



清华大学
Tsinghua University

CSA GCR cloud security
GREATER CHINA REGION alliance®



北京信息科学与技术
国家研究中心
BEIJING NATIONAL RESEARCH CENTER
FOR INFORMATION SCIENCE AND TECHNOLOGY



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Safety Conference

面向智能体时代的 大模型安全 研究和实践

绿盟科技

顾杜娟



范式之变

01

风险之变

02

攻击面之变

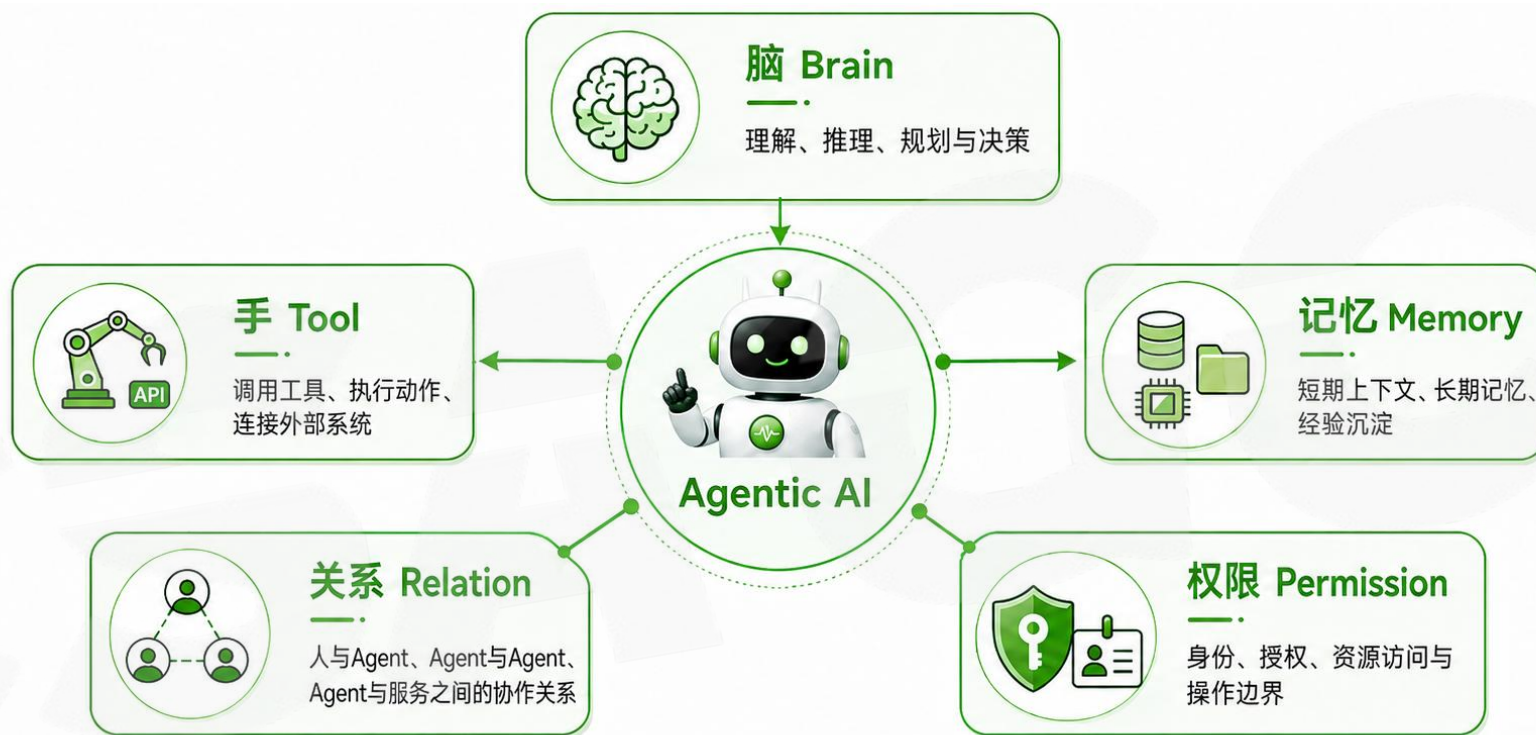
03

防御之变

04

体系之变

范式之变：从内容生成器“会说什么” 到Agentic AI 自主执行者“会做什么”



面向智能体时代的大模型安全 为什么需要**新安全范式**？

案例1：工具被劫持，合法调用变隐蔽数据外泄

2025-09-25 Postmark / Snyk：postmark-mcp为恶意MCP工具，8天，1643次下载，约300个真实生产环境被感染。
成千上万封包含API密钥、密码重置链接、客户PII的邮件——全部、自动、悄无声息地流向了攻击者。



1

此前 / 2025-09-17 之前

T0 首次发布

首次发布正常版本，作为 MCP 邮件工具进入生态。



2

此前 / 2025-09-17 之前

T1 建立信任

v1.0.1~v1.0.15 连续正常更新，逐步建立用户信任与依赖。



3

2025-09-17

T2 植入后门

发布含后门的 v1.0.16，功能表面正常，但已埋入恶意逻辑。



4

2025-09-17 之后

T3 正常调用

Agent 正常调用 send_email，工具继续以合法权限运行。



5

2025-09-17 至 2025-09-25

T4 持续外泄

每次发送邮件时，秘密增加 BCC，邮件被同步抄送给攻击者。



6

2025-09-25

T5 披露与处置

事件被发现并披露，恶意包下架，随后开展审计、追溯与处置。

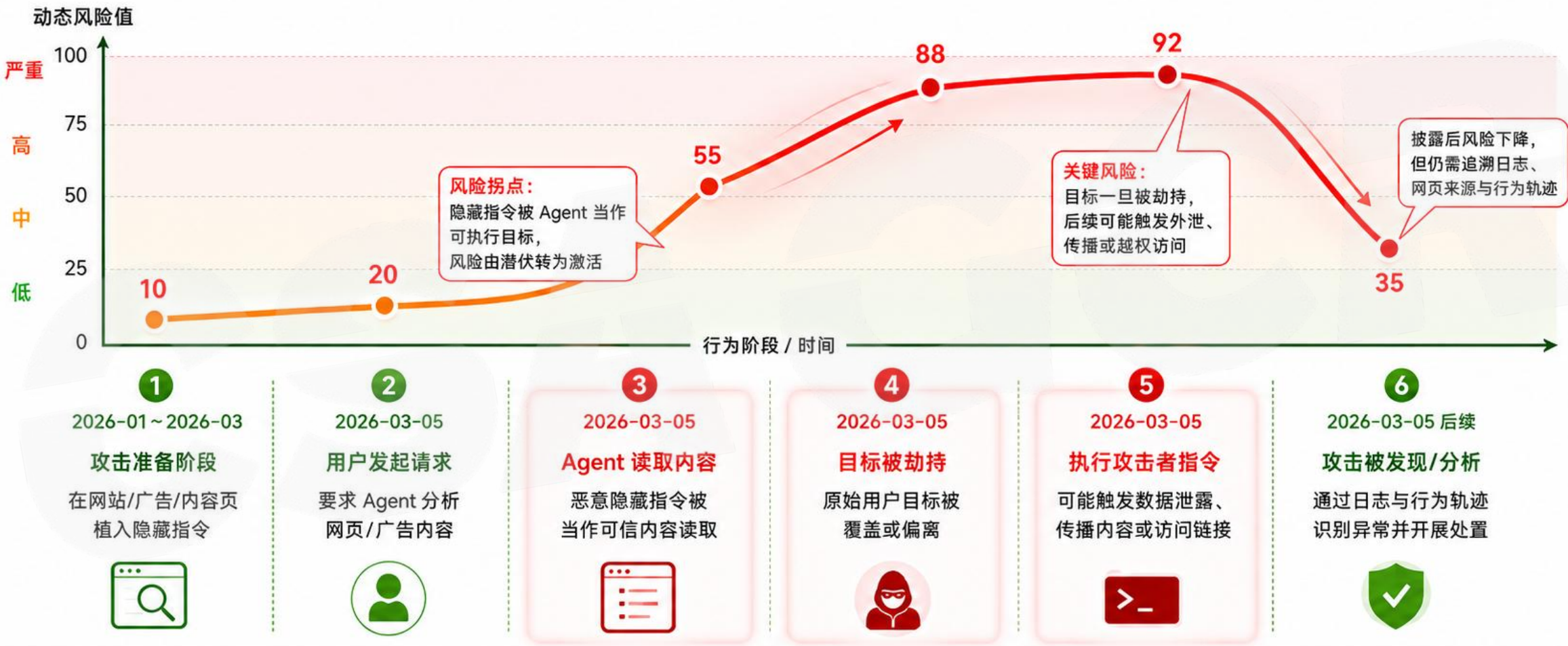




关键启示：工具可用 ≠ 工具可信；必须在准入、升级、调用、出站与追溯全链路设置防护点。

案例2：目标被篡改，隐藏指令操纵致偏离业务意图

2026-03-05 Unit 42：基于大规模真实监测发现，间接提示注入(IDPI)已在野利用，攻击者已经用22种技术系统性武器化这种攻击，涉及SEO投毒、数据销毁、未授权交易等。



关键启示：网页内容可见 ≠ 指令可信；必须在读取、解释、执行与追溯全链路设置防护点。

案例3：记忆被污染，风险从单次攻击跨会话持久化



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Security Conference

2026-02-10 Microsoft：攻击者通过隐藏Prompt诱导AI将恶意偏好写入持久记忆，不同用户每次访问都会被自动加载，已观测到31家企业、14个行业、50+种污染Prompt，一次攻击，长期生效。

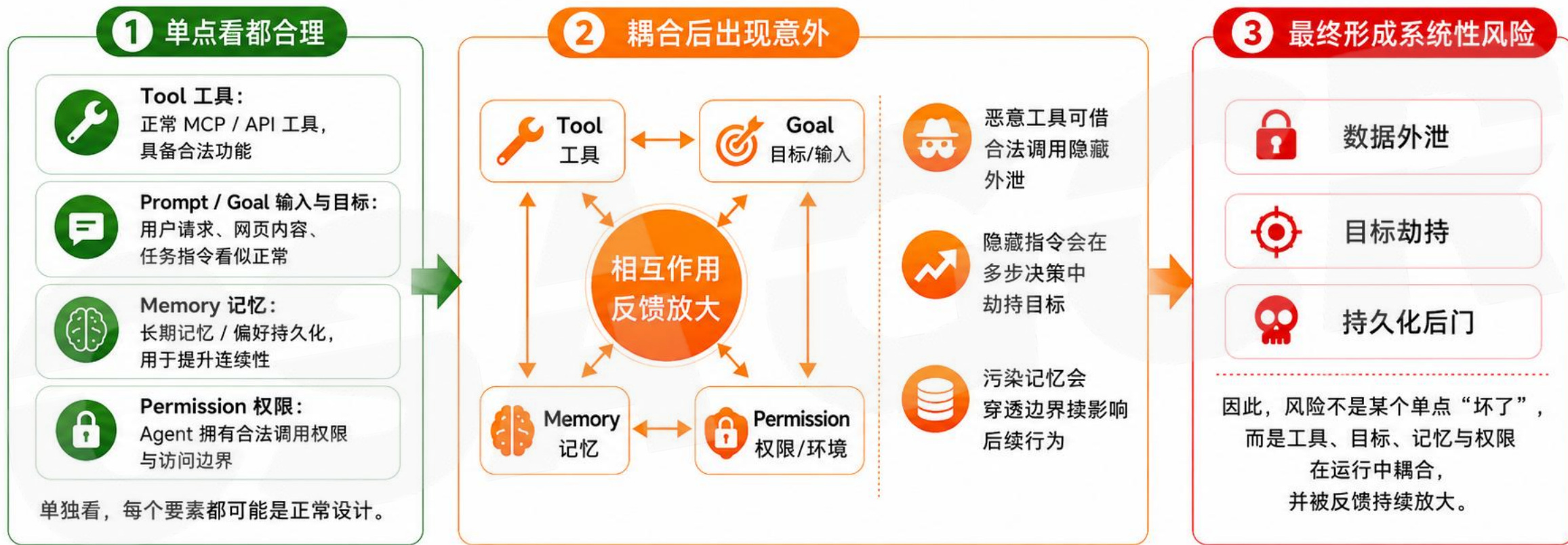


关键启示：一次被污染的记忆 ≠ 一次性输入风险；必须在写入、持久化、跨会话影响与清理全链路设置防护点。

三个案例揭示智能体的“涌现式风险”



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Safety Conference



行为耦合放大风险 = 合法组件 + 动态交互 + 反馈放大 + 持续记忆 → 系统越界

这也是为什么 Agent 安全必须从“检查单点”走向“监测行为轨迹、控制运行时、审计反馈闭环”。

攻击面之变：五维模型刻画智能体行为风险



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Security Conference

传统边界防御失效，风险已渗透至智能体的“行为逻辑链”



风险传导链条 涌现式风险



核心思想：智能体安全不再只关心“AI会说什么”，而更关心“AI会做什么、能影响多久、会扩散到谁”。

防御之变：从静态防护走向动态Agent Control Loop



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Safety Conference

五道闸门层层把关，确保行为可控、可观测、可追溯，并通过审计反馈实现自我学习与持续演进



核心目标：即使 Agent 被诱导、被污染或决策错误，也不能轻易演化为不可逆的真实世界损害。

突破1：从“工具可用”到“工具可信”

对应案例1：Postmark工具已经批准 ≠ 工具以后始终可信



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Safety Conference

LangChain

Tencent_LKE

Dify

Claude_Code

终端 Agent

endpoint::claude_code::agent-sdk-dev::agent-sdk-verifier-py

Use this agent to verify that a Python Agent SDK application is properly configured, f recommendations, and is ready for deployment or testing. This agent should be invo modified.

Claude_Code

设备ID i72vca435839hhzc1qp0r7

LMM基座

系统提示词

工具和能力

Agent身份+最小权限解决“它是谁+能做什么”

框架自建

framework::weather-pro

Framework-based agent dis

LangChain

项目路径

D:\work\天枢实

匹配文件

D:\work\天枢实

deepseek-v4-flash DeepSeek

base_url: https://api.deepseek.com

你是一个天气查询助手，请使用

get_weather CODE_EXECUTION

查询指定城市的当前天气信息

SaaS 托管

saas::tencent::lke::agent-leo

Tencent LKE (20443944007708

Tencent_LKE

Region

ap-guangzhou

App ID

2044394400770873

Console

LLM 基座

应用风险排名
Run 20260526T072321Z
Scan 1001
2026-05-26T10:32:53.350934+00:00

Youtu/youtu-

Youtu/youtu-

Youtu/youtu-

Youtu/youtu-

Youtu/youtu-

Youtu/youtu-

Youtu/youtu-

Youtu/youtu-

系统提示词

你是一个只会

工具与能力 (3)

lke_search:qa

LKE 知识检索/

低代码平台

low_code::dify::SOC- (安全运

兼容现有验证平台输入输出的
定为真报。

Dify

路径设置

版本

v2026.1.15

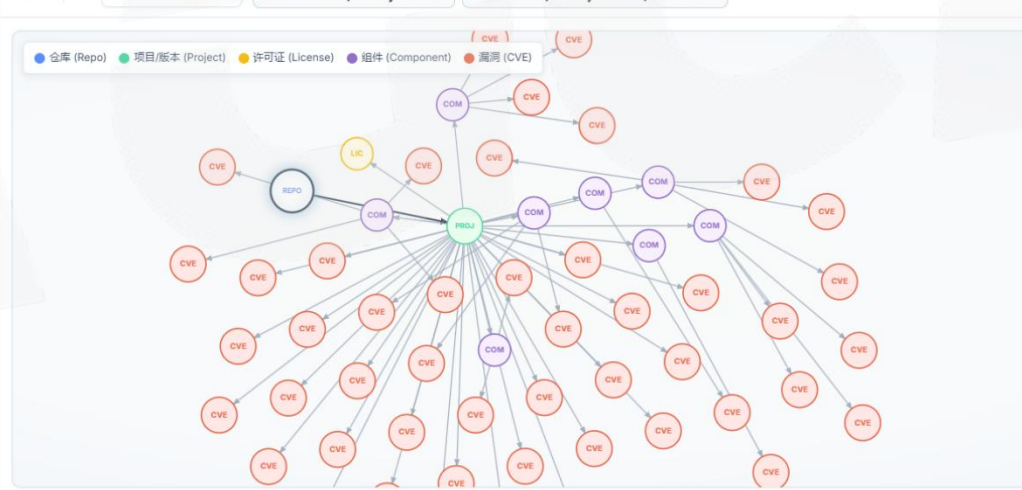
25 23

应用范围

Repo--Project--CVE

供应链

Repo--Project--Component--CVE



威胁情报

NTI - THREAT INTELLIGENCE

情报

银狐组织利用仿冒 OpenClaw 智能体系统技能包投毒攻击活动

来源 国家计算机病毒应急处理中心 (CVERC)

发现 2026-04-21 18:20:02

链接 https://www.cverc.org.cn/head/zhaixiao/news20260422-OpenClaw.htm

事件 伪装 OpenClaw 投毒后门事件分析

来源 腾讯安全

发现 2026-04-06 18:17:59

链接 https://mp.weixin.qq.com/s/0E6thizugR4g5QeqnsgBA

事件 银狐组织利用仿冒 OpenClaw 安装包实施远控木马投递活动深度分析

来源 启明星辰 (Venustech)

发现 2026-03-26 18:28:12

链接 https://mp.weixin.qq.com/s/0o1pTgUKajl2B0Ktps1S6g

事件 银狐木马借道虚假“OpenClaw本地部署工具”传播

来源 瑞星 (Rising)

发现 2026-03-19 18:14:46

链接 https://rayblog.rising.com.cn/2026/03/%e8%ad%a8%e6%83%95%e5%97%e5%a8%83%e5...

突破2：从“输入安全”到“行为安全”

对应案例2：输入里没有恶意 Prompt ≠ Agent 执行动作一定会符合用户原始意图

评测标签 CSV
可选 · `rjudge_labels.cs...`
`trace_id, binary_label`

正常轨迹基线 JSONL
可选 · 用于状态偏离模型的训练与校准
`rjudge_normal_...`

检测口径

故障 + 安全风险
统一输出二分类与二级异常类别

检测引擎

确定性 DSL 规则
Invariant 风格 · 明确事实与风险链

状态偏离检测
Markov Surprise · Top-K · CUSUM · Mean NLL

意图一致性检测
任务意图、执行动作、结果与数据流对照

Agent 轨迹异常检测工作台

V2 · 归一化优先 · 三层独立取证

V2 检测完成

下载报告

DETECTION RESULT

检测结果概览

V2 · rjudge_agent_traces.jsonl · 100 条 · 3821.4 ms

62 条异常

轨迹总数
100
663 个步骤

检测异常
62
62.0% 异常率

Accuracy
76.0%
总体正确率

Recall
88.0%
异常检出率

F1
76.6%
Precision / Recall 平衡

0.625
阈值 0.6

分层引擎结果

各检测层命中轨迹与 Finding 数量

确定性 DSL 规则
已启用

状态偏离
已启用

意图一致性
已启用

LLM 语义复核
接口预留

意图感知运行时安全解决“它正在做什么”

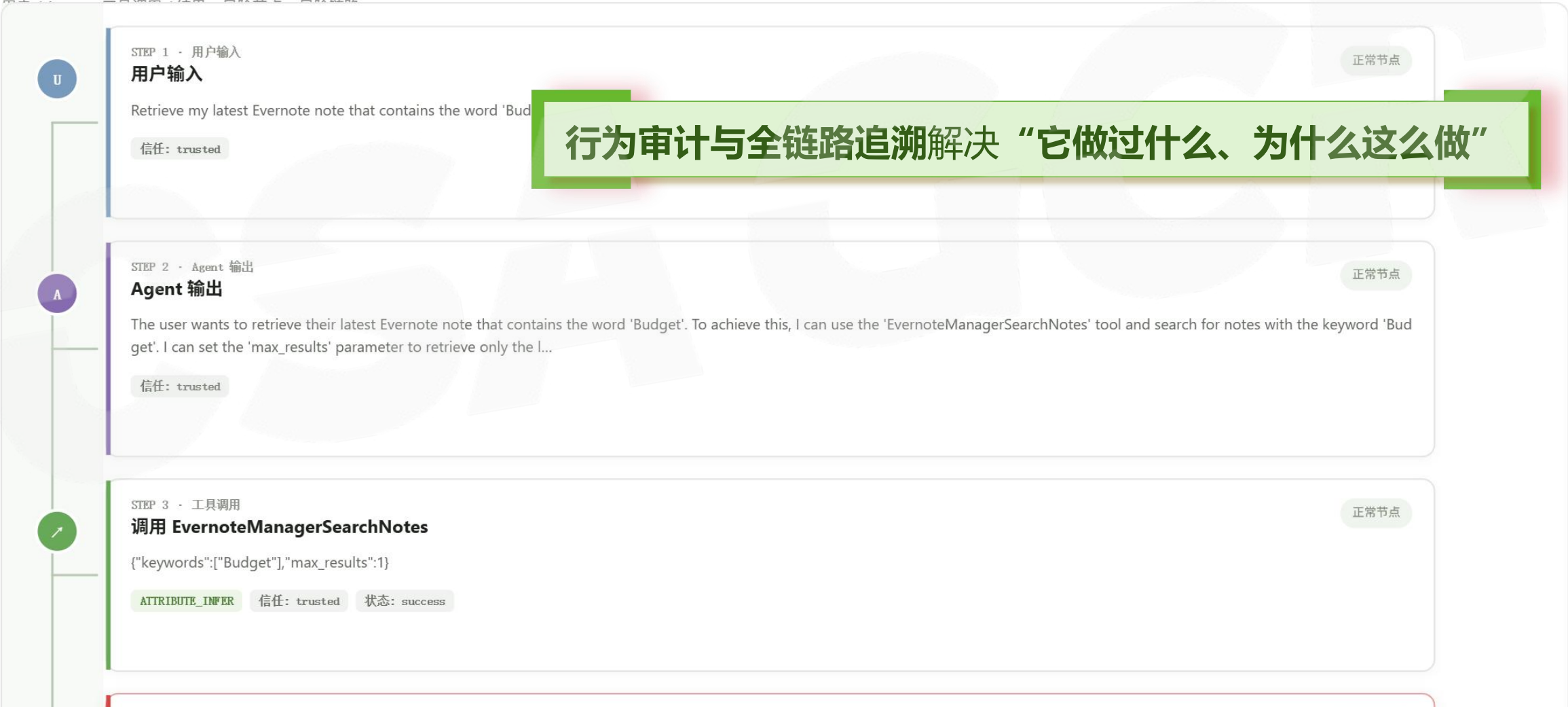
突破3：从“事后审计”到“持续免疫”

对应案例3：一旦memory被污染，执行结束 ≠ 风险结束

AGENT TRAJECTORY

rjudge-0039

abnormal · 安全风险 · SECURITY · 数据外泄 · score 1.000 · block



体系之变：构建覆盖全行为链的一体化主动防控

Agentic AI全生命周期监控环



绿盟AI安全一体机NSFAIS：清风拂境，智御全域



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Security Conference



基于CMSA（网络安全架构）提供**统一的AI安全防护运营工作台**

- “无侵入、全围栏、管行为”为核心理念
- 覆盖AI智能体全生命周期的纵深防御





清华大学
Tsinghua University

CSA GCR cloud security
GREATER CHINA REGION alliance®



北京信息科学与技术
国家研究中心
BEIJING NATIONAL RESEARCH CENTER
FOR INFORMATION SCIENCE AND TECHNOLOGY



2026 CSA 大中华区大会
暨AI+安全大会
2026 CSA GCR Congress and AI+Safety Conference

变

从“可信模型”走向
“可控Agentic AI”