由 Codex 自动生成

AI前沿发展日报 | 2026-07-25（Asia）

日期：2026-07-25；覆盖窗口：2026-07-24 07:00 至 2026-07-25 ai），纳入窗口内发布或完成公开确认的重要新增。

今日要点

• Meta AI 开始在部分市场上线可连接邮箱与日历、持续执行定时任务并生成幻灯片的行动能力。消费级助手正从一次性回答转向长期任务关系。

• Microsoft 把高保真图像与低延迟语音模型推入公开预览，并披露自研模型已在 PowerPoint、OneDrive 和 Dynamics 365 中显著降低 GPU 成本。模型竞争进一优化的生产经济性阶段。

• Anduril 在两个月前以 610 亿美元估值融资后，又讨论约 1,000 亿美元的新估值。主系统、软件和硬件一体化的国防科技，正在获得接近传统军工巨头的资本定价。

• NVIDIA、Microsoft、Meta 等 25 家公司联合要求美国政策制定者避免过早限制重模型。开放模型之争已经从技术许可问题升级为产业政策与国家竞争议题。

今日结论

1. 个人 AI 助手的竞争点正在从“回答得更好”移向“持续替用户办事”，但真正的护城河将是跨应用权限、任务记忆、可中断性和稳定交付，而不是单次演示效果。

2. 自研模型的商业价值开始用生产数据证明：当模型能针对具体产品降低 80% 以上 GPU 成本或提高保存率，模型所有权就从研发偏好变成毛利与产品迭代速度的控制点。

3. AI 产业的开放与封闭路线正在形成利益联盟；企业不应把开放权重仅理解为低价替代，而应把可本地运行、供应商切换、合规责任和安全运维能力一起纳入采购决策。

AI 产品与应用

Meta AI 从对话入口转向持续执行任务

Meta 开始在部分市场为 Meta AI app 和 meta.ai 推出行动能力。由 Muse 驱动的助手可以连接邮箱与日历、规划多步骤任务、生成研究报告和幻灯片，并按设定时间持续提供每日简报、训练计划或趋势更新；用户可在任务执行过程中实时改变方向。Meta 表示，相关能力将在未来数周扩展到更多国家和 WhatsApp。Meta 官方发布（https://ut.fb.com/news/2026/07/meta-ai-muse-spark-doesnt-

新增变量不是又一个研究或幻灯片功能，而是 Meta 把社交内容、个人偏好、外部应用和周期性任务放进同一助手。对品牌、电商和本地服务商，这可能形成从灵感发现到计划与购

买的新入口；对用户和企业，则必须重点验证连接账户的授权粒度、任务失败通知、敏感动作确认和撤销机制。官方尚未公布首批市场名单、连接器范围、任务成功率或使用上限。

模型与技术进展

Microsoft 用两档专用模型争夺图像与语音生产负载

Microsoft 将 MAI-Image-2.5-Pro 与 MAI-Voice-2-Flash 保真图像模型，图像输出定价为每百万 tokens 106 美元；后者面向高并发语音场景，官方称速度是 MAI-Voice-2 的两倍、价格低 32%，定价为每百万字符 15 美元。Micr AI 官方发布（https://microsoft.ai/?post_type=new）

更重要的是生产验证：Bing Image Creator 已默认全量使用自研图像模型；PowerP 图像任务的 GPU 成本较 GPT-Image-2 最多降低 84%；OneDrive 关键图像编 率提高 26%、P95 延迟约降 25%；MAI-Voice-2-Flash 在 Dynamics r 中最多降低 89% GPU 成本。这些数字均为 Microsoft 自测或自有产品数据，尚缺独 横向复核，但已说明“通用最强模型”并非每个工作负载的经济最优解。

投融资与商业动态

Anduril 两个月内再次讨论抬升估值至约 1,000 亿美元

Reuters 引述两名知情人士称，Anduril 正与投资者讨论新融资，目标估值约 1,000 元；融资金额尚未确定，曾被讨论的方案要求投资者同时承诺一年内、在达到财务指标后参与第二轮更高估值融资。公司回应称尚未对未来融资作出决定。Reuters 报道转引（https://www.investing.com/news/economy-news/exclusived-in-talks-to-raise-funding-at-about-100-billion-v

这距离 Anduril 以 610 亿美元估值完成 50 亿美元融资仅约两个月。公司披露 2025 收入为 22 亿美元，并在扩张无人机、导弹和自主作战软件。约 1,000 亿美元仍是谈判信号而非已完成交易；若落地，意味着资本市场愿意把 AI 驱动的自主系统平台按接近大型传统军工企业的战略稀缺性定价，同时也把政府采购周期、冲突需求和硬件量产风险一并推高。

政策、伦理与安全

美国科技产业形成开放权重政策联盟

NVIDIA、Microsoft、Meta、IBM、Palantir、Hugging Face 等 pen Weights and American AI Leadership》联名信，要求美国议员 实施会压制竞争或把创新推向海外的过早限制。信中主张，开放模型有利于安全研究、网络防御、产业创新和各国建立自主 AI 能力。Reuters 报道（https://sa.marke

com/news/nvidia-microsoft-and-other-tech-giants-b
e7f51dfd189f727）联名信原文 PDF（https://images.nvidia
nd-American-AI-Leadership.pdf）

值得注意的不是立场本身，而是签署方结构：芯片、云、企业软件和开放模型平台站到同一侧，而 OpenAI、Anthropic 与 Google 未列名。政策博弈将围绕能力门槛、发布责游滥用和对中国模型的限制展开。对企业采购者，开放权重带来的部署自主性不会自动消除责任；模型来源核验、补丁维护、滥用监测和高风险场景验收会更多转移到部署方。

X 平台高信号

1.

信号标题：Claude Opus 5 进入 GitHub Copilot，并按供应商 API 标价

类型：模型接入与计费变化

来源或链接：GitHub Changelog（https://github.blog/chang
-opus-5-is-now-available-in-github-copilot/）

核心观点：GitHub 于 7 月 24 日开始把 Claude Opus 5 逐步提供给 Cop
x、Business 和 Enterprise 用户，覆盖 VS Code、CLI、云端 Ager
E；用量按供应商 API 标价计费，企业管理员必须主动启用对应策略。

为什么重要：前沿编码模型的分销越来越由开发平台完成，模型可用性、企业开关和按量成本在同一入口被绑定。

影响：工程团队应按任务难度设置模型路由与预算，不能把“已进入模型选择器”误认为默认开放；安全测试还需覆盖其对网络安全相关请求的额外拦截。

验证状态：支持计划、接入端、计费方式和管理员开关均由 GitHub 官方核验；实际到账时间为渐进式，GitHub 未公布独立生产基准。

2.

信号标题：Runway 为视频、图像和音频上线可设硬价格上限的模型路由

类型：生成媒体平台规则变化

来源或链接：Runway 官方发布（https://runway.com/news/compar
-runway-media-router）

核心观点：Runway Media Router 已在 Runway Dev 上线。客户可按质量、置偏好，并用供应商白名单、黑名单及单次生成价格上限作为硬约束；系统会返回实际采用的模型及选择原因。

为什么重要：生成媒体开始复制语言模型的多供应商路由模式，但把主观质量与每秒视频成本纳入生产决策。

影响：创意平台可用一个端点管理频繁变化的模型目录，同时通过 dry-run 在不产生费用的情况下验证路由；路由质量仍取决于 Runway 的评测与评分方法。

验证状态：支持模态、约束项、dry-run 和正式可用状态已由 Runway 核验；具体模型覆

盖与路由评测数据未完整公开。

３．

信号标题：Ｍｉｃｒｏｓｏｆｔ　用　ＯｎｅＤｒｉｖｅ　保存率和　ＧＰＵ　成本披露自研模型回报

类型：硬采用与单位经济性数据

来源或链接：Ｍｉｃｒｏｓｏｆｔ　ＡＩ　官方数据　（ｈｔｔｐｓ：／／ｍｉｃｒｏｓｏｆｔ．ａｉ／？ｐｃ

核心观点：ＭＡＩ－Ｉｍａｇｅ－２．５　成为　ＯｎｅＤｒｉｖｅ　关键图像编辑任务的默认模型后，保存率２６%、Ｐ９５　延迟约降　２５%，中等利用率下效率提升　２．５　倍；Ｄｙｎａｍｉｃｓ　３６５　Ｃｏｐｒ　使用　ＭＡＩ－Ｖｏｉｃｅ－２－Ｆｌａｓｈ　后，ＧＰＵ　成本最多降低　８９%。

为什么重要：厂商首次把自研多模态模型的价值同时落到用户行为、延迟和算力成本，而不只报告排行榜名次。

影响：企业评估专用模型时应把完成率、用户采纳、尾延迟和　ＧＰＵ　成本联动观察；这些数据来自　Ｍｉｃｒｏｓｏｆｔ　自有环境，迁移到外部负载前需复测。

验证状态：指标及部署场景由　Ｍｉｃｒｏｓｏｆｔ　官方核验；未披露样本量、对照周期和完整测试条件。

４．

信号标题：Ｍｅｔａ　ＡＩ　的周期任务先在独立应用上线，ＷｈａｔｓＡｐｐ　延后扩展

类型：产品访问与渠道变化

来源或链接：Ｍｅｔａ　官方发布　（ｈｔｔｐｓ：／／ａｂｏｕｔ．ｆｂ．ｃｏｍ／ｎｅｗｓ／２０２６／０ｒｋ－ｄｏｅｓｎｔ－ｊｕｓｔ－ｔｈｉｎｋ－ｉｔ－ａｃｔｓ／）

核心观点：新增行动能力于　７　月　２４　日先在部分市场的　Ｍｅｔａ　ＡＩ　ａｐｐ　与网页端推出，用户一次设定后可持续接收定时结果；ＷｈａｔｓＡｐｐ　等更多入口将在未来数周接入。

为什么重要：Ｍｅｔａ　选择先在可控入口验证长期任务，再把能力带入高频通信渠道，降低一次性大范围扩张的运营风险。

影响：面向　ＷｈａｔｓＡｐｐ　生态的服务商获得明确但尚无具体日期的渠道信号；当前不应假设所有地区、账户或连接器已经可用。

验证状态：分阶段上线与　ＷｈａｔｓＡｐｐ　后续扩展由　Ｍｅｔａ　核验；首批国家、配额、定价和连接器清单未披露。

５．

信号标题：２５　家公司在美国开放权重争论中形成跨层联盟

类型：产业政策立场变化

来源或链接：Ｒｅｕｔｅｒｓ　报道　（ｈｔｔｐｓ：／／ｓａ．ｍａｒｋｅｔｓｃｒｅｅｎｅｒ．ｃｏｍ／ｎ－ａｎｄ－ｏｔｈｅｒ－ｔｅｃｈ－ｇｉａｎｔｓ－ｂａｃｋ－ｏｐｅｎ－ｓｏｕｒｃｅ－ａｉ－ｍｏｄｅｌｓ

核心观点：ＮＶＩＤＩＡ　ＣＥＯ　Ｊｅｎｓｅｎ　Ｈｕａｎｇ　在其首条　Ｘ　帖文中发布联名信，签署方横云、企业软件、国防科技与模型社区；新增变量是产业链多个受益方以共同政策诉求公开结盟，而非单个模型厂商自行辩护。

为什么重要：开放权重的政策影响者已从模型开发者扩展到依赖模型扩散来销售算力、云服

务和企业工具的公司。
影响：未来限制若按发布方式而非能力与用途划线，可能直接改变整条供应链的竞争格局；
签署并不等于已形成具体立法文本。
验证状态：联名信、签署方及 X 发布动作经原文与 Reuters 交叉核验；政策制定者是否
采纳尚无定论。

前沿研究速递

AI Agent 能优化推理服务，但搜索策略仍过于单一

做了什么：InferenceBench 要求 Agent 在两小时内、使用一张 H100 部署兼容
API 的推理服务器，并分别优化预填充延迟、解码延迟、并发吞吐与综合表现；研究比较
了 15 种前沿 Agent 配置。arXiv （https://arxiv.org/abs/26

新在哪里：Agent 相对朴素 PyTorch 基线最高提速 8.08 倍，并可超过默认 vLLM
但仍落后于同等时间预算下最高 11.53 倍的简单超参数搜索。轨迹分析显示，Agent 知道
许多优化方法，却常过早收敛到同一种推理软件。

潜在应用方向：推理引擎自动调优、GPU 成本优化、模型服务部署验收，以及评估编码 Ag
ent 是否真正具备开放式工程搜索能力。

一句话判断：Agent 已能执行复杂性能工程，却还不擅长系统探索；企业应要求它保留候
选方案、对照实验和最优解选择证据。

医疗文本水印可能引入临床语义错误

做了什么：研究在 11 个语言模型、7 个视觉语言模型和多类临床推理任务上测试 5 种文
本水印方法，并加入医学专家验证的质量、术语精度与幻觉审计流程。arXiv （https://ar
xiv.org/abs/2607.20462）

新在哪里：水印不仅影响通用文本质量，还会造成词汇破坏、虚构医学术语，以及放大影像
所见的错误归因或遗漏；常用汇总指标可能掩盖这些临床相关失败。

潜在应用方向：医疗 AI 内容溯源、临床文书生成验收、影像报告安全测试，以及高风险
行业的水印政策设计。

一句话判断：可追踪性不能以临床准确性为代价；医疗场景采用水印前必须做领域级语义审
计，而不能沿用通用基准。

重复采样不能替代多模型不确定性评估

做了什么：研究比较同一模型在温度 1 下重复运行 100 次，与 24 个不同模型在温度 0
下各运行一次，并在 MMLU、HellaSwag、GSM8K 上用随机矩阵检验区分真实结构与采样噪

声。arXiv（https://arxiv.org/abs/2607.20464）

新在哪里：单模型重复采样可提供逐题不确定性，但跨问题最多只有一个高于噪声的维度；多模型组合则识别出四个显著维度，能揭示哪些问题会以相关方式共同失败。

潜在应用方向：高风险问答的置信度估计、多模型路由、评测集诊断，以及采购时比较“多次调用一个模型”和“使用异构模型组合”的价值。

一句话判断：自一致性适合判断一道题是否稳定，不足以描绘模型知识边界；需要系统性风险视图时，多样模型比多次抽样更有信息量。