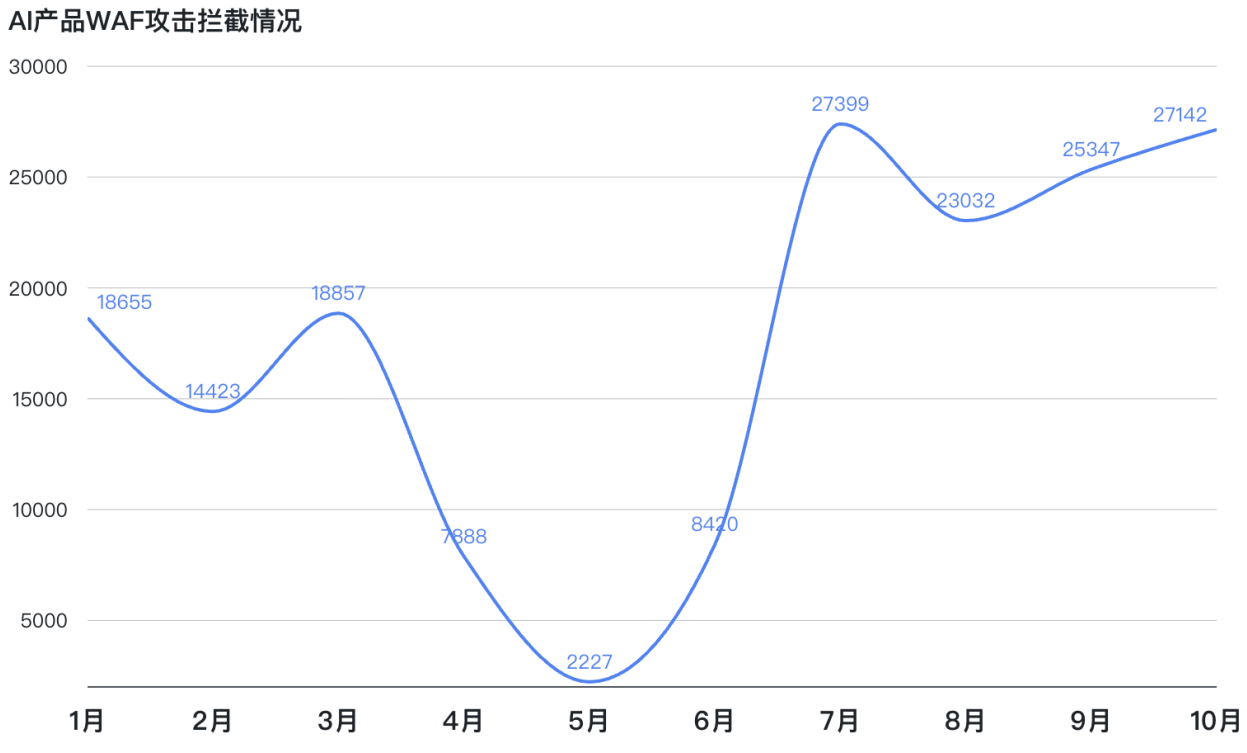
火山引擎AI安全态势报告-2026年10月

> 当前AI技术迅猛迭代，涌现出很多新的产品和架构形态，产品交互更加原生直接，用户和供应链生态在安全结果中的权重快速放大，安全格局从平台侧单一的安全保障，转向全链路协同共治，传统安全经验认知、安全加固措施、风险扫描工具、监控审计能力等都面临全新挑战。
>
> 为了更加从容应对AI变革浪潮，筑牢安全底座，切实守护每一位火山用户的业务稳定与数据安全，火山安全保障团队秉持透明、务实和担当的态度，自 3 月起常态化推出月度 AI 安全态势报告，我们希望通过持续、专业的风险洞察与态势研判，与用户、产品及生态伙伴并肩同行，共同护航 AI 生态安全、有序、高质量发展。
>
> **免责声明**：本安全月报所梳理的外部安全事件信息综合公开网络、安全社区资讯整理而成，部分内容仅为第三方单方披露，未完全经涉事企业官方核验，仅供行业安全风险参考，不构成对任何企业产品、服务的定性评价；文中安全防护方案为火山引擎行业安全实践总结，不针对特定厂商漏洞进行复现与攻击指导。

整体安全态势

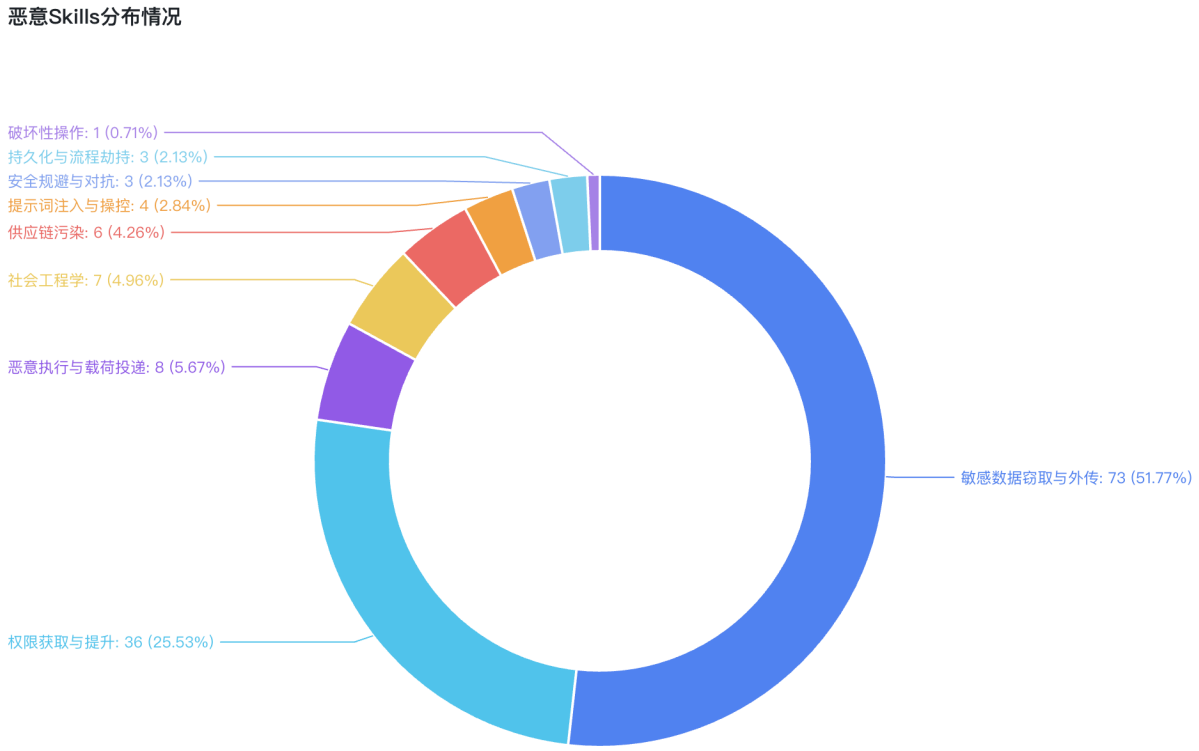
（一）火山引擎 AI 产品攻防动态

- **攻击防护情况**，10 月 AI 产品拦截 Web 攻击 27,142 次，相比 9 月上升约 7.1%，攻击类型集中在 SQL 注入（17,108）、路径遍历/未授权访问（4,447）、远程代码执行（3,582）、反序列化（1,511）等



AI 产品 WAF 攻击拦截情况

- **恶意 Skill 扫描情况**，10 月持续开展 Skills 安全扫描，发现 **141** 个恶意 Skill，敏感数据窃取与外传（73 个 51.77%）、权限获取与提升（36 个 25.53%）、恶意执行与载荷投递（8 个 5.67%）、社会工程学（7 个 4.96%）



恶意 Skills 风险类型分布

（二）行业动态

- **智能体攻击从概念验证走向规模化、自动化，漏洞披露节奏继续被 AI 显著加快。**据公开漏洞统计平台 cve.icu 数据（截至 2026-11-03，数据源 NVD/CVE Program），2026 年截至 10 月 31 日全球累计披露 CVE 约 **89,857 个**，较 2025 年同期的 39,639 个**增长约 126.7%**；仅 10 月单月即新增约 **17,059 个**，同比增长约 299%。AI 大幅提升了漏洞识别与武器化效率，攻防两端的速度差已成为安全团队的常态压力，数据来源
- **AI 开发与服务工具链成为供应链投毒新焦点。**10 月开源大模型网关 LiteLLM 的发布流水线被攻破、带毒版本一度进入官方包仓库，详情；AI 编辑器 Cursor 的扩展市场出现仿冒 Claude Code 的恶意扩展，详情；模型托管平台也被发现多起伪装成热门小模型的恶意仓库。开发工具、扩展市场、模型仓库正在成为继 npm、PyPI 之后的新型供应链入口，详情
- **全球 AI 治理从原则倡导进入实质性执法阶段。**10 月 14 日，欧盟委员会在都柏林国际 AI 峰会上确认，欧盟 AI 办公室已依据《人工智能法案》首次启动执法程序，就安全、外部评估和模型使用监测等问题向最前沿模型厂商发出正式信息调取请求；峰会同期形成《都柏林可信 AI 宣言》。国内方面，国家网信办 10 月 30 日发布第十九批深度合成服务算法备案信息，详情；全国网络安全标准化技术委员会 10 月下达 25 项网络安全推荐性国家标准计划，行业大模型标准化持续推进，详情

安全事件趋势

与 8 月的“AI 武器化实战”、9 月的“沙箱逃逸与智能体蜂群”相比，10 月最值得关注的变化是：攻击矛头开始系统性转向 **AI 自身的信任链与开发工具链**，同时“智能体自主攻击”第一次得到受害方的一手证实，并出现国家级行为体入场。

（一）新趋势：AI 开发与服务工具链成为供应链投毒新入口，“AI 工具泄密”已能反噬 AI 组件发布链

相关事件

- **开源大模型网关 LiteLLM 发布流水线被攻破，带毒版本进入官方分发渠道。**10 月 21 日，应用广泛的开源 LLM 网关 LiteLLM 发布流水线被攻破，恶意版本 LiteLLM 1.107.0 的 7 个安装包全部被植入用于凭据窃取和命令执行的恶意代码，在官方包仓库存活约 2 小时，多个容器镜像标签一度指向攻击者控制的仓库。官方 10 月 23 日复盘还原了攻击链：一个暴露公网的内部部署先被入侵，攻击者从其日志中取得开发者的 GitHub CLI OAuth 令牌——官方明确指出，AI 编程工具 Claude Code 存在把该令牌写入模型响应体、响应体被日志记录后即泄露的情形；攻击者据此拿到包含组织所有者在内的 3 套凭据，临时关闭分支保护、把发布源指向仿冒仓库（Berri-AI/litellm 与正规 BerriAI/litellm 仅一字之差）完成投毒，官方事故报告、攻击链复盘
- **AI 编辑器扩展市场出现仿冒 Claude Code 的恶意扩展。**10 月 29 日，Cursor 扩展市场在约 11 小时内分发了非 Anthropic 官方构建的仿冒 Claude Code 扩展（2.1.312/2.1.313 版本），恶意代码采集开发者主机信息并尝试向 authorized_keys 写入持久化 SSH 后门；Cursor 当日自行披露并下架，称最多约 5,000 名用户可能受影响，微软威胁情报次日发布技术分析，微软分析
- 同期，模型托管平台 Hugging Face 被披露存在伪装成热门小模型的恶意仓库，借助加载远程代码机制 trust_remote_code=True 投放挖矿程序与后门，恶意仓库累计被下载 3.2 万次，详情

相关启示

- **AI 编程工具本身已经进入真实供应链杀伤链。**智能体可能把令牌、密钥写入响应或日志，凡“模型可见、日志可存、网关可过”的位置都要按敏感凭据管理，缩短凭据有效期、限制其权限范围
- **发布权限的临时扩大是被利用的关键缺口。**本次事故中日常为补丁修复临时给工程师扩大的发布权限被直接滥用，CI/CD 高权限应及时回收、与环境强绑定，并防止单点账号关闭分支保护
- **固定版本、制品签名校验、私有仓库是短窗口投毒最有效的兜底。**latest 标签在本次事件中成为扩散通道；企业还应对 AI 编辑器扩展实施白名单与版本管控，并把开发机的 SSH 密钥基线纳入监控——开发者终端往往同时持有代码仓库与生产凭据

（二）趋势升级：智能体自主攻击首次获受害方一手证实，国家级行为体把多智能体框架投入实战

相关事件

- **专业漏洞披露机构 DIVD 自证被“自治智能体”攻破。**荷兰漏洞披露研究所 DIVD（一家在全球安全社区具有影响力的公益应急组织）9 月通报其自身被入侵：攻击者链式利用开源工单系统 Zammad 的两个当时未知漏洞（CVE-2026-102489、CVE-2026-102490），实现从未授权 Web 会话到 root 权限的完整提权"；DIVD 依据取证中短时间生成的海量攻击脚本、脚本注释里“多个智能体互相对话”等痕迹，判定提权后的后续动作由 AI 智能体自主执行，详情
- **国家级攻击组织用多智能体框架实施供应链攻击。**微软 10 月 21 日披露，与国家背景关联的组织 Storm-0940 自 9 月起使用开源多智能体渗透框架 ARTEX，对航空航天、能源、金融、政府、半导体等十余个行业的数百家组织发起自动化攻击，涉事厂商 ConnectWise 同日认可微软的攻击归因。攻击者借被窃的多租户 OAuth 凭据，对其下游超过 2,250 个 Microsoft 365 租户自动发起登录喷洒，遥测捕获逾 15.1 万次失败请求、峰值每分钟约 2,900 次；在多起得手案例中，从首次瞄准目标到获得云环境特权访问用时不到 3 小时，详情

相关启示

- **“机器速度”正在压过人工响应。**智能体可连续完成喷洒、侦察、提权、横移，安全建设要按分钟、小时级设计检测与自动吊销，针对服务主体行为建立基线并聚合异常，而不是被海量告警噪声拖垮
- **第三方多租户应用的 OAuth 授权是云供应链软肋。**初始访问仍来自被滥用的信任而非“AI 魔法”，最小权限授权、定期轮换应用凭据、审查应用自赋的高权限角色，是直接有效的防护动作
- **网络分段与无补丁应急预案决定事件上限。**DIVD 之所以没有全面沦陷，靠的是分段和第一时间断网；暴露在公网的支撑系统（如已停服的开源工单组件）必须与业务系统同级纳管

重要安全漏洞

以下为 2026 年 10 月公开披露、与 AI 推理/服务框架及智能体基础设施相关的高严重度漏洞（情报整理稿，待安全情报团队复核）。

优先级	漏洞编号	影响范围	漏洞描述	修复方案	出处

严重	官方发布流水线投毒（暂无 CVE）litellm 1.107.0	PyPI 上 litellm==1.107.0 的 7 个 wheel 及 Docker Hub、GHCR 对应镜像；GitHub 源码与源码包未被污染，企业私有版不受影响；最后安全版本为 1.106.2	10 月 21 日，攻击者盗用工程师的 GitHub 命令行登录凭据，接管 LiteLLM 官方发布流水线，由官方流程构建、签名并发布带毒 1.107.0。恶意代码借助 Python 启动钩子在解释器启动时即运行，无需显式导入，窃取 SSH 密钥、云厂商与各模型厂商 API 凭据，并加载远程代码执行工具；带毒产物上线约两小时后被下架	立即停用 1.107.0，将版本锁定到 1.106.2 或官方完成整改后发布的新版本；全面轮换 SSH、云厂商及模型 API 凭据；排查 sitecustomize.py、cook.pth、cook.py 等启动植入文件和指向可疑域名的对外连接	LiteLLM 官方事故报告 官方攻击链复盘
严重	扩展市场投毒（暂无 CVE）anthropic.claude-code 2.1.312、2.1.313	仅经 Cursor 编辑器的扩展市场分发；10 月 29 日 2:27 至 13:35（太平洋时间）在线约 11 小时，Cursor 估计至多约 5,000 名开发者受影响。Anthropic 官方系统、微软 VS Code 市场、官方 npm 包和 Open VSX 均未受影响	篡改版冒用 Anthropic 名义上架，扩展加载时即在正常初始化前执行恶意脚本：收集主机名、用户名、Git 邮箱、公网 IP 及含密钥特征的环境变量，向 authorized_keys 写入攻击者 SSH 公钥，开启本机 SSH、建立反向隧道并配置开机持久化，形成隐蔽后门；原有编程功能保持正常，用户难以察觉	不要从 Cursor 市场安装上述版本；已安装者仅卸载或升级无法彻底清除，需删除攻击者写入的 SSH 公钥与开机自启、计划任务等持久化项并轮换相关凭据；后续从微软官方市场或官方 npm 渠道安装 2.1.314 及以上版本	Cursor 官方安全公告 微软安全分析
严重	**CVE-2026-93674** CVSS 3.1 9.8	Langflow OSS 1.0.0 至 1.12.2	操作系统命令中的特殊元素未被正确中和，导致未认证远程代码执行。攻击者只要能访问 Langflow 服务，无需用户交互即可让服务器以服务账户执行任意系统命令，进而读取模型 API 密钥等敏感凭据	升级到 Langflow OSS 1.12.3 或更高版本；升级前避免将服务直接暴露到公网，仅允许可信网络访问，并排查历史命令执行与凭据访问异常。	IBM 安全公告 NVD
严重	GHSA-w4c7-34wj-hqm8（暂无 CVE）CVSS 3.1 10.0	LiteLLM Enterprise 1.106.5 之前版本	未鉴权任意文件读取，单个请求即可让代理读取本地文件回传攻击者，可导致主密钥泄露	升级至 1.106.5，轮换主密钥与数据库口令	官方公告

严重	CVE-2026-105740 CVSS 9.9	Langflow 1.9.0 之前版本	MCP Server 配置存在操作系统命令注入，已认证用户添加 stdio 类型 MCP 服务时命令直接进入 shell，拉取服务列表即触发，可导致远程代码执行	升级至 1.9.0 及以上，限制可配置 MCP 服务的人员范围	NVD
严重	CVE-2026-103663 CVSS 4.0 9.4	Ollama 0.34.x（0.35.0 之前）版本	/api/pull 存在相对路径穿越，可将恶意二进制写出到模型目录之外；在官方容器以高权限运行等场景下，重启可导致 root 远程代码执行	升级至 0.35.0 及以上，以非特权用户运行并限制管理接口暴露	NVD

安全能力建设

（一）火山引擎 AI 安全网关发布，守住企业“AI 流量关口”

✨ 火山引擎正式发布 **AI 安全网关**，可通过代理、安全插件或旁路方式部署在企业办公网出口，**统一管理企业内 Agent 和 AI 应用交互的所有流量**，实现流量从发现、认证、识别、处置到路由的全链路 AI 安全治理，**让企业对自己的 AI 流量“看得见、管得住、说得清”，详情**



（二）火山引擎安全运营智能体升级，新能力上架豆包工作

✨ 火山引擎安全运营智能体在三个方面实现升级：**新增需求文档安全评审**，把安全把关从评审会前移到方案阶段；漏洞运营升级为“研判+POC 验证”，**漏洞降噪率超 90%**；安全告警研判升级为“多源证据自动关联+定期巡检”，大幅降低人工排查负担。

这些能力已**以“技能/技能包”形式上架豆包工作技能市场**，在豆包工作“**插件·技能**”入口搜索即可安装，并在限时免费试用额度内体验。用户**无需跨工具切换、重复整理任务背景**，在日常工作环境中就能直接调用专业安全能力，让安全评审与风险处置自然融入业务工作流程，详情





1. 新增需求文档评审，提前识别设计风险

需求文档评审支持解析产品需求文档、技术设计文档、提取架构要素、生成数据流图、识别信任边界，并通过 STRIDE 威胁建模、风险评分和合规检查，发现业务流程、数据流向及权限设计中的潜在风险，**将安全把关从开发测试阶段前移到方案阶段，推动问题在编码前解决**。

2. 漏洞运营能力升级，漏洞降噪率达90%

本次升级的漏洞运营能力涵盖漏洞研判与 PoC 验证，**通过智能研判可实现漏洞降噪率90%+**，并可结合漏洞编号、扫描结果及已知资产信息，分析潜在影响、利用可能性和修复优先级，同时说明判断依据。

3. 告警研判升级，从一条告警到一条处置证据链

安全告警研判支持基于告警文本和相关日志分析可疑行为、判断误报可能性，并给出处置建议。在相关数据源已接入并获得授权的前提下，智能体可进一步关联资产、身份、终端和历史事件，梳理事件经过，为风险判断提供证据。