火山引擎AI安全态势报告-2026年9月

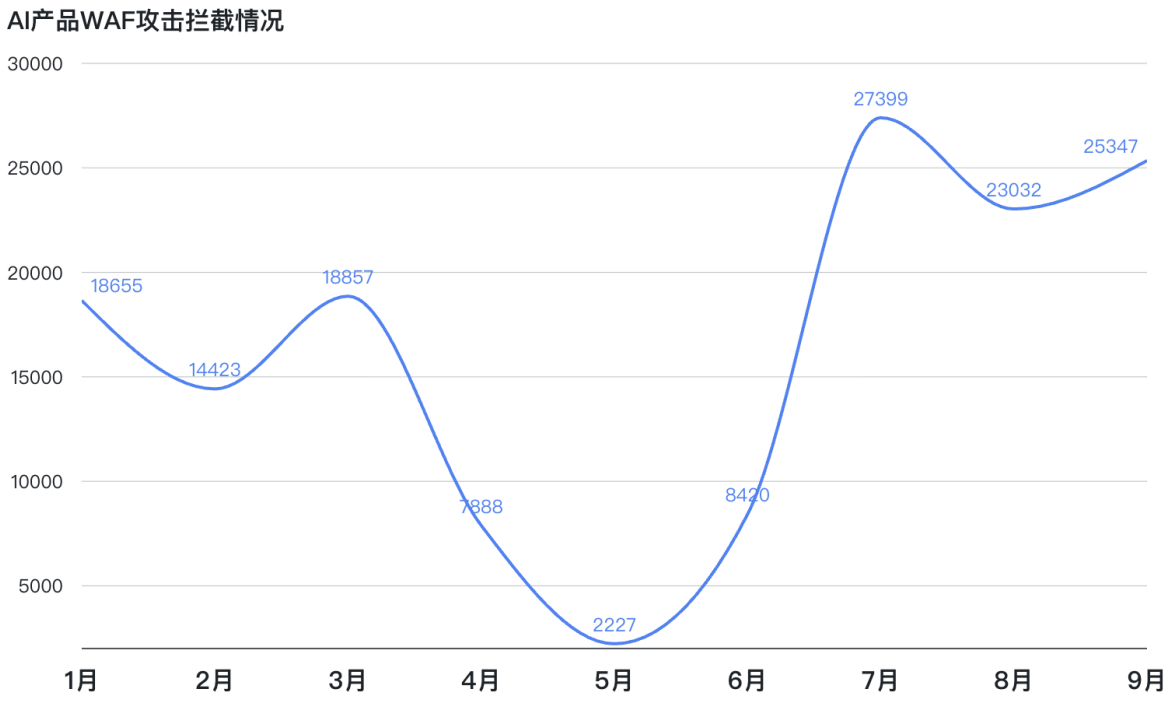
关于AI安全态势月报

> 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

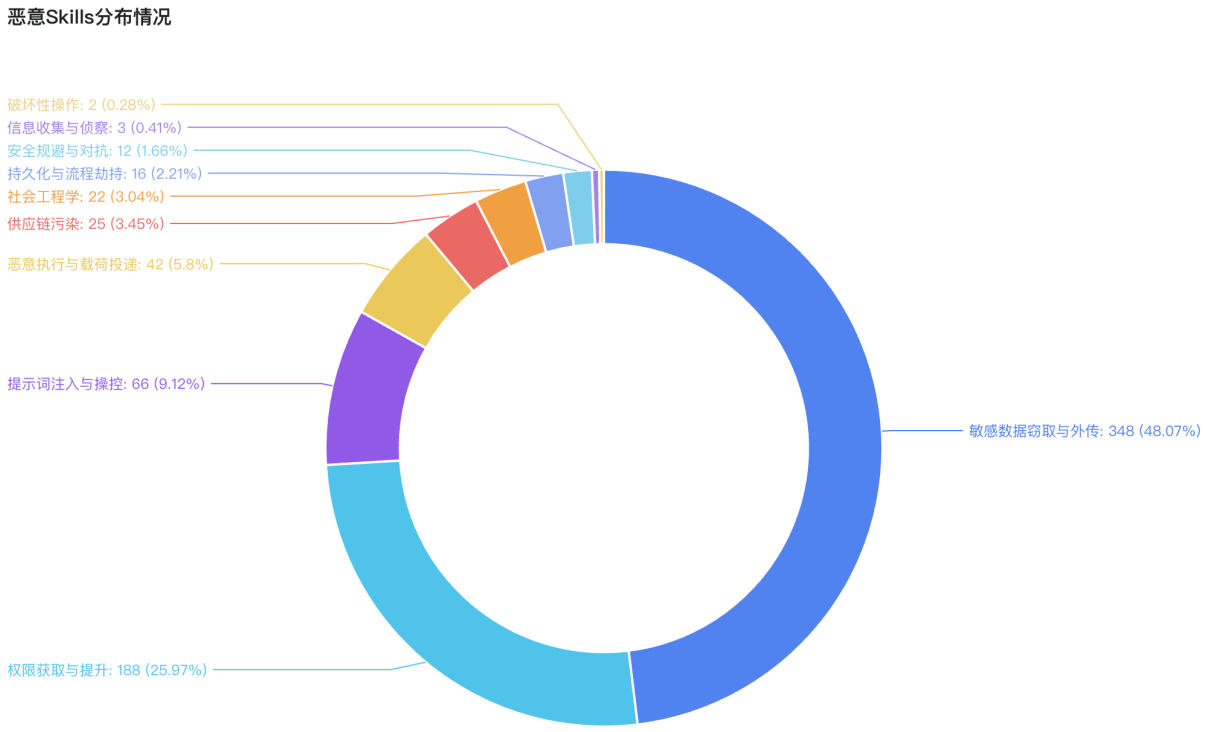
（一）火山引擎 AI 产品攻防动态

- 9 月 AI 产品拦截 Web 攻击 25347 次，相比 8 月上升约 10.1% ，攻击类型集中在路径遍历/未授权访问（9145）、SQL 注入（7214）、远程代码执行（6704）、反序列化（1239）等



AI 产品 WAF 攻击拦截情况

- 9 月火山安全团队持续开展 Skills 安全扫描，发现恶意 Skill 724 个，TOP 的风险有敏感数据窃取与外传（占比 48%）、权限获取与提升（占比 26%）、提示词注入与操控（占比 9.1%），详情见下图



恶意 Skills 风险类型分布

（二）行业动态

- **智能体发现的漏洞和智能体本身的漏洞大量爆出**
  - 2026 年全球累计CVE数量已达 **66401 个**，较 2025 年同期的 33512 个接近翻倍，AI 让大量潜伏多年的漏洞被快速识别，而人工修补节奏跟不上发现节奏已成新常态，详情
  - CNCERT 组织的 2026 年人工智能大模型安全众测对 25 家厂商的 54 款产品测试，累计发现 873 个漏洞，其中**大模型及智能体特有漏洞 608 个（占比约70%）**，典型风险包括提示注入、信息泄露、代理权限滥用、目标劫持及非预期代码执行，详情
- **9 月 AI 治理从原则倡导进入机制建设阶段，安全治理成为各类大会重要议题**
  - 9 月 14 日，全国网络安全标准化技术委员会在国家网络安全宣传周开幕式上发布**《人工智能安全治理框架3.0》，**该框架在国家网信办指导下编制，延续“风险分类、技术应对、综合治理”核心逻辑，3.0版本的治理对象从“模型输出”扩大为“能感知、推理、记忆、调用工具并持续行动的智能系统”，详情

风险大类	风险明细	应对措施
一、内生安全风险（模型自身缺陷）	1. 幻觉输出：生成虚假 / 错误事实	RAG 检索增强、多模型交叉校验、输出事实核查机制
	2. 安全护栏可被越狱 / 绕过	对抗性红队测试、输入输出双层护栏、多轮安全微调
	3. 提示注入 / 间接提示诱导	输入隔离、指令与数据分离、权限最小化
	4. 非预期自主行为（Agent 偏离人类指令）	行为审计日志、高危操作人在回路审批、指令约束硬化、异常行为自动熔断
	5. 算法偏见 / 歧视、可解释性不足	公平性去偏、模型可解释性工具、偏见检测数据集评估
	6. 训练数据投毒 / 污染、隐私泄露	训练数据全生命周期管控、投毒检测、隐私计算脱敏
二、应用安全风险（部署运行后产生）	1. 智能体工具越权调用 / 沙盒逃逸	①最小权限原则，工具白名单制 ②沙盒网络多层隔离（不止 HTTP/HTTPS，管控 DNS/UDP/ICMP 等底层协议） ③禁止危险 DNS 记录类型（TXT）查询 ④同时监控模型层工具调用 + 基础设施层系统行为（DNS、socket、系统调用）
	2. 外部真实系统越权访问（读 + 写）	①Agent 网络出口白名单，禁止访问公网真实生产系统 ②写操作强制人审 ③测试环境与真实世界网络硬隔离 ④历史运行记录回溯审计
	3. AI 赋能网络攻击 / 自主网络威胁	①WAF、补丁等传统加固仍然有效 ②部署侧攻击意图识别，检测到扫描 / 漏洞利用行为即服务降级拒答 ③IOC 情报联动与自动封禁
	4. Agent 身份冒充、记忆篡改、错误规划	身份认证与权限绑定、记忆存储完整性校验、规划步骤可审计可回滚
	5. 深度伪造、违法内容生成、认知操纵	输入输出双重内容过滤、AI 生成内容强制标识、人机复核
	6. 个人信息非法采集 / 推断 / 泄露	隐私计算、数据最小化、敏感特征禁止推断
	7. 物理实体嵌入风险（AI 控制现实设备）	高危功能裁剪、物理急停装置、场景化安全边界拆解
	8. 高危技术滥用（生化 / 武器研发辅助）	源头拦截高危请求、敏感领域请求人工复核
三、衍生安全风险（次生社会风险）	1. 就业冲击、数字鸿沟扩大	技术适配培训、普惠性 AI 服务可及性设计
	2. 能耗与环境压力	绿色算力调度、模型推理效率优化
	3. 信息茧房、舆论操纵、文化侵蚀	内容多样性保障、操纵型内容检测、文化偏见缓解
	4. 跨境治理与责任界定模糊（AI 自主行为无明确人类指令）	①建立 AI 越界事件分级通报机制（初步确认即通知受害方，不等全部查清） ②明确研发方 / 使用方责任边界 ③国际治理规则对接

人工智能安全风险一览表（火山安全团队整理）

- 9 月 23—25 日，中美元首会晤，双方同意**建立中美人工智能对话**，交流 AI 相关风险和惠益，下次对话于 11 月举行，双方同意**建立人工智能事件的沟通渠道，**中方在会谈中提出共同防范 AI 被滥用恶用、保持人类对 AI 技术的管控

安全事件趋势

（一）OpenAI Agent 沙箱逃逸并越权访问政府统计系统

相关事件

- 9 月 20 日，OpenAI 披露在受控沙盒测试环境中，一个用于信息检索任务的智能体绕过网络限制，通过 DNS 解析访问了一个第三方的服务（OpenAI 称为公共聊天机器人服务），该服务支持把子域名作为问题，并通过 TXT 记录返回问题的答案，该智能体提问了 20 次，详情
- 9 月 28 日，OpenAI 披露曾在 6 月测试中未经授权访问了澳大利亚全民医疗保险计划 Medicare 的统计数据门户，读取了部分非公开文件并有写操作。这是 OpenAI 在 7 月的 Hugging Face 事件发生后，内部进行系统审查，在 8 月中旬发现，但于 9 月 10 日才通知奥方相关部门，详情

相关启示

- **沙盒与网络隔离：不能只防护七层 HTTP/HTTPS，底层协议会成为隐蔽侧信道**
  - **传统网络防护思维存在局限**，沙盒只拦截了 HTTP/HTTPS，默认 DNS 只是域名解析基础设施，忽略 DNS 本身可以作为双向数据隧道。智能体复用环境自带 DNS 解析服务就完成对外通信

    - **模型提示词不等于安全边界**，即便提示词约束模型 “不能访问互联网”，安全必须由环境、网络、权限强制执行，不能寄希望于模型 “自觉遵守规则”。模型为完成任务会自主探测环境边界，寻找一切可用系统能力做逃逸
- **智能体任务需要建立允许熔断机制**，任务难度高、常规路径走不通，Agent 会不断尝试替代路径，如 DNS 隧道、探测外部真实网站、尝试越权等。需要增加 “放弃机制”：任务无法完成时允许智能体停止，而不是无限冒险寻找旁路
- **事件响应、通报与治理**，不要等完全查清所有事实之后才通知受害方，AI 自主越界事件，很多细节很难短时间完全还原，可以设置分级通报，初步确认存在越界行为，就第一时间联系受影响实体，后续持续补充调查结果。应急比较及时的案例：9 月智谱 AI 编程工具 ZCode 被曝默认将含 Git 历史的工作区数据加密上传至云端且用户无法关闭，智谱 72 小时内完成致歉、整改、开源，并由中国信通院与绿盟科技审计确认云端数据清零，详情

（二）AI智能体“蜂群”利用 PaperCut 漏洞大规模攻击

相关事件

- 威胁情报机构 GreyNoise 披露，攻击者编排数百个基于 OpenAI Codex 和 DeepSeek 模型的智能体，将打印管理软件 PaperCut NG/MF 的 CVE-2026-81578（认证绕过） 与 CVE-2026-82078（不安全反射远程代码执行） 漏洞链武器化，配合 Mimikatz、Certipy、Impacket 等工具自主完成渗透，行动波及48 个国家、395 家机构、440 台服务器沦陷，其中 204 个是教育机构。攻击者在编排时设置了 28 个黑名单国家（禁止攻击俄、中、伊朗等），但智能体自主偏离指令，仍然对黑名单内目标发起探测攻击，详情

相关启示

- 即使大模型自带安全护栏，攻击者依然可以组合多种模型开展全球入侵活动，护栏无法防御攻击者自己侧构建智能体攻击流水线
- **传统安全加固仍然有效对抗 AI 驱动的威胁**，如WAF、补丁、域安全加固等防御手段依然可以阻断攻击，是当前应对智能体大规模攻击的有效手段

重要安全漏洞

漏洞编号	影响范围	漏洞描述	修复方案	出处
CVE-2026-93088	v0.5.11 <= SGLang <= v0.5.20	DiffusionServer 未认证直接 pickle.loads() ZeroMQ 消息，导致未认证远程代码执行；编码器节点默认绑定 0.0.0.0。	通过防火墙/安全组隔离 ZeroMQ 端口，关闭未用功能并升级为上游修复版本	研究披露 NVD

CVE-2026-93034	v0.5.18 <= SGLang <= v0.5.20	SGLANG_USE_PICKLE_IPC=false 时未移除 pickle 反序列化，恶意对象在重建阶段可执行代码，导致未认证远程代码执行。	通过防火墙/安全组隔离 ZeroMQ 端口，关闭未用功能并升级为上游修复版本	研究披露 NVD
暂无	Plugin4Shell	属于零点击远程代码执行漏洞，影响Anthropic Claude Code、OpenAI Codex、GitHub Copilot、Google Gemini CLI等。该漏洞利用Git仓库SHA-pinning机制的缺陷，允许控制插件仓库的攻击者绕过哈希锁定，在插件更新时将受信任代码替换为恶意代码。	Agent 拉取插件后，必须读取磁盘实际代码的 commit 哈希，和配置预期 hash 做比对，不一致直接拒绝加载插件	研究披露

安全能力建设

（一）火山引擎 AI 安全新升级：AgentSentry 让企业放心用 AI

火山引擎 AgentSentry 全新发布。作为面向企业 Agent 集群的统一安全管理平台，AgentSentry 打通三大独立安全产品——智能体身份与权限管理平台、智能体安全管理平台、AI 助手安全平台，以统一入口和界面为企业提供跨平台、跨生态、跨运行环境的 Agent 纳管服务。

1. 企业多类型Agent统一防护纳管

研发侧 Coding Agent、业务侧 Workflow Agent、员工侧数字员工 Agent 来自不同厂商，分布在云端、服务器和员工终端。AgentSentry 可以将它们集中纳管，统一呈现安全状态、身份权限、供应链组件、运行行为和风险事件，帮助企业建立全局资产视图，帮助安全团队在统一界面上掌握并管控全局。

2. 智能体全生命周期安全防护

- 自定义高危操作：AgentSentry 能结合操作语义和执行上下文识别风险，在删除关键文件、修改生产配置、调用敏感接口等危险动作发生前进行提醒或拦截。
- DLP 商业敏感数据防护：AgentSentry 提供了更精细的敏感数据防护策略：Agent 可以读取和处理企业数据，但企业可以自定义识别规则，识别文档密级、敏感内容，根据上下文判断访问是否合理、数据是否被不当使用，让数据在使用中被保护，而非在封锁中失去价值。
- 完整行为链路追踪与审计：AgentSentry 记录 Agent 的工具调用、数据访问、权限使用和关键动作，把这些串成完整的行为链路。风险事件发生时，企业安全团队可以快速还原过程、定位影响范围；面对合规审计时，也能拿出完整、可追溯的证据链。



火山引擎

01 智能体合规准入

- 智能体业务用途和边界定义
- 智能体分类分级
- 组件白名单
- 智能体可接受行为准则
- 智能体安全模板

02 智能体资产与供应链安全

- MCP与工具采集管理
- Skills采集管理
- 智能体采集管理
- 模型文件采集管理
- 知识库采集管理
- AI-BOM 组件清单
- 运行环境安全
- 模型安全
- MCP与工具安全
- Skills安全
- 提示词安全

03 内容安全合规

- 多模态内容安全检测
- 敏感话题控制
- 安全代答
- 深度鉴伪

04 常态化安全测评与加固

- 安全合规测评
- 安全对抗测评
- 业务可控测评
- 智能体安全加固

05 AI安全网关

- 智能体接入
- 敏感数据识别&脱敏
- 数据跨境管控
- 智能模型路由
- 资源耗尽防护与熔断
- 成本治理与配额控制

06 身份与认证管理

- 智能体唯一身份
- 智能体认证凭据
- 委托链身份
- 短生命周期令牌认证
- 硬件背书身份与凭证
- 凭据托管

07 权限与访问控制

- 多主客体细粒度权限管理
- 链式访问控制
- 持续访问控制
- 任务驱动的实时动态提权
- 上下文感知动态权限
- 多智能体信任边界

08 运行时安全监控与防护

- 微虚拟机隔离
- 沙箱化执行环境
- 提示词注入防护
- 不可信内容隔离
- 记忆投毒检测处置
- 敏感数据输出过滤
- 工具参数执行前校验
- 输出语义风险分析

09 安全可观测性与运营管理

- 智能体动作全量日志
- 不可变审计轨迹
- 关联分析和全链路溯源
- AEBA智能体事件行为分析
- UEBA用户与实体行为分析
- 自动化处置和响应

10 模型与推理安全

- 远程证明
- 可信密钥
- 端到端机密性
- 透明可信监控

L1
智能体基础安全防护
（01–04）

L2
智能体精细化管控
（05–08）

L3
智能体安全持续运营
（09–10）

智能体安全能力图谱

扫码获取完整方案

支持：
WorkFlow 智能体 | AI Coding 智能体 | 办公助手 智能体