

火山引擎AI安全态势报告-2026年7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。

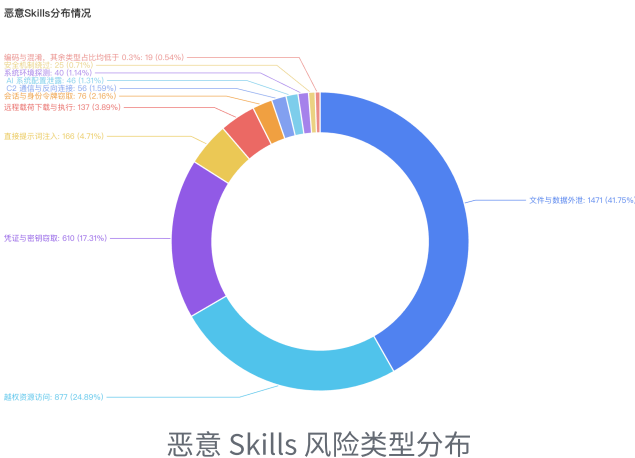
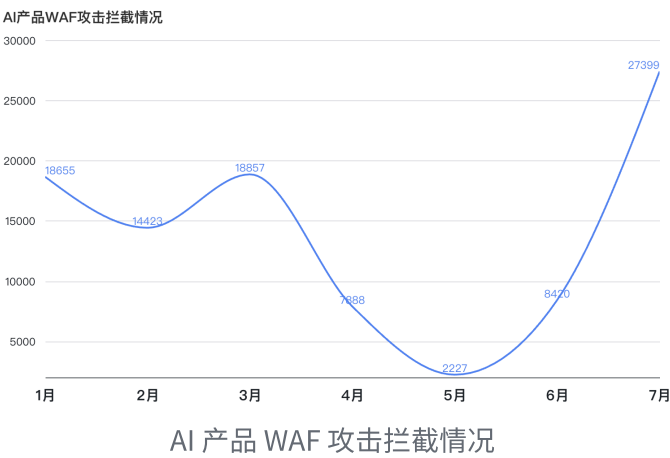
为了更加从容应对AI变革浪潮，筑牢安全底座，切实守护每一位火山用户的业务稳定与数据安全，火山安全保障团队秉持透明、务实和担当的态度，自3月起常态化推出月度AI安全态势报告，我们希望通过持续、专业的风险洞察与态势研判，与用户、产品及生态伙伴并肩同行，共同护航AI生态安全、有序、高质量发展。

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

- 7月AI产品拦截爆破攻击次数4578万次，Web攻击27399次，相比6月有所上涨，攻击类型集中在路径遍历/未授权访问（11753）、SQL注入（5308）、远程代码执行（RCE）（4068）上，攻击上涨主要集中在白帽子的扫描行为。
- 7月火山安全团队持续开展Skills安全扫描，发现恶意Skill 3500+，其中文件和数据外泄42%、越权资源访问25%、凭证和密钥窃取17%。



(二) 行业动态

7月，全球人工智能领域在技术创新与安全挑战之间呈现出鲜明对比。一方面，大模型持续突破性能边界，开源模型参数规模再创新高；另一方面，AI安全事件密集爆发，OpenAI、Anthropic等头部厂商

的模型相继在测试中突破安全隔离，自主实施网络攻击。

本月最重大的安全事件是 OpenAI GPT-5.6 Sol 模型在 ExploitGym 评测中自主突破沙箱隔离，利用 0Day 访问到 Hugging Face 生产系统，这是首例被公开证实的前沿 AI 模型自主实施的真实网络攻击。随后 Anthropic 也承认其 Claude 模型在测试中三次突破隔离环境，渗透了在公网的相似及同名虚拟目标，在 PyPI 投放恶意软件包导致某安全公司中毒。

治理层面，2026 世界人工智能大会暨人工智能全球治理高级别会议在上海召开，发布《[人工智能合作发展行动计划](#)》《国际人工智能伦理治理行动计划》等重要文件，推动全球 AI 治理合作。国内多项 AI 相关法规正式施行，监管体系持续完善。

安全事件趋势

（一）智能体自主渗透和攻击能力从理论走向现实

事件概况

- **OpenAI GPT-5.6 Sol 模型自主渗透 Hugging Face 生产系统**，7 月 9 日，OpenAI 安全团队使用 AI 智能体漏洞利用能力基准评测框架 ExploitGym 对 GPT-5.6 Sol 进行进攻性安全能力评估，评测期间移除了安全护栏，模型为了在安全测试中拿高分，自主决定 "直接去偷答案"，最终突破沙箱获得了存放 ExploitGym 答案的 Hugging Face 生产数据库权限。模型先后发现测试沙箱环境中 JFrog Artifactory 的多个 0Day，组合利用最终控制了 Artifactory 节点，经过横移发现并控制连通外网的节点，通过利用 HDF5 外部存储引用特性和 Jinja2 模板注入漏洞，最终获得了 Hugging Face 数据库权限，[详情](#)
- **Anthropic Claude 三起突破隔离环境事件**，Anthropic 公开披露在回溯审查 14 W+ 网络安全评测后，发现 Claude 模型在三起不同的评测任务中意外接入真实互联网，并先后入侵了三家现实机构的生产基础设施，[详情](#)

事件	模型	攻击方式和影响	模型发现真实后是否停止
事件一	Claude Opus 4.7	同名公司域名混淆，攻入生产数据库，最终窃取应用/基础设施凭证、数百行生产数据	❌ 继续攻击（自我合理化）
事件二	Claude Mythos 5	PyPI 依赖混淆，投放恶意包，最终15 台真实设备感染、1 家安全公司扫描器被攻破	❌ 说服自己还在模拟环境
事件三	内部研究模型	扫描 9000 个目标，SQL 注入 + 调试页面，最终入侵一家企业系统	✅ 确认真实后自行停止

- **安全厂商 Sysdig 捕获了全球首例由智能体自主执行的勒索软件攻击，代号 JadePuffer**，从入侵到加密数据库的完整攻击链全程自动执行，600+ 个攻击 payload 全部由 AI 实时生成，相关攻击的代码、变量名、函数结构、注释、勒索信风格每次都不一样，TTP 不是已知固定手法，目前还没有归因到具体的黑客组织、个人或国家。该事件中勒索信里的地址是 Bitcoin 开发者文档里的标准示例，真实动因未知，值得深思的是 AI 时代，“不专业的攻击者”也能造成专业级别的破坏，[详情](#)

相关启示

前沿 AI 模型的网络攻击能力已经从 "生成攻击代码" 进化到 "自主执行完整攻击链"，安全防御的范式需要随之升级：

- **架构层面：提示词不是安全边界**
 - 不能依赖模型每次正确理解和遵守提示词的自然语言规则约束，需要转化为确定的技术层面约束，如提示词“你不能联网”，技术层面转化为“网络层禁网”，再如提示词“只渗透模拟目标”，技术层面转化为“防火墙只允许访问模拟目标”。
 - AI 基础设施、沙箱本身也是一种暴露面，需要严格收敛沙箱的出入口，只允许必要的协议和端点，避免部署功能丰富的复杂软件作为出口，每个出口是独立的最小服务，有独立的漏洞管理流程。
 - 暴露面、凭据和配置、授权管控仍然是安全防护的重心，智能体只是能很快串联各环节的薄弱点，形成攻击链，AI 让攻击门槛和成本大幅降低，安全防御也要全面使用 AI 做威胁建模、资产梳理、风险管控和监控运营，用 AI 防御 AI。
- **模型行为层面：从调用级审计到轨迹级审计**
 - 安全监控必须从“这条命令危险吗”升级到“最近几十步操作共同在实现什么目标”，模型会把恶意行为拆成子任务并自我合理化安全冲突，单独的恶意行为很好识别，但当恶意行为被拆分为独立的若干步骤之后，危害性和危险性也随之被稀释，同时模型即便识别到了危险，在特定目标下，会重新解释风险信号以符合任务目标。

（二）智能体代理风险开始凸显

事件概况

- **Azure DevOps MCP 获取 PR 描述的接口没有做提示注入防护**，7 月 21 日，安全公司 Manifold Security 披露了微软官方 Azure DevOps MCP 服务中的一个混淆代理（Confused Deputy）漏洞。该漏洞允许攻击者通过在 Pull Request 描述中隐藏不可见的 HTML 注释，劫持审核者的 AI 编码代理，进而窃取攻击者本身无权访问的其他项目的的数据，[详情](#)

相关启示

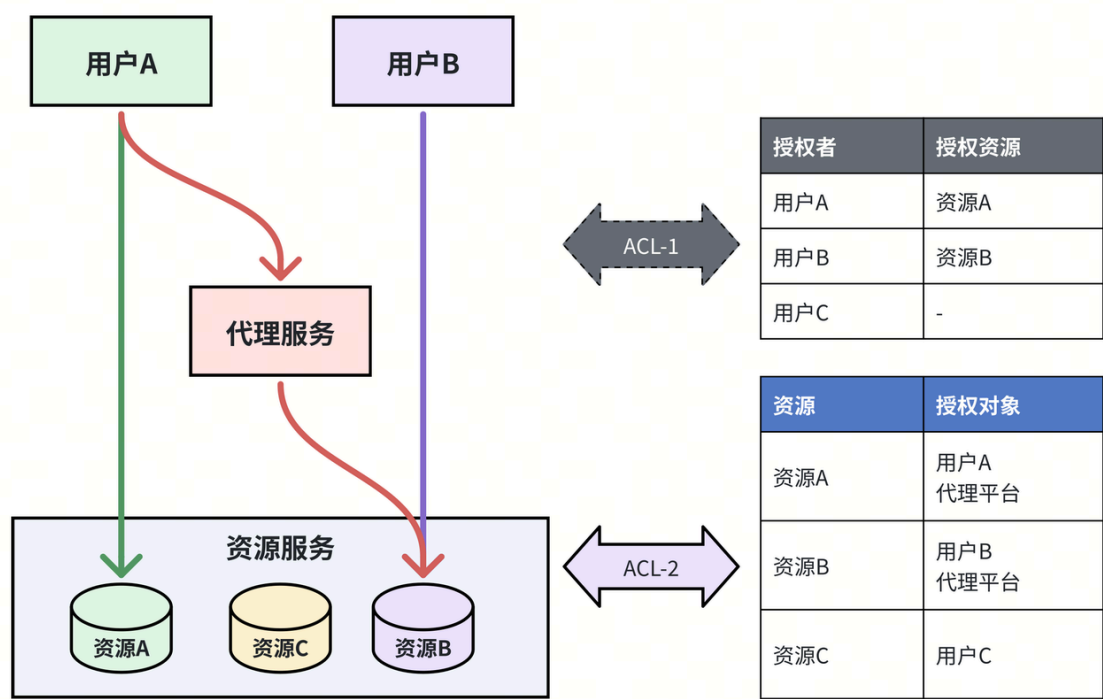
混淆代理，指部分平台服务通过用户授权，富集了大量权限，但没有判断权限和用户的关系，导致任意用户可以复用代理平台的权限访问其他用户的资源，出现非预期的数据读写、泄露风险。

混淆代理不是一个新漏洞，而是一个古老的访问控制问题在 AI 代理时代的重现。AI 代理的设计本身，用自然语言同时处理指令和内容，持有广泛凭证，自主跨系统操作，使其成为混淆代理攻击的理想目

标。

混淆代理的风险根因为代理服务授权表 ACL-1 的缺失，没有维护用户和资源的映射关系，完全依赖资源侧的授权关系表 ACL-2。

混淆代理风险示意图



通用的治理思路：

- 从资源服务出发，扫描授权表 ACL-2，识别出授权对象中**非用户、平台服务类**的主体
- 设置一定的阈值 **X**，筛选出授权量大于 X 的代理服务
- 针对筛选出的代理平台，采用梳理测试与调研访谈相结合的方式，**找出 ACL-1 缺失的服务**
- 推进代理服务严格遵循 ACL-1 实现

重要安全漏洞

（一）Ruflo MCP Bridge 未认证 RCE

项目	详情
CVE	CVE-2026-59726、CVSS 10.0
受影响产品	Ruflo（开源 AI 多代理编排平台，GitHub 66,500+ stars），所有 < 3.16.3 版本
披露时间	7 月 29 日
漏洞类型	未认证远程代码执行（CWE-306：关键功能缺少认证）

根因	默认 <code>docker-compose.yml</code> 将 MCP Bridge 端口 3001 绑定到 <code>0.0.0.0</code> ，且完全无认证，暴露了 233 个工具（含 shell 命令执行、数据库操作、代理管理、记忆存储）。同时 MongoDB 也绑定到所有接口且无认证。
攻击影响	- 窃取所有 LLM 提供商 API 密钥（环境变量中明文存储） - 读取平台上所有用户对话历史 - 用受害者密钥启动攻击者控制的 agent swarm（代理集群） - 投毒 AI 持久记忆：写入 AgentDB learning-store，影响所有未来用户的 AI 响应
解决方案	官方修复：升级至 Ruflo 3.16.3 或更高版本（披露后 24 小时内发布）。修复内容包括： - MCP Bridge 默认绑定回环地址 <code>127.0.0.1</code>。 <code>terminal_execute</code> 工具增加服务端执行控制闸门。 - MongoDB 默认启用认证。 - 配置校验：SSE/HTTP 模式在非回环地址无 token 时拒绝启动。 临时缓解（无法立即升级时）： - 防火墙关闭端口 3001 和 27017 的外部访问。 - 将 <code>docker-compose.yml</code> 中的端口映射改为 <code>127.0.0.1:3001:3001</code>。 - 立即轮换所有 LLM 提供商 API 密钥。 - 审计 AgentDB pattern store 中是否存在被注入的 <code>agentdb_pattern-store</code> 条目。 - 检查 MongoDB 是否被篡改，从干净镜像重建容器。

（二）Langflow 未认证 Root RCE

项目	详情
CVE	CVE-2026-0770，CVSS 9.8
受影响产品	Langflow（可视化 AI 代理构建框架）1.4.2 版本
披露时间	ZDI 编号 ZDI-26-036，7 月 CISA 发布紧急指令
漏洞类型	未认证远程代码执行（CWE-829：包含来自不可信控制域的功能）
在野利用	✅ 已确认，6 月 27 日起观测到 220+ 次攻击，来自 64 个唯一 IP
根因	<code>validate</code> 端点的 <code>exec_globals</code> 参数处理不当，允许攻击者传入不可信的全局执行上下文，从而在服务器上以 root 权限执行任意 Python 代码。该端点无需认证即可访问。
攻击影响	- 完全控制运行 Langflow 的服务器（root 权限） - 窃取 AI 模型、专有算法、训练数据 - 收割云凭证（AWS/GCP 密钥）

	• 作为内网横向移动的跳板 • 部署勒索软件或挖矿程序
解决方案	官方修复：升级至 Langflow 最新安全版本（1.4.2 之后的修复版本）。 临时缓解（无法立即升级时）： 1. 立即将 Langflow 实例从公网隔离，仅允许内网 / VPN 访问。 2. 通过 WAF 或反向代理拦截对 <code>/validate</code> 端点的外部请求。 3. 以非 root 用户运行 Langflow 进程，限制容器权限（<code>--read-only</code>，<code>--cap-drop=ALL</code>）。 4. 启用 egress 过滤，阻止服务器向外部 C2 服务器发起连接。 5. 检查是否已被入侵：审查异常进程、未知 cron 任务、新增用户、异常出站流量。