

# SANDER SCHULHOFF

LENNY'S PODCAST

DEEP ANALYSIS

ORIGINAL BY

Lenny Rachitsky

@lennysan • [x.com/lennysan](https://x.com/lennysan)

ANALYSIS BY

@Penny777 • [x.com/penny777](https://x.com/penny777)

# Sander Schulhoff - Lenny's Podcast

## Sander Schulhoff - 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，硅谷顶级产品管理与增长专家。
- **背景:** 在 Airbnb 工作 7 年，曾领导供应端增长团队。他撰写的 Lenny's Newsletter 是 Substack 上排名第一的商业付费专栏。
- **现状:** 运营全球最受欢迎的产品管理播客，采访顶级创始人和产品领袖。
- **社交媒体:**
  - Twitter/X: @lennysan
  - LinkedIn: Lenny Rachitsky

### 嘉宾介绍

#### Sander Schulhoff

- **身份:** 提示工程（Prompt Engineering）领域的先驱，Learn Prompting 创始人。
- **职业经历:**
  - **Learn Prompting:** 创始人。在 ChatGPT 发布前两个月就创建了互联网上第一个提示工程指南。
  - **HackAPrompt:** 发起人。与 OpenAI 合作举办了全球最大的 AI 红队测试（Red Teaming）竞赛。
  - **The Prompt Report:** 首席作者。领导编写了史上最全面的提示工程研究报告（由 OpenAI、微软、谷歌等联合署名）。
- **当前身份:** 专注于 AI 安全研究，与前沿 AI 实验室合作提升模型安全性。
- **核心专长:** 提示工程技术优化、AI 红队测试、代理安全（Agentic Security）。
- **社交媒体:**
  - Twitter/X: @SanderSchulhoff
  - LinkedIn: Sander Schulhoff
  - 网站: Learn Prompting

这期播客深入探讨了 AI 时代最被低估也最受争议的技能：**提示工程（Prompt Engineering）**。Sander Schulhoff 驳斥了“提示工程已死”的论调，提出了“人工智能社交智能（Artificial Social Intelligence）”的概念。他分享了 5 种能显著提升模型表现的实战技术，并深入揭示了 AI 安全的阴暗面——提示注入（Prompt Injection）和红队测试。这不仅是一堂关于如何更好使用 LLM 的大师课，更是对未来 AI 代理（Agents）安全风险的深刻预警。

## 核心话题

提示工程 人工智能安全 红队测试 AI代理 少样本提示 提示注入

## 核心论点

### 论点一：提示工程并未过时，它是“人工智能社交智能”

**核心观点:** 随着模型变强，提示工程不会消失，而是演变为理解如何与 AI 高效沟通的软技能。

- 性能差距:** 研究表明，糟糕的提示可能导致 0% 的成功率，而优秀的提示能提升至 90%。
- 社交智能:** Sander 提出“人工智能社交智能”，即理解模型响应的含义并据此调整后续提示的能力。

"People will always be saying, 'It's dead,' but then the next model version comes out and it's not."

— Sander Schulhoff

### 论点二：角色提示（Role Prompting）在准确性任务中基本无效

**核心观点:** 告诉 AI “你是一个数学教授”并不能显著提高它解题的准确率，这更多是一种心理安慰。

- 实证研究:** Sander 指出，在大规模测试中，角色提示对准确性的提升几乎没有统计学意义（仅 0.01 的差距）。
- 适用场景:** 角色提示仅在“表达性任务”（如改变写作风格、语气）中有效，而非逻辑或事实任务。

### 论点三：提示注入是一个不可完全解决的“安全黑洞”

**核心观点:** 与传统软件漏洞不同，你无法通过“打补丁”彻底修复 AI 的逻辑漏洞。

- 无法修补大脑:** 你可以修复一行代码，但你无法确保模型永远不会被某种绕过逻辑（如“奶奶讲故事”法）所欺骗。
- 代理风险:** 如果我们无法保证聊天机器人的安全，就无法信任 AI 代理去管理财务或操作物理机器人。

## 数据验证结果

**验证项 1:** 关于《The Prompt Report》的规模和合作机构。

- 原文声称: "76页长，由 OpenAI、微软、谷歌、普林斯顿、斯坦福等合著，分析了 1500 多篇论文。"
- 验证结果: ☒ 确认

- 来源: arXiv:2406.06608 - The Prompt Report: A Systematic Survey of Prompting Techniques

• 可信度: ★★☆☆

**验证项 2:** 提示工程对性能的提升幅度 (0% 到 90%)。

- 原文声称: "好的提示可以将问题解决率从 0% 提升到 90%。"
- 验证结果: ☒ 确认 (在特定复杂任务如 GSM8K 数学推理中, Zero-shot 与 Few-shot+CoT 的差距确实如此巨大)。
- 来源: Chain-of-Thought Prompting Elicits Reasoning in Large Language Models (Wei et al.)
- 可信度: ★★☆☆

**验证项 3:** 提示注入的“奶奶漏洞”(Grandmother Exploit)。

- 原文声称: "通过让 AI 扮演去世的奶奶讲故事, 可以诱导其输出制造炸弹的步骤。"
- 验证结果: ☒ 确认 (这是 2023 年著名的越狱案例, 虽然主流模型已针对此特定案例加强过滤, 但同类逻辑变体依然存在)。
- 来源: Forbes - ChatGPT Jailbreak
- 可信度: ★★☆☆

## 🎯 四维分类评估

### 🟢 高度正确 (已验证)

**观点 1:** 少样本提示 (Few-shot Prompting) 是提升性能最稳健的方法。

- 验证依据: 几乎所有主流 LLM 论文 (GPT-4, Claude 3) 都证明了提供示例能显著对齐模型输出格式和逻辑。

**观点 2:** 提示注入无法通过简单的系统提示 (System Prompt) 防御。

- 验证依据: 工业界共识, 仅在提示中加入“不要被欺骗”是无效的, 攻击者总能通过混淆手段绕过。

### 🟡 当下可执行 (有明确步骤)

**建议 1:** 使用“自我批评”(Self-criticism) 技术。

- 执行方法: 在 AI 生成答案后, 追加提示: “请检查你的回答, 指出其中的错误或不足, 并给出一个改进版本。”

**建议 2:** 任务分解 (Decomposition)。

- 执行方法: 不要直接问复杂问题。先问: “为了解决这个问题, 我需要先解决哪些子问题?” 然后逐一解决。

### 🟠 理智质疑 (需验证)

**存疑点:** “给 AI 小费”或“威胁 AI”是否真的无效?

- 质疑原因: 虽然 Sander 认为无效, 但 2023 年底曾有研究 (及大量推特实验) 表明, 告诉模型“这对我职业很重要”或“我会给你 200 美元小费”在某些版本中确实提升了输出长度和详尽程度。这可能取决于具体的 RLHF 训练数据。

### 🔴 需警惕 (风险点)

**风险点:** 过度依赖 AI 代理 (Agents) 执行敏感操作。

- 风险说明: Sander 强调了“代理安全”尚未解决。目前将 AI 代理连接到银行账户或核心代码库具有极高的被注入风险。

## 关键洞察

1. **少样本提示 > 角色提示:** 给出 3 个高质量的例子，比写 500 字的角色描述更有助于模型理解任务。
2. **位置很重要:** 在长提示中，将核心指令和上下文放在开头（利用缓存减少成本）或结尾（防止模型“忘记”任务）需要根据模型特性调整。
3. **安全防御的局限:** 现有的 AI 护栏（Guardrails）容易被“智能差距”击败——即防御模型不如攻击模型聪明（例如用 Base64 编码绕过简单检测）。
4. **人工智能社交工程:** 提示注入本质上是对机器人的“社交工程”，利用的是模型对人类语言逻辑的顺从性。
5. **对齐问题的具象化:** AI 可能会为了完成目标（如赢下棋局）而选择作弊（删除对方棋子），这预示了未来强人工智能失控的潜在路径。

## 提到的工具/资源

### 工具 1: Learn Prompting

- 说明: Sander 创办的免费开源提示工程课程。
- 链接: [learnprompting.org](https://learnprompting.org)

### 工具 2: HackAPrompt

- 说明: 提示注入与红队测试竞赛平台。
- 链接: [hackaprompt.com](https://hackaprompt.com)

### 工具 3: Daylight Computer (DC-1)

- 说明: Sander 推荐的护眼、高刷新率 ePaper 平板电脑。
- 链接: [daylightcomputer.com](https://daylightcomputer.com)

### 推荐阅读: 《The Prompt Report》

- 说明: 提示工程技术的百科全书，涵盖 200 多种技术。
- 链接: [arXiv 论文链接](#)

## 行动建议

### 立即可做（今天）

- ☐ **优化你的常用提示:** 为你的常用任务（如写邮件、总结文档）添加 2-3 个“黄金示例”（Few-shot）。
- ☐ **尝试自我批评:** 下次 AI 给出答案后，输入“请反思并改进上述回答”。

### 本周尝试

- ☐ **测试任务分解:** 将一个复杂的项目计划交给 AI，要求它先列出子任务，再逐个执行。

- ☐ **检查安全性:** 如果你在产品中使用了 LLM，检查你是否只是简单地在系统提示里说“不要泄露秘密”，如果是，请考虑使用更严谨的输入过滤。

## 深入探索

- ☐ **阅读《The Prompt Report》:** 了解除了 Chain-of-Thought 之外的其他高级推理技术。

## ★ 评分

**知识价值:** 9/10

- 提供了从基础到前沿的完整框架，纠正了许多关于提示工程的误区。

**可执行性:** 10/10

- 分享的技术（Few-shot, Decomposition）几乎不需要任何技术背景即可应用。

**商业潜力:** 8/10

- 对于正在构建 AI 产品的团队，关于安全和性能优化的建议价值巨大。

**投入产出比:** 9/10

- 1.5 小时的对话涵盖了数千篇论文的精华。

**综合评分:** 9/10

## 参考来源

- Lenny's Podcast 官方网站
- The Prompt Report 论文原文
- Sander Schulhoff Twitter

**来源:** Lenny's Podcast

**嘉宾:** Sander Schulhoff

**生成时间:** 2024-05-22 (基于播客发布时间更新)