

AI前沿发展日报 | 2026-08-26 (Asia/Shanghai)

日期：2026-08-26；覆盖窗口：2026-08-25 06:45 至 2026-08-26 06:45 (Asia/Shanghai)，以窗口内正式上线、进入量产、完成融资或出现可核验规则与执法变化的事实为主。

今日要点

- Claude 将聊天与 Cowork 的记忆合并，并把逐主题查看、编辑和删除交给用户；长期上下文开始从单一对话能力变成跨工作入口的产品层。
- OpenAI 首次公布自研推理芯片 Jalapeño 的实测结果：在三类开放模型上，峰值能效提高 1.5—1.9 倍、端到端延迟降低 1.7—3.6 倍；模型公司正在把单位电力可交付的推理量变成核心经营指标。
- Stability AI 获得 7,600 万美元 B 轮，EA、索尼音乐、环球音乐和华纳音乐同时入股；创意模型的资本来源正在与版权、分发和专业生产场景更紧密地绑定。
- 澳大利亚 ARIA 将纯 AI 生成录音排除在榜单和奖项之外，台湾则对绕道出口 130 台 B300 服务器的案件提起公诉；AI 产业规则已同时深入内容资格与硬件流向。

今日结论

1. 长期记忆的竞争重点已从“记得更多”转向“跨入口连续工作且可被用户逐项纠错”；企业采购时应把可见性、删除权和敏感信息默认设置列入验收条件。
2. 推理基础设施开始按任务类型分化，通用加速器之外将出现更多低延迟专用路径；云与模型厂商的优势会越来越取决于每千瓦吞吐、单用户时延和软件适配速度的组合。
3. 创意 AI 的商业通行证正在由资本、版权方和行业资格规则共同决定；合法训练数据、人类主创证明与可审计制作记录会直接影响融资、分发和榜单准入。

AI 产品与应用

Claude 打通聊天与 Cowork 记忆，用户可逐主题修订

Anthropic 于 8 月 25 日将 Claude 聊天与 Claude Cowork 合并为同一套记忆：聊天中积累的项目、偏好和人员信息可直接用于 Cowork 云端任务，Cowork 产生的新上下文也会返回聊天。记忆会在对话进行时更新，并以主题文件形式向用户展示，可逐项阅读、编辑或删除。Free、Pro 和 Max 默认开启；Team 与 Enterprise 由管理员决定是否提供、个人仍需自行开启。健康、

宗教信仰等敏感主题默认不保存，用户可另行选择开启；身份证件号码、犯罪史和移民身份等内容即使开启也不写入。Anthropic 官方说明 (<https://claude.com/blog/claudes-memory-works-everywhere-and-you-decide-whats-in-it>)

这次变化把个人上下文从聊天便利功能推进到执行型 agent 的工作状态。对企业而言，真正的收益是减少反复交代项目背景；主要风险则是错误记忆会跨任务传播。上线前应验证管理员开关、用户删除后的实际生效范围、敏感主题策略，以及记忆能否与项目级权限清晰分离。

模型与技术进展

OpenAI 用 Jalapeño 把推理能效与交互时延同时拉进竞争区

OpenAI 公布首款自研推理芯片 Jalapeño 的首批结果。在 InferenceX 基准上，针对 GPT-OSS 120B、DeepSeek R1 670B 和 Kimi K2.5 1T，官方称其峰值每瓦工作量是对比系统的 1.5—1.9 倍，端到端延迟低 1.7—3.6 倍，高交互负载性能高 2.1—4.1 倍。芯片额定功耗 700W，测试中的持续功耗不高于 550W；OpenAI 计划在 2026 年底前开始内部部署，第二代已深入开发。OpenAI 工程说明 (<https://openai.com/index/jalapeno-first-results/>)

这里的新增变量不是单一跑分，而是模型公司开始直接控制芯片、内存、网络 and 编译软件的协同优化。数据均来自 OpenAI 自测，比较系统、模型量化方式与生产良率仍需独立核验；但若规模部署后保持优势，推理毛利、agent 响应速度和电力约束会同时改善，也会迫使云厂商提供更细的工作负载路由。

投融资与商业动态

Stability AI 获 7,600 万美元，娱乐产业从采购方变为股东

Stability AI 宣布完成 7,600 万美元 B 轮融资；在 Prem Akkaraju 2024 年就任 CEO 后，公司累计取得的股权与可转债融资达到 2.32 亿美元。本轮新旧投资者包括 Electronic Arts、Sony Music Group、Universal Music Group、Warner Music Group、AMD Ventures、Coatue 和 Greycroft。资金将用于创意生产产品、应用研究和专业服务；公司近期发布的 Stable Audio 3.0 强调使用完全授权的数据训练。Stability AI 官方公告 (<https://stability.ai/news-updates/stability-ai-latest-funding-backed-by-entertainment-industry-biggest-names>)

唱片公司和游戏厂商同时入股，说明创意模型的竞争正在从“生成质量”扩展到数据授权、专业工作流和渠道进入。战略投资者能提供版权与客户入口，但官方没有披露估值、收入或本轮各方出资比例；后续应重点看授权成本是否形成毛利压力，以及专业服务能否转化为可重复的软件收入。

政策、伦理与安全

ARIA 为 AI 音乐设置“人类主创”门槛

澳大利亚唱片业协会 ARIA 宣布，从下周一开始，纯 AI 生成录音不得进入 ARIA 官方榜单，也无资格参加 ARIA Awards；AI 仅作为辅助工具的作品仍可参与，但歌曲、主唱和主要乐器须由人类创作或演奏，使用的 AI 服务也须合法。规则出台前，一首使用 AI 声音和鼓点改编的歌曲已连续 16 周进入澳大利亚前 20。AP 报道 (<https://apnews.com/article/australia-ai-generated-music-charts-ban-aria-9bfb0c91166ae4405a6df1a3c4891687>)；IFPI 全球原则 (<https://www.ifpi.org/ifpi-rolls-out-global-principles-for-the-eligibility-of-recordings-developed-using-ai-in-official-music-charts-worldwide/>)

这不是对 AI 辅助制作的全面禁令，而是把“实质由人完成”变成行业分发与荣誉资格。唱片公司、平台和创作者将需要保存主创、演奏、模型与授权记录；难点在于混合制作的边界如何核验，以及榜单规则会否进一步影响推荐、版税和广告资格。

台湾起诉 B300 服务器绕道出口案，内部合规人员被纳入追责

台湾检方对九人提起公诉，其中包括一名 Nvidia 员工和两名前 Super Micro 员工。检方称，涉案方以虚假机房和终端用户资料订购 130 台搭载 B300 的高端 AI 服务器；74 台已通过直接运输或经印度尼西亚、日本和香港抵达中国，另 56 台被拦截。检方并称，一名 Nvidia 经理是批准放行 B300 GPU 的关键人物；现阶段为检方指控，法院尚未定罪。AP 报道 (<https://apnews.com/article/taiwan-china-us-nvidia-ai-server-chip-illegal-export-511e9ae69d517c49f19512d45b1a8b0c>)

这起案件把出口管制风险从经销商尽调推进到供应商内部授权、现场核查和物流追踪。芯片与服务器企业不能只留存客户声明，还需验证机房电力、机架容量、网络条件和最终流向；否则形式完备的文件也可能被认定为明知或协助规避。

X 平台高信号

1.

信号标题：Salesforce 把 100 多项企业技能开放给授权 agent

类型：平台访问 / 企业软件

来源或链接：Salesforce 官方发布 (<https://www.salesforce.com/ap/news/press-releases/2026/08/25/salesforce-turns-enterprise-applications-into-enterprise-capabilities/>)

核心观点：Salesforce 扩展 Headless 360，Headless 360 MCP Server 已进入开放测试，Data 360 MCP Server、100 多项可复用技能、Slackbot MCP Client 与 Server 已正式可用；授权 agent 可继承既有身份、权限、验证规则和业务逻辑，动态调用 Marketing、Sales、Service、Commerce 等能力。

为什么重要：企业 agent 不再需要为每个界面重写一套集成，既有 CRM 逻辑开始成为可复用的机器执行能力。

影响：平台团队可减少点对点连接，但应逐项验证技能的写权限、审批点和撤销机制；Salesforce 引用的客户案例称 EVA 已自动解决一半聊天请求，这一数据仍需客户侧独立复核。

验证状态：功能范围和当前可用状态由 Salesforce 一手页面确认；客户成效数据来自厂商案例，未见独立审计。

2.

信号标题：NVIDIA Groq 3 LPX 进入量产，Nebius 成为首个云端采用者

类型：硬件可用性 / 推理性能

来源或链接：NVIDIA 官方公告 (<https://investor.nvidia.com/news/press-release-details/2026/NVIDIA-Groq-3-LPX-Now-in-Full-Production-With-World-Class-Speed-for-Agentic-AI/default.aspx>)

核心观点：Groq 3 LPX 已进入全面量产，并将作为 Vera Rubin 平台的低延迟推理扩展；NVIDIA 称其在 10 万 token 上下文运行 Gemma 4 31B 时达到每秒 3,400 个输出 token，Nebius 将首先通过 Token Factory 提供访问。

为什么重要：长链 agent 的总时延会被每一步生成速度放大，专用解码硬件因此成为独立采购类别。

影响：云客户可为交互型任务单独选择低延迟实例，但产品价格、区域、正式上线日和真实并发性能尚未公布，当前不能据此计算生产成本。

验证状态：量产状态、测试数字与首个采用者由 NVIDIA 一手公告确认；性能来自 Artificial Analysis 测试，云端价格与交付节奏仍待 Nebius 公布。

3.

信号标题：英国成为首个接入乌克兰 Avengers AI Labs 的国际伙伴

类型：政府合作 / 数据准入

来源或链接：英国政府公告 (<https://www.gov.uk/government/news/new-partnership-set-to-see-the-uk-and-ukraine-develop-battle-winning-technology-as-britain-secures-access-to-ukraines-avengers-ai-labs>)；双方意向声明 (<https://www.gov.uk/government/news/joint-declaration-of-intent-between-the-united-kingdom-of-great-britain-and-northern-ireland-and-ukraine-on-a-uk-ukraine-artificial-intelligence-partn>)

核心观点：两国签署 AI 合作意向，英国将成为首个接入 Avengers AI Labs 的国际伙伴；数据来自乌克兰战场的日光与红外传感器。双方已启动三个英国初创企业试点，并计划研究面向无人机、机器人和自主系统的低功耗 AI 芯片。

为什么重要：真实、持续更新的极端环境数据比单纯模型权重更稀缺，政府间协议正在成为此类数据的准入通道。

影响：合作可加快防务与关键基础设施产品验证，同时会把数据分级、模型用途限制、误识别责任和自主系统人工控制推到采购前端。

验证状态：协议、数据来源与试点均由英国政府一手确认；具体数据规模、访问权限和部署效果未公开。

4.

信号标题：OpenAI 封禁一组用 AI 包装伪研究机构的俄罗斯来源账户

类型：平台执法 / 影响行动

来源或链接：OpenAI 调查报告 (<https://openai.com/index/disrupting-malicious-uses-of-ai-influence-campaign-russia/>)

核心观点：OpenAI 称已封禁一组很可能来自俄罗斯、通过 VPN 使用 ChatGPT 的账户；相关人员用模型生成跨 X、LinkedIn、Facebook、Substack 与 Telegram 的推广内容，并将流量导向一个含大量复制、错署名学术文章及亲俄“主权指数”的伪智库网站。抽查的 36 篇文章中有 34 篇复制自其他网站。

为什么重要：影响行动的新增做法不是大规模生成帖子，而是先制造看似可信的机构、专家与指标，再用 AI 扩散其权威外观。

影响：平台与品牌安全团队需要联查内容来源、机构身份、作者履历和跨平台分发网络；只检测单篇 AI 文本会漏掉主要风险。

验证状态：账户处置、样本审查和传播范围来自 OpenAI 一手调查；官方称典型帖文触达较低，外部机构尚未独立归因全部参与者。

5.

信号标题：AI 安全公司 Alice 融资 1.4 亿美元并披露接近 1 亿美元 ARR

类型：融资 / 企业采用数据

来源或链接：公司融资公告 (<https://www.newswire.com/news/alice-raises-140m-to-prevent-ai-from-misbehaving-in-the-real-world>)

核心观点：原 ActiveFence 更名后的 Alice 完成由 Apax Digital 领投的 1.4 亿美元融资，累计融资达 2.8 亿美元；公司称 ARR 接近 1 亿美元、AI 业务两年增长超过 500%，并服务十大领先模型实验室中的八家。

为什么重要：红队、提示注入防护和运行期安全正在从实验室成本中心变成可独立规模化的企业软件类别。

影响：大型客户会更愿意购买跨模型的安全层，但也应要求按攻击类型披露检出率、误报率、修复时长和生产事故覆盖，而非只看客户数量。

验证状态：融资额、收入和客户覆盖均来自公司新闻稿；融资参与方已列明，但 ARR、增长率和客户口径未见第三方审计，因此作为公司披露使用。

前沿研究速递

MobilePA-Bench：让手机 agent 在真实运行约束下接受测试

做了什么：论文 (<https://arxiv.org/abs/2608.23035>) 构建可执行、带状态的移动端沙箱，覆盖 13 个功能领域与 212 个真实工具，同时测试工具调用、子 agent 协作、长期记忆和组合技能调用。

新在哪里：它不只比较界面点击或离线 API 匹配，而是让任务面对严格的调用顺序、权限限制、实时数据库变化和意外运行错误，并用执行证据判定完成情况。

潜在应用方向：手机个人助理、跨应用自动化、端侧 agent 评测、权限策略与交互式强化学习。

一句话判断：移动 agent 的难点已不是找到按钮，而是在权限、状态和失败恢复同时存在时仍能完成闭环。

EchoWM：让世界模型同时生成可导航画面与同步声音

做了什么：论文 (<https://arxiv.org/abs/2608.23189>) 提出 EchoWM，以连续六自由度轨迹控制第一或第三人称场景，并联合生成 720p 视频、环境声、音乐和语音。

新在哪里：离散导航指令与连续姿态被映射到统一的度量尺度轨迹，模型在长序列中同时保持视角移动和多类音频同步，而不是先做视频再外挂声音。

潜在应用方向：可进入的生成式游戏、虚拟拍摄、机器人仿真、沉浸式培训和空间内容制作。

一句话判断：生成媒体正在从“播放一段结果”转向“用户行动会持续改变视听世界”，但物理一致性和长时稳定性仍需独立测试。

Apodex 1.1：把长任务能力拆成环境扩展与多 agent 协同

做了什么：论文 (<https://arxiv.org/abs/2608.23283>) 通过可执行的文件、搜索与代码环境训练 agent，并让系统维护跨工具状态、出处、失败恢复和异步子任务整合；同时发布 35B 参数的开放权重 Mini 版本与开源工作台。

新在哪里：研究目标不是单轮推理分数，而是能否在长时间任务中持续产生可验证进展，并把环境轨迹与协同记录继续用于训练。

潜在应用方向：深度研究、财务分析、科学计算、复杂代码维护和需要人工中途介入的异步任务。

一句话判断：长任务模型开始把“协作与恢复能力”当作可训练对象，但官方称进入领先性能区间，仍需在统一成本和人工介入条件下复现。