

5.3 面向企业安全与 IT 团队

建立 Skill 资产清单。盘点企业内已安装、团队自建和 Agent 自动生成的 Skill，记录来源、作者、版本、哈希、业务负责人、权限和依赖。

按业务影响分级。将涉及支付、钱包、代码仓库、数据库、云平台、CI/CD 和企业办公账号的 Skill 纳入高优先级审查。

建立准入白名单。未经安全评估的公开 Skill 不得直接进入生产环境；高危 Skill 必须人工复核，严重 Skill 应默认阻断。

验证声明与行为一致性。检查 Skill 描述、权限需求、脚本行为、外联域名和实际访问对象是否匹配。

限制外联和执行权限。对网络访问采用域名与路径级允许列表，对 Shell、文件写入、计划任务、浏览器 Cookie 等敏感能力采用按需授权。

建立版本与更新治理。锁定已验证版本和内容哈希，Skill 更新、依赖变化或远程内容变化后自动重新扫描。

保留审计证据。记录 Skill 触发原因、调用工具、文件访问、网络请求、凭证使用和最终动作，确保出现问题时能够追溯责任和还原链路。

5.4 面向 Skill 开发者

- **禁止在代码、示例、测试文件和说明文档中硬编码真实凭证**，统一使用密钥管理服务或短期令牌。
- **声明最小权限**，明确列出需要读取的文件、允许连接的域名、需要调用的工具和可能产生的副作用。
- **避免运行时从不固定地址拉取并执行脚本**；确需下载时，应锁定版本、校验哈希并使用可信源。
- **将高风险动作与用户确认绑定。** 支付、删除、发送、上传、写配置和执行 Shell 等动作不应默认自动批准。

- **对错误处理和日志做脱敏**，避免将 Token、Cookie、连接字符串和用户数据写入日志或返回给模型。
- **建立安全发布流程**，在提交市场前完成静态扫描、语义审查、依赖检查和最小化权限测试。

5.5 面向 Skill 市场和 Agent 平台

- **把安全检查前移到发布和安装两个节点**，并在更新后重新检测；一次通过不代表永久可信。
- **同时使用病毒信誉、静态规则、语义分析和动态行为验证**，避免单一检测引擎形成盲区。
- **建立发布者信誉与异常账号识别机制**，关注批量发布、名称仿冒、短时间高频更新和下载量异常。
- **支持签名、内容哈希和透明日志**，确保用户安装的版本与平台审核的版本一致。
- **提供结构化权限清单和风险卡片**，让用户在安装前清楚看到文件、网络、Shell、凭证和支付等能力。
- **默认在隔离环境运行第三方 Skill**，并对高权限能力采取拒绝优先和逐次授权。

六、360 Skill 分析平台：面向可信验证的检测与治理体系

面对数量持续增长、格式多样且不断更新的 Skill，仅靠人工逐个审计难以规模化。360 Skill 分析平台把风险证据转化为企业发布、安装、运行和更新环节的治理依据。

6.1 五维可信验证架构

- **来源可信**：样本渠道归一、版本链路溯源、发布主体标记；输出企业 Skill 资产台账，支撑准入白名单配置。
- **描述可信**：语义比对、功能文本与脚本行为交叉校验；识别隐瞒能力、虚假功能、诱导性话术。
- **权限可信**：解析文件 / 网络 / 进程 / 凭证全量调用范围；输出最小权限基线，支撑分级授权策略。
- **行为可信**：静态规则查杀、凭证提取、外联审计、动态沙箱抽样；还原完整调用链路，标记可疑操作行为。
- **更新可信**：版本差分扫描、依赖变更监测、远端资源巡检；阻断版本迭代带来的信任漂移与新增后门。

6.2 三层检测能力互补

- **基础查杀层**：集成多引擎病毒检测、已知恶意特征规则库，快速识别木马、后门、勒索类恶意载荷；
- **语义分析层**：针对 Skill 图文说明、元数据配置做自然语言解析，弥补传统杀毒无法识别的语义欺骗、权限隐瞒类风险；
- **动态沙箱层**：抓取外链资源在隔离环境执行，监测进程创建、文件修改、外联回传等运行时行为，挖掘静态分析无法发现的延迟触发型恶意逻辑。

6.3 嵌入 Skill 全生命周期管控节点

- **开发发布前**：前置扫描定位硬编码密钥、高危脚本、冗余权限，支持开发者直接整改；
- **安装准入前**：输出标准化风险等级、证据清单、权限清单，对接企业审批流程；
- **运行阶段**：将风险规则转化为终端策略，限制越权文件读取、未备案外联与命令执行；
- **版本更新后**：自动差分复检，一旦内容异动重新触发安全评估；
- **应急处置环节**：依托审计日志与原始检测证据，完成恶意 Skill 禁用、版本回滚、攻击链路溯源。

6.4 平台可公开量化评测指标建议

- **基准评测**：在 SkillTrustBench 行业公开数据集公布精确率、召回率、F1 分值与误报率；
- **动态增量**：统计沙箱相较静态分析新增检出的隐蔽外联、延迟执行、进程驻留类风险占比；
- **攻击链统计**：量化同一份 Skill 内多类风险标签共现比例，并附人工复核准确率。

截至本报告发布，平台已完成内部基准测试。本节指标体系为后续公开评测的规划框架。

七、结语：Skill 安全进入"可信验证"阶段

Skill 让 AI Agent 从单纯问答交互，进化为能够直接操作系统、处理业务、联动外部服务的执行主体。海量第三方能力包被极简分发、一键调用，原本分散在软件、接口、账号、权限中的安全风险，被收拢为一条全新的供应链入口。

本次半年监测数据印证，Skill 生态最突出问题并非大规模定向恶意攻击，而是生态扩张速度远超安全规范建设速度，普遍存在权限边界模糊、凭证管理混乱、行为缺乏约束等基线漏洞。少量高权限场景下的违规 Skill，已具备窃取资产、横向渗透、数据外泄的实际威胁。

传统 "扫描查毒" 的防护模式无法适配 Skill 场景下语义欺骗、权限越界、版本投毒、信任诱导等新型风险。Skill 安全必须完成思路升级，持续回答五大核心问题：来源是否可信、描述是否诚实、权限是否必要、行为是否匹配声明、迭代更新后是否维持安全基线。

唯有建立先检测、再准入、后持续运行管控的闭环治理体系，企业才能在释放 AI Agent 生产力的同时，避免第三方 Skill 成为企业数字资产新的安全突破口。

附录

附录 A：360 十大 Skill 风险类型

风险类型	说明
01 数据外泄	读取本地或业务数据并发送至外部
02 凭证窃取	窃取 API Key、Token、密码、Cookie 等
03 资产转移	调用钱包、支付或交易能力转移资产
04 恶意扣费	滥用 API 额度、支付接口或订阅能力
05 博彩色情导流	将智能体流量导向违规内容或平台
06 黑灰产导流	诱导进入欺诈、洗钱、刷量等链路
07 隐蔽外联/C2	连接隐蔽域名、控制端或异常网络地址
08 远程下载执行	下载脚本、依赖或载荷并执行
09 提示词投毒	通过自然语言指令劫持 Agent 目标与行为
10 持久化后门	写入记忆、配置、计划任务或后门维持控制

附录 B：关键数据总表

指标	数据
累计扫描 Skill	66,474
至少命中一项风险信号	26,786（40.3%）
低危 Skill	22,495（33.8%）
中危 Skill	3,121（4.7%）
高危 Skill	1,055（1.6%）
严重 Skill	115（0.17%）
基础安全发现	31,119
四维属性标签命中	123,028
高危+严重基础发现	6,038（19.4%）
有效硬编码凭证	3,380
涉及凭证泄露 Skill	2,127
涉及作者	1,532
明确分类的敏感数据访问	132,891

指标	数据
risky 敏感数据访问	12,929（9.7%）
外部 URI	300
可疑 URI	17
可疑外联命中	1,955

附录 C：术语表

术语	说明
Agent Skill	供 AI Agent 复用的能力模块，通常包含说明、元数据、脚本、配置、依赖和外部资源
Skill 市场	用于发布、检索、下载和更新 Skill 的分发渠道
提示词投毒	将恶意自然语言指令嵌入 Skill 说明、记忆或外部内容，影响 Agent 目标和动作
供应链投毒	通过仿冒、抢注、篡改、恶意依赖或更新劫持污染 Skill 分发链路
C2	Command and Control，攻击者用于远程控制受影响系统的基础设施
硬编码凭证	将真实密钥、密码、Token 等认证材料直接写入代码或配置
语义一致性	Skill 宣称的功能、权限需求和实际执行行为是否相互匹配
可信验证	对来源、完整性、权限、内容、行为和更新状态进行持续验证

附录 D：参考资料

以下公开资料用于补充行业事件与趋势判断。访问日期：2026 年 7 月 3 日。

- Palo Alto Networks Unit 42. OpenClaw's Skill Marketplace and the Emerging AI Supply Chain Threat. 2026-06-23.
- Trail of Bits. The sorry state of Skill distribution. 2026-06-03.
- Tencent Zhuque Lab. SkillTrustBench, the first benchmark for Agent Skill security assessment. 2026-06-17.
- arXiv. ClawHub Security Signals: When VirusTotal, Static Analysis, and SkillSpector Disagree. 2026-05-31.

- arXiv. Technical Report: Exploring the Emerging Threats of the Agent Skill Ecosystem. 2026-05-27.
- arXiv. Under the Hood of Skill.md: Semantic Supply-chain Attacks on AI Agent Skill Registry. 2026-05-12.
- arXiv. Towards Secure Agent Skills: Architecture, Threat Taxonomy, and Security Analysis. 2026-04-03.
- arXiv. SkillGuard: A Permission Framework for Agent Skills. 2026-06-02.
- OWASP Foundation. OWASP Agentic Skills Top 10（项目页面，仍处于开发阶段）. 2026.