

AI 技术每日分析：99.1% 零样本叠衣之后，具身智能开始比拼可靠性

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 7 月 19 日

摘要

今天的 AI 技术动态集中在可靠性、安全、推理芯片和终端智能体四个方向。Sunday Robotics 发布 ACT-2 预览，宣称其机器人基础模型在陌生家庭和未见衣物上实现 99.1% 的零样本叠衣成功率，并提出用 “Solve” 标准披露测试条件和人工介入程度。Capital One 开源 VulnHunter，把攻击路径推演、反证机制和代码修复建议组合成面向开发者的智能体安全工具。AI 推理芯片创业公司 Etched 据报正洽谈按约 200 亿美元估值融资，但首款产品仍处在客户测试阶段。TechCrunch 对 Vertu 手机内置 Hermes Agent 的实测则显示，智能体越愿意自主执行，越需要可靠的时间理解、上下文保持和确认机制。Ai2 近期公开的 Shippy 架构进一步指出，高风险智能体需要确定性工具、隔离沙箱和系统级评测，而不能只看底层模型分数。

Contents

一、ACT-2 把机器人竞争焦点转向“泛化后的可靠性”

Sunday Robotics 发布 ACT-2 预览。公司称，该模型能够把少量内部微调带来的性能提升迁移到未见环境；在不同陌生家庭和不同衣物组成的测试中，叠衣任务取得 99.1% 的零样本成功率。其核心主张是，一条新的微调示例即可教会预训练模型一种能够泛化的行为，从而减少为每个家庭、物体和布局重新采集大量机器人数据的需求。

公司同时提出“一个 Solve”的测试表达方式，要求说明机器人完成了什么任务、在哪些环境测试、需要多少额外训练、是否依赖人工远程协助。这个提议针对机器人演示长期存在的可比性问题：单次成功视频无法说明失败率，也无法判断环境是否经过特别布置。

ACT-2 的 99.1% 目前主要来自公司披露，仍需更多独立测试，但其评价方法值得关注。具身智能要进入家庭和服务场景，决定产品价值的不是偶尔做对，而是面对长尾变化时持续做对。

二、Capital One 开源 VulnHunter：让 AI 先证明漏洞“确实可利用”

Capital One 发布并开源 VulnHunter。它不是只寻找危险函数或代码模式，而是从 API、网络消息和文件上传等攻击者可触达入口出发，沿应用逻辑向前推演，判断攻击是否能够穿过数据转换和安全检查，最终形成可执行的攻击路径。

工具加入了反证引擎。发现问题后，系统会主动寻找不成立的前提、攻击链中的逻辑缺口以及阻止漏洞利用的条件，尽量在告警到达开发者之前排除误报。对于未被推翻的结果，VulnHunter 会汇集跨文件证据、解释攻击者可能获得的能力，并生成供工程师审查的定向修复建议。项目采用 Apache 2.0 许可证，初始实现围绕 Claude Opus 4.8 和 Claude Code

构建。

这类工具代表 AI 代码安全从“生成更多告警”转向“验证真实可利用性”。企业安全团队常被大量低质量告警拖慢，智能体的价值取决于能否给出证据链、降低误报，并把发现、复核和修复连接到同一工作流。

三、Etched 估值洽谈升至 200 亿美元：资本押注推理专用芯片

《华尔街日报》7 月 18 日报道，AI 芯片创业公司 Etched 正在洽谈由 Jane Street 等投资方参与的新融资，目标估值约 200 亿美元；公司同时还在推进另一笔按 100 亿美元估值计算的融资。相关交易尚未完成，条件仍可能变化。

Etched 专注于大模型推理芯片。公司此前披露，台积电已经制造其首款芯片，面向客户提供包含芯片、机架和软件的推理集群，并称已获得约 10 亿美元合同订单。不过，首款产品仍处于客户测试阶段，尚未经过大规模商业部署验证。

估值快速上升说明推理成本已经成为 AI 产业竞争的核心。训练支出高度集中，而推理会随着每一次用户调用持续发生。专用 ASIC 如果能在特定模型架构上提高吞吐量和能效，就可能切入 GPU 之外的市场；但架构适配范围、软件生态、良率和真实负载稳定性仍决定其商业上限。

四、Hermes Agent 实测暴露终端智能体的“自主性悖论”

TechCrunch 对 Vertu Alphafold 手机内置 Hermes Agent 进行了多日测试。该智能体基于开源 Hermes 项目，可分析本地文件、操作多个应用、记忆对话，并在必要时转交人工礼宾服务。测试中，Hermes 更愿意直接执行跨应用任务，但也出现提醒时间错误、行程日期错误、工作流未完成以及数日后无法继续读取此前文档上下文等问题。

与之对比，Gemini 在部分任务中会先追问机场、提醒工具等细节，执行步骤较少，却得到更准确的结果。这说明终端智能体并非越少确认越好。涉及日程、消息、支付、导航和企业数据时，系统需要区分低风险自动执行与高风险确认，保持状态，并允许用户看到和撤销即将发生的操作。

智能手机为智能体提供了通讯录、日历、文件、定位和应用权限，是最接近真实行动入口的终端。产品竞争将从聊天界面转向权限编排、上下文连续性和错误恢复。

五、Ai2 复盘 Shippy：高风险智能体应评测整个系统

Ai2 近期公开海事智能体 Shippy 的技术复盘。Shippy 面向海洋保护和执法分析，错误答案可能导致巡逻船偏离目标、浪费资源甚至影响人员安全。团队因此把可靠性拆成确定性工具、沙箱隔离和系统级评测，而不是只根据基础模型排行榜选择方案。

其架构思路是让模型负责理解意图和规划步骤，把可验证的计算、地理查询和数据处理交给确定性工具；代码及外部调用运行在隔离环境中；评测对象则是完整智能体，包括工具选择、引用证据、失败恢复和最终答案。

这条经验适用于企业智能体。模型升级可能改善推理，却不会自动解决权限、工具错误、状态丢失和数据污染。真正可上线的智能体需要把非确定性推理包在可观察、可限制和可回放的工程框架内。

参考文献

- Sunday Robotics: 《ACT-2 Preview: Generalizing Reliability》, 2026-07-17; 用途: 核验 ACT-2 架构主张、99.1% 零样本成功率和单样本微调泛化。 <https://www.sunday.ai/blog/act-2-preview>

- Business Insider: 《Sunday Robotics says its robot can fold clothes it has never seen in unfamiliar homes》, 2026-07-16; 用途: 补充陌生家庭测试、“Solve”标准和传感手套训练方式。<https://www.businessinsider.com/sunday-robotics-memo-home-robot-fold-laundry-99-success-2026-7>
- Capital One Tech: 《Announcing VulnHunter》, 2026-07-16; 用途: 核验开源工具、反证引擎、前向攻击分析和修复流程。<https://www.capitalone.com/tech/open-source/announcing-vulnhunter/>
- GitHub / Capital One: 《VulnHunter》, 持续更新; 用途: 核验许可证、安装方式、架构文档和已知限制。<https://github.com/capitalone/vulnhunter>
- VentureBeat: 《Capital One releases VulnHunter, an open-source AI tool that finds software flaws before hackers do》, 2026-07-17; 用途: 补充行业评价和企业开源安全背景。<https://venturebeat.com/technology/capital-one-releases-vulnhunter-an-open-source-ai-tool-that-finds-software-flaws-before-hackers-do>
- The Wall Street Journal: 《AI Chip Startup Etched Is in Talks for \$20 Billion Valuation》, 2026-07-18; 用途: 核验最新融资洽谈、估值和交易未完成状态。<https://www.wsj.com/tech/ai/ai-chip-startup-etched-is-in-talks-for-20-billion-valuation-caf1787d>
- TechCrunch: 《Nvidia competitor Etched hits \$5B valuation, \$1B in sales for AI chip》, 2026-06-30; 用途: 补充首款芯片测试、合同订单和推理集群产品背景。<https://techcrunch.com/2026/06/30/nvidia-competitor-etched-hits-5b-valuation-1b-in-sales-for-ai-chip/>
- TechCrunch: 《Vertu wants executives to pay \$6,880 for an AI agent—here’s how it actually performs》, 2026-07-17; 用途: 核验 Hermes Agent 跨应用测试、错误类型和与 Gemini 的对比。<https://techcrunch.com/2026/07/17/vertu-wants-executives-to-pay-6880-for-an-ai-agent-heres-how-it-actually-performs/>

- Ai2 / Hugging Face: 《What building Shippy taught us about building agents》, 2026-07; 用途: 核验确定性工具、沙箱隔离和系统级智能体评测思路。 <https://huggingface.co/blog/allenai/shippy-tech-blog>
- Allen Institute for AI: 机构官网, 持续更新; 用途: 补充 Shippy 及环境智能平台研发背景。 <https://allenai.org/>

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址：<https://gyznsw.cn>