

火山引擎AI安全态势报告-2026年6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。

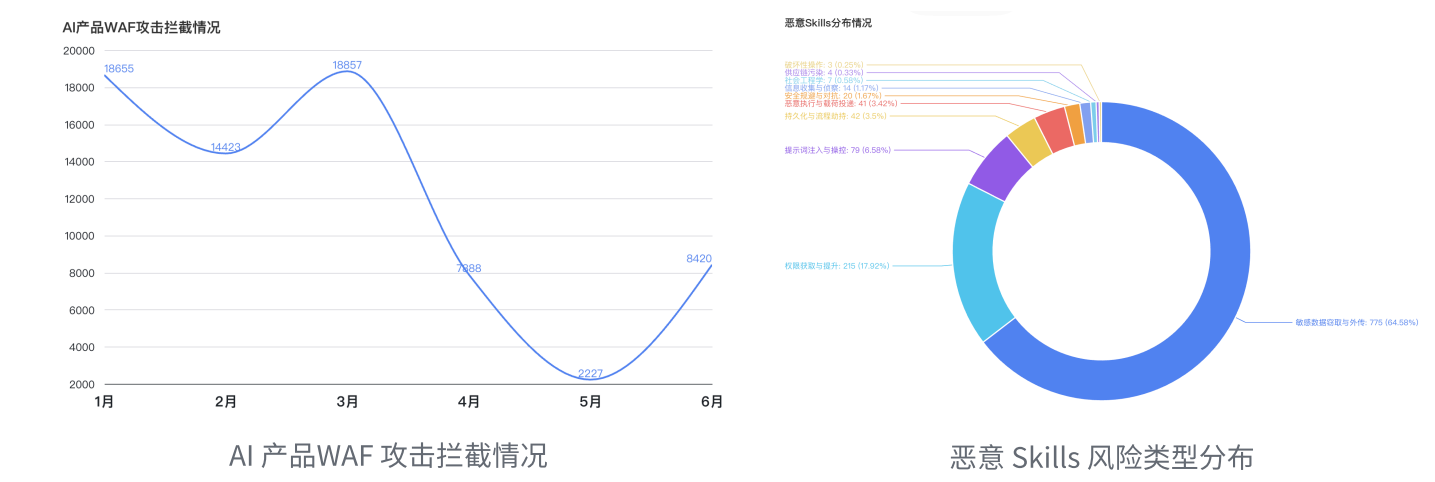
为了更加从容应对AI变革浪潮、筑牢安全底座，切实守护每一位火山用户的业务稳定与数据安全，火山安全保障团队秉持透明、务实和担当的态度，自3月起常态化推出月度AI安全态势报告，我们希望通过持续、专业的风险洞察与态势研判，与用户、产品及生态伙伴并肩同行，共同护航AI生态安全、有序、高质量发展。

免责声明：本安全月报所梳理外部安全事件信息综合公开网络、安全社区资讯整理，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

- 6月AI产品拦截Web攻击8420次，相比5月大幅增加，主要来自白帽子的渗透测试。攻击主要集中在路径遍历/文件包含（3.2k）、SQL注入攻击（1.8k）、Log4j漏洞攻击（940）、敏感文件探测（680）、反序列化攻击（576）
- 6月火山安全团队持续开展Skills安全扫描，发现恶意Skill 1200+，其中65%是敏感数据窃取与外传、18%是权限获取与提升、7%是提示词注入与操控



(二) 行业动态

6月AI安全的主战场整体迁移到“智能体（AI Agent）安全”，攻、防、监管三方同时围绕Agent的身份、运行时与供应链重新布局，AI安全正从“模型对齐与内容安全”，进入“Agent全链路攻防”

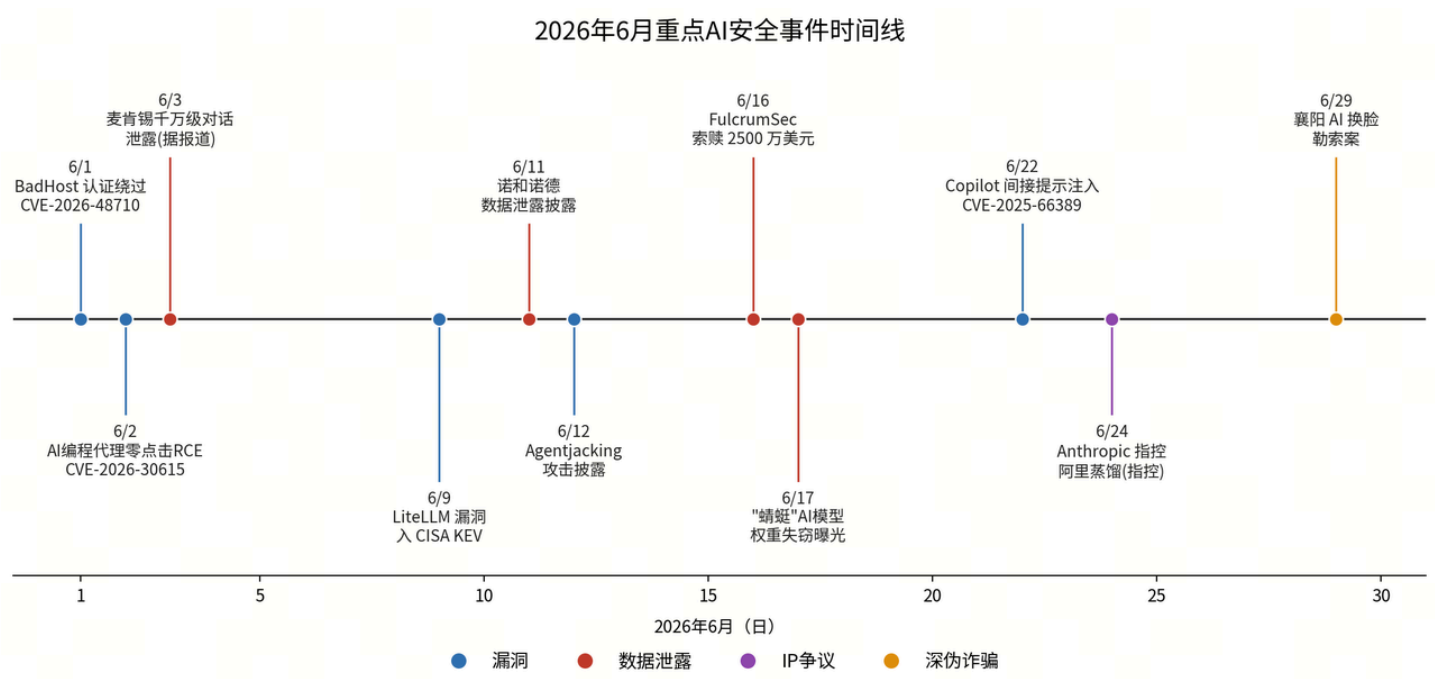
阶段。

6月是安全大会的密集窗口，各大安全会议喊出的口号高度趋同——“智能体正在颠覆安全”。从北京到伦敦、从拉斯维加斯到日内瓦，议程主线都从“如何用AI”转向“如何在AI被滥用前守住防线”。

6月火山引擎 Force 大会在北京顺利落幕，大会指出 Agent 智能体规模化商用，安全是前置底座，并重磅发布“可信、可控、合规”的 AI Trust 智能体安全体系，是国内唯一同时拿下智能体可信、安全防护双信通院认证的产品解决方案，以期解决企业“想用 AI 但不敢上核心数据”的行业痛点。

安全事件趋势

本月安全事件不在于某一起的烈度，而是沿着“智能体攻击面”持续铺开——从底层框架漏洞，到AI网关与编程代理，再到模型权重本身被当作核心资产窃取，这启示我们：在大模型安全领域，纵深防御已经从传统加分加固手段变成不可替代的核心安全架构，需要通过“多层独立、异构防护、一层失效、多层兜底”机制，层层压缩风险爆炸半径。6月安全事件脉络如下图：



(一) 智能体风险爆炸半径持续扩大

事件概况

- Instagram 智能体客服缺陷导致大量账号被劫持

3月Instagram上线AI智能客服机器人，AI客服拥有直接修改账号绑定邮箱、下发密码重置链接的高权限接口，缺少对Owner的校验，无二次核验，攻击者伪装成目标账号常用登录地区、绕过基础地域风控，通过问答如“原邮箱已丢失，账号@xxx需要到新邮箱”，AI即向黑客邮箱发送校验码、直接执行换绑请求，4月17日漏洞首次被攻击者利用、5月底达到顶峰、6月1日官方确认漏洞已修复，[详情](#)

- ChatGPT 官方 Google Sheets 插件间接提示词注入

5月5日，OpenAI 官方发布 ChatGPT for Google Sheets 正式版，5月8日安全厂商 PromptArmor 向 OpenAI 提交间接提示词注入高危漏洞，受害者打开投毒表格、并在侧边栏调用 ChatGPT（汇总、翻译、公式生成、数据清洗任意指令即可触发，无需点击可疑链接、下载程序）即受影响。安全厂商 31 日公开漏洞和完整攻击链。6月1日，OpenAI 紧急上线热修复，关闭脚本生成功能。多家海外企业用户批量撤销该插件 Google Drive 授权、禁止员工使用相关插件处理敏感业务数据，[详情](#)

相关启示

1. 智能体身份管控

围绕智能体存在多种身份：**任务触发者身份**（发起身份，**Subject-S**）、**智能体主体身份**（Agent 本体身份，**Agent-A**）、**CLI / 后端接口执行身份**（执行身份，**Executor-E**），上述安全事件的根源是三层身份的失衡。其中 Subject-S 身份直接映射资产所有者、Agent-A 身份预绑定高权限 Executor-E 凭证、Subject-S 交互请求无需二次核验、直接复用 Agent-A 身份绑定的 Executor-E 执行操作都是非常高风险的操作。相关安全实践：

◦ 身份传导阻断

解决 Subject-S 伪造穿透到 Executor-E

- i. 不能将任务触发者 Subject-S 的浅层特征（IP、地域、会话、设备）作为签发高危 Executor-E 执行身份的唯一依据，需要叠加独立资产所有权校验（原邮箱、人脸、实名、短信多因子）
- ii. 智能体禁止预绑定长期高权限 Executor-E 凭证，所有接口调用必须实时申请单次临时执行身份
- iii. 区分“只读类”和“权属变更 / 写操作类” Executor-E，写操作强制脱离智能体自动执行链路、引入人工确认

◦ 中间层隔离

解决指令投毒冒用 Executor-E

- i. Agent 内部做输入源隔离：区分「用户手动输入可信指令」和「外部文件 / 网页 / 对话第三方不可信内容」，两类输入分配两套隔离执行上下文，不可信输入禁止调用高权限 Executor-E
- ii. 第三方插件必须使用最小权限临时代理 Executor-E，禁止完整继承 Subject-S 的身份和授权
- iii. 智能体仅做语义理解，无权限决策能力，所有权限由独立的鉴权网关（安全决策中心）管控

◦ 日志与溯源分层

审计日志必须独立记录三层身份字段，缺一不可

- **Subject-S**：会话 ID、设备、IP、发起主体标识
- **Agent-A**：智能体实例 ID、服务类型、插件编号
- **Executor-E**：接口凭证 ID、授权范围、执行操作、操作目标资源；出现越权行为可快速定位：是 Subject-S 伪造、Agent-A 权限配置失误，还是 Executor-E 管控缺陷

2. 智能体纵深防御

◦ 架构与权限

- i. 最小权限硬约束：禁止给智能体分配超出任务必需的高危权限（账号改绑 / 密码重置、脚本执行、跨资源全量访问等），账号找回、凭证变更等核心安全流程禁止通过 AI 对话入口触发
- ii. 决策与执行强分离：LLM 仅负责意图识别，所有敏感操作的关键参数（用户 ID、目标邮箱、操作范围）必须来自服务端可信上下文（登录态、数据库绑定记录），绝对不采信 LLM 从用户输入中解析的敏感参数
- iii. 权限隔离：智能体权限与用户权限做切割，禁止默认继承用户全量权限（如表格插件仅能访问当前打开文件，不得授权整个云盘访问权）

◦ 输入防护

- i. 输入源分级隔离：严格区分用户主动指令（可信）与外部数据（导入表格 / 文档 / 第三方接口返回，不可信），不可信数据仅作为处理对象，禁止解析为 AI 指令，阻断间接提示词注入
- ii. 工具调用硬校验：LLM 输出工具调用请求后、执行前，增加确定性规则校验层，不依赖 System Prompt 软约束，直接拦截与用户原始意图不符的调用（如用户要求总结数据却触发脚本生成、跨文件读取）

◦ 身份认证

- i. 高危操作强校验：改绑、密码重置、数据外发等操作，必须触发独立于 AI 对话流程的多因子验证（原手机 / 邮箱验证码、2FA 动态码），禁止仅靠 IP 地域等易伪造弱特征判断身份
- ii. 会话强绑定：禁止在未登录状态下处理任何账号相关请求，AI 所有操作必须与用户已认证会话强关联

◦ 运行时防护

- i. 工具沙箱：智能体可调用工具默认运行在受限沙箱，非必要不开放代码执行、外发网络请求能力；即使开放脚本能力，也需在无高权限、无外网的隔离环境运行
- ii. 异常熔断：实时监控 AI 行为，对短时间批量敏感操作、意图不符的工具调用、向未知域名外发数据等异常行为，直接中断操作转人工审核
- iii. 失效安全设计：防护机制失效时默认拒绝操作，AI 无法判断请求合法性时直接转人工，禁止“默认放行”

◦ 兜底防线

- i. 保留传统安全底线：2FA 双因素认证为账号安全最后防线，AI 能力不替代传统安全机制（本次 Instagram 事件中开启 2FA 的账号全部未被攻破）
- ii. 快速降级开关：为 AI 高危能力设计一键关停能力，漏洞爆发时无需发版即可快速切断攻击路径

iii. 权限透明：明确告知用户智能体 / 插件的真实权限范围，企业端支持配置 AI 敏感操作禁令

◦ 审计与响应

- i. 全链路结构化日志：完整记录 AI 意图识别、工具调用、参数传递、执行结果全流程，关联真实用户身份，支持攻击溯源与特征提取
- ii. 快速漏洞响应：建立 AI 安全专属响应通道，缩短修复窗口（AI 漏洞从利用到大规模爆发通常仅 1-2 个月）

（二）诺和诺德数据泄露事件

事件概况

3 月黑客组织 FulcrumSec 利用诺和诺德的安全漏洞入侵其内部系统，6 月 1 日黑客联系诺和诺德高管勒索，诺和诺德明确拒绝支付赎金，于 6 月 11 日发布公告承认内部 IT 系统遭未经授权访问、部分数据被窃取，6 月中旬黑客组织开始在暗网公开部分数据样本、并私下拆分出售完整数据集，[详情](#)。

黑客单方爆料（官方未核实），共窃取 1.3 TB、超 70 万份文件，包括企业源代码、已上市及在研药物（包括 GLP-1 类重磅产品管线）的专有研发数据、临床试验完整数据，**内部“蜻蜓”AI 药物研发项目全量资产：多模态模型权重、专有训练数据集、模型源代码、训练日志、容器镜像（相当于其 AI 制药整条管线被窃取）**，全球工厂生产设施的机密运营文件等。

相关启示

此次事件导致 AI 模型和在研管线数据泄露可能导致其在 GLP-1 领域的技术领先优势被削弱，AI 模型本身已成为网络攻击的高价值目标，且黑客已开始使用 AI 自动化攻击工具，医药行业网络攻防对抗烈度显著升级。

AI 研发环境成为关键突破口，黑客不再随机扫描通用漏洞，而是针对企业高性能计算集群、AI 训练服务器、研发内网实施持续性横向渗透；AI 研发环境普遍存在大量开放算力接口、共享容器、批量数据传输通道，边界防护薄弱，极易成为入侵内网的跳板。

企业 AI 数据资产的保护指南：

• AI 资产分级隔离与全域台账管控

- 建立完整 AI 资产清单，分级标注涉密等级：
- 分级部署隔离架构：核心 AI 训练集群独立安全域，与办公、测试网物理 / 逻辑隔离，跨域访问强制安全网关、多因子认证
- 离线算力、本地模型存储介质闭环管控，禁止私自拷贝、外传完整模型文件

• AI 专属数据防泄露与模型知识产权保护

- 训练数据侧：原始临床、商业机密数据训练前全脱敏，敏感字段不可复原；部署 AI 专用 DLP，识别模型权重、海量样本批量传输行为，实时拦截外带
- 模型资产侧：核心模型加密存储、部署时混淆处理，添加不可移除数字水印，被盗后可溯源；禁止完整模型对外导出，对外交付仅开放 API 推理能力，不交付权重文件

- 外部输入隔离：区分可信内部实验数据与第三方、外部导入文件，外部数据禁止写入一级核心训练数据集，防范数据投毒
- **AI 场景专属应急与持续风险运营**
 - 针对 AI 资产设计独立勒索 / 泄露应急预案：算力集群隔离切断、模型备份异地加密存储、涉密资产快速冻结导出权限
 - 常态化 AI 安全红队演练，模拟 APT 横向渗透、模型批量窃取、容器漏洞利用攻击
 - 区分“公有第三方 AI 工具”与“企业私有自研 AI”两套管控规范，严禁研发人员使用未审批公共 ChatGPT 等工具处理未脱敏管线、实验数据，规避影子 AI 泄露风险

重要安全漏洞

（一）SpringAI 向量存储模块查询注入漏洞（CVE-2026-47835）

SpringAI 是 Spring 生态中的开源 AI 应用开发框架，用于在 Java 应用中集成大语言模型、Embedding 模型、向量数据库等 AI 能力，并为 RAG 应用提供统一的向量存储抽象。本次漏洞影响 SpringAI 的向量存储过滤表达式转换逻辑，涉及 Elasticsearch、OpenSearch、GemFire 等向量存储实现。

受影响版本中，AbstractFilterExpressionConverter 类的 emitLuceneString() 方法在处理元数据键时，未对制表符、换行符、回车符和全角空格等特殊空白字符进行转义。攻击者如果能够控制传入向量检索元数据过滤表达式的字段名或键值，就可以借助这些未转义字符拼接 OR、AND 等查询操作符，改变原本的查询语义，从而在向量数据库后端执行非预期查询。

修复版本中，官方在 emitLuceneString() 的 switch 分支中新增对 \t、\n、\r 和 和 的处理，输出前统一添加反斜杠转义，使这些字符只能作为普通查询内容参与匹配，不能再作为查询语法边界或操作符注入点使用。

漏洞明细

代码块

```
1 // AbstractFilterExpressionConverter.java
2 // 修复前：未覆盖 \t、\n、\r、 等特殊空白字符
3 switch (c) {
4     ...
5     context.append('\').append(c);
6 }
7
8 // 修复后：新增特殊空白字符转义分支 case '\t':
9 case '\n':
10 case '\r':
11 case ' ':
12     context.append('\').append(c);
```

修复建议

- 将 org.springframework.ai:spring-ai-elasticsearch-store 升级到1.0.9、1.1.8或以上版本

（二）LiteLLM 认证绕过漏洞（CVE-2026-49468）

LiteLLM 是开源的大模型API代理网关，用于统一代理和管理 OpenAI、Anthropic、Azure、Bedrock 等多个模型服务接口，常被部署在企业内部作为模型调用入口、鉴权层和额度管理层。

受影响版本中，LiteLLM 代理鉴权逻辑在判断请求路由时使用 request.url.path。Starlette 在构造 request.url 时会受到 Host 请求头影响，攻击者可通过构造包含/?或/#的恶意 Host 头，使 request.url.path 被解析为/。由于/属于LiteLLM 的公开路由，认证中间件会误认为当前请求无需鉴权，从而绕过认证访问受保护的管理接口。

修复版本中，官方将路由判断依据从 request.url.path 改为直接读取 ASGI scope["path"]。该字段是 FastAPI 实际用于路由匹配的权威路径，不会被攻击者控制的 Host 头重写。同时，修复逻辑还处理了 root_path 前缀剥离，保证代理挂载在子路径时仍能正确识别真实请求路径。

漏洞明细

代码块

```
1  # litellm/proxy/auth/auth_utils.py
2  # 修复前：路由判断依赖 request.url.path, 可能受到 Host 头影响
3  route = request.url.path
4  # 修复后：改为读取 ASGI scope 中的真实 path
5  scope = request.scope
6  raw_path = scope.get("path")
7  root_path = scope.get("root_path")
8  if root_path and root_path != "/" and raw_path.startswith(root_path):
9      return raw_path[len(root_path):]
10 return raw_path
```

修复建议

- 将LiteLLM 升级到1.84.0及以上版本

安全防护指南

（一）火山引擎 SkillScan 智能体技能安全防护体系

智能体技能安全防护从**全维度安全检测能力、全生命周期安全保障、多层级纵深防御方案、持续化安全运营闭环**四个层级系统构建 SkillScan 安全防护体系，形成覆盖“事前预防、事中检测、事后响应”的闭环防护能力，确保智能体技能生态全程安全可控。

全维度安全检测能力

覆盖五大风险场景，提供多维度安全检测：

1. **包体文件合规风险**：检测可执行文件、硬编码凭证、目录穿越、内容安全等基础风险
2. **声明层安全风险**：检测提示词注入、敏感行为声明、高风险依赖、声明与代码一致性问题
3. **代码层安全风险**：检测恶意代码、已知漏洞、容器逃逸、敏感路径访问、日志敏感信息泄露等风险
4. **网络与资源使用安全风险**：检测未加密通信、恶意网络连接、C2通信特征、资源耗尽等问题
5. **开源合规与供应链安全风险**：检测开源许可证合规性、依赖组件漏洞、供应链威胁等风险。以上检测能力均为默认配置，无需额外开启

全生命周期安全保障

覆盖技能从开发到上线全生命周期的安全能力：

1. **开发阶段**：提供安全开发最佳实践指导，包括最小权限、输入验证、依赖安全、敏感信息保护等规范，从源头降低风险。
2. **发布阶段**：支持离线批量扫描、在线实时扫描、CI/CD集成、OpenAPI 接入四种模式，实现“不扫描、不上架”的安全卡点。存在以下高风险项的技能禁止准入（一票否决）：内容安全风险、提示词注入、未通过评估的敏感行为声明、恶意代码、容器逃逸/敏感信息收集、敏感路径访问、C2 特征网络连接、许可证严重不合规。
3. **上线阶段**：通过持续运营保障技能安全稳定运行。

多层次纵深防御方案

以“静态扫描 + 动态分析 + 多引擎协同”为核心，构建“静态检测 + 动态验证 + 智能研判”的多层防护体系：

- **核心技术能力**：
 - a. 静态扫描：无需运行技能即可发现大部分安全风险，平均耗时 116.85 秒
 - b. 动态分析：隔离沙箱中运行检测静态扫描遗漏的风险，平均耗时 271.51 秒，成功率 91.97%
 - c. 证据约束的大模型语义分析：高危严重风险准确率 97.72%，召回率 92.39%
 - d. 多源检测结果统一映射：整合多引擎输出，结构化呈现风险
- **分场景定制防护**：
 - 云上技能市场：采用“入口层+内容层+运营层”三层检测策略
 - 内场技能共享：额外增加内网敏感信息检测、数据外发检测、权限合规检查能力

持续化安全运营闭环

建立风险发现-定级-处置-复测的完整运营流程：

1. **风险发现**：扫描生成结构化风险报告

- 2. 风险定级：按严重/高危/中危/低危定级
- 3. 通知处置：推动技能负责人修复
- 4. 复测验证：修复后重新扫描验证

安全能力建设

(一) 火山引擎智能体一体化安全解决方案

6月23日，火山引擎举办2026 Force大会，在AI Trust安全论坛上，火山引擎云安全团队分享了智能体一体化安全解决方案。这个“一体化安全解决”设计为三层架构：

- **上层是接入与调度层**——企业内的智能体五花八门，AI Coding类、智能体平台类、AI助手类，接入方式各不相同，调度层对不同智能体做统一对接和接入管理，在智能体之间、安全能力之间做总控调度
- **中间层是安全能力综合层**——安全攻防、行为权限、供应链和环境安全、行为护栏、审计溯源五大能力被重新组合，形成面向智能体场景的综合性防御体系
- **底层是安全模型能力层**——用AI对抗AI，通用大模型和安全垂类小模型组合，为上层安全能力提供检测、处置、响应、分析的智能支撑



附录

六月重点大会一览

大会	时间 / 地点	主办方	AI安全关键看点
Gartner 安全与风险管理峰会	6月1-3日 / 美国 National Harbor	Gartner	强调"在智能体失控前先管住它"，并点名四大关键威胁[20]
	6月2日 / 北京	奇安信	

北京网络安全大会 BCS 2026			主题"AI时代，攻防先行"；齐向东主张"三位一体"内生安全[21]
Infosecurity Europe 2026	6月2-4日 / 伦敦 ExCeL	Infosec urity	聚焦Agentic AI；现场演示开源LLM驱动的自适应蠕虫[22]
CVPR AdvML@CV 工作坊（第6届）	6月4日 / CVPR	学术界	主题为视觉-语言智能体（VLA）的安全[23]
SANS ICS 安全峰会	6月8-10日 / 美国 Orlando	SANS	工业控制系统防护的实战技术训练[24]
Identiverse 2026	6月15-18日 / 拉斯维加斯	SailPoi nt	聚焦"智能体经济"下非人类身份的爆发式增长[25]
ISA OT 网络安全峰会	6月16-18日 / 布拉格	ISA	工业安全标准与供应链风险[26]
UNIDIR AI、安全与伦理全球会议	6月18-19日 / 日内瓦	UNIDIR	从国际和平与安全视角审视AI影响[27]
第十四届互联网安全大会 ISC.AI 2026	6月24日 / 北京	360集团	主题"智能体颠覆安全"；发布"倚天屠龙"攻防双能力[28]