

SANDER SCHULHOFF 2 0

LENNY'S PODCAST

DEEP ANALYSIS

ORIGINAL BY

Lenny Rachitsky

@lennysan • x.com/lennysan

ANALYSIS BY

@Penny777 • x.com/penny777

Sander Schulhoff 2.0 - Lenny's Podcast

Sander Schulhoff 2.0 - Lenny's Podcast 深度分析报告

主持人介绍

Lenny Rachitsky

- **身份:** 前 Airbnb 产品负责人，全球顶级产品管理专家。
- **背景:** 在 Airbnb 工作 7 年，见证了平台从早期增长到全球巨头的过程。
- **现状:** 运营全球排名第一的产品管理播客及 Newsletter，拥有超过 60 万订阅者。
- **社交媒体:**
 - Twitter/X: @lennysan
 - Newsletter: Lenny's Newsletter

嘉宾介绍

Sander Schulhoff

- **身份:** AI 安全研究员，对抗性鲁棒性 (Adversarial Robustness) 专家。
- **职业经历:**
 - **Learn Prompting:** 创始人。编写了互联网上第一个也是最大的提示词工程 (Prompt Engineering) 指南。
 - **HackAPrompt:** 创始人。发起了全球首个生成式 AI 红队 (Red Teaming) 竞赛，吸引了 OpenAI、Google、Anthropic 等巨头参与。
- **学术成就:** 其关于提示词注入的研究论文在 20,000 多篇投稿中脱颖而出，荣获 **EMNLP 2023 最佳主题论文奖**。
- **当前身份:** 专注于 AI 安全教育与红队测试研究，运营 hackai.co。
- **核心专长:** 提示词注入 (Prompt Injection)、越狱 (Jailbreaking)、AI 红队测试、对抗性攻击。
- **社交媒体:**
 - Twitter/X: @sanderschulhoff
 - LinkedIn: Sander Schulhoff

这期播客是一次关于 AI 安全现状的“警世钟”。Sander Schulhoff 抛出了一个极具争议且深刻的观点：**当前的 AI 防护栏（Guardrails）根本不起作用**。他指出，随着 AI 从简单的聊天机器人进化为拥有执行权限的“智能体（Agents）”和机器人，安全风险正呈指数级增长。Sander 详细解释了为什么“修补 AI 的大脑”比“修补传统软件漏洞”难得多，并揭露了 AI 安全行业中存在的过度营销现象。这期内容不仅适合技术人员，更是每一位正在将 AI 集成到业务中的产品经理和高管的必听课。

核心话题

AI安全 提示词注入 智能体风险 对抗性鲁棒性 AI红队测试 网络安全

核心论点

论点一：防护栏（Guardrails）的失效与“无限攻击空间”

核心观点: AI 防护栏无法提供真正的安全，因为攻击者的搜索空间几乎是无限的。

- 无限可能性:** 对于 GPT-5 级别的模型，潜在的提示词组合是 1 后面跟着 100 万个 0。即使防护栏能拦截 99% 的攻击，剩下的 1% 依然是天文数字。
- 自适应攻击:** 人类攻击者是自适应的。Sander 的研究显示，人类红队成员几乎可以在 10-30 次尝试内突破任何现有的防御系统。
- 行业谎言:** 许多 B2B AI 安全公司声称其产品能“完全拦截攻击”，这在科学上是不成立的，更多是销售话术。

"You can patch a bug, but you can't patch a brain. If you find a bug in software, you can be 99.99% sure it's solved. In AI, you can be 99.99% sure the problem is still there."

— Sander Schulhoff

论点二：从“聊天风险”到“智能体灾难”

核心观点: AI 安全问题的严重性取决于 AI 拥有的权限，而非模型本身。

- 低风险场景:** 如果只是一个只读的 FAQ 机器人，被注入攻击（如让机器人夸奖希特勒）虽然有公关风险，但没有实质性破坏。
- 高风险场景:** 当 AI 变成智能体（Agent），拥有读取邮件、操作数据库、调用 API 的权限时，提示词注入将导致数据泄露、资金盗取甚至物理伤害（如工业机器人）。
- 间接注入:** 攻击者不需要直接与 AI 对话，只需在网页或邮件中埋入恶意指令，当 AI 智能体去读取这些内容时，就会被“远程操控”。

论点三：传统安全与 AI 安全的交汇

核心观点: 解决 AI 安全问题的短期方案不在于 AI 本身，而在于传统的权限管理。

- 权限隔离:** 不要给 AI 超过其任务所需的权限。

- **CAMEL 框架:** 借鉴 Google 的思路，根据用户请求动态限制智能体的权限（例如：如果用户只要求总结邮件，则临时剥夺 AI 的“发送邮件”权限）。
- **沙盒化:** 运行 AI 生成的代码时，必须在完全隔离的容器中进行。

✅ 数据验证结果

验证项 1: Sander Schulhoff 的论文是否获得 EMNLP 2023 最佳论文奖？

- 原文声称: "That paper went on to win the best theme paper at EMNLP 2023 out of about 20,000 submissions."
- 验证结果: ✅ 确认
- 来源: EMNLP 2023 Awards List。论文题目为《Ignore This Title and HackAPrompt: Exposing Vulnerabilities in Large Language Models》。
- 可信度: ★★☆☆

验证项 2: ServiceNow Assist AI 是否存在二阶提示词注入漏洞？

- 原文声称: Lenny 提到有人发现了 ServiceNow 智能体可以被诱导执行 CRUD 操作。
- 验证结果: ✅ 确认
- 来源: 安全研究员 Johann Rehberger 的博客 Embrace The Red 详细记录了此漏洞。
- 可信度: ★★☆☆

验证项 3: Comet 浏览器是否存在 AI 泄露数据的风险？

- 原文声称: Sander 提到 AI 浏览器在访问恶意网页时可能泄露用户数据。
- 验证结果: ✅ 确认
- 分析: 多个安全研究（如 Sayak Saha 的研究）已证明，具备 AI 侧边栏的浏览器若能访问页面内容并拥有 API 访问权，极易受到间接提示词注入攻击。
- 可信度: ★★☆☆

🎯 四维分类评估

🟢 高度正确（已验证）

观点 1: 提示词注入目前没有完美的算法解决方案。

- 验证依据: OpenAI 和 Anthropic 的系统提示词（System Prompts）至今仍能被复杂的越狱技巧突破。

观点 2: 智能体（Agents）的权限管理是当前最紧迫的安全防线。

- 验证依据: 遵循“最小权限原则（Principle of Least Privilege）”是网络安全界的长期共识。

🟡 当下可执行（有明确步骤）

建议 1: 实施“读写分离”权限控制。

- 执行方法: 如果 AI 只需要读取数据进行总结，不要给它数据库的写入权限或 API 的 POST 权限。

建议 2: 使用 CAMEL 框架逻辑。

- 执行方法: 在系统架构中引入中间层，根据当前 Task 的上下文动态分配 Token 权限。

🟡 理智质疑（需验证）

存疑点: “AI 防护栏公司完全没用” 这一说法是否过于绝对？

- 质疑原因: 虽然无法拦截 100% 的攻击，但防护栏可以过滤掉 90% 以上的低级、常见攻击，降低系统被大规模自动化扫描的风险。

🔴 需警惕（风险点）

风险点: 过度依赖 AI 自动红队测试。

- 风险说明: Sander 指出，自动红队工具往往只能发现已知模式。真正的威胁来自于人类攻击者的自适应策略。如果 CISO 看到自动测试报告显示“安全”就掉以轻心，将面临巨大隐患。

🔑 关键洞察

1. **AI 的本质是概率而非逻辑:** 传统软件是确定性的，而 AI 是概率性的。这意味着你无法通过逻辑补丁彻底消除 AI 的错误行为。
2. **安全重心转移:** AI 安全的战场正在从“模型层”转向“应用架构层”。与其试图训练一个完美的模型，不如构建一个安全的运行环境。
3. **间接注入是“隐形炸弹”:** 随着 AI 能够自主浏览网页和读取文档，互联网上的任何内容都可能成为攻击载体，这打破了传统的“可信输入”边界。
4. **AI 安全人才荒:** 市场急需既懂大模型原理（Transformer 架构）又懂传统网络安全（渗透测试、权限管理）的复合型人才。
5. **能力与安全的博弈:** 越强大的模型（如 GPT-5）通常越难被简单越狱，但一旦被突破，其造成的破坏力（由于其高权限和高智能）也越大。

🔧 提到的工具/资源

工具 1: Learn Prompting

- 说明: 全球最广泛使用的免费提示词工程学习资源。
- 链接: learnprompting.org

工具 2: HackAPrompt Dataset

- 说明: Sander 团队开源的全球最大提示词注入攻击数据集，用于模型基准测试。
- 链接: [Hugging Face - HackAPrompt](https://huggingface.co/HackAIPrompt)

框架: CAMEL (Communicative Agents for "Mind" Exploration)

- 说明: Google 提出的多智能体通信框架，Sander 强调了其在权限限制上的应用潜力。
- 链接: CAMEL-AI.org

课程: Hack AI (Maven Course)

- 说明: Sander 亲自授课的 AI 安全实战课程，面向非技术背景和安全从业者。
- 链接: hackai.co

🚀 立即可做（今天）

- ☐ **权限审计:** 检查公司目前部署的所有 AI 插件或聊天机器人，确认它们是否拥有不必要的 API 访问权或数据库写入权。
- ☐ **输入隔离:** 确保 AI 处理的用户输入不会直接拼接在系统提示词（System Prompt）中。

📅 17 本周尝试

- ☐ **红队演练:** 尝试用“越狱”思维攻击自家的 AI 产品，看看是否能让它绕过核心业务逻辑。
- ☐ **阅读论文:** 阅读 Sander 的获奖论文《Ignore This Title and HackAPrompt》，了解最新的攻击向量。

🔍 深入探索

- ☐ **架构重构:** 研究如何引入“中间人”机制（如 CAMEL 逻辑），在 AI 执行敏感操作前增加一层非 AI 的逻辑校验。

★ 评分

知识价值: 9.5/10

- 揭示了 AI 行业最不愿面对的真相，具有极高的前瞻性。

可执行性: 8/10

- 虽然模型层难以修复，但架构层的建议非常具体且符合工程实践。

商业潜力: 9/10

- 随着智能体普及，AI 安全将成为一个巨大的蓝海市场。

投入产出比: 10/10

- 听一小时播客可能帮你避免未来数百万美元的数据泄露损失。

综合评分: 9.2/10

📖 参考来源

- Lenny's Podcast 官方网站
 - Sander Schulhoff Twitter
 - EMNLP 2023 官方获奖名单
-

来源: Lenny's Podcast

嘉宾: Sander Schulhoff

