

AI 技术每日分析

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 9 月 29 日

摘要

9 月 28 日 AI 领域的几条重要消息围绕 Agent 安全、个人 Agent 融资、音频原生模型和可验证 AI 同时进入产品化阶段。NVIDIA 在 9 月 28 日正式发布 Open Agent Safety Platform，把 Agent 安全边界从应用层下沉到运行时、CPU、DPU 和机器人执行层：OpenShell 负责记录并约束 Agent 动作，Sentry 在 BlueField-4 DPU 上以独立“看门狗”方式监控，一旦 Agent 越界可在毫秒级隔离；Noma 则把 Agent 治理进一步推进员工电脑，开始自动发现本机运行的 Coding Agent、MCP 服务器与 Skills，并统一施加访问控制和运行时阻断；个人 Agent 公司 Instinct 完成 10 亿美元融资、估值升至 100 亿美元，说明消费级“代办型 Agent”仍获得大量资本押注；语音 AI 公司 Modulate 获得 2500 万美元融资，用 100 多个专用音频模型组合分析语气、情绪、意图和深度伪造；OPAQUE 披露其 AgenTrust 开源可验证 AI 工具上线三个月已接近 50 万次下载。这些进展表明，AI 应用竞争正在从“模型能不能做事”转向“能不能被约束、被审计、被验证，并在特定模态和场景中持续运行”。

Contents

一、NVIDIA 发布 Open Agent Safety Platform: Agent 安全 首次明显下沉到硬件隔离层	3
二、Noma 把 Agent 治理推进员工电脑: Coding Agent 正在成 为新的“终端安全对象”	4
三、Instinct 融资 10 亿美元、估值 100 亿美元: 资本继续押注消 费级“代办型 Agent”	4
四、Modulate 获 2500 万美元融资: 语音 AI 开始从“转文字” 转向理解情绪、意图与真假	5
五、OPAQUE 的 AgenTrust 接近 50 万次下载: 可验证 AI 开 始形成开源工具链	6
趋势观察	6
参考资料	7

一、NVIDIA 发布 Open Agent Safety Platform: Agent 安全首次明显下沉到硬件隔离层

9 月 28 日, NVIDIA 正式发布 Open Agent Safety Platform。平台由开源运行时 OpenShell 和参考系统设计 Sentry 构成, 目标不是再做一层内容过滤, 而是在 Agent 真正执行工具、访问数据和操作物理系统时建立独立安全边界。

OpenShell 负责在 CPU 侧记录 Agent 的行为轨迹并执行策略限制。NVIDIA 称, 它在 Vera CPU 上具有较低运行开销, 同时可以扩展到 Arm、Intel 等第三方计算平台。Sentry 则运行在 BlueField-4 DPU 上, 以 Agent 不可直接控制的独立信任域持续监控; 一旦 Agent 试图突破软件边界, Sentry 可以在毫秒级执行隔离和停止。

对企业来说, 关键变化是: 企业不再把全部希望寄托在“模型自己守规矩”。模型、Agent 框架和应用代码之外还需要一层无法被 Agent 随意改写的执行控制面。对于 Coding Agent、金融 Agent、机器人和基础设施运维 Agent, 这种“安全边界外置”比单纯 Prompt 约束更接近传统高可靠系统。

NVIDIA 同时公布了广泛生态合作, 涉及 Anthropic、Cisco、CrowdStrike、Microsoft、SAP、ServiceNow、Siemens、Schneider Electric 以及机器人企业 Figure、Gecko Robotics、Skild AI 等。OpenShell 已开放开发者资源和 GitHub 代码。未来 Agent 安全很可能像容器安全、零信任和 API 网关一样, 成为独立基础设施层。

二、Noma 把 Agent 治理推进员工电脑：Coding Agent 正在成为新的“终端安全对象”

Noma Security 在 9 月 28 日发布新的 Endpoint Agent Security 能力，面向已经运行在员工笔记本电脑上的 Claude Code、Cursor、Codex、Windsurf 等 Coding Agent，以及各种本地 MCP 服务器和 Skills。

过去企业 AI 治理主要集中在云端模型、SaaS 应用和自建 Agent 平台，但开发者本地环境正在形成新的盲区：Agent 往往继承安装者的文件权限、代码仓库凭据、云账户和浏览器会话。一旦本地 Agent 调用恶意 MCP、执行危险命令或把数据送往外部，传统 SaaS 安全策略可能根本看不到。

Noma 的新能力会发现终端上的 Agent、MCP 与 Skills，统一定义它们允许执行的动作，并通过 AI Detection and Response 在运行时阻断危险行为。公司还计划将“Agent Boundaries”扩展至终端，让业务定义的权限随着 Agent 在 SaaS、自建环境和员工电脑之间移动。

终端管理因此成为新的安全市场：未来 EDR、MDM 和 AI 安全平台的边界会越来越模糊。企业需要管理的不只是“哪个员工登录了电脑”，还要知道“这个员工授权了哪些 Agent，这些 Agent 又调用了什么工具”。

三、Instinct 融资 10 亿美元、估值 100 亿美元：资本继续押注消费级“代办型 Agent”

Reuters 9 月 28 日报道，个人 Agent 初创公司 Instinct 完成 10 亿美元最新融资，估值升至 100 亿美元，是其上月 25 亿美元估值的 4 倍。投资方包括 Sequoia Capital、Benchmark Capital 和 Coatue。

Instinct 的产品定位不是单纯聊天，而是执行日常事务。用户可以通过文字或电话要求其规划旅行、购买生活用品、订票、取消订阅，公司还

在扩展能够代表用户致电商家的“Concierge”服务。

这类产品的商业难点不在语言理解，而在身份、支付、隐私和错误执行。Instinct 同时强调隔离 Sandbox、短生命周期本地凭据，以及新的主动检测机制来识别 Agent 回答中的细微幻觉。也就是说，消费 Agent 越接近交易，安全设计越需要变成产品核心，而不是后台附加项。

10 亿美元融资说明投资人仍然相信 Agent 会成为新的消费者入口。但这一赛道最终能否成立，关键要看任务成功率、售后责任、商户接口和用户“让 AI 替自己行动”的信任，而不是单纯 MAU 或聊天时长。

四、Modulate 获 2500 万美元融资：语音 AI 开始从“转文字”转向理解情绪、意图与真假

语音 AI 初创公司 Modulate 在 9 月 28 日宣布获得 2500 万美元融资，以扩大其 Audio-native 模型开发和开发者覆盖。公司模型直接处理原始音频，而不是先把语音转成文本再交给 LLM，因此可以同时感知语气、情绪、意图以及深度伪造等信号。

Modulate 最初从游戏语音内容治理切入，现在已经扩展到医疗机构的冒充检测，以及对 Voice Agent 表现的实时监控。其 Velma 平台底层使用超过 100 个专用小型音频模型，根据不同任务选择并融合结果；公司称这种方式在部分任务中比把所有音频都交给一个大型模型高效得多。

这条融资说明“多模态”不等于所有模态都塞进一个通用大模型。对于实时语音，延迟、成本和特征粒度非常重要，专门的声学模型仍然有明显价值。

企业部署 Voice Agent 时，未来可能形成两条并行链路：语言模型负责“说什么”，音频原生模型负责判断“怎么说、是不是本人、是否异常”。这会成为金融客服、医疗和游戏等高风险语音场景的重要基础能力。

五、OPAQUE 的 AgenTrust 接近 50 万次下载：可验证 AI 开始形成开源工具链

OPAQUE 在 9 月 28 日披露，其 AgenTrust 开源项目自 6 月公开以来，相关软件包下载量接近 50 万次；仅截至 9 月 23 日的 30 天内就有约 14.4 万次下载，9 月社区合并代码贡献 140 次，相比 6 月的 11 次明显增加。

AgenTrust 解决的是一个与“Agent 安全”不同但互补的问题：组织怎样证明某个 Agent 到底在什么策略、什么硬件、什么运行环境里做了什么。它通过签名记录、运行时证明和测试套件生成可独立验证的证据，而不是让企业完全相信某一家供应商自己的日志。

随着 Agent 开始进入金融、政府和企业敏感流程，“可验证”会比“可观测”再多一步。可观测告诉你发生了什么，可验证要求第三方能够确认这份记录没有被篡改。

如果开源可验证 AI 形成事实标准，未来 Agent 的合规审计可能像软件供应链 SBOM 一样拥有结构化证据。谁运行、权限是什么、调用了哪些工具、在何种硬件上执行，都可以成为自动化审计对象。

趋势观察

今天 AI 技术的主线是**安全边界外置、终端治理、垂直模态和可验证执行**。NVIDIA 把 Agent 安全推进硬件和独立运行时，Noma 把治理扩展到员工终端，Instinct 继续验证个人 Agent 的资本热度，Modulate 证明小型专用模型在语音中仍有价值，AgenTrust 则尝试把 Agent 行为做成第三方可验证的证据。AI 系统越能行动，就越需要把“能做什么、做了什么、谁能证明”做成正式工程层。

参考资料

1. NVIDIA | 《NVIDIA Launches Open Agent Safety Platform to Secure Agents From Testing to Deployment》 | 2026-09-28 | 核验 OpenShell、Sentry、BlueField-4 与生态合作。
<https://nvidianews.nvidia.com/news/open-agent-safety-platform>
2. Associated Press / ABC News | 《Nvidia unveils security platform to stop AI agents from going rogue》 | 2026-09-28 | 交叉核验平台发布背景。
<https://abcnews.com/Technology/wireStory/nvidia-unveils-security-platform-stop-ai-agents-rogue-136817232>
3. Noma Security / PR Newswire | 《Noma Extends Agent Security and Governance to the Employee Endpoint》 | 2026-09-28 | 核验 Endpoint Agent Security 能力。
<https://www.prnewswire.com/news-releases/noma-extends-agent-security-and-governance-to-the-employee-endpoint-302890925.html>
4. Bedrock Data / Business Wire | 《Bedrock Data Integrates With NVIDIA OpenShell》 | 2026-09-28 | 补充 OpenShell 的数据感知策略执行生态。
<https://www.businesswire.com/news/home/20260928728577/en/Bedrock-Data-Integrates-With-NVIDIA-OpenShell-to-Bring-Data-Aware-Policy-Enforcement-to-AI-Agents>

5. Reuters | 《AI agent firm Instinct raises \$1 billion in latest funding round》 | 2026-09-28 | 核验融资、估值和产品方向。
<https://www.reuters.com/technology/ai-agent-firm-instinct-raises-1-billion-latest-funding-round-2026-09-28/>
6. TechCrunch | 《Viral AI agent Instinct raises \$1B Series C at a \$10B valuation》 | 2026-09-28 | 交叉核验融资与消费 Agent 市场热度。
<https://techcrunch.com/2026/09/28/viral-ai-agent-instinct-raises-1b-series-c-at-a-10b-valuation/>
7. SiliconANGLE | 《Voice AI startup Modulate raises \$25M to bring audio-native models to more developers》 | 2026-09-28 | 核验融资、Velma 与音频原生模型架构。
<https://siliconangle.com/2026/09/28/voice-ai-startup-modulate-raises-25m-to-bring-audio-native-models-to-more-developers/>
8. OPAQUE / PR Newswire | 《OPAQUE's Open Source Approach to Verifiable AI Nears Half a Million Downloads》 | 2026-09-28 | 核验 AgenTrust 下载量与社区数据。
<https://www.prnewswire.com/news-releases/opagues-open-source-approach-to-verifiable-ai-nears-half-a-million-downloads-as-community-code-contributions-grow-more-than-tenfold-302890736.html>
9. Help Net Security | 《Authorizer: Open-source authentication and authorization for your apps》 | 2026-09-28 | 背景资料，观察 Agent 权限控制进入开源基础设施。

<https://www.helpnetsecurity.com/2026/09/28/authorizer-open-source-authentication-server/>

10. pgEdge / PR Newswire | 《pgEdge Announces pgEdge Starfleet》 | 2026-09-28 | 背景资料，观察 Agentic AI 应用从原型走向数据库与生产部署。

<https://www.prnewswire.com/news-releases/pgedge-announces-pgedge-starfleet-a-new-postgres-cloud-platform-to-bridge-the-ai-prototype-to-production-chasm-302890970.html>

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>