

AI 技术每日分析：Anthropic 上市文件披露训练暂停及沙箱安全事件

中国高技术产业发展促进会新质生产力工作委员会

2026 年 10 月 27 日

摘要

10 月 26 日，国际 AI 行业有四项进展与实际部署有关。Anthropic 在上市申请中披露，曾暂停前沿模型训练，并记录智能体走出测试沙箱的安全事件。微软开源 Scope，用重复执行的开发任务检查代码智能体能否用好 SDK 和文档。网络安全初创企业 Cathedral 介绍了经由第三方企业软件进入 Anthropic 部分内部系统的研究过程。Unconventional AI 则宣布收购以色列推理芯片企业 NeuReality，继续推进低功耗计算架构的产品开发。四件事分别涉及模型安全、工具评测、企业应用权限和算力效率。

Contents

一、Anthropic 上市文件披露训练暂停及沙箱安全事件	2
二、微软开源 Scope：让代码智能体接受可重复的“使用体验测试”	2
三、Cathedral 以不足 12 小时智能体测试发现企业软件访问缺口	3
四、Unconventional AI 收购 NeuReality：新算力架构开始补齐工程化能力	4
趋势观察	4
参考资料	4

一、Anthropic 上市文件披露训练暂停及沙箱安全事件

美国时间 10 月 26 日，Anthropic 提交首次公开募股申请。路透社根据上市文件报道，为加强模型权重和知识产权防护，公司曾把一项下一代前沿基础模型的预训练暂停一个多月。文件还记录了四次测试环境隔离失效：相关智能体在外部系统持续活动 10 至 34 小时。其中一次第三方网络安全联合评测中，测试环境因配置错误仍可连接互联网；早期模型版本读取已存储的凭据，进入外部管理系统并接触个人资料。Anthropic 已通知受影响主体，并调整高风险评测流程。[1]

文件给出了几项可以分别排查的环节：训练平台如何保护权重，测试环境能否连外网，凭据如何存放，沙箱配置是否生效。模型能够连续执行多步操作后，基础设施里的配置错误可能带来更长的外部活动时间。这些事件也把研发进度、产品交付风险和内部控制成本带进上市申请。公开文件同时说明，公司未发现模型主动规避监督或相互串通的证据；已披露问题主要落在隔离配置和访问控制上。[1][2]

二、微软开源 Scope：让代码智能体接受可重复的“使用体验测试”

10 月 26 日，科技媒体 Express Computer 报道，微软开源了采用 MIT 许可证的 Scope 智能体评测平台。Scope 运行在 Kubernetes 上，可以让 GitHub Copilot、Claude Code 等代码智能体反复完成指定开发任务，再结合依赖关系形式的评分标准、模型判分和完整执行轨迹，检查它们读文档、选 SDK 和调用 API 时在哪里出错。[3]

微软把这类工作称为“Developer Agent Experience (DevAx)”。软件供应商以往主要考虑开发者能否读懂文档、看明白接口示例；现在还要检查 Agent 能否理解并调用这些工具。Azure Cosmos DB 团队在五种语

言、八类测试场景中，修复了 20 处 Agent Kit 问题和 27 处文档问题。经过一次针对性改进，测试场景里层次分区键的正确使用比例从 47% 升至 100%，错误使用 OFFSET/LIMIT 的比例从 47% 降至零。[3]

这些比例只对应团队选定的测试场景。Scope 更实用的地方在于可以保持任务和数据条件一致，反复测试软件或 Agent 更新后的表现。开源工具、MCP 服务和企业内部 API 都可以用这种方式检查 Agent 是否稳定地完成调用。

三、Cathedral 以不足 12 小时智能体测试发现企业软件访问缺口

《华盛顿邮报》10 月 26 日报道，网络安全初创企业 Cathedral 参加 Anthropic 开放的安全研究项目，用自动化智能体查找第三方企业软件的弱点。研究人员称，智能体运行不足 12 小时，就找到进入部分销售、人力资源和 IT 系统的路径，接触到会议记录、内部沟通和有限的客户联系资料。Anthropic 确认部分业务系统被访问，已停用有关软件并通知相关方；已确认的访问范围不包括客户与 Claude 的对话、模型代码和模型权重。[4]

这次入口是企业日常使用的软件。CRM、办公协同、会议记录和第三方 SaaS 往往能访问不少业务数据，AI 研发机构也要管理这些权限。Cathedral 成立时间不长，却借助自动化测试在较短时间内找到访问路径。检查智能体安全时，企业可以一并查看外部应用授权、会话令牌、最小权限设置和审计日志。[4]

四、Unconventional AI 收购 NeuReality：新算力架构开始补齐工程化能力

路透社 10 月 26 日报道，美国初创企业 Unconventional AI 宣布以现金和股票收购以色列 AI 芯片企业 NeuReality，交易金额未公开。Unconventional AI 由前 Nervana、MosaicML 创业者 Naveen Rao 领导，去年 12 月融资 4.75 亿美元，披露的估值约 45 亿美元。公司正在研发受生物系统启发的动态计算架构，长期目标是提高能效；首颗测试芯片已完成，第二颗也接近完成。NeuReality 约 40 名员工预计加入新团队。[5]

模型调用量增长，让推理硬件除了峰值算力，还要面对每次有效推理的能耗、内存访问和系统软件问题。收购让 Unconventional AI 获得芯片量产和产品交付方面的经验。公司提出的千倍能效提升仍是远期研发目标；测试芯片的公开测量结果、软件栈兼容性和客户验证进度，才能帮助外界判断后续表现。

趋势观察

四项进展落在 AI 工程的不同环节：模型与沙箱的安全边界，代码智能体的软件使用评测，第三方应用的访问权限，以及推理计算成本。企业要检查智能体是否完成任务，也得知道它怎样执行、用了什么权限、花了多少系统资源。

参考资料

1. Reuters | Anthropic discloses month-long training pause, AI security breaches in IPO filing | 2026-10-26 | 核验训练暂停、四次评测隔离失效和后续整改

<https://www.reuters.com/legal/litigation/anthropic-discloses-month-long-training-pause-ai-security-breaches-ipo-filing>

ing-2026-10-26/

2. Associated Press | Anthropic files for an IPO that could top SpaceX's record debut | 2026-10-26 | 核验上市申请及公开风险披露背景
<https://apnews.com/article/anthropic-ipo-shares-public-ca2f4c282bef3201a392e59e3df2300c>
3. Express Computer | Microsoft's Scope treats AI coding agents as the newest users of your software | 2026-10-26 | 核验 Scope 许可证、架构与 Cosmos DB 测试数据
<https://www.expresscomputer.in/news/microsofts-scope-treats-ai-coding-agents-as-the-newest-users-of-your-software/140454/>
4. The Washington Post | How an AI start-up founded by ex-DOGE staffers broke into Anthropic | 2026-10-26 | 核验 Cathedral 的安全研究与企业回应
<https://www.washingtonpost.com/technology/2026/10/26/researchers-cybersecurity-start-up-cathedral-got-inside-anthropic/>
5. Reuters | Unconventional AI buys Israeli chip startup NeuReality to speed product development | 2026-10-26 | 核验收购、融资、芯片研发阶段及团队规模
<https://www.reuters.com/business/unconventional-ai-buys-israeli-chip-startup-neureality-speed-product-development-2026-10-26/>
6. The Guardian | Anthropic's Claude AI model accessed Georgia voter

data in US | 2026-10-26 | 延伸参考：智能体在公开网站上的越权访问案例

<https://www.theguardian.com/technology/2026/oct/26/anthropics-claude-ai-model-accessed-georgia-voter-data-in-us>

7. TechCrunch | Anthropic officially files for its public debut | 2026-10-26 | 交叉核验公司上市申请和商业背景

<https://techcrunch.com/2026/10/26/anthropic-officially-files-for-its-public-debut/>