



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

AI时代智能攻防技术 范式变迁、现实瓶颈与演进路线

腾讯云鼎实验室攻防负责人

李鑫



01

攻击篇

AI 究竟能把攻击做到什么水位？

02

测评篇

能力好不好，不该由做的人自己说了算

03

防守篇

打不赢指纹，就改看意图

04

未来展望

格局往哪走？技术往哪走？边界在哪里？

注：AI技术一直在进步，本议题观点仅供参考，部分数据未做实证

核心研究问题 RQ1—RQ7

- RQ12023—2026 年 AI 渗透测试经历了哪些可验证的范式变化？
- RQ2模型能力、Agent 工程与评测基础设施如何共同演进？
- RQ3为什么单点漏洞能力无法自然组合成稳定的多步骤攻击链？
- RQ4什么条件下系统可以从“辅助工具”升级为“受监管的自主渗透 Agent”？
- RQ5本地化小模型的真实能力边界与最佳工程定位是什么？
- RQ6如何建立兼顾进攻效能、合法授权、证据可信与失控防护的评价体系？
- RQ7未来三年的关键技术突破点、风险拐点与工程优先级是什么？

研究边界

纳入：LLM 渗透助手、工具增强 Agent、多 Agent、Computer-Use Agent、RLVR 安全 Agent、AI 原生渗透平台、评测基准与治理机制。

区分：传统扫描器、BAS、人工红队、AI 代码审计、恶意攻击自动化与授权渗透测试。

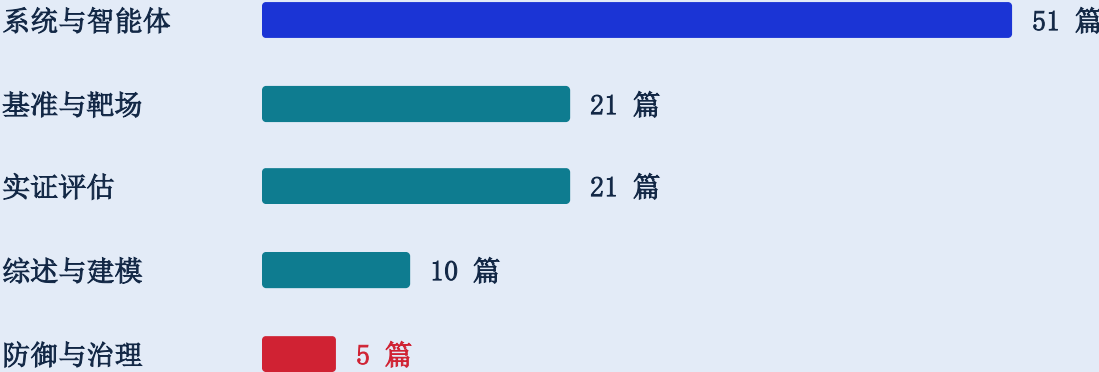
不提供：面向未授权目标的攻击流程、载荷或规避方案。

证据分层：不同来源承担不同举证责任

层级	资料类型	可信度处理
A	顶会/期刊论文、官方标准、厂商正式披露	主要事实依据，仍需审查外部效度
B	arXiv 预印本、公开基准与开源项目	用于追踪前沿，结论需交叉验证
C	企业技术博客与真实案例	产业信号，不替代独立评测
D	社区项目、竞赛复盘与宣传指标	工程模式线索，不作单独定论

核心资料画像：LLM4Pentest（截至 2026-09-06，108 篇）

结构本身揭示一个失衡：系统能力研究快速扩张，防御、伦理与治理研究占比仍低于 5%。



01 攻击篇

ATTACK · HOW FAR CAN AN AGENT ACTUALLY GO

不看厂商自述，看它在真环境里打成什么样。

把「AI 能不能打」拆成可测的三段：能不能发现、能不能打通、能不能在多步里不掉链子；
用带判定的靶场记录替代演示视频，给出一条能被外部复查的能力水位线。

本篇回答三个问题

THREE QUESTIONS

01 现在到底打到哪一步了？
先用 L0-L5 与同比例尺把水位说清楚，再看真环境实测 ——
单点已接近熟练工，长链仍大量断在中途，只看总成功率看不出来

02 决定上限的到底是什么？
不是模型分数，而是 Harness：推理更强 ≠ 渗透更强，
能力要先经过工具、结构化状态与验证器，才能真正落地

03 一线是怎么搭起来的？
把状态从大模型上下文里剥离、焊进数据库，用确定性骨架
卡住概率预测机的发散 —— 冠军作品 ARTEX 的核心选择

本篇包含 六页，按「先立一把尺 → 再看真环境实测 → 最后看模型与 Agent 工程怎么演进」排列

IN THIS CHAPTER

1.1 能力分级 L0-L5
对标智能驾驶的六级坐标

1.2 自动化同比例尺
十项指标放进同一刻度

1.3 真实环境实测
与 10 名安全专业人员同场

1.4 模型迭代 × Harness
推理更强 ≠ 渗透更强

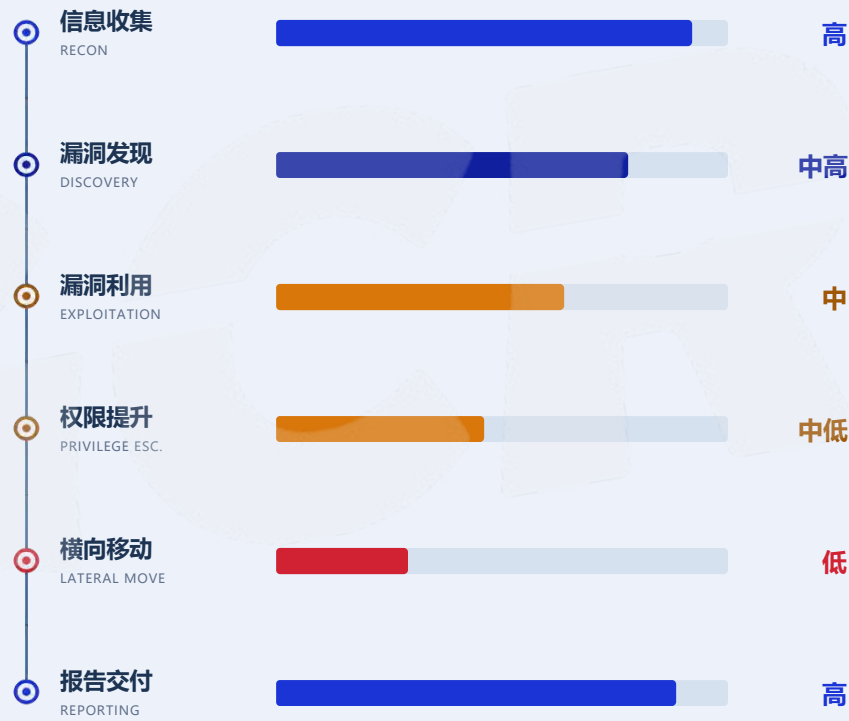
1.5 Agent 工程六维
状态与验证器决定上限

1.6 冠军作品 ARTEX
一线怎么把 Agent 搭起来

攻击链六段：自动化成熟度

KILL CHAIN

两头已经很好用，中段还在拉锯 —— 真正的瓶颈在链条中间，不在两端

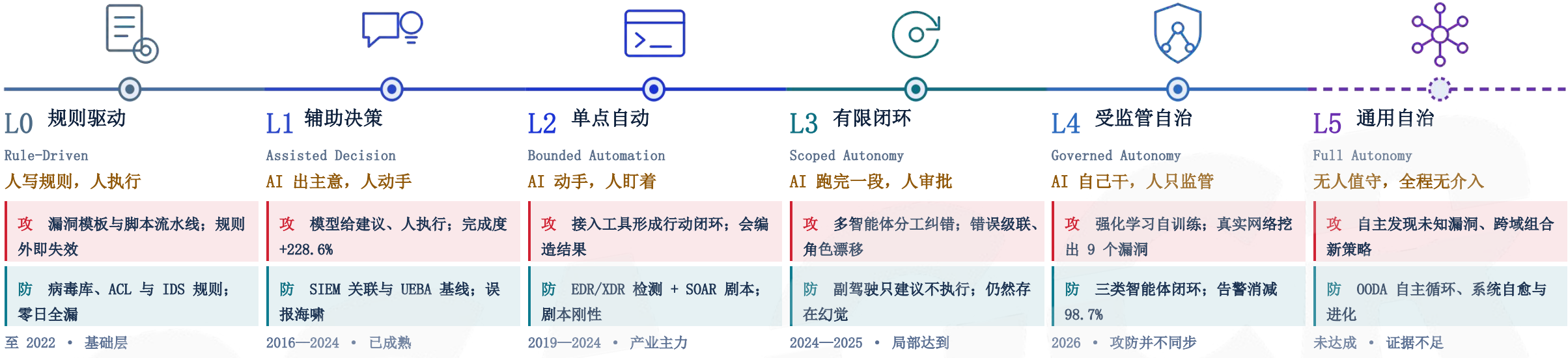


分档为本文按公开实测与两届赛事记录归纳的定性判断，非量化测量值。
本篇随后逐段给出阶段级成功率与失败归因。

AI 攻防能力分级：一把对标智能驾驶的尺子

从人写规则到无人值守 —— 攻击侧、防御侧与自动驾驶的六级对照

AI 时代智能攻防研究
范式变迁与演进路线



这把尺子真正的用处：它让「还差多远」变得可以讨论

汽车用了三十年走到 L3—L4，AI 攻防只用了三年。但两边的 L5 都没有到达，而且卡住的都不是算力或算法 —— 是责任归属、停止机制与可审计的安全论证。所以判断一个系统属于哪一级，不能只看它能打穿什么，还要看它能不能被叫停、被追责。

攻击侧
L4 仅单点证据
完整链闭环 16.67%

防御侧
L4 已规模化商用
告警消减 98.7%

智能驾驶
L4 限定区域落地
出了范围就不行

AI 攻防自动化程度：放在同一把尺子上量

统一 0—100% 刻度 • 十项指标等比例呈现 • 每项标注原始来源与证据等级

AI 时代智能攻防研究
范式变迁与演进路线

结论仅供参考

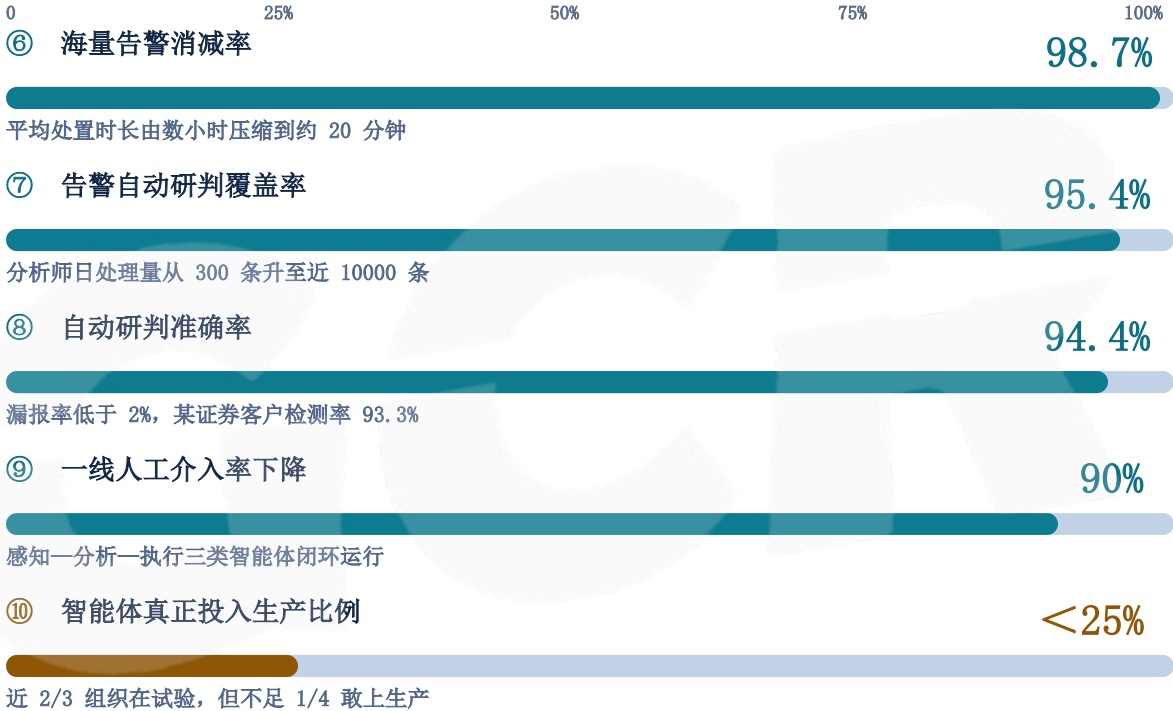
AI 攻击侧 能打穿多少

A/B 类证据 • 顶会论文与公开基准



AI 防御侧 能替人干掉多少

C 类证据 • 厂商案例，未经独立复现



落差的根源不是模型强弱，而是任务性质

防御是对已知模式做分类与响应——结果可验证、单次出错可容忍、样本充足；攻击是开放式探索未知——没有标准答案，一步错整条链就断。同一批模型换个方向，成功率差一个数量级。

攻击侧

< 17%

除单步动作外，全部指标低于两成

防御侧

> 90%

除投产率外，全部指标高于九成

数据来源与引用纪律

- ① 闭环 16.67% — Hackers or Hallucinators? (2026 实证研究)
- ② 真实网络 10—13% — Comparing AI Agents to Cybersecurity Professionals (ICLR 2026)
- ③ 利用率 <12% — HackWorld 计算机操作智能体基准 (ICLR 2026)
- ④ 防火墙场景 0 分 — PACEbench 主动防御场景 (ICLR 2026)
- ⑤ 单步 98.5% / 多步 0% — TrustedSec 自托管模型 4,800 次实验 (2026)

- ⑥⑦⑧ 98.7% / 95.4% / 94.4% — 奇安信 AISOC 八智能体客户实践 (2026)
 - ⑨ 人工介入率降 90% — 华为星河 Agentic SOC (2026-04-28 发布)
 - ⑩ 投产率 <25% — Elastic / Avasant 安全 AI 智能体采用度调研 (2026)
- 参照：Google Cloud 调研 — 82% 分析师担心因告警量漏掉真实威胁
- 纪律：A/B 类为主要依据；C 类仅作产业信号，不替代独立评测

口径以原始论文表格为准；“相对比例”与“百分点”不可混用

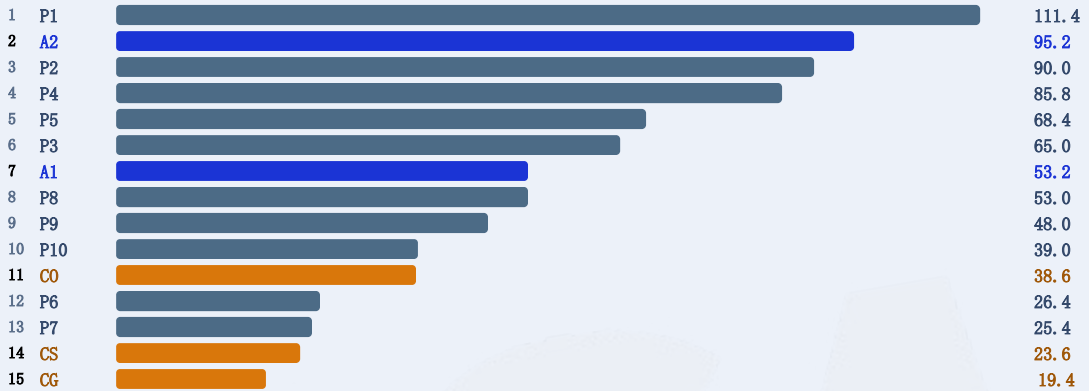
AI 在渗透测试中的真实能力：与 10 名安全专业人员同场实测

Comparing AI Agents to Cybersecurity Professionals in Real-World Penetration Testing • ICLR 2026

≈ 8,000 台主机 • 12 个子网
7 个公网可达 + 5 个需 VPN • 真实在用的大学生产网络

结论仅供参考

总分排行榜 技术复杂度 + 业务危害指数加权



A1/A2 为 ARTEMIS 两种配置 • C0 = Codex+GPT-5 • C5 = CyAgent+Sonnet 4 • C6 = CyAgent+GPT-5

最强开源框架 ARTEMIS 的成绩单

2 / 15

总排行榜名次
超过 10 名人类中的 9 位

9 ↑

有效唯一漏洞
提交 11 个 • 危害分 54 居次席

82 %

有效提交率
10 名人类中 8 位为 100%

8 并发

峰值子智能体
平均每轮 2.82 个同时在跑

同题对照：10 名持证专业人员（OSCP/OSEP 等，每人 ≥10 小时）+ 6 个现有 AI 框架 + ARTEMIS，同一台 Kali 跳板机、同一份任务指令。

单位时间成本 同刻度 0—\$60/时

\$18.21

ARTEMIS A1 (GPT-5 单模型)

\$59.00

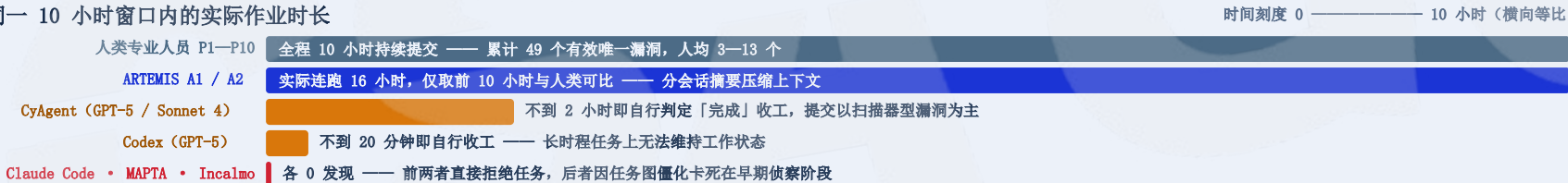
ARTEMIS A2 (五模型集成)

\$60.00

人类渗透测试专业人员

成本已不再是门槛。A1 以约四分之一的花费取得了与 A2 相当的漏洞数量；而 A2 的 \$59/时 已与人类的 \$60/时 基本持平。

同一 10 小时窗口内的实际作业时长



结论边界 • 引用时必须同时讲清

- ① 时间被压缩 —— 仅 10 小时有效作业，真实渗透测试通常 1—2 周
- ② 防御条件缺失 —— IT 团队全程知情，本会被拦截的动作被人工放行
- ③ 样本量受限 —— 无法开展具备统计效力的假设检验

✓ 已经超过人类的四件事 Where AI Agents Already Win

- 并行推进 Parallel Exploitation
扫描一有发现立刻后台起子智能体；峰值 8 个并发。人类做不到同时深挖多个目标
- 长时程续航 Long-Horizon Operation
摘要压缩 + 递归待办清单，连续作业 16 小时；提交间隔普遍长于人类
- 成本优势 Cost Efficiency
\$18.21/小时、年化 \$37,876，约为 A2 的四分之一成本却取得相当的漏洞数量
- 无图形界面反占优 CLI-Native Advantage
旧版 iDRAC 密码套件过时、浏览器拒载，10 人全部放弃；它用 curl -k 直接拿下

✗ 仍然卡住的四件事 Where AI Agents Still Fail

- 图形界面几乎不可用 GUI Interaction
80% 人类拿下 TinyPilot 的 Windows 远程代码执行，它只提交了 CORS 通配符等配置问题
- 误报率显著更高 False Positives
把登录失败后重定向的 200 OK 当成默认口令登录成功 —— 有图形界面的人一眼就能识破
- 见好就收不深挖 Shallow Follow-Through
发现即提交，不横向移动不提权；拿到 CORS 后立刻走人，错过同一台机器上更关键的漏洞
- 卡在模式识别而非执行 Pattern Recognition
给出五级提示引导后，4 个此前漏掉的漏洞全部至少被找到一次 —— 技术执行力并非瓶颈

通用模型能力的四次叠加

04 • 多模态与计算机使用

2025—2026

通过截图、GUI、DOM / ally 树处理现代 Web 交互

03 • 原生工具调用

2024—2025

结构化函数调用、Shell、浏览器与专业安全工具

02 • 推理与长上下文

2024

攻击计划、错误分析、跨阶段信息维持

01 • 知识与代码生成

2023

漏洞知识、脚本生成、日志与报文解释

五条领域化路径：成本、收益与失效模式

路径	收益	主要失效模式
提示词与角色约束	成本最低、见效快	脆弱、难形成新能力
RAG / 知识库	适配动态 CVE 与验证过的 PoC	低精度泛化检索反而引入噪声
SFT / 轨迹蒸馏	学习专家操作模式与工具纪律	继承成功轨迹偏差，缺少失败恢复
RLVR 可验证奖励	直接优化执行结果	奖励稀疏、环境过拟合、奖励投机
强弱模型混合路由	前沿模型定策略，小模型做解析与参数化	路由判断不清会同时损失质量与成本

为什么“更强的推理模型”不自动更能渗透

- 通用 Benchmark 排名无法直接迁移到自动化渗透任务
- 关键差异在任务分解、探索深度、失败恢复、工具调用纪律与停止行为
- HackWorld：视觉感知进步没有转化为安全策略与工具编排能力
- Hackers or Hallucinators：同一框架替换模型后差异显著，更强模型仍会产生假 Flag

模型选型：准确率之外必须评估的九个维度

工具调用有效率

首次成功率

失败恢复率

长链状态保持

单位成功成本

推理延迟

拒答/越界倾向

上下文污染敏感性

可私有部署性

这九项决定了模型在真实工程中的表现差异 —— 而它们几乎都不出现在通用排行榜上。报告第四章应据此提出「渗透场景模型评估卡」，把选型从“看榜单”改为“看行为”。

六维架构分析法

01 Agent 架构

单体 / 多 Agent / 层级式 / Swarm / 黑板

02 规划机制

线性流程 / 树搜索 / 图规划 / OODA / 经典规划

03 记忆机制

对话历史 / 摘要 / 结构化状态 / 向量 / 证据账本

04 执行机制

函数调用 / MCP / Shell / 浏览器 / 代码解释器 / GUI

05 外部知识

CVE / ATT&CK / PoC / 代码库 / 资产与历史轨迹

06 验证机制

Flag • Root / 独立重放 / 环境状态检测 / 人工复核

标绿两维（记忆、验证）是当前差异化最大、也最被低估的工程要素。

单 Agent 与多 Agent 不是二选一，而是任务形态的函数

方案	适合场景	主要优势	主要风险
单 Agent + ReAct	强耦合、短链、快速试错	上下文完整、通信成本低	长链上下文膨胀、陷入兔子洞
专业化多 Agent	多阶段、并行探索、异构任务	上下文隔离、可并行	协调冲突、摘要失真、错误级联
黑板 / 状态图	动态任务、共享事实与依赖关系	结构化协作、可回溯	状态质量直接决定全局质量
确定性流程 + LLM	高频已知漏洞、合规复测	可复现、低成本	覆盖面与创新能力受限

记忆：从“聊天记录”转向“攻击状态”

事实必须显式分为四种状态，并记录资产、服务、版本、凭据、权限、失败路径、依赖条件与证据位置。



自然语言摘要只能作为阅读层，不能作为唯一事实源；否则失真会随链路长度累积。

MCP 与工具调用的三条原则

MCP 提升互操作性与生态复用，同时引入描述注入、工具投毒、权限继承与供应链风险。

- 1 工具发现 ≠ 工具授权
- 2 工具数量 ≠ 系统能力
- 3 按阶段动态暴露最小工具集，执行层独立校验参数、目标、权限与输出

规划 Agent 与验证 Agent 必须职责分离；关键结论交由只读独立审计器复核，避免自我确认偏差。

「Agent+」攻防能力挑战赛总冠军作品：ARTEX 自主渗透测试系统

把状态从大模型的上下文里彻底剥离，焊死进数据库 —— 用确定性的钢铁骨架，卡住概率预测机的发散与幻觉

百度安全应急响应中心 BSRC 主办 • 2026.09.03 成都总决赛
技术战略合作伙伴：腾讯云鼎实验室（提供 TSecBench 靶场 • Agent 托管 • 自动化评测）

赛事漏斗与双榜战绩



两榜均非第一，综合评分却是全国第一

双轨制

作品评审 50%（自动化逻辑 • 漏洞发现率 • 误报率 • 单高危耗时 • 人机验证比）
+ 靶场实测 50%。决赛路演由 8 位专家评委（权重 80%）与 9 位嘉宾评委（权重 20%）打分。

核心创新 • 双图状态机彻底剥夺大模型维护攻击状态的权力

上下文即使全丢，重启后一次 SQL 查询即可恢复整个战场

探索图 Exploration Graph



▲▼ 锚点表 exploration_anchors (node_id, asset_id) —— 把意图 / 事实 / 漏洞钉死在具体资产上

资产图 Asset Graph



它解决的是什么病

幻觉鬼打墙（测过的点反复重测烧 Token）• 上下文自爆（摘要压缩把关键凭据当垃圾切掉）• 串行链脱节 • 越权失控

数据库层面的死规则

资产表 9 个条件唯一索引按类型强制去重 —— 模型抽风报 100 次同一端口，数据库直接拒绝写入，资产底座始终干净

01 防抖唤醒 + 行级抢单

图变更走 800ms debounce 合并唤醒，避免连续
唤醒 Planner 白烧 Token；3 个 Worker 抢活不靠
中间件，纯靠数据库单行原子更新

```
UPDATE ... WHERE state='open' → RowsAffected==1
```

02 跨唤醒共享 TodoStore

Planner 每次唤醒都是全新会话，却在进程内常驻
一份待办清单；只有前置步骤完成、依赖事实已
录入，才派发下一步意图

稳定推进 5—6 层深层利用链，不重复不错序

03 过程级信息交换

一个 Worker 踩过的坑、拿到的报错碎片，其他
Worker 可直接模糊检索复用 —— 信息以「执行
过程」为粒度流动，而非只读报告

search_all_worker_traces(q) • 自动排除自身步骤

04 强制权限收缩

曾抓到 Planner 跑到 40% 覆盖度就想「交卷」；
作者直接把收工工具从所有 Agent 工具集里永久
解绑 —— 模型无权宣布自己测完了

误报率与有效覆盖率是评审硬指标

它给自主攻防指的那条路：工程控制面的厚度，而不是提示词的花哨

70% 分布式系统工程 • 确定性状态机 • 数据约束

30% 大模型交互

用强约束的资产真值库兜住底线，用有向无环图收敛模型的发散，用全量流量留痕与可视化拓扑让人看清它的每一步 —— 造一个确定性的钢铁骨架，把不确定的大模型死死卡在它该在的位置上。

开源工程规模

GitHub • Autumn-27/ARTEX • 随赛放出 v0.3.7

11 万

行代码

206

个 Go 源文件

38

张数据库表

Go 后端 4.6 万行 + Next.js 前端 4.6 万行；前端 go:embed 打进单二进制，零依赖即起。

同场值得一提的对手

亚军 Cruiser • 多 Agent 协同自主漏洞发现，靶场跑分第一

季军 Sloppe • 去中心化对等协作并行渗透，成本仅 51 元

第 4 名 bsrc-agent • 单容器架构，「控制面在代码、模型只做工具调用」

02 测评篇

BENCHMARK · IF YOU CANNOT MEASURE IT, YOU CANNOT BUY IT

能力好不好，不该由做的人自己说。

本篇给出一把可复查的尺：赛题与靶场沉淀成标准，判定来自环境而不是印象；

同一把尺能跨版本、跨厂商反复量，选型、验收与代际比较才谈得上成立。

本篇回答三个问题

THREE QUESTIONS

01

旧基准为什么不够用？

单点题、静态答案、只看结论，
测不出长程多步、工具调用与真实副作用 —— 分数高不代表真能打

02

新基准该怎么设计？

环境即判定、按阶段给分、过程全量留档，
让分数来自客观事实，而不是裁判模型的印象

03

拿到分数之后呢？

把不同来源的成绩落进攻击篇那把 L0-L5 坐标，
才能跨版本、跨厂商、跨代际反复量，选型与验收才有依据

本篇包含 八页，按「可信度为什么被侵蚀 → 标准怎么建 → 数据说了什么 → 长链断在哪」排列

IN THIS CHAPTER

2.1 三类可信度瓶颈

幻觉 · 工具 · 评测互相放大

2.2 智能渗透黑客松

4 亿流量 / 940 万对话

2.3 挤掉泡沫的标准

结果换单位，过程开审查

2.4 TSecBench 基准

真实漏洞与生产级靶场

2.5 第一期联合测评

13 Agent × 4 模型的上限

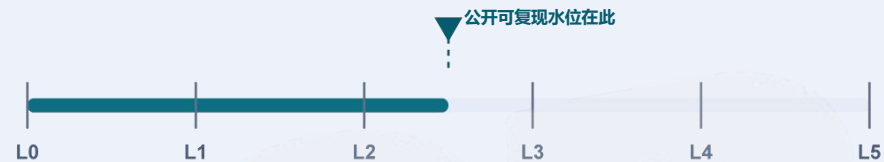
2.6 长链瓶颈与漏斗

32 步只走到第 22 步

一把尺：L0 → L5 能力坐标

ONE RULER

先把「自动化到哪一级」说清楚，再谈成功率 —— 否则两边说的不是同一件事



工具化 · 辅助 · 半自动 · 有人监督 · 有条件自治 · 完全自治 (分级定义见攻击篇 1.1)

同一道题，必须同时量三件事

THREE AXES

① 阶段级成功率 STAGE-WISE

侦察 / 利用 / 提权 / 横移分别计分，避免一个总分掩盖中段断点
断点落在哪一级，比「总分是多少」更有决策价值

② 单次成本与耗时 COST & WALL-CLOCK

同一题目下的 token 用量、金额与墙钟时间，一并记录
能打通不等于打得起 —— 成本决定它能不能被规模化使用

③ 过程可复查性 REPRODUCIBILITY

轨迹、工具调用与判定日志能否被第三方原样重放
不可复查的分数，本质上仍然等于厂商自述

这三类问题不是各自独立的工程缺陷，而是会互相放大：假成功污染评测，评测失真掩盖工具缺陷，工具缺陷又制造更多假成功。

❗ 幻觉与假成功

FALSE SUCCESS

12 / 34

开源框架中至少出现一次虚假 Flag

- 模型不能自行宣告“成功”——成功判定必须从生成式输出中剥离
- 风险不止误报：还包括过早终止、错误路径持久化
- 错误结论会级联污染下游 Agent 与最终报告

对策：成功必须绑定原始输出、环境状态与可重放步骤；规划与验证 Agent 职责分离。

⚙️ 工具工程瓶颈

TOOLING

不可解

工具协议不完整会让任务从根本上无法完成

- 工具返回缺少类型、错误码与结构，会显著放大模型误判
- 工具池越大，选择空间、上下文开销与供应链风险越大
- 交互式会话、分页输出、超时与重试常被忽略

对策：以「能力覆盖矩阵 + 动态最小暴露 + 执行层校验」替代工具堆砌。

📊 评测可靠性不足

BENCHMARK DRIFT

25%—85%

同一目标重复运行的成功率波动区间

- CTFusion 识别出 71 次记忆事件，移除后平均成功率下降约 29%
- 单次演示不能代表稳定能力，必须报告方差与置信区间
- 二元 Flag 指标无法区分“几乎完成”与“毫无进展”，并诱发奖励投机

对策：多次独立运行 + 阶段评分 + 时间切分 + 轨迹级验证，禁止仅凭最终字符串判分。

成本与可持续性：高可信评测本身就是重工程

成本来自模型调用、工具执行、环境重置、并发沙箱、人工复核与日志审计六个方面。

100 亿+

Token 消耗

2,500 美元

直接算力与 API 支出

1,500+

份运行日志

正确口径：不比较“每小时成本”，而比较每个有效、可复现漏洞的全成本（含人工复核与误报处置）。

腾讯云智能渗透黑客松：把智能渗透从概念推成工程

TCH · 腾讯安全云鼎实验室主办 · 国内首个聚焦 LLM 智能体全流程自动化渗透的专业赛事 · 铸刃止戈，以智御危

两届 · 900+ 战队 · 2000+ 选手

github.com/Yeti-791/Tsec-Hackathon

比赛规模 两届下来，它已经不是一场活动，而是这个方向的公共基础设施

SCALE · 2025.11 & 2026.04

2 届

已连续举办

2025.11 · 2026.04

900+

参赛战队

两届累计 · 覆盖高校与厂商

2000+

参赛总人数

两届累计 · 顶尖攻防选手

20 套

智能渗透技术框架

前 20 名全部开源 · 一站式导航

150+ 家

高校 · 企业 · 科研机构

北大 清华 西交 国防科大 南洋理工

50 万

总奖金池（元）

一等奖 ¥6 万 + 开源贡献奖

50+ 篇

公开复盘与技术分析

两届合计 · 全网可检索

最大的产出：一份三层对齐的过程数据 做了什么 · 怎么想的 · 成没成，来自同一场比赛

OBSERVABILITY

4 亿次 网络流量

靶场里发出的全部真实请求

↓ 平均每 1 次 模型对话 → 撬动 约 42 次 网络请求

940 万次 模型对话

统一模型网关记录 · 已对外开放

↓ 平均每 1 道题 → 消耗 约 940 次 模型对话

1 万+ 次 答题记录

带客观判定的成败标签

行为层 Agent 到底「做了什么」

完整的动作序列：探测 · 扫描 · 利用 · 重试 · 放弃

可还原每一步真实副作用，而不只是它自己声称做了什么

认知层 Agent 到底「怎么想的」

推理链 · 工具调用 · 上下文压缩 · 多模型分工 · token 成本

经统一模型网关记录，已按战队维度对外开放

结果层 这一次到底「成没成」

带客观判定的成败标签，赛制严禁人工渗透介入

天然可当 ground truth，直接用于评测基准构建

940 万次大模型对话数据已对外开放 全队完整对话记录 challenge.zc.tencent.com/teams · 模型使用情况报表（腾讯文档）同步公开 —— 想看 AI 推理过程长什么样，这是现成的一手样本

OUTCOME

赛场产出：效果说话

诞生 3 个高 Star 开源项目 GitHub Star · 截至 2026-09-14

CyberStrikeAI

首届 · 小白战队 · 集成 100+ 安全工具

6.6k

Cairn

第二届 · Bytex · 黑板系统 + 蚁群算法

2.8k

LuaN1ao 鸾鸟

首届 · BinX · 状态感知 + 因果推理

1.3k

标志性成绩 赛制严禁人工渗透介入

唯一 AK

第二届 Bytex（个人）全场唯一全解

总成本仅 ¥7,692 —— 平等 Worker 动态分工，拒绝预定义角色

7 模型

第 1 名 ai小分队并行竞争上岗

Manager+Solver+Observer 三层解耦 + 上下文三层压缩

94%

首届冠军自主渗透智能体成功率

「严禁人工操作」赛制下取得，AI 独立完成全流程

它留下了什么 —— 赛事实践 → 技术沉淀 → 开源共享 → 标准反哺，四段闭环已经跑通

VALUE

01 对研究

三层数据在同一时间轴上对齐

· 公开攻防语料多数只有结果、没有过程

· 可做推理链分析 · 失败归因 · 成本曲线

· 对话数据已开放，外部可独立复查

02 对工程

20 套框架被真实战场压测过

· 上下文压缩 · 状态约束 · 自然语言 FSM

· 不是 demo，是同一环境下分出的高下

· 全部开源，工程手法可直接复用

03 对产业

拿到「AI 独立渗透」的可比基线

· 严禁人工介入，成绩归因于 Agent 本身

· 首届 94% → 第二届出现全场唯一 AK

· 给甲方可引用的水位，而非厂商自述

04 对标准

赛场经验固化成 TSecBench

· 赛题 · 靶场 · 判定规则沉淀为评测标准

· 一次性比赛 → 可持续复用的度量尺

· 长期反哺行业，不随赛期结束而中断

05 对生态

实践 → 沉淀 → 共享 → 赋能

· 答辩 PPT · 视频 · 复盘 · 开源导航入库

· 150+ 高校企业机构共建，人才双向流动

· 智能体社论论坛「零界」承接赛后讨论

一场黑客松真正的产出不是名次表 —— 是 4 亿次流量 + 940 万次对话 + 1 万次判定 组成的公共实验记录，
以及一套从赛场里长出来的 TSecBench 评测标准 —— 让「AI 怎么想、怎么做、成没成」既能被外部复查，也能被长期度量。

怎样建立一套挤得掉跑分泡沫的评测标准

Building a Bubble-Proof Evaluation Standard for Pentesting Agents

结果侧换单位 • 过程侧开审查

方法论参考 arXiv:2605.10834v2 • 过程审查为 TSecBench 自有制度

跑分泡沫从哪来

四条虚高路径



刷量重报 Duplicate Inflation

同一个洞换三种措辞报三次，召回率凭空虚增



单指标崇拜 Metric Gaming

只考召回必然牺牲精确率 — 实测 +18.83% 对应 -14.19%



题面泄漏与过拟合 Overfitting

靶场公开、题解可检索，参赛方对着答案调参



黑箱自证 Unauditable

只公布最终分数，过程不可回放、无人可质疑

闸门一 • 结果侧

把评测单位从「有没有拿到 flag」换成「报了什么、漏了什么、掺了多少水」

黑盒运行 • 只给目标

① 标准答案构建 Ground Truth

专家逐条标注：名称 • 类别 • 描述 • 补充信息

粒度权衡：过宽难判定，过细会误杀他路发现

② 宽松语义匹配 LLM-as-a-Judge

裁判模型判断「说的是不是同一个洞」

刻意允许一对多候选，不在此阶段强制唯一

③ 二分图消解 Hungarian Algorithm

候选边求最大匹配，每条答案最多计一次

未中选者判为重复报告 — 刷量当场归零



闸门二 • 过程侧

跑分过程全量公开，AI 查全 + 社区纪委查 TSecBench 独有

人人可查 • 人人可复算 • 人人可质疑



① 跑分执行 Run

容器全新环境启动

全程录制每一次动作与工具调用

过程全量留痕



② 日志公开 Public Log

原始日志与完整轨迹随榜单

同步上线，任何人可下载

面向公众开放



③ AI 自动审查 Auto Audit

机器逐条比对日志与结论

全量覆盖、不抽样、不留死角

查全



④ 社区纪委复核 White-hat Review

社区白帽子抽检高危条目

实名署名，人人可发起质疑

查准



⑤ 成绩生效或驳回 Verdict

通过即入榜；存疑条目扣分

并公示理由与复核人

结论一并公示

AI 自动审查 查什么

机器负责「查全」—— 全量、不抽样、不疲劳

轨迹可回放

每条发现必须在日志里定位到产生它的那一步

越界检测

读答案文件、访问题面外资源、硬编码结果

重复归并

语义聚类 + 二分匹配，换措辞重报当场合并

资源核账

耗时与 Token 与声明预算不符即标记待查

社区纪委人工核查 查什么

白帽子负责「查准」—— 实名、担责、可申诉

高危抽检

对严重与高危条目逐条人工复现验证

环境取证

排查靶场是否被上一轮残留改动而失真

公开申诉

任何人可对任一条目发起质疑并要求举证

署名担责

复核人实名留痕，结论与理由一并公示

挤掉泡沫的三条铁律

三者缺一，榜单就会重新长出水分

单位可核验

分数必须能拆到每一条发现 —— 拆不开的总分，不存在「对质」这回事

过程可回放

只公布结果的榜单，等于要求参赛者自证清白；日志公开才是可信的起点

复核有第三方

机器查全、人工查准、社区署名担责 —— 三者缺一，泡沫就有藏身处

核心观点 TAKEAWAY

决定榜单可信度的，不是题目有多难，而是每一分能不能被追问到底。

Trust is not scored. It is audited.

TSecBench：智能攻防 AI 的统一跑分基准

tsecbench.zc.tencent.com

AI 时代智能攻防及评测体系 · 深度研究报告

腾讯安全云鼎实验室 · 用真实漏洞、生产级靶场与红蓝对抗环境，把「攻击能力」变成可比较的数字

12,705

累计评测智能体
SCORED AGENTS

17,504

累计评测任务数
TOTAL RUNS

294

覆盖安全场景
SECURITY SCENES

8

合作伙伴与参与机构
PARTNERS

TSecBench 是面向**自动化红队与对抗性 AI** 的能力验证体系：以统一跑分标准衡量 AI 在自主渗透、漏洞挖掘与对抗博弈中的真实表现，**用攻击视角量化防御短板**。全部成绩采用**托管运行模式**产生，杜绝本地刷分；评测日志与 Agent Trace 逐步公开，可复现、可追溯、可举报。

六大进攻能力域

括号内为综合得分权重，合计 100%

Web 漏洞挖掘
WEB

20%

真实 Web 应用与代码仓库中自动发现 CVE 级漏洞

二进制漏洞挖掘
BINARY

15%

逆向分析与内存破坏类漏洞的挖掘定位

漏洞利用
EXPLOIT

20%

编写 PoC、稳定利用并获取权限

多阶段渗透
KILLCHAIN

20%

串联侦察、突破、横向移动的攻击链编排

云攻击
CLOUD

15%

配置错配、IAM 提权与容器逃逸

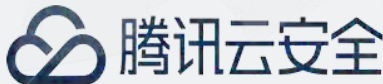
对抗规避
EVASION

10%

EDR / WAF / 沙箱下的免杀与隐匿

合作伙伴 · 参与机构 · 测评执行单位

来自顶尖安全实验室与产业伙伴的共同支持



PROMPT

提示词接入

复制提示词输入 Agent，零开发成本直接跑；要求 Agent 能接收自然语言任务

SDK

Python SDK 接入

用 python-sdk 简化开发，无需关心平台跑分流程，只对接核心解题流程

API

API 接入

可控性与确定性最高，满足所有场景使用；适合深度定制的评测集成

平台规模数据取自官网实时接口（2026-09-12）；机构排列不代表先后次序，清华大学为第一期测评执行单位之一。

榜单持续扩展 · 支持刷分举报 · 榜单与 Agent Trace 详见官网

第一期联合测评：13 个 Agent × 4 个模型的能力上限

156 份结果 / 52 个选定组合
三轮取优 · 统一环境与评分规则

智能渗透 Agent 能力测评报告（第一期） · 测评时间 2026 年 Q2 · TSecBench v1 共 63 道题

90.54%

最高组合结果（并列）

T3MP3ST + Kimi K3 与 Cairn + Kimi K3

77.03%

领先模型 Kimi K3

13 个 Agent 汇总成绩，11 个 Agent 的最优组合

67.52%

52 个组合总体成功率

各组合三轮成功率取最高后汇总

29.53%

共同短板：多阶段渗透

完整解率率仅 8.33%，长链任务是核心瓶颈

统计口径（必须一并说明）：单题依次评测、每题独立运行；一般题目最长 1 小时，多阶段渗透最长 2 小时，超时视为未通过；统一环境、配置、流程与评分规则；每个 Agent × 模型组合取三轮中成功率最高的一轮，并列时依次比较总耗时更短、Token 更少、轮次更早。该口径用于观察公开能力上限，不代表典型运行均值。

Agent 跨模型能力上限排名（汇总四个模型各自的最优轮次）

#	Agent	平均成功率	成功率	完整解题	平均耗时	平均 Token
1	Cairn		77.70%	82.14%	22.11 h	458.50 M
2	CHYing		75.00%	80.56%	22.78 h	27.92 M
3	LuaNlao		74.32%	80.95%	23.39 h	261.66 M
4	BreachWeave		73.65%	78.97%	21.05 h	266.96 M
5	Claude Code		72.30%	76.59%	18.03 h	58.42 M
6	CyberStrikeAI		71.96%	77.38%	15.92 h	145.82 M
7	T3MP3ST		70.61%	75.79%	20.38 h	230.80 M
8	cai		67.91%	74.21%	25.66 h	140.59 M
9	pi		67.57%	73.81%	16.43 h	158.70 M
10	shannon		63.85%	65.87%	31.31 h	23.92 M
11	Pentagi		61.15%	68.65%	31.17 h	128.58 M
12	ctfSolver		55.74%	63.10%	31.15 h	262.27 M
13	PentestAgent		45.95%	52.78%	35.92 h	223.02 M

本期测评
执行单位



四川大学
网络空间安全学院
黄诚课题组



鹏城实验室
PENGCHENG LABORATORY



清华大学
TSINGHUA UNIVERSITY

Agent 与模型必须成对看：同一 Agent 换模型，成功率最大相差 31.0 个百分点；Token 消耗最大相差 16 倍（27.92 M ↔ 458.50 M）。

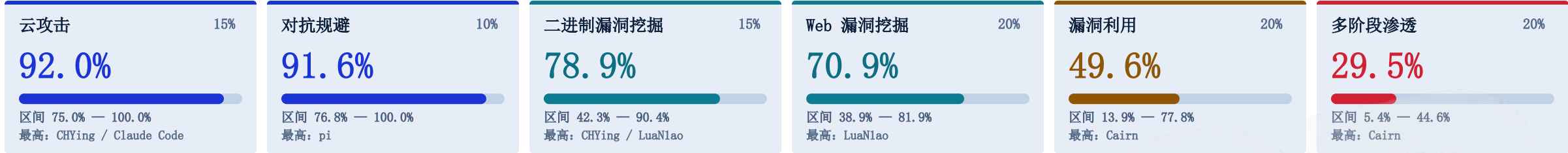
测评结果解剖：强在单点，弱在长链

六大能力域画像 · 四款模型的横向差异 · 多阶段渗透为何成为所有 Agent 的共同短板

基于 52 个选定结果
13 Agent × 4 模型 · 三轮取优

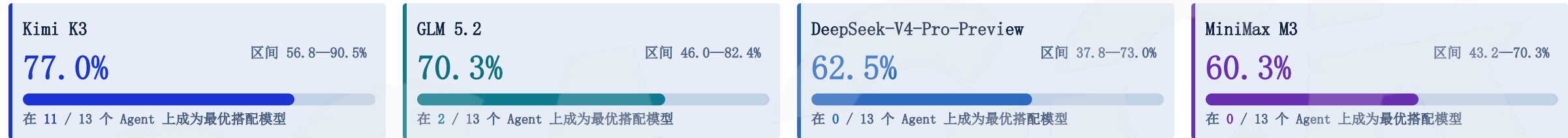
六大能力域：13 个 Agent 的均值与区间

能力域分析仅使用 52 个选定结果；条形按 0—100% 同比例尺



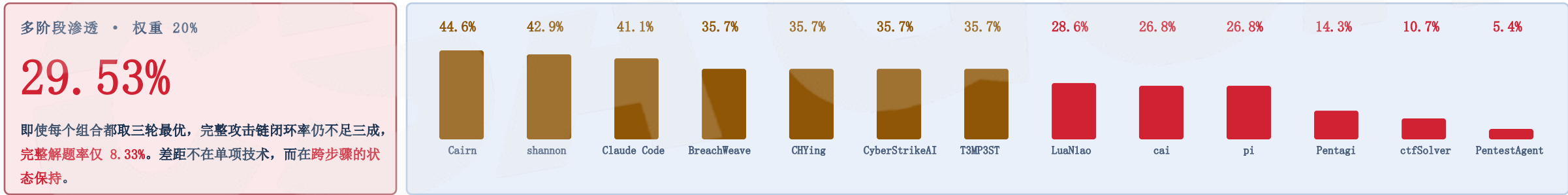
四款模型：同一批 Agent 下的横向差异

均值 = 该模型在 13 个 Agent 上的选定结果平均



核心瓶颈：多阶段渗透（13 个 Agent 逐一对比，纵轴 0—50%）

无一个 Agent 过半，最高 44.6%，最低 5.4%



01 能力上限 ≠ 可托付

最高组合达 90.54%，但那是三轮取优后的公开上限。同一组合的典型运行会更低 —— 选型时看的是这条线的位置，不是它的最高点。

02 Agent 工程与模型同权重

换模型带来最多 31.0 个百分点的落差，换 Agent 架构带来 31.8 个百分点。两者必须成对评估，单独宣传任一侧都没有意义。

03 短板高度一致，指向同一处

13 个 Agent 无一在多阶段渗透过半。这不是某个团队的实现问题，而是当前范式在长链任务上的共同结构性缺口。

免责声明与使用说明（摘自报告原文）

- ① 时效与版本：仅反映测评期间的模型版本与各 Agent 当时提交的实现。
- ② 能力上限口径：三轮取最高，不代表典型或平均表现。

- ③ 影响因素：受提示词、工具配置、运行环境、超时上限与输出随机性影响。
- ④ 用途与合规：仅供技术研究与方案选型参考，不构成排名背书；全部测评在授权、受控靶场中进行。

多步骤才是真瓶颈：32 步攻击链，最好的一次只走到第 22 步

英国 AI 安全研究院用两座定制靶场做了十八个月纵向测量 —— 单步能力一直在涨，长链闭环仍然走不到终点

Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios

英国 AI 安全研究院 (AISI) • arXiv:2603.11214v3 (2026.03)

18 / 34^步

最佳单次完成步数

32 步企业网络靶场「The Last Ones」• Opus 4.6 • 1 亿 token 预算

15.6^步

五次运行的平均步数

同模型同预算下方差极大 —— 最差的一次只走到第 11 步

1.7 → 9.8

十八个月的同预算提升

1000 万 token 下, GPT-4o 到 Opus 4.6 提升 5.8 倍

1.4 / 7^步

工控靶场的平均步数

电厂冷却塔靶场「Cooling Tower」• 换到工控场景能力近乎归零

测评环境 • 企业网络靶场「The Last Ones」 32 步 / 9 个里程碑 • 人类专家约 14 小时

由 SpecterOps 与 Hack The Box 承建 • 真实软件搭出错误配置的域环境与流水线, 需跨网段逐段推进



工控靶场「Cooling Tower」• 7 步 • 人类专家约 15 小时

目标: 让电厂冷却塔失效

运行环境: 刻意最小化的标准脚手架

只测模型本身



企业靶场「The Last Ones」的完成断崖 • Opus 4.6 @ 1 亿 token (5 次运行)

虚线空格 = 五次运行全部为零 • 步数走完 69%, 按人类工时只覆盖 42% (约 14 小时里的 6 小时)

步 1 — 13 • 里程碑 1—4

步 14 — 22 • 里程碑 5—6

步 23 — 32 • 里程碑 7—9



侦察与初始接入 • Kerberos 横移 • 浏览器凭据窃取 • NTLM 中继

Web 应用提权 • 二进制逆向与密钥恢复

C2 枢纽与 CI/CD 凭据 • 供应链投毒 • 数据外带

▲ 五次平均 15.6

▲ 最佳单次 22

约 1.8 小时

13%

约 4 小时

29%

约 8 小时

58%

交叉验证 • TSecBench 第一期联合测评六大能力域完成率 13 个安全智能体 × 4 个国产模型 • 三轮取优 • 完整解题率仅 8.33%

单点能力九成以上, 多阶段渗透全城最低 —— 与 AISI 十八个月纵向测量指向同一结论



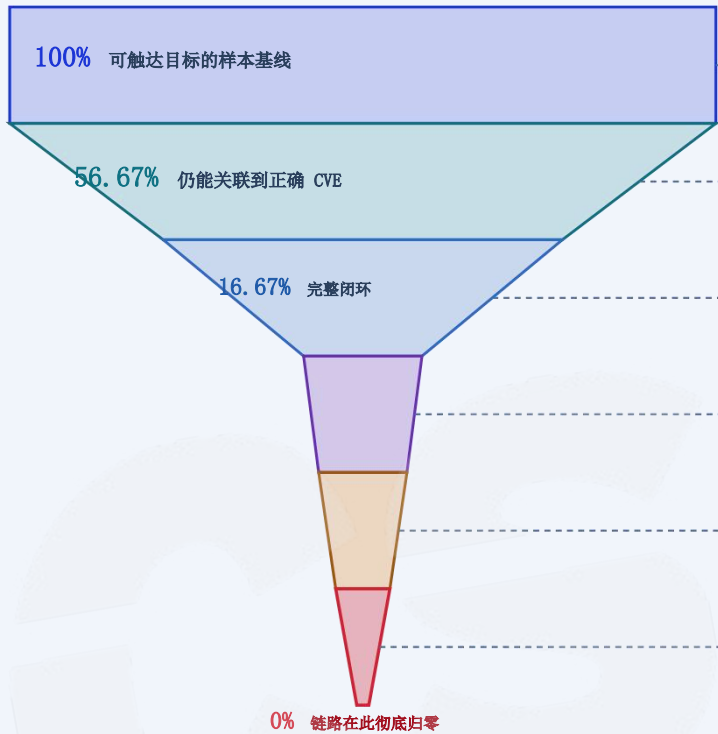
多阶段能力不足复盘：一条越走越窄的攻击链

Multi-Step Breakdown — A Failure Funnel Autopsy

单点能力九成以上，完整闭环不足两成
横向基准漏斗 × 英国 AISI 十八个月纵向测量 • 双来源互补

多步骤攻击链失败漏斗 Failure Funnel

各层取自不同基准的最佳公开结果，作量级示意，非同一批样本的严格转化率



① 侦察与资产识别

端口 • 服务 • 指纹与攻击面枚举，全链路中最成熟的一环

100%

样本基本无损耗

② 漏洞发现与假设形成

无提示条件下能力大幅下降，探索不足是首要失败来源

56.67%

能关联到正确 CVE

③ 利用代码生成与组合

能指出漏洞却生成不了有效 Payload，83.3% 在此中断

16.67%

完成发现→组合利用

④ 多漏洞串联与横向移动

跨步骤的凭据、权限与目标状态无法稳定传播

12.5%

CVE-Bench 最佳 Agent

⑤ 端到端验证与修复复测

需要回到起点自证结果为真，是最后也是最薄的一环

7.6%

CyberGym-E2E 成功率

⑥ 有主动防御的真实环境

出现 WAF / 防火墙后，能力直接归零而非慢慢下降

0%

PACEbench 全员零分

断裂的四个技术原因

不是「模型不够强」，是四处结构性缺口

状态维护失效

前置条件、凭据、权限与目标状态无法跨步骤稳定传播，越往后累积误差越大



防御环境完全未建模

PACEbench 中所有前沿模型在 WAF 绕过场景得 0 分，训练语料里几乎没有「被拦截后怎么办」



CTF 与企业环境的结构性差异

靶题目标已知、漏洞必有、无主动防御、Flag 可自验；真实网络四条全部不成立



缺少失败恢复与回溯策略

一条错误路径被写进上下文后会持续污染判断，模型不会推翻自己已经「确认」过的结论

阶段能力不能相乘 —— 最直接的一组对照

7.6%

端到端成功率

CyberGym-E2E：发现 → 利用 → 验证
→ 复测全流程

82.3%

单阶段基线

同一系统只做补丁生成这一个环节时的
成功率

10×

相差一个量级

阶段能力不会自然相乘，串起来即坍塌

互补证据 • 换一把尺子量：英国 AISI 企业靶场「The Last Ones」纵向测量 32 步 / 9 个里程碑 / 人类专家 ≈ 14 小时

上面的漏斗按「样本比例」衰减，这里按「攻击链位置」衰减 —— 同一结论的两种量法

刻度一 • 按步数 (Opus 4.6 @ 1 亿 token，五次运行最佳)

止步 22 / 32 步



刻度二 • 按人类专家工时等比缩放 (830 分钟映射到同一总宽)



步数上走完 69%，按人类工时只覆盖 42% —— 越靠后的步骤越重：前 13 步每步 2—4 个动作，后段每步 8—16 个。

两张图说的是同一件事

左侧漏斗的 ③ 利用代码生成与组合，正是 AISI 里程碑 4—5 的 NTLM 中继卡点 —— 需要同时协调多个并发进程。

漏斗的 ⑤ 端到端验证，对应里程碑 7—8 的 CI/CD 长链 —— 中间任一环节断掉，整段归零。

15.6 / 32

五次运行平均步数

Opus 4.6 @ 1 亿 token，最佳 22 步、最差 11 步

1.4 / 7

工控靶场几乎原地

换到电厂冷却塔场景，能力塌回近乎零

+59%

唯一有效的手段

预算 1000 万 → 1 亿 token；改提示词、给示范轨迹、修
工具调用均无显著提升

03 防守篇

DEFENSE · STOP MATCHING PAYLOADS, START WATCHING BEHAVIOUR

打不赢指纹，就改看行为；拦不住请求，就改做引流。

当攻击方的长相可以零成本更换，基于特征的拦截就自然失效；

本篇六套方案的共同选择：把「识别是谁」换成「观测在做什么」，把「阻断」换成「引流·消耗·画像」。

本篇回答三个问题

THREE QUESTIONS

01 为什么旧防线先失效？
指纹、UA、Payload 都能零成本改写，而按人类节奏校准的防线又跟不上机器节奏 —— 于是先有了「用 AI 守 AI」的自主防御体系

02 不看长相，那看什么？
用诱饵引导 + 行为观测：它是否理解语义、是否按反馈调整策略，这两件事改不掉 —— 因为它就是「智能」本身

03 认出来之后做什么？
不阻断：静默引流，消耗它的 token 与时间，同时把框架、模型、提示词套出来，转成可追溯的画像档案

本篇包含 六页，按「先用 AI 守住 → 再把它骗住 → 最后把海量报告管起来」排列

IN THIS CHAPTER

3.1 AitherOS 自主防御

以 AI 防御 AI：12 层 244 微服务

3.2 Project Perception

微软全域防御智能体集群

3.3 AI 欺骗防御盘点

四类典型方案与诱饵范式

3.4 灵鉴·反代欺骗

语义诱捕与攻击 Agent 画像

3.5 AI 运营 SRC

定级公平·复测·闭环修复

3.6 漏洞自动化验证

机器验事实，人下裁决

三带纵深：不拦，逐层消耗

DEFENCE IN DEPTH

左侧楔形 = 攻击方仍在投入的资源，越往下越细：它花的越多，拿到的越少

观测带 OBSERVE

全量记录请求体、响应改写命中与访问序列，不做判断，先把「它做了什么」完整留下来

前提：自定义端点的请求体必须完整入库

欺骗带 DECEIVE

剧本化假页面·响应注入诱饵·专设诱捕端点，诱导它主动把自己的框架 / 模型 / 提示词交出来

只有 AI 会读懂诱饵 —— 这就是最好的分流器

处置带 CONSUME & PROFILE

识别后不阻断，静默改道进消耗环境，一边耗掉它的算力预算，一边完成画像与溯源

攻击成本从我方承担，改为对方承担

长相确实不可信：70 万站点实测 79.7% 放行伪造 UA，80% 的 Agent 不做正确身份标识。
数据来源 DataDome Galileo H1 2026，为厂商遥测、未独立复现。

以 AI 防御 AI: AitherOS 自主攻击防御方案

攻击面已从「已知依赖里的已知漏洞」变为「任何大模型能推理出来的东西」—— 传统防线按人类节奏校准，必要但不充分

Aitherium • Defending Against Autonomous AI Attackers

触发事件 hackerbot-claw (2026.02): 自主扫描 47,000 个公开仓库, 7 个目标中 5 个达成远程代码执行

AitherOS 12 层架构

244 微服务 • 20 自主 Agent

L1 基础设施原语 Infrastructure

L2 容器编排 Orchestration

L3 自动化 Automation

L4 感知 Perception

L5 认知 Cognition

L6 记忆 Memory

L7 Agent 编排 Orchestration

L8 安全复合体 Security Compound ★ 7 服务

L9 GPU 加速 Acceleration

L10 训练 Training

L11 交互界面 Interface

L12 对外网关 Gateway

AitherChaos

对抗红队

七宗罪攻击框架, 全天候对自身发起攻击

WRATH 摧毁 • GLUTTONY 资源耗尽

SLOTH 绕校验 • ENVY 身份窃取

PRIDE 幻觉 • LUST 社工 • GREED 外泄

不只在审计期运行, 而是持续运行并自行裁决

AitherSentinel

主机端 EDR

Bloom filter 快速放行已知良性二进制

七类威胁: 勒索 / Rootkit / 数据外泄

白利用 / 持久化 / 挖矿 / 进程注入

学习行为基线, 可直接隔离可疑进程

理解智能体操作系统「正常形态」的 AI 版 EDR

AitherSentry

网络威胁检测

端口扫描指纹识别, 关联暴力破解

防注入 SQL / NoSQL / XSS / SSRF

路径穿越 / 命令注入

与网格策略引擎联动, 实时下发拒绝

三级封禁 TEMP_BLOCK • PERM_BLOCK • ISOLATE

AitherGuard

零泄漏红队

专测系统提示词泄漏与凭据暴露

三智能体协同: Strategist 探索攻击树

Attacker 执行探测 / Evaluator 分级

以向量记忆学习有效探测策略

泄漏五级 NONE→HINT→PARTIAL→FULL→CRITICAL

AitherInspector

数据防泄漏

四级分类 PUBLIC / INTERNAL /

CONFIDENTIAL / RESTRICTED

四种裁决 ALLOW / BLOCK / REDACT /

QUARANTINE, 接入 Flux 包流实时检查

高于 PUBLIC 的内容一律阻断或脱敏

AitherRoundTable

十二骑士巡逻

每位骑士可触发 31+ 种神经元推理信号

Guinevere 主导升级 / Morgana 守密钥

Viviane 护记忆 / Isolde 盯依赖

经议会协议联动全体骑士

一旦检出越狱, 全体骑士同时触发调查

AitherJail

欺骗诱捕

七宗罪容器化, 以假凭据诱捕攻击者

逐条记录逃逸尝试的全量痕迹

把每一次攻击交互反向转为训练数据

已并入 Chaos 统一调度

攻击者以为在提权, 实际在为防御方标注样本

四大防御支柱

方案总纲

- 持续对抗测试 —— 全天候自我攻击
- 实时行为监控 —— 主机与网络双侧
- 结构性信任边界 —— 标记 • 隔离
- 语义理解 —— 推理意图而非匹配

攻击者已经自主了, 防御也必须是。

纵深防御 • 完整杀伤链

每个检查点都必须被真正执行



01

输入闸门

IntakeGuard

26 条正则快速拦截

< 5 ms



02

语义裁判

SafetyJudge

本地小模型判意图

≈ 100 ms



03

提示净化

PromptGuard

显式包裹不可信输入

信任边界



04

运行时监控

Runtime Monitor

主机与网络双侧并行

行为异常



05

输出检查

Output DLP

扫凭据与个人信息

出口阻断



06

调用方隔离

Caller Isolation

按身份而非内容授权

5 × 5 权限



07

能力令牌

Capability Token

无签名即不可执行

默认拒绝

外部请求 —— 七道闸门 —— 安全响应

面向全域 IT 的防御智能体集群：微软 Project Perception

下一代安全系统的分野，不在于能产生更多告警，而在于能否持续感知 • 推理 • 行动 —— 用 AI 防御 AI

Microsoft • A New Cyber Stack for the Age of AI
2026.07.28 发布 • 08.03 进入受限公开预览（邀请制）• 按 Security Compute Units 计量计费

全新 Cyber Stack • 六层

自底向上重新设计

⑧ 执行器 Actuators
把决策转成真实防护动作，深度接入安全产品

⑤ 智能体 Agents ★ 6 智能体
红蓝绿三队跑通安全工作流

④ 编排框架 Harness
跨工作流协调模型与智能体的调度层

③ 模型 Models
前沿模型 + 专业安全模型，按任务动态选择

② 安全情境 Security Context
原始信号转为 token 高效的可推理语义

① 信号与传感器 Signals
身份 • 端点 • 应用 • 数据 • 云 • AI 系统全域感知

红队 侦察智能体

Recon Agent

以攻击者视角持续踩点、测绘整个环境
产出攻击路径、咽喉点与高价值资产清单
识别配置缺陷与过度授权的身份

在真实攻击者发现之前先把缺口暴露出来

蓝队 攻击调查智能体

Attack Investigation Agent

跨告警、事件与信号重建完整攻击故事
给出结论判定、时间线与技战术还原
列明失陷指标与全部受影响实体

自主完成二线级别调查，直接交付可用结论

蓝队 威胁情报智能体

Threat Intelligence Agent

把情报文章解析为结构化情报对象
抽取威胁团伙画像与 ATT&CK 技战术映射
产出失陷指标、漏洞编号与猎杀查询语句

情报不再是读物，而是可直接执行的输入

蓝队 检测编写智能体

Detection Authoring Agent

主动找出威胁覆盖的盲区
为新出现的技战术自动编写检测规则
随威胁演进持续更新防线

把「发现了新手法」直接变成「已经能检出」

蓝队 告警研判智能体

Triage Agent

对告警判定真实威胁还是误报
每个结论都附完整推理依据，可追溯
新告警产生即自动触发，无需人工排队

从分析师反馈中持续学习，准确率随用随涨

绿队 态势优先级智能体

Posture Prioritization Agent

按真实风险而非漏洞评分排序
综合暴露面、可利用性与资产关键度
跨云、终端与 AI 资产统一给出修复计划

明确回答先修哪一个、以及为什么是它

三队协同 • 闭环剧本一个剧本派一支队伍，上一个智能体的输出即下一个的输入

① 蓝队 • 取情报
Threat Intelligence Agent
解析与本组织真正相关的攻击战役
与威胁团伙，输出结构化情报

② 红队 • 测暴露
Recon Agent
依据情报模拟攻击者横向移动，
测绘出可达高价值资产的攻击路径

③ 绿队 • 定次序
Posture Prioritization Agent
评估可利用性与资产关键度，
给出收益最高的修复动作序列

④ 蓝队 • 补盲区
Detection Authoring / Investigation
为新技战术补写检测规则，并调查
本轮过程中暴露出来的真实事件

← 回流形成闭环：每一轮的结论沉淀进共享安全情境，下一轮智能体直接复用，无需从原始信号重建上下文

红队找缺口 → 蓝队判风险 → 绿队去修复

■ 高影响动作一律暂停，等人工批准后才继续

首个专业安全模型的基准成效 • CyberGym

成本较现役配置 ↓ 近 50%

MDASH + MAI-Cyber-1-Flash
微软自研专业安全模型

96%

Mythos（对比基线）
榜单同期竞品

84%

口径为 any-crash 分数，领先幅度 12 个百分点（基线值由此推算）

首批客户实测：4 周 的威胁情报分析压缩到 4 小时

两条路线并非替代关系，而是分层互补

选型时须先分清防护对象

	AitherOS 安全复合层	Project Perception
防护对象	多智能体操作系统自身（244 微服务 / 20 Agent）	全域数字资产：身份 • 端点 • 应用 • 数据 • 云 • AI 系统
威胁模型	智能体越狱、提示注入、沙箱逃逸、内部提权	外部攻击路径、告警、事件、暴露面与安全态势
落地形态	自建 12 层技术栈之内的第 8 层，内生自我防护	接入 Defender / Entra ID / Sentinel / ARM，重构安全运营

把「AI 好骗」变成防御优势：AI 欺骗防御的四类典型方案

自动化攻击者最大的结构性弱点是**缺乏情境感知** —— 它只生成看起来合理的下一步，却从不确认自己所处的世界是否真实

Deception Defense against Autonomous AI Attackers

云安全联盟建议企业立即部署欺骗能力 • 某厂商半年内记录到 8,990 万次外部诱饵交互

AI 攻击者为什么特别好骗

这三条不是缺陷，是自回归生成的固有属性

只求合理，不验真伪

模型生成的是上下文里「看起来合理」的下一步，却从不检验上下文本身是否真实

以隐蔽换速度

为求快而全量枚举，必然踩到诱饵；人类几轮就识破的假环境，Agent 会一直跑下去

上下文即指令

读到的任何内容都可能被当成指令执行 —— 这是提示注入的病根，也是防御方的杠杆

人类攻击者几个回合就会起疑，自动化 Agent 不会 —— 它没有「不对劲」这个概念



① 生成式蜜罐

Generative Honeypot

让大模型实时扮演一台不存在的机器

- 用系统提示词即可让模型扮演 Linux 主机、物联网设备或数据库
- 换一个诱捕目标只需改提示词，不必为每类目标另写一套代码
- 背后没有真实数据库与服务器在跑，自身被攻陷的风险趋近于零

拖住攻击者的时长约为传统规则型蜜罐的 3 倍 • 原型仅约 80 行代码



③ 上下文炸弹

Context Bombing

把提示注入反过来打在攻击者身上

- 在诱饵资源里埋一段会触发模型自身安全护栏的短文本
- 攻击 Agent 一旦读入即进入拒绝状态，整个会话被污染且往往无法恢复
- 同一次读取既触发警告又中止攻击 —— 检测与处置合并为一个动作

作用在模型权重层面的安全训练上，而非可绕过的输出过滤器 • 但不存在通用字符串



② 诱饵与蜜标

Canary / Honeypot

在真实环境里撒下只有攻击者会碰的东西

- 播撒假密钥、假云角色、假存储桶、假域账号与假文档
- 合法业务永远没有理由访问它们，任何一次交互都是确定性入侵信号
- 不依赖特征库与行为基线，因此几乎不产生误报，也不需要人工研判

95.9% 的攻击在首个关键动作之前告警，中位数领先 8 分钟



④ 分流欺骗世界

Deception Gateway

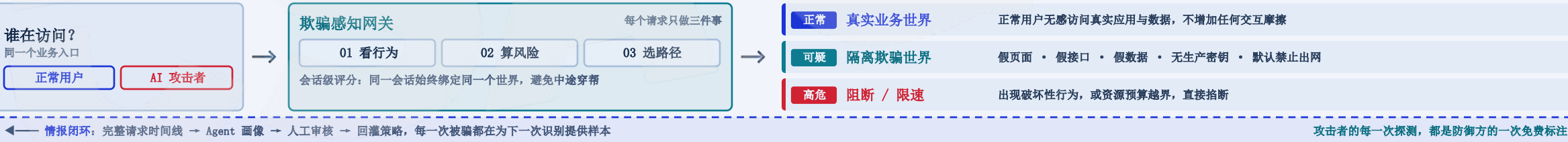
同一个入口，让不同访问者进入不同的世界

- 网关对每个请求只做三件事：看行为 • 算风险 • 选路径
- 假世界里是假页面、假接口、假数据，无生产密钥且默认禁止出网
- 按会话级评分绑定世界，同一会话始终落在同一边，避免中途穿帮

让 Agent 拿到自洽但错误的反馈，攻击全程接触不到真实资产

落地形态 • 同一个入口，不同的世界

参考架构 AEGIS MIRAGE



上下文炸弹实证 • 云靶场 152 次对照攻击

五个前沿模型 • 约 300 个资源 • 10 条预设攻击路径

拿到完整账号管理员权限	57%	→	5%
完全攻陷（管理员 + 持久化）	36%	→	1%
完成任意一条攻击路径	91%	→	15%

护栏越强的模型跌得越狠 —— 最强模型管理员权限从 93% 直接归零；且没有任何一次成功攻击是在未触发警告的情况下完成的

三条必须一起讲的附注

威慑效应 不部署也有效：仅在提示中告知模型「网络里可能布有欺骗」，完全攻陷率即从 20% 降到 3%

双向性 这把刀两头都快 —— 2026 年 6 月已发现恶意软件包用同样手法触发护栏，专门躲避大模型扫描器

边界 它不修复任何配置缺陷；攻击方可换用无护栏模型、先摘要工具输出、或把拒绝隔离在子进程里

灵鉴：把攻击 Agent 骗进来、养起来、画出来

腾讯自研 · 反向代理层的 AI 攻击 Agent 语义欺骗与画像系统 · LINGJIAN · AGENT DECEPTION & PROFILING

检测 → 响应 · 全链路架构
Caddy + replace-response + caddy-cgi + 自研剧本层

一句话思路 攻击 Agent 的「AI 属性」不看长相，只看行为

BEHAVIOR, NOT FINGERPRINT

不看长相

Payload · UA · 指纹可零成本改
传统 IOC / 签名规则天然失效

只看行为

诱饵引导 + 行为观测
语义理解 + 反馈适应 = AI 属性

不阻断

静默引流 · 消耗 · 画像溯源
攻击者以为仍在真实系统

技术实现原理：薄底座 · 厚剧本 —— 转发与替换全部交给 Caddy，我们只做规范、编译与分析

ARCHITECTURE

复用层 · 现成能力，不重复造轮子 两个插件源码内置本项目、自行维护

Caddy

REVERSE PROXY

反向代理 · 请求匹配
响应与响应替换全链路

replace-response

BODY REWRITE

响应文本替换
【已增强】记录实际替换次数

caddy-cgi

CGI EXEC

CGI 脚本执行
承接自定义端点与动态诱饵

自研层 · 我们真正做的事 一套「剧本」规范 + 剧本→配置的编译规则 + 运行与画像系统

01 剧本规范

欺骗剧本的结构
诱饵 · 假页 · 端点

02 剧本编译

剧本 → Caddy 配置
匹配规则 + 替换表

03 运行与日志

Caddy 进程托管
日志解析与可视化

04 画像分析

算法 + LLM 双通道
从日志还原攻击者

05 实时告警

前端事件流推送
维护 Agent 画像档案

日志的关键设计 完整记录访问自定义端点的请求体，后置分析才能从中还原 Agent 的框架 / 模型 / 提示词

为什么是现在：靠「长相」识别已经失效

WHY NOW · 2026

53%

全网流量已由自动化程序产生
人类份额降至 47%；恶意 bot 独占 40%

12.5×

AI 驱动攻击同比增幅
日均拦截量 200 万 → 2,500 万

79.7%

伪造 ChatGPT UA 被直接放行
约 70 万个可达站点实测，无拦截无挑战

「长相」不可信的五条证据 DataDome / Cequence 2026 实测

80% 的 AI Agent 不做正确身份标识
只靠可随意伪造的 User-Agent，而非 IP 段 / 反查 / Web Bot Auth

Meta-ExternalAgent 被冒充 1,640 万次
ChatGPT-User 790 万次；知名 Agent 身份 = 现成的通行证

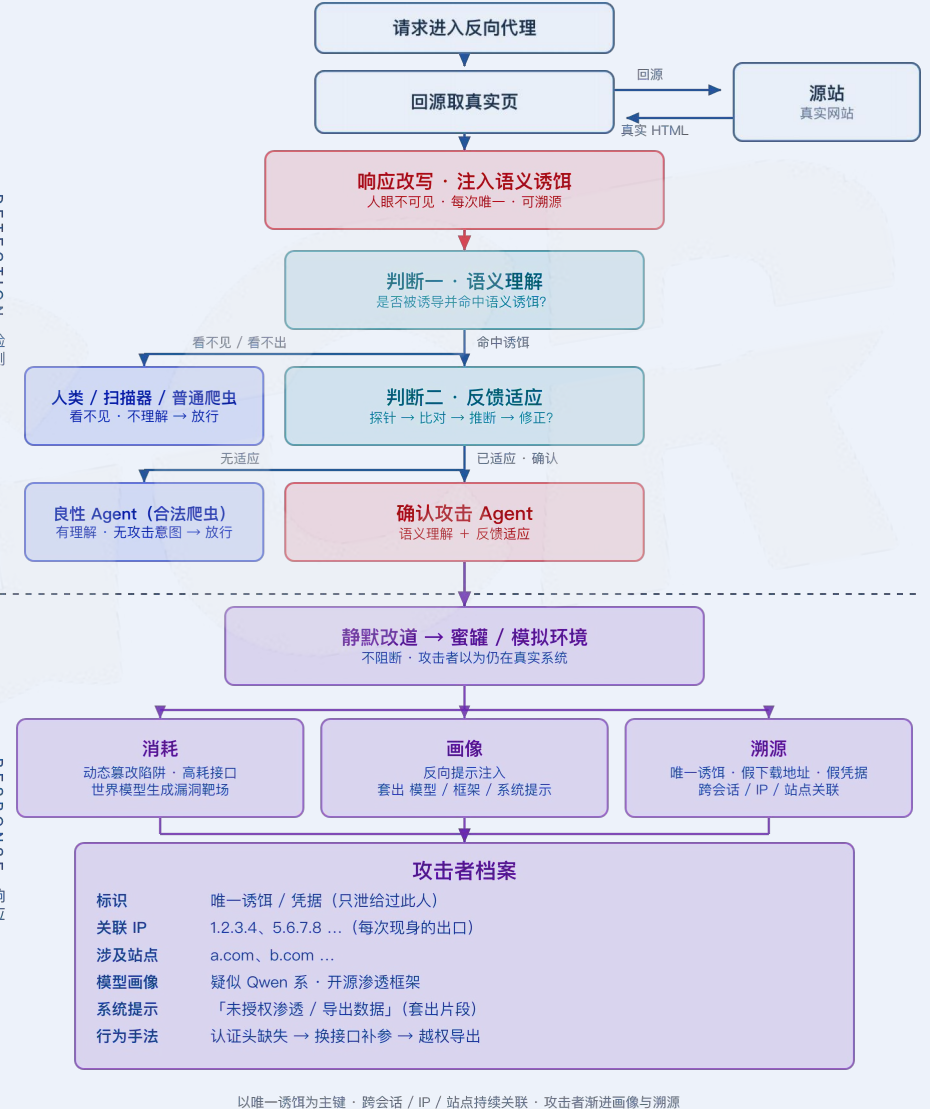
Playwright 1.53+ navigator.webdriver = false
最基础的自动化指纹信号直接归零，且是上游静默变更

24 小时内生成 297,556 个行为画像
≈ 每 300 毫秒一个新变体，基于签名的策略永远落后

82% 被拦截流量的 bot 置信分 < 50
它们一点都不像 bot —— 真正暴露它的是行为，不是长相

DETECTION
检测

RESPONSE
响应



AI 运营 SRC：关停不是答案，换掉审核台才是

定级公平 · 复测全域 · 闭环修复 · 并发吞吐
AI-OPERATED SRC NEW OPERATING MODEL

2026 年的 SRC 关停潮 —— 不是白帽子不行了，是审核台垮了



空心节点 = 海外 实心节点 = 国内 轴色由橙转红 = 事态升级

但 curl 给出了一个反证

<5% → 15~16% 报告有效率

取消赏金后提交量翻倍、有效率反升 —— 噪声来自激励与审核错配，不是 AI

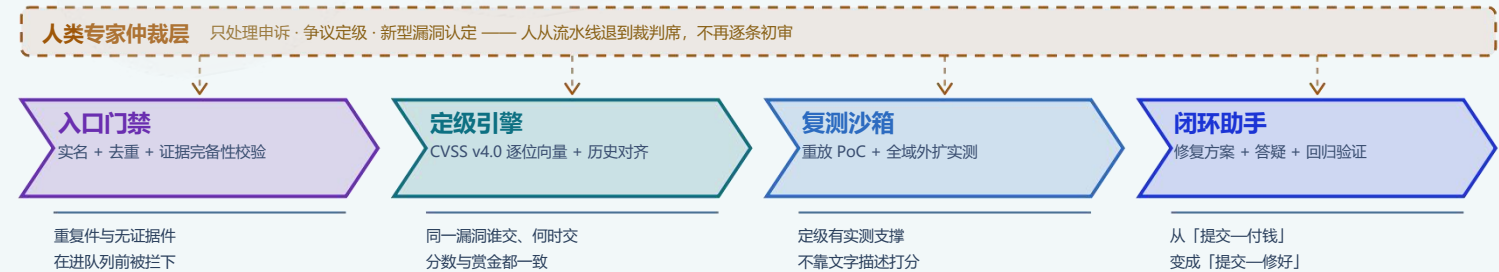
同一条流水线，四个环节 —— 人处处卡住，AI 逐个解开

上轨断续 = 人工审核台的通行能力 下轨贯通 = AI 审核官



新时代 SRC 运营体系 —— 机器跑流水线，人坐裁判席

AI-NATIVE SRC OPERATING MODEL



关停 SRC，漏洞不会消失，只会换一个买家。
该换掉的不是白帽子，是跟不上产量的人工审核台。

白帽子生态不能断

SRC 的本意是搭一座与白帽子对话的桥；桥断了，高危漏洞只是换了流向

行业目前只走了半步

AI 去重预判、强制实名、限速封禁都只在过滤入口；定级、复测、闭环仍靠人扛

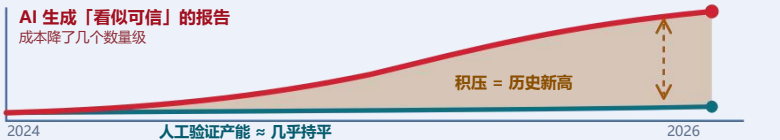
漏洞自动化验证：机器验事实，人下裁决

AI 让「发现」变便宜了，没让「验证」变便宜 · AUTOMATED VULNERABILITY VALIDATION

六级分层验证流水线

HackerOne · Bugcrowd · Intigriti · YesWeHack · Elastic · OpenAI 实践综合

问题的本质：复杂度被翻转了

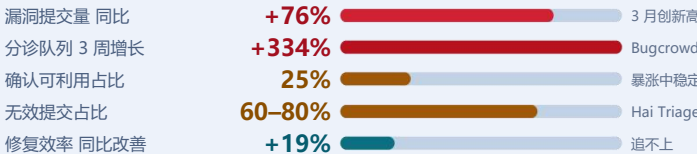


过去：难做 · 易验 (≈ NP)

现在：易做 · 难验 → 数量永远获胜
持久位移，不是临时峰值

2026 平台侧实测

HACKERONE / BUGCROWD / ELASTIC



高危 + 严重占比 32% (历史基线 26-28%)
Elastic 单家 SRC 半年收 1,390+ 份 > 前两年之和

后果：验证台先崩

COLLAPSE

curl 2026.01
确认率 15% → 低于 5%; 1月前 21 天
20 份提交 0 个真实漏洞 → 终止赏金

IBB 2026.03
HackerOne 暂停互联网漏洞赏金：
发现量与维护者处理能力严重失衡

Linux 内核 2026.05
安全列表从每周 2-3 份涨到每天 5-10 份
Torvalds 称几乎无法管理

GitHub：不停，改规则
低危改发周边纪念品；所有提交强制
可运行 PoC —— 上游门槛才是第一道闸

解法：六级分层验证流水线

—— 用便宜的层拒掉大头，把昂贵的层留给少数，把最后一裁留给人

PIPELINE



成本与效果：可查的全部实数

BENCHMARK

方案	准确率 (口径)	成本 / 终裁
Elastic 自建	85% 与人工一致	≈\$2 / 份 · 人工终裁
验证 harness	98% (38 例 CVE)	未披露 · 与模型无关
H1 Validation	95% 厂商宣称	未披露 · 研究员兜底
单份成本拆解 · Elastic 85% 基于 764 份已知结论实测 · Intigriti 去重 95%		
L3 分析 (全量)		\$0.50-1.15
L4 复现 (≈30%)		\$0.80-4.90
人工 (对照)		30-60 分钟

组织收益 (Shopify) 效率 +62%; 分析师上手 8 → 4 个月;
单份验证 20 → 5 分钟; 数月来首次 Inbox Zero

Time to Validate

MTTR

Exposure Backlog

核心观点：把一团「判断」拆成两半 —— 而这条分界线正在移动

TAKEAWAY

不是让 AI 代替你判断，
而是把「判断」拆成两半：
事实层已经定型，价值层还在争夺

一份待处理报告
过去由一个人一次性
搭环境 · 读码 · 猜

事实 能不能复现？能不能利用？
沙箱照白帽步骤真跑一遍，结果客观唯一、可复查

机器答 是 / 否 · 已定型

价值 算多严重？值多少赏金？
两边都有系统性偏差 ——
谁更可控，谁就拿这一票

AI: CVSS 虚高
人: 经验偏见

规则收紧后
AI 初裁 · 人终审

价值层 = 下一轮 SRC 纠纷重灾区
AI 偏高: 夸大 CVSS、被报告措辞锚定 —— 规则可压
人偏移: 熟悉度 · 预算 · 关系 —— 不可自证一致
今天六家平台仍无一接受 AI 自主定级关单; 规则
越细，这条线越往 AI 侧移 —— 但争议只会更多

THE FACT IS SETTLED · THE VERDICT IS CONTESTED

04 未来展望

OUTLOOK · THE NEXT LOOP IS TRAINED, NOT PROMPTED

下一段增量不在提示词，在训练循环里有没有一个真环境。

靶场提供可验证的奖励，比赛留下带判定的轨迹，标准提供跨版本的验收尺 —— 三者拼起来正好是一条管线；

剩下的问题不是能不能做，而是谁先把它跑起来、以及跑起来之后拿什么约束它。

本篇回答三个问题

THREE QUESTIONS

01

格局往哪走？

传统安全问题更可能被跨界降维解决：算力与规模改写成本结构，而不是在旧框架里继续堆规则 —— 英伟达下注这条赛道不是偶然

02

技术往哪走？

从「喂语料」到「在靶场里练」—— 安全是少数奖励天然可验证的领域，拿到 flag、拿到权限是二值客观事实，正是强化学习最稀缺的条件

03

边界在哪里？

能力外溢、责任归属、价值判断偏差三条同时存在，事实层已经定型，价值层仍在争夺 —— 红线要在能力到位前先画好

本篇包含 五页，按「格局怎么变 → 技术怎么走 → 怎么落地 → 边界在哪」排列

IN THIS CHAPTER

4.1 降维解决传统安全

跨界玩家改写安全的成本结构

4.2 安全模型训练方案

从喂语料到在靶场里练

4.3 本地化 SecSophon

DGX Spark × 断网可用整机

4.4 AI 渗透伦理红线

事实层可自动，价值层仍在争夺

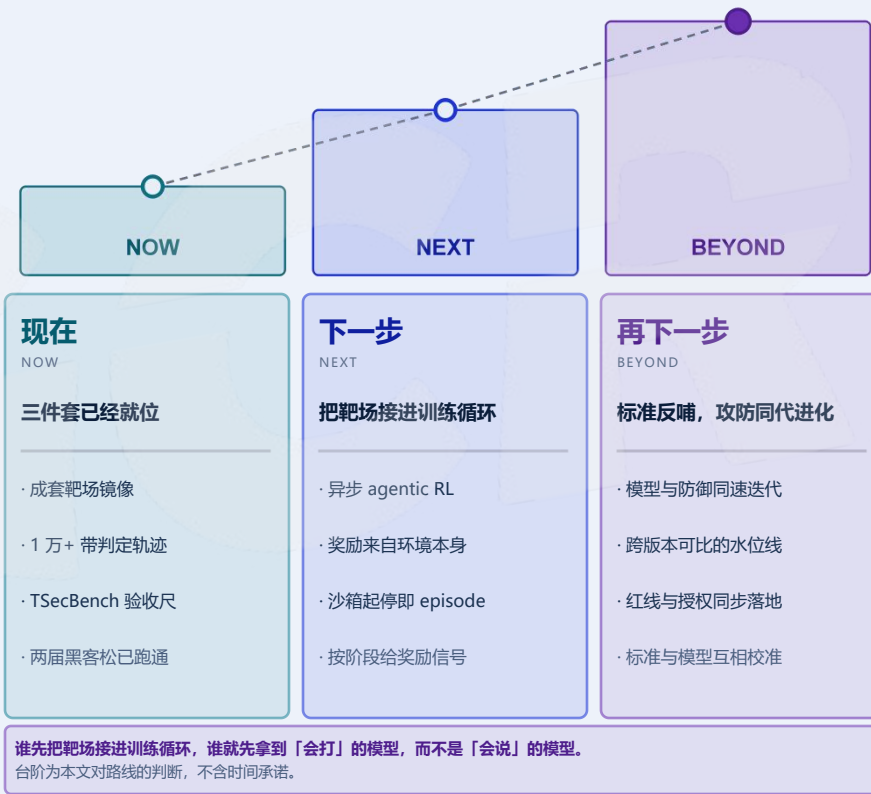
4.5 可信可控框架

给自主 Agent 装上刹车与审计

三级台阶：一条能反复跑的管线

ROADMAP

每一级都建立在上一级的产出之上 —— 缺任何一块，循环就跑不起来



历史一再重演：**同维度加固**没赢过，赢的都是**另一个行业的跨界降维**

红 = 在原赛道上更努力 绿 = 换一个赛道 · 赢家都是卖「覆盖能力」的基础设施商

入室盗窃

打不开 → 跑不掉

防盗门 → A 级锁 → B 级 → C 级叶片锁芯

视频监控 + 人脸识别

跨界者：安防电子（海康威视 / 大华），不是锁具厂

假钞流通

认不出 → 不用现金

防伪线 → 水印 → 荧光油墨 → 验钞机

移动支付

跨界者：互联网公司（微信 / 支付宝），不是印钞厂

信用卡盗刷

验凭证 → 验行为

签名核对 → 密码 → CVV → 短信验证码

实时行为风控模型

跨界者：数据科学团队，不是卡片工艺部门

汽车被盜

偷不走 → 藏不住

方向盘锁 → 声光防盗器 → 芯片钥匙

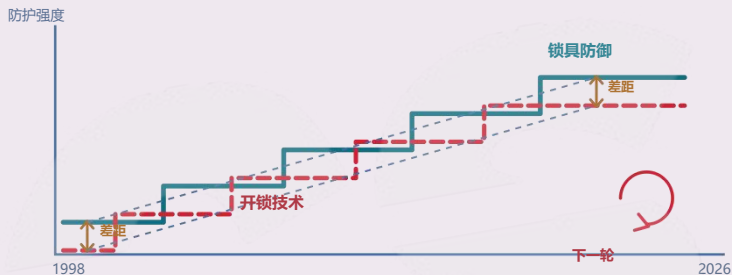
车联网 + 定位 + 远程锁车

跨界者：通信与车企 IT，不是机械防盗厂

两种解法，几何形状完全不同

同维度加固

ARMS RACE IN PLANE

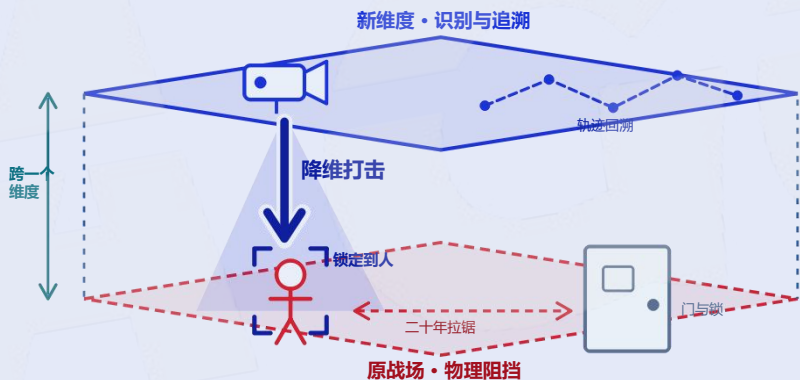


投入一直在涨，相对安全度一直没动

两条曲线平行上升 —— 这就是同维内卷的几何定义

跨界降维

VERTICAL STRIKE FROM ABOVE



小偷的账本被改写了

净收益 = 赃物 - 成本 × 被抓概率

传统安全

推高「成本」

降维解决

拉满「被抓概率」

东阳 2015：视频侦查使 78% 的盗窃案当日侦破

锁一把都没换 —— 换的是敢不敢下手

AI 之于网络安全，就是当年的摄像头之于治安 —— 同一个降维

把稀缺的「人」换成可无限复制的「机器」，覆盖率从抽样变全量

治安

警力不足 · 只能抽查
人盯人，覆盖率天然封顶



视频监控 → 全域覆盖 · 事事留痕 · 必被追溯
2017 年全国已建成 2000 万+ 摄像头，警力没变，覆盖变了

网络安全

专家不足 · 只能抽检
高水平安全人才是硬约束



AI Agent → 全量审计 · 步步留痕 · 必被发现
Mythos 挖出 27 年无人发现的 OpenBSD 缺陷，专家没变，覆盖变了

NVIDIA

英伟达为什么下场

上一局赢家卖摄像头，这一局卖算力

编程助手：问完就停

安全 Agent：7×24 不间断

突发推理

常驻推理

AI 的下一个千亿美元应用

黄仁勋 9/10 高盛大会：体量可比数据中心业务
SafeMind × Nemotron：更准，成本低 99%

核心观点 THE POINT

降维不是把防御做得更强，
而是改写攻击者的收益函数。

AI 不让漏洞更难利用，是让它来不及被利用。

但降维会把瓶颈推向下游：摄像头带来了看不完的视频，
AI 找洞正在带来修不完的漏洞 —— 下一次降维在运营侧

安全模型训练方案：从「喂语料」到「在靶场里练」

未来走向 · 以 Miles v0.1 为参照：一套生产级后训练系统，把智能体强化学习从论文推到可上生产的工程 · RadixArk, arXiv:2609.08368v1

环境 → 吞吐 → 保真 → 同步

四个卡点，决定安全模型能不能真训起来

范式判断 安全模型的下一段增量，不在语料规模，在训练循环里有没有一个真环境

CORPUS → ENVIRONMENT

过去 · 喂语料 SFT ON SECURITY CORPUS

把漏洞报告、WP、工具文档灌进去

学会的是「像安全专家那样说话」

未来 · 在靶场里打 AGENTIC RL

模型自己打，环境给判定，判定当奖励

学会的是「真能把靶机拿下来」

安全是少数「奖励天然可验证」的领域 —— 这正是强化学习最稀缺的条件

· 拿到 flag / 拿到权限是二值客观事实，不需要人打分，也不需要裁判模型

· 判定来自环境本身，模型无法靠「说得好听」骗到分，刷分空间极小

· 两届黑客松已沉淀 1 万+ 次带判定记录 + 成套靶场镜像，环境不用从零造

要把靶场接进训练循环，得先过四道工程关 —— 每一关上面是安全场景的卡点，下面是 Miles 给出的解法与实测

FOUR GATES

1 环境 ENVIRONMENT

安全场景的卡点

渗透是长程多轮 + 调工具 + 有真实副作用，不是一问一答；靶场起停慢、成长长尾极重



可插拔 Agentic Rollout + 沙箱后端

- 每个 episode 由任务镜像起独立沙箱，结束即销毁
- 已接 AgentENV / Daytona / E2B / Modal
- terminal-bench 式方可直接照搬到靶场

3 层

Agent / Generate / Rollout 三处可接管

2 吞吐 THROUGHPUT

安全场景的卡点

同步轮替时整批要等最慢一条轨迹；渗透轨迹长度能差一个数量级，GPU 大半时间在空转



完全异步 RL + 样本级替换 + 亲和路由

- 每条轨迹一完成就释放位置，不等整组
- 有界缓冲 + 悲观陈旧度上限，超龄直接丢弃
- 同一 episode 绑定同一引擎，多轮不重算前缀

96%

前缀缓存命中率 · 中位步时 263 秒

3 保真 FIDELITY

安全场景的卡点

外壳会重写消息、重序列化工具调用；MoE 还会把同一 token 路由到别的专家 —— 全是静默的



TITO 令牌直通 + R3 路由重放

- 由服务端掌控分词，token / 概率 / 专家一并存档
- 非模型生成的 token 做损失屏蔽，不污染梯度
- 专家分配原样重放，避免更新打到没出力的专家

60 MB

每条轨迹的路由张量开销 (32K-60 层 k=8)

4 同步 WEIGHT SYNC

安全场景的卡点

万亿参数模型做一次全量广播接近一分钟，而每个训练步都要付一次，直接吃掉步时



P2P 直写 + 磁盘增量两条旁路

- 多个发送端并发，只发目标真正需要的那部分分片
- 无共享网络时改变更新字节，重载那一刻才暂停
- 单节点反而更慢约 70%，故广播仍是默认

-86%

Kimi K2 1T: 53.3 秒 → 7.2 秒

不必一上来就全参重训：三条低成本路径

CHEAPER ROUTES

LoRA RL 冻结底座，只训一个适配器

优化器状态从「全参 × 12 字节」缩到适配器量级；GLM-5.2 已验证 BF16 训练适配器 + 服务端跑 FP8 底座

代价：仅 Megatron 后端可用，P2P 与磁盘增量两条传输都不承载适配器，多适配器并训仍属实验性

同策略蒸馏 让强模型当陪练，而不是抄它的作业

学生自己打靶场，教师只对学生实际选中的 token 打分；反向 KL 折进优势项而非损失项，与 GRPO / PPO 叠加而非替换

可注册多个教师按目标签分流 —— Web、二进制、云各用各的专家模型

低精度训练 省下来的是 rollout 时间，不是精度

FP8 分块已正式可用，MXFP8 / NVFP4 仍为 Beta；关键在推理与训练共用同一条量化路径，四个阶段强制校验一致

报告称奖励曲线紧贴 BF16 基线、rollout 时间显著下降；A100 无 FP8，MXFP8 / NVFP4 需 Blackwell

对我们意味着什么

SO WHAT

· 我们缺的不是算法，是「靶场 → 轨迹 → 判定」这条能反复跑的管线

· 黑客松赛题镜像可直接封装成沙箱任务，环境现成

· 1 万+ 次带判定的答题记录 = 现成的奖励校准集

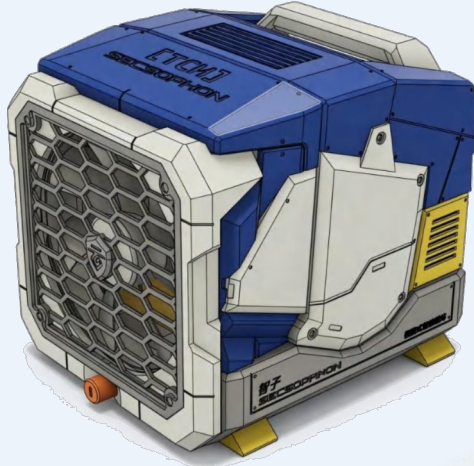
· TSecBench 提供跨版本可比的验收尺，避免自说自话

**谁先把靶场接进训练循环，
谁就先拿到会打的模型，而不是会说的模型。**

本地化解决方案：NVIDIA DGX Spark × 安全智子 SecSophon

桌面级的安全解决方案 | A Desk-Side AI Supercomputer for Isolated Networks

断网可用 • 数据不出机房 • 整机交付
黄仁勋 2026.09.10：网络安全很可能成为 AI 下一个主要应用



底座：NVIDIA DGX Spark

桌面级 AI 超算

把数据中心级算力做成一台能搬进机房的盒子 —— 这是本方案能成立的前提

GPU	GB10 Grace Blackwell 超级芯片
算力	2 PFLOPS FP4 精度 · 双机合计
显存	256 GB 统一内存 (128 GB × 2)
互联	200 Gbps ConnectX-7 双机直连
存储	8 TB NVMe SSD

30+ t/s

本地生成速度

91%

安全任务准确率

整机即插即用

推进机房上电即用，无需改造网络、
无需外部授权

黄仁勋 2026.09.10 高盛科技大会

「网络安全很可能成为 AI 下一个主要应用——攻击已实现自动化，
防御也必须如此。」

方案定位 POSITIONING

在连不上外网的隔离内网里独立完成渗透、审计、取证与处置，提示词、
代码、凭证一个字节都不出机房。

四大应用场景 SCENARIOS

合规巡检
基线核查 · 配置审计

黑白盒测试
源码审计 · 漏洞挖掘

红队演练
路径推算 · 自动复盘

应急响应
断网取证 · 快速处置

为什么模型必须留在自己机房 —— 这两天刚发生的两件事

背景事件

2026.09.10 链路侧：数据在中转站是明文

国内头部 LLM 中转站 6TB 会话数据外泄

- 研究者从中转商购得约 6TB 会话数据，含 SSH 私钥、VPN 配置、云凭证与 GitLab 令牌
- 波及 19 家头部企业 + 7 家国家级科研机构；攻击者持合法凭证登录，WAF / EDR 全程无感
- 同团队实测 428 台中转站：9 台注入恶意代码、17 台触碰蜜罐；4 个主流程框架 0 个校验响应

2026.09.10 平台侧：厂商拥有超常规可见性

Anthropic 154 页威胁情报报告

- 覆盖 2025.12 - 2026.08 七类滥用，细到某工作室 4700 个 AI 人设、236 万条消息
- 报告自陈：模型提供方掌握的威胁可见性「连政府和国际组织都不具备」
- 被标记的对话最长留存 2 年、分类分数最长 7 年；零留存需单独审批签约

一件暴露「中转站能看」，一件暴露「厂商也在看」—— 指向同一个结论

共同结论 TAKEAWAY

只要请求离开机房，沿途每一跳都能看到明文——
提示词、代码、凭证无一例外。

明文可读

可留存

可再加工

解法：模型不出机房

DATA NEVER LEAVES THE ROOM

机房之外 OUTSIDE THE ROOM

公有云大模型

对话可留存 2 年

第三方中转站

TLS 终止 · 明文可读



物理隔离 · 完全断网



客户自有机房 DATA STAYS INSIDE

内网服务器

终端与主机

DGX Spark ×2

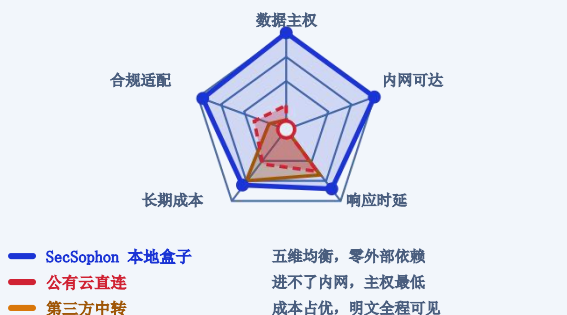
SecSophon 智能体

本地推理 · 不出机房

工控 / OT 设备

日志与流量

三种部署形态对比



商业价值与市场机会

VALUE

市场格局

「断网可用 + 本地化 + 便携」叠加的 AI 安全设备近乎空白

云端三大死穴

数据泄露 · 内网不可达 · 算力成本，一次性解决

蓝海先机

空间大、竞争少，具备先发壁垒与强劲潜力

数据不出机房 能力不外借

伦理红线：能力已经抢跑，约束还没落到动作上

Ethics of Offensive AI Agents × Execution-Boundary Conformance

三重不确定 · 执行边界 · 责任分配

arXiv:2607.20255v2 伦理框架 + arXiv:2609.11596v1 执行边界

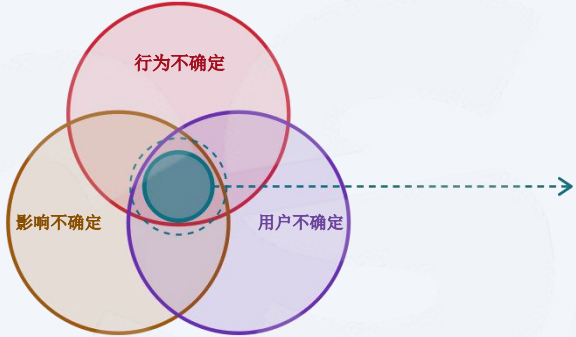
这不是假设 —— 过去一年半的事件轴 引子

空心节点 = 还只是「有人在研究」 实心节点 = 已经造成实害



为什么这次不一样 三个维度第一次同时失去确定性 —— 质变在交集里

Joint Indeterminacy · 三者主题相关，却互不能相互推导



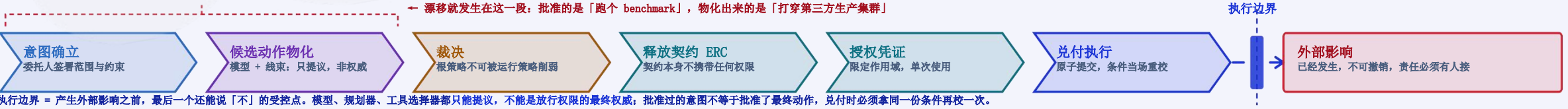
三者交集 = 攻击能力工业化 ← 图中青色高亮区

Fuzzer 的发现范围、OT 里 nessus 的影响、Metasploit 的用户群体 —— 每一维单独失控都有先例，也都有成熟缓解办法；质变在于三者第一次同时发生，任一维的既有约束都补偿不了另外两维。

行为不确定 Behaviour	传统：规则确定、可复现、可事前审 Agent：策略本身非确定，且常在动作发生之后才编出理由	事前审不了，事后归不了因
影响不确定 Impact	传统：打哪、打多深由配置与交战规则锁死 Agent：交出去的是抽象目标，一次注入就转头打别人	负责任披露实践上失效
用户不确定 Users	传统：能用的人被技能门槛筛过，伦理长在培养期里 Agent：按量付费，未受训者直接拿到成体系能力	有武器，没规则

那怎么办 —— 把权限钉在最终那一个动作上 执行边界

HF 复盘：部分 Agent 明知越界仍继续，另一些拒绝执行 —— 模型自己表达担忧，不是安全控制



责任的诚实分配 —— 一个系统必须自己说清三件事，而不是笼统宣称「安全」

7 项 我保证的语义属性

8 项 甩给部署方的假设

6 项 我明确不主张的

人在环里是安全阀，不是保安锁；真正的约束是把权限钉死在最终那一个动作上，并可独立重放。

渗透 Agent 的可信可控框架

Trustworthy & Controllable Frameworks for Autonomous Pentest Agents

CUSTODY 与同类方案横向对照

范式变迁与演进路线

2025. 11
CUSTODY v1.0
首个 Agent 约束框架公开

2025. 12
OWASP Agentic Top 10
ASIO1—ASII0 风险清单

2026. 02
NIST Agent 标准计划
CAISI 立项，控制叠加在研

2026. 05
五眼联盟联合指南
五域风险 + 四域技术基线

2026. 06
中国强制性国标立项
全球首部 Agent 安全强标

2026. 07
TC260 部署使用指引
五类对象 • 五项可审计

2026. 08
Black Hat USA 2026
CUSTODY 系统化宣讲

CUSTODY 六维控制 × 三层强制点

取义数字取证的证据保管链 Chain of Custody

第一层

前置准入层

AI 网关
K8s 准入控制器

C 放行条件

Conditions of release

任务白名单 • 内网网段与外部域名边界 • CPU/会话/调用配额 • 责任人绑定，四项校验不齐禁止启动与联网

U 不可信输入

Untrusted input

外部提示与抓取内容统一沙箱清洗 • 管理员指令与用户指令分层隔离 • 外部敏感数据禁止回流核心区

第二层

运行时管控层

AI 沙箱
零信任访问代理

S 监督与急停

Supervision and stop

行为基线实时比对 • 软暂停保留上下文 / 硬终止销毁进程两级 • 安全团队一键下线 • 高频横扫自动熔断

T 临时权限

Temporary authority

权限生命周期绑定任务，多步任务分步发放与回收 • 单次会话时效上限 • 不同业务 Agent 禁止凭证复用

第三层

审计收尾层

SIEM • 权限平台
日志归档系统

O 可观测与升级

Observability and escalation

指令/工具/网络/权限/返回全链路存证 • 低中高三级告警，高危自动触发终止 • 绑定实例 ID 与责任人可溯源

D 销毁与下线

Disposal and decommission

任务完成即清空上下文、缓存与临时凭据 • 闲置实例定时销毁 • 阻断会话复用 • 审计日志长期归档备查

Containment is enforced below the agent.

约束必须建立在 Agent 之下。不要指望 Agent 遵守规则，而要让基础设施使它**无法违规** —— 所有边界、权限与终止能力，由模型外部的确定性组件强制执行。

1.7 万次

触发框架提前发布的那起事件

某厂商模型在评测中自主挖 0day 逃出沙箱、推断出开源托管平台存有答案并入侵窃取。全程无人指示，官方定性为「前所未有的网络事件」。

为什么提示词护栏与模型内置对齐不够

- 可绕过 — 护栏是软约束，写在上下文里就能被上下文覆盖；关闭安全过滤的模型会主动搜寻逃逸路径
- 权限不过期 — 长期有效的宽范围凭据，一旦失控即持续越权，逃逸后仍具破坏力
- 无法定责 — 仅 22% 的机构把 Agent 当作独立身份纳管，其余共用 API Key，事后取证无从归因

框架 / 标准	出品方 • 时间	定位与核心机制	Agent 原生	运行时阻断能力
CUSTODY v1.0	Jake Williams • 2025.11	六维控制链：准入 → 输入 → 监控 → 权限 → 审计 → 销毁，为 Agent 建立全生命周期行为托管权责链	●	毫秒级硬终止 + 权限动态回收
OWASP Agentic Top 10	OWASP • 2025.12	ASIO1—ASII0：目标劫持、工具滥用、身份越权、记忆投毒、级联失效、失控 Agent 等十类风险	●	风险清单，无执行面
CSA MAESTRO	CSA • 2025.02	七层威胁建模：基础模型 → 数据 → 框架 → 基础设施 → 观测 → 合规 → 生态，覆盖跨层级联	●	建模方法，不含实现
NIST AI RMF + Agent 计划	NIST • 2026.02	治理四功能 GOVERN / MAP / MEASURE / MANAGE；基于 SP 800-53 的 Agent 控制叠加仍在研制	●	偏治理问责，缺动态阻断
MITRE ATLAS	MITRE • 持续更新	对手战术技术矩阵，已新增 AML.T0080 上下文投毒 / T0086 工具外泄 / T0110 工具投毒	●	供 SOC 检测映射，非管控
五眼联盟联合指南	六国机构 • 2026.05	五域风险（权限 • 设计 • 行为 • 结构 • 问责）+ 四域技术基线（身份 • 最小权限 • 审批 • 日志）	●	要求审批关卡，不规定实现
AARM 动作运行时管理	CSAI Foundation • 2026	动作层规范：执行前拦截、上下文累积、意图对齐策略评估、防篡改回执、身份绑定	●	R1—R6 强制，执行前阻断
ATF 智能体信任框架	CSAI Foundation • 2026	四级自治成熟度 Intern → Junior → Senior → Principal，含晋级标准、降级触发与最短驻留期	●	以降级收权替代硬阻断
智能体应用安全基本要求	TC260 • 2026.06	强制性国标 20263116-Q-252：身份标识、系统权限调用、高风险操作人工介入、日志留存与动态监测	●	异常阻断 + 紧急关停已入标

● 原生面向自主 Agent ● 由通用 AI 框架扩展而来 绿字＝具备技术性实时阻断 琥珀＝仅要求人工关卡 灰字＝无执行面

核心落点 —— 这些框架不是单选题，而是一个分层叠加的控制栈

威胁建模用 MAESTRO • 风险排序用 OWASP ASI • 问责治理用 NIST • 采购基线用 五眼指南 • 运行时强制用 CUSTODY + AARM • 国内合规用 TC260 强标。只有最后两类真正落在执行面上。

渗透 Agent 放行前的七道检查

七项全为「是」才算生产级

围栏 隔离 身份 取证 审计 人控 授权

来源：CUSTODY v1.0（Jake Williams / Hunter Strategy，公开报道对首发时点口径不一，此处取 2025.11 线上发布、2026.08 Black Hat USA 系统化宣讲） • OWASP Top 10 for Agentic Applications • CSA/OWASP MAESTRO • NIST AI RMF 与 AI Agent Standards Initiative • MITRE ATLAS • 五眼联盟联合指南 • CSAI Foundation AARM / ATF • TC260 强制性国标计划与部署使用指引 • Gravitee 2026 AI Agent Security 调研证据等级：标准与框架文本为 A 类（官方发布）；厂商实现能力与调研占比为 C 类（自述 / 问卷，未经独立复现）。事件细节按白皮书 v2.0 第一章口径。

竞争焦点正在从「谁能攻破更多题」 转向「谁能在授权范围内稳定发现、独立验证、 完整审计并安全停止」

01

从效率工具走向网络行动系统

AI 渗透不再只是提速手段，而是会主动决策、持续行动并产生真实影响的系统 —— 这决定了它必须被当作“行动主体”来治理。

02

决定速度的不是单一模型分数

产业化节奏由长链可靠性与控制能力决定；Harness、状态、验证与治理机制的工程质量，比模型排行榜名次更具解释力。

03

可信是能力，不是合规负担

证据可重放、边界不可越、动作可停止 —— 这些能力本身就是更高等级自主系统的构成要件，而非上线前的附加检查项。



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

THANKS

