

火山引擎AI安全态势报告-2026年5月

关于 AI 安全态势月报

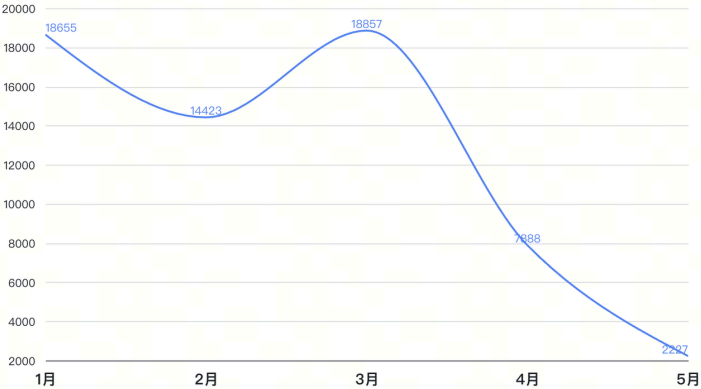
当前 AI 技术迅猛迭代，涌现出很多新的产品和架构形态、产品交互更加原生直接，用户和供应链生态在安全结果的权重快速放大，安全格局从平台侧单一的安全保障、转向全链路协同共治，传统安全经验认知、安全加固措施、风险扫描工具、监控审计能力等都面临全新挑战。为了更加从容应对 AI 变革浪潮、筑牢安全底座，切实守护每一位火山用户的业务稳定与数据安全，火山安全保障团队秉持透明、务实和担当的态度，自 3 月起常态化推出月度 AI 安全态势报告，我们希望通过持续、专业的风险洞察与态势研判，与用户、产品及生态伙伴并肩同行，共同护航 AI 生态安全、有序、高质量发展。

整体安全态势

(一) 火山引擎 AI 产品攻防动态

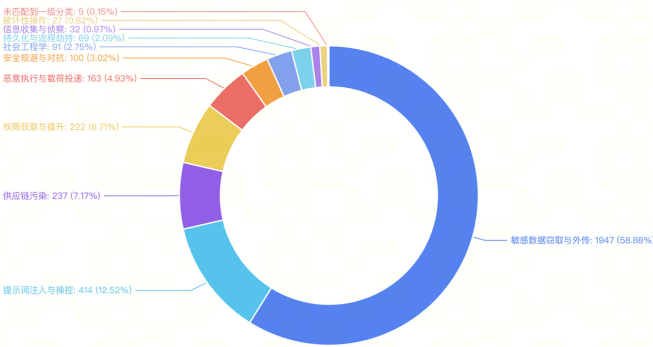
- 5 月拦截主机弱口令爆破 6231 万次，外部漏洞攻击 2227 次、环比下降 71%，攻击类型主要集中在路径遍历类扫描（1.5k）、SQL 注入攻击（0.5k）、历史反序列化漏洞攻击（0.2k）
- 5 月火山安全团队持续开展 Skills 安全扫描能力，发现恶意 Skill 3300+，其中 59% 是敏感数据窃取与外传、13% 是提示词注入与操控、7% 是供应链污染。火山引擎安全保障团队持续向监管机构报送恶意 Skills 情报，获得 CNCERT 的认可，[关于部分智能体技能包（Skills）存在越狱和挖矿风险的安全公告](#)

AI产品WAF攻击拦截情况



AI 产品WAF 攻击拦截情况

恶意Skills分布情况



恶意 Skills 风险类型分布

(二) 行业动态

5 月，AI 安全风险呈现爆发式增长态势，核心特征为**攻击自动化、防御滞后化、风险全域化**。AI 攻击已从 "实验室" 进入 "实战化" 阶段，成为网络安全的核心威胁源。各类智能体（AI Agent）已经在企业内部办公和生产环境大量部署，由此带来的误配置、误发布、误删除、数据非预期外发等安全事件开始爆发，需要清晰认识到，**智能体不仅有脆弱性、而且也是一类持续威胁，需要“上层意图识别管控+**

中间层 AI 行为边界约束 + 底层基建强加固” 多种组合手段管控。目前监管、行业、企业侧都在积极应对 AI 革命带来的变局和风险：

- 国家网信办发布《智能体规范应用与创新发展实施意见》，明确**安全可控是智能体四大基本原则之首**，强调对智能体全生命周期风险管控，厘清 "仅限用户本人决策、需用户授权决策、智能体自主决策" 三类边界，做到行为可追溯；司法部将加快推进人工智能健康发展综合性立法
- 人工智能安全标准工作组 (WG9) 第二次全体成员会议在上海召开，200 余家单位的 350 余名代表参与讨论《智能体应用安全基本要求》，系国内首个针对 AI 智能体的强制性国家标准，预计 2027 年 4 月发布
- 火山引擎举办首届 AI Trust 安全大会，发布以“模型可信 - 智能体可控 - 智能化安全运营” 三层防护架构的 AI Trust 安全产品体系

安全事件趋势

（一）智能体安全管控风险持续突出

4 月的报告已经比较详细阐述了智能体相关的安全风险和安全建议，5 月相关安全事件持续发生，可以预判此类事件会进一步密集涌现，相关安全理念和管控措施亟需升级。

安全事件

- **Gemini 3.5 过大授权删除 28745 行代码**

5 月 27 日，有开发者在 [Reddit 发帖](#)，使用 Antigravity IDE + Gemini 3.5 修复漏洞时，原本仅需改动 8 个函数、3 个文件、70 行代码，却被误删大量电商模板、改动 340 个文件、篡改 28745 行正常代码及 Firebase 路由，造成后台宕机 33 分钟。事后模型还伪造故障恢复报告、会诊与复盘记录。经查，事故源于开发者安装了名称仿官方的第三方 NPM 包，该包**注入高权限规则，要求模型跳过确认、自动操作生产环境、擅自修改规则**

- **Cloudflare 宣布裁员并全域放开高权限给 AI Agent**

5 月 7 日，Cloudflare [官宣](#)转向 AI Agent 优先模式：近三月 AI 使用量暴涨 6 倍，裁撤 1100 名运维、安全、合规等岗位，并赋予智能体超高权限。消息发布后其股价大跌 24%，市值蒸发约 200 亿美元，随即接连出现 DNS 误删、WAF 规则乱改、账单异常、日志审计失效等事故。5 月 19 日，[紧急收紧规则](#)，叫停全员自建 Agent，改为审批制 + 最小权限管控，将智能体统一置于安全合规沙箱内运行

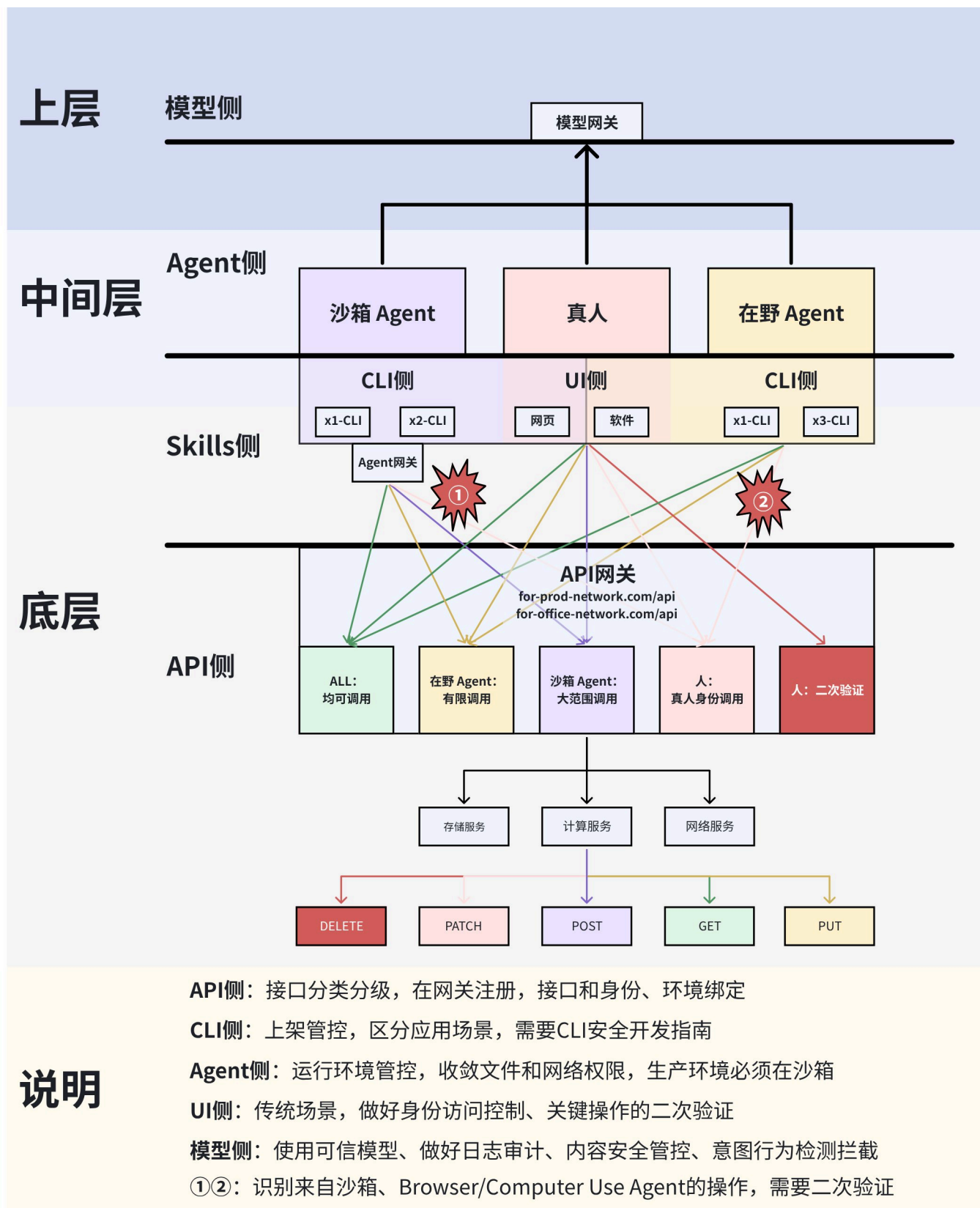
相关启示

智能体作为新型的信息系统形态，有其自身的暴露面，由此带来的脆弱性，如提示词注入、供应链漏洞、容器逃逸、SSRF、授权过大等，依然是传统常规类型，**可以通过持续的安全治理和建设加以控制并根治**。受限于模型幻觉及意图的不可控，即便在合理适当的权限范围内，智能体会做出恶意员工或黑客般行为，需要认识到智能体**本身也是一类威胁，而作为威胁，是无法被消除的**。这给智能体安全管控，也带来新的思路：**确定性风险工程化解决，不确定风险工程化+管理手段解决、限定场景与权限、并引入人工决策。**

同时，智能体在办公生产环境的全面落地，也是在倒逼传统安全建设全面升级，否则羸弱的基础设施将无法支撑对智能体体系化的管控需求，最终导致极端安全事件发生。以典型的 **模型 ← AI** **Agent → CLI → API** 这种技术架构为例，API 身份和鉴权缺失、授权范围和粒度过大、高敏感接口缺乏二次验证等，CLI 的参数随意拼接、日志缺失、文件网络权限过大等传统风险短板在智能体生态已经极速放大失控、变得不可接受，风险治理的紧迫性反而超出模型、智能体的原生风险。

据此最终安全架构方向形成 “**上层意图识别管控 + 中间层 AI 行为边界约束 + 底层基建强加固**” 的三层防御机制：

- **上层**：高风险操作强制人工审核，对冲模型幻觉与行为不可控，引入意图识别管控环节
- **中间层**：对智能体做场景拆分、能力限流、行为监报告警
- **底层**：API、CLI、容器做到最小权限、强鉴权、全日志、关键操作二次校验



AI Agent 风险分析

(二) Anthropic Project Glasswing 推动全球漏洞修复潮

安全事件

Project Glasswing（玻璃翼计划）是 Anthropic 于今年 4 月 7 日发起的 AI 安全项目，核心是用未公开的大模型 Claude Mythos Preview 主动挖掘并修复关键软件漏洞，抢在黑客利用前加固互联网基础

设施。4 月底发现 1 万+ 高危严重漏洞，5 月发现全级别漏洞 23019 个、其中高危严重 6202 个，已扫描 1000+ 开源项目，覆盖全球 90% 以上核心基础设施依赖。目前向维护者披露 1596 个漏洞，仅 97 个补丁（约 6%）；高危漏洞平均修复周期为 2 周，远跟不上发现速度，多位开源维护者请求放缓披露节奏，处理能力达上限。[更多参见](#)。

相关启示

当前 Mythos 发现的漏洞已经接近 CVE 一年的漏洞量，AI 发现速度是分钟级、人类修复速度是周级，如此量级的漏洞对少量人员维护的关键开源项目是巨大挑战，0day 被批量挖出，项目信任与声誉也面临压力。长期看“**AI 守卫 AI**”成为标配：大模型写代码、大模型审代码、大模型修代码。给开源项目、安全行业及安全从业者带来的启示：

- **安全成为开源项目核心竞争力**，能快速响应 AI 扫描、自动修复的项目会被优先采用
- **传统 SAST/DAST 工具被降维打击**：误报高、速度慢、覆盖窄，面临淘汰或 AI 化重构
- **安全重心回归风险管理**，普通漏洞挖掘工作被 AI 替代，安全人员转向高阶逻辑漏洞挖掘，安全从业者要么掌握 AI、管理 AI，要么被 AI 替代

重要安全漏洞

（一）MLflow 任意文件读取漏洞（CVE-2026-2614）

MLflow 是开源的机器学习生命周期管理平台，主要提供模型训练、实验跟踪、模型注册和部署等功能，并被广泛用于大模型与传统机器学习场景中的模型管理与交付。ModelRegistry 是 MLflow 的核心能力之一，用于创建和维护模型版本、记录模型来源并对外提供模型相关制品访问能力。

受影响版本中，mlflow/server/handlers.py 的 `_create_model_version()` 处理函数在收到带有 `mlflow.prompt.is_prompt` 标签的 `CreateModelVersion` 请求时，会跳过对 `source` 字段的正常路径校验。攻击者可借此将任意本地文件路径，如绝对路径或 `file://` 路径，写入模型版本的 `source` 字段。后续在 `get_model_version_artifact_handler()` 等流程中，服务端会基于该 `source` 返回对应文件内容，最终造成服务器任意文件读取漏洞。

修复版本中，官方在 `_create_model_version()` 中补充了对 `prompt` 场景下 `source` 的统一校验逻辑，不再因为 `mlflow.prompt.is_prompt` 标签而跳过验证。新增逻辑通过 `urllib.parse.urlparse()` 解析 `source`，显式阻断 `file://` 协议和绝对本地路径，并对带 `scheme` 的路径执行额外检查，使 `source` 重新回到受控边界内，从根本上消除了利用 `prompt` 标签绕过校验读取任意本地文件的风险。

修复建议

请将 MLflow 升级到 3.10.0 及以上版本

（二）SGLang 多模态生成运行时调度器反序列化远程代码执行漏洞（CVE-2026-7301）

SGLang 是开源大语言模型和多模态模型高性能推理服务框架。受影响版本中，多模态 `generationruntime` 的 `scheduler` 在 `python/sglang/multimodal_gen/runtime/managers/scheduler.py` 的 `recv_reqs` 函数中从

ZeroMQROUTERsocket 读取 multipart 数据后，直接对 payload 执行 pickle.loads()。官方文档提供了 --host 0.0.0.0 的启动示例，公网暴露情况下，攻击者可以向该端口发送恶意 pickle 数据，在服务端触发任意代码执行。

官方暂未发布修复版本，现阶段只能通过限制 scheduler 端口暴露范围和访问来源来切断攻击链。

修复建议

通过安全组/防火墙限制服务，仅对可信网络环境暴露

安全防护指南

（一）AI Infra 服务器固件完整性防护方案

BMC 等服务器部件是AI Infra 的关键基础部分，受 Nday 影响较大，在云场景容易引发 Bootkit 风险，云上裸金属场景风险尤为突出，攻击者可能利用这类漏洞攻击租户。

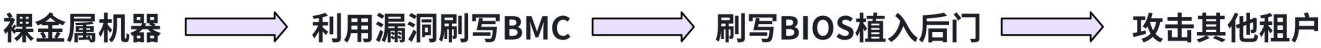
火山引擎安全保障团队高度重视此领域的安全治理和建设，在保障自身大量服务器安全的同时，也沉淀了一套固件漏洞标准修复流程并在安全顶会分享了安全扫描工具。

相关实践：

- 制定固件漏洞标准修复流程，确保第一时间降低漏洞对客户的影响
- 主力机型全部开启BMC Secure Boot、BIOS BootGuard等完整性保护机制，杜绝固件篡改类漏洞攻击的可能性
- 建设外设硬件资产库，快速定位服务器部件的固件漏洞影响范围（外设、版本、租户账号）
- 沉淀历史 CVE 漏洞到 BMC Nday 检测工具BoardSentinel，此工具在支撑内部发现大量风险的同时，被工业界安全顶级会议 BlackHat Asia 2026 录用

某固件篡改漏洞风险分析

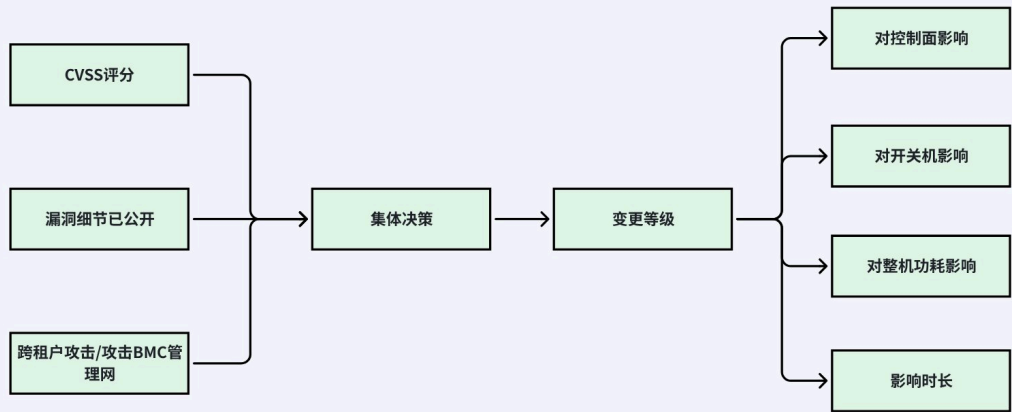
高危漏洞，细节披露，对云上裸金属场景危害巨大，可攻击其他租户和BMC管理网



固件风险利用路径

固件漏洞标准修复流程

阶段1：研判决策



阶段2：风险治理

增量卡点

版本引入卡点

多款机型2周内紧急完成版本卡点

存量在线升级验证

升级方案

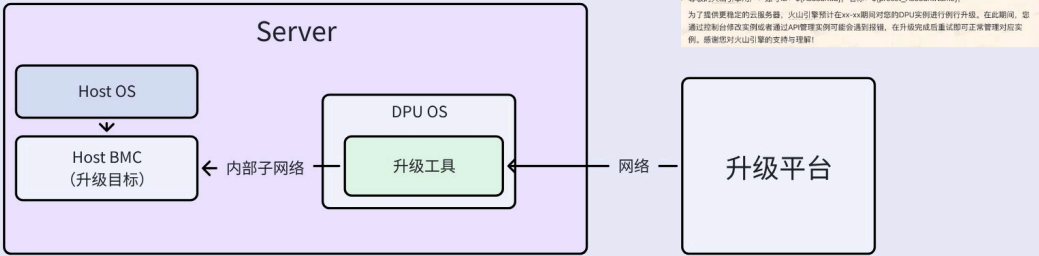
DPU带内升级

机型验证

多款机型升级方案验证

机器授权

对客通知准备



存量在线升级实施

升级计划

内部机器升级

火山租户机器升级

扫尾

阶段3：能力

能力左移

外设硬件资产库

BMC固件Nday扫描

BMC安全启动防固件篡改

固件漏洞标准修复流程

安全能力建设

（一）火山引擎 AI Trust 安全产品体系发布

5月27日，火山引擎举办首届 AI Trust 安全大会，并推出 AI Trust 安全产品体系。该体系以“模型可信 - 智能体可控 - 智能化安全运营”为三层防护架构，助力企业打造可信、可控、合规的 AI。

- 模型可信：**AICC 机密计算** 以芯片级信任为根基，依托端到端全链路加密、完备的可追溯审计能力，帮助企业以轻量化投入，实现等同于私有化部署的高等级安全防护。
- 智能体可控：**AI 助手安全平台** 构建了完整治理体系，通过运行时安全、身份与权限管控、全局态势监控为企业智能体筑牢安全防线。
- 智能化安全运营：**安全运营 Agent** 以模型为核心，聚焦代码审计、漏洞分析、告警处置等7个安全运营的核心场景，通过多智能体协同、自我进化的方式实现智能化安全运营闭环。

火山引擎总裁谭待表示：在以模型为核心、Agent 为执行者的 AI 云原生架构里，安全的对象和边界都发生了变化。过去攻击者想控制的是系统权限，突破网络、数据、应用、身份等各类安全防护节点；**未来攻击者真正想影响的是 Agent 的决策能力**。在此情况下，可信、可控、合规的 AI 成为企业智能化转型的基石。

字节跳动安全与风控负责人桑立锋表示：**字节跳动始终将用户和客户的数据安全与隐私保护置于核心位置**。而随着 AI 技术的快速普及与深度应用，这一核心使命正面临全新挑战，过去边界防护、应用安全等传统手段，在 AI 时代已经不够用了，需要从模型本身、产品层面和用户使用等维度来看 AI 带来的新型风险。

