

AI 技术每日分析：OpenAI 首次披露 Jalapeño 推理芯片数据

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 8 月 26 日星期三

摘要

今天 AI 技术方向的主线，集中在推理芯片、端侧 Agent、面向 Agent 的搜索索引、跨工具记忆与企业安全治理。OpenAI 首次披露 Jalapeño 推理芯片在公开模型负载上的测试数据，显示模型公司正把竞争延伸到自研推理基础设施；Perplexity 推出 Portable Computer，尝试把持续运行的个人 Agent 带到本地优先设备；Keenable 完成 2600 万美元融资，建设面向 Agent 可执行任务的实时网页索引；Anthropic 开始把 Claude 与 Cowork 的记忆连接起来；AI 安全公司 Alice 完成 1.4 亿美元融资。AI 平台的竞争正在由单一模型能力，转向芯片、设备、索引、记忆与安全治理的系统能力。

目录

一、OpenAI 披露 Jalapeño 推理芯片数据：模型公司开始纵向整合推理基础设施	2
二、Perplexity 推出 Portable Computer：本地优先 Agent 尝试长期执行个人任务	3

三、Keenable 融资 2600 万美元：网页索引开始为 Agent 执行而设计	3
四、Claude 与 Cowork 连接记忆：跨工具连续性成为 Agent 基础能力	4
五、AI 安全公司 Alice 融资 1.4 亿美元：企业为 Agent 攻击面管理投入更多预算	4
趋势观察	5
参考资料	5

今天 AI 技术方向的主线，集中在推理芯片、端侧 Agent、面向 Agent 的搜索索引、跨工具记忆与企业安全治理。OpenAI 首次披露 Jalapeño 推理芯片在公开模型负载上的测试数据，显示模型公司正把竞争延伸到自研推理基础设施；Perplexity 推出 Portable Computer，尝试把持续运行的个人 Agent 带到本地优先设备；Keenable 完成 2600 万美元融资，建设面向 Agent 可执行任务的实时网页索引；Anthropic 开始把 Claude 与 Cowork 的记忆连接起来；AI 安全公司 Alice 完成 1.4 亿美元融资。AI 平台的竞争正在由单一模型能力，转向芯片、设备、索引、记忆与安全治理的系统能力。

一、OpenAI 披露 Jalapeño 推理芯片数据：模型公司开始纵向整合推理基础设施

OpenAI 在 8 月 25 日首次公开 Jalapeño 推理芯片的部分测试结果。官方披露的 1.5—1.9 倍峰值能效提升和 1.7—3.6 倍端到端延迟改善，来自其公开的特定模型、批量和系统配置对比，不应理解为对所有模型、所有 GPU 和所有生产负载都能稳定复现的通用结论。

这项进展真正重要的地方，是 OpenAI 开始把模型、编译与系统软件、芯片和数据中心推理服务放在同一条优化链上。过去推理成本下降主要依靠蒸馏、量化、KV Cache 和调度优化；当模型公司进一步控制芯片的数据流和内存结构，单位 Token 成本、响应延迟和供给稳定性都有可能继续改善。

Jalapeño 仍处于早期披露阶段，量产规模、适用模型范围和长期可靠性还需要后续数据验证。但它已经释放出明确信号：头部模型公司的竞争边界，正在从模型训练扩展到完整推理基础设施。

二、Perplexity 推出 Portable Computer: 本地优先 Agent 尝试长期执行个人任务

Perplexity 在 8 月 25 日发布 Portable Computer。官方将其描述为本地优先的个人计算设备，核心卖点是让 Agent 持续处理搜索、计划、任务队列和工具调用，并在可用时更多利用设备侧环境与用户数据。

这里的“本地优先”不等于所有推理都完全离线，也不等于设备天然消除了隐私风险。产品仍可能依赖云端模型或联网服务；本地文件、浏览器和系统权限一旦开放给 Agent，也会带来新的越权、误操作和提示注入风险。

它的产品意义仍然明确：个人 AI 正在从一次性问答入口转向可持续运行的任务系统。下一阶段竞争会落在权限可见、任务可暂停、动作可确认、状态可恢复和数据可删除，而不只是回答是否流畅。

三、Keenable 融资 2600 万美元：网页索引开始为 Agent 执行而设计

Keenable 在 8 月 25 日宣布完成 2600 万美元融资。公司定位并不是再做一个面向人的搜索页面，而是构建面向 Agent 的实时网页索引与执行基础设施，使 Agent 能够理解页面结构、表单、按钮、商品和跨页面关系。

需要注意的是，“可操作索引”不等于已经覆盖整个互联网，也不代表 Agent 可以绕开登录、权限、验证码和网站规则。它更接近一种为机器任务设计的数据和语义层，目标是减少 Agent 在网页变化、页面定位和信息抽取上的脆弱性。

如果这类基础设施成熟，搜索价值链会进一步分层：人类使用的品牌入口之外，网页语义、变化监测、结构化抽取和任务执行接口会形成新的技术市场。

四、Claude 与 Cowork 连接记忆：跨工具连续性成为 Agent 基础能力

Anthropic 在 8 月 25 日宣布推进 Claude 与 Cowork 之间的记忆连接，使用户在不同工作界面中的部分上下文能够连续使用。官方强调的是用户可控的上下文延续，而不是默认无限期保存所有工作数据。

对企业用户而言，记忆真正有价值的地方，是让 Agent 知道项目进展、已确认决定、协作对象和待办状态，从而减少每次从头解释。但长期记忆也必须有明确边界：哪些信息能写入、保存多久、谁能访问、何时过期、如何删除，都要进入治理设计。

记忆因此不再只是“更懂用户”的体验功能，而是权限、合规和数据生命周期问题。跨工具连续性越强，越需要清晰的可见性与撤销能力。

五、AI 安全公司 Alice 融资 1.4 亿美元：企业为 Agent 攻击面管理投入更多预算

AI 安全公司 Alice 在 8 月 25 日宣布完成 1.4 亿美元融资。公司业务覆盖 AI 资产发现、模型与 Agent 风险测试、提示注入防护、身份权限和运行监测等方向。融资额本身说明，企业安全预算正在从传统终端和云安全向 Agent 治理延伸。

Agent 连接邮件、代码库、知识库、财务和办公系统后，风险不再只是“输出一句错误答案”。提示注入可能诱导工具调用，过宽权限可能造成越权访问，缺乏审计则会让企业无法还原 Agent 做过什么。

因此，企业采购 AI 会越来越同时考察准确性与可控性：身份是否明确、权限能否最小化、关键动作是否需要确认、行为能否审计、异常能否立即阻断。

趋势观察

今天五条动态共同指向一个变化：AI 平台竞争正从模型层扩展到完整数字基础设施。自研芯片负责降低推理成本，端侧设备承载长期任务，Agent 索引连接开放网页，跨工具记忆维持上下文，安全平台负责权限和审计。未来真正有竞争力的 AI 系统，需要把模型、芯片、数据、工具、记忆和安全稳定地组合在一起。

参考资料

1. OpenAI | Jalapeño 推理芯片技术披露 | 2026-08-25。
2. Perplexity | Portable Computer 产品发布资料 | 2026-08-25。
3. Keenable | 2600 万美元融资及 Agent 网页索引介绍 | 2026-08-25。
4. Anthropic | Claude 与 Cowork 记忆连接更新 | 2026-08-25。
5. Alice | 1.4 亿美元融资公告及企业 AI 安全产品资料 | 2026-08-25。
6. 企业 Agent 权限、审计与提示注入相关公开研究资料 | 2026 年 8 月。

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>