由 Codex 自动生成

AI前沿发展日报 | 2026-06-25（Asia）

日期：2026-06-25；覆盖窗口：截至 2026-06-25 12:00（Asia/Shang
信号，重点纳入美国与欧洲 2026-06-24 发布、在北京时间 2026-06-25 进入亚洲工
的信息；信息基座：官方发布、一级媒体、TOP500 / 研究源与高信号公开观点交叉核验。

今日总览

今天的主线是 AI 竞争继续向“算力栈、软件栈、组织落地”三层同时下沉。OpenAI 与 B
roadcom 公布 Jalapeño，说明头部模型公司正在把推理成本控制变成核心战略，而不只
是采购更多 GPU。Qualcomm 收购 Modular、TOP500 的 LineShine
植性和国家级高性能计算两端说明：AI 基础设施正在从单一芯片竞争变成全栈生态竞争。
应用侧，ByteDance Seed2.1 和 Thomson Reuters 专业服务报告给出同
已经不满足于“能用 AI”，开始要求 AI 进入高价值工作、质量控制和客户交付。

今日要点

- OpenAI 把自研推理芯片公开化，frontier model 竞争开始更深地绑定硬件与服务
。
- Qualcomm 收购 Modular，说明 AI 基础设施竞争已经从硅片性能扩展到软件可移植
和开发者入口。
- 企业 AI 落地的现实瓶颈不再是“员工愿不愿用”，而是组织是否能把 AI 纳入可审计
交付流程。

AI 产品与应用

- ByteDance Seed2.1 明确押注高价值专业工作，而不是停留在通用问答。
- ChatGPT 默认模型更新继续向决策、研究、购物等高频实用任务靠拢，产品层正在追求
更高日活入口密度。

模型与技术进展

- Jalapeño 代表模型公司开始直接定义推理芯片的负载形态、内存路径和服务约束。
- 长上下文优化与 agent 失败归因研究同步推进，说明 2026 年的技术焦点已从“做更
大模型”转向“让模型更便宜、更可控地工作”。

投融资与商业动态

• Qualcomm 拟以约 39 亿美元收购 Modular，交易重点不是收入并表，而是抢占 C
之外的软件层。
• Thomson Reuters 的专业服务调查显示，AI 投资已开始影响客户流失风险、人才保留
和服务定价。

政策、伦理与安全

• Thomson Reuters 报告量化了 shadow AI 风险，说明企业若不给出合规工具
自行绕过组织边界。
• 国家级算力、企业级权限治理和 agent 审计正在同时进入管理层议程，安全问题已经
从附加项变成主线。

今日结论

• 2026 年 AI 竞争的差异化来源，越来越多来自硬件、软件栈与组织能力的联动，而不
是单点模型分数。
• 企业真正需要采购的是“可交付的 AI 系统”，不是孤立模型能力。

今日三条结论

1. 推理成本正在成为模型公司的战略边界：OpenAI 做自研推理芯片，意味着未来模型价
格、延迟、容量和产品体验会更多由垂直整合能力决定。
2. AI 基础设施的第二战场是软件可移植性：Qualcomm 收购 Modular 不是补一个工
而是在争夺 CUDA 之外的开发者默认层。
3. 企业 AI 的瓶颈已从“员工是否尝试”转向“组织是否能把 AI 变成可审计、可交付、
可计价的工作系统”。

今日 Top 5 大事件

1. OpenAI 与 Broadcom 发布 Jalapeño，自研推理芯片进入

发生了什么： OpenAI 与 Broadcom 在 2026-06-24 发布 Jalapeño
一款 Intelligence Processor，面向 LLM inference 设计，是双方
款 AI 加速器。OpenAI 表示该芯片围绕 ChatGPT、Codex、API 和未来 age
s 的真实服务模式设计，Broadcom 负责芯片实现、网络与量产系统，Celestica 参与板
和机架系统集成。来源：OpenAI 官方发布（https://openai.com/index/
m-jalapeno-inference-chip/）、Broadcom 投资者新闻稿（http
m.com/news-releases/news-release-details/openai-a
mized-intelligence-processor）

关键信息： OpenAI 强调 Jalapeño 是 blank-slate LLM infere

AI 加速器改造成通用推理卡；芯片设计参考了模型 roadmap、kernel、serving s、内存移动、网络和服务模式。OpenAI 还称该项目在九个月内完成 tape-out，并使用 OpenAI 模型加速设计流程。

为什么重要： 头部模型公司的最大成本正在从训练扩展到持续推理。谁能把硬件、模型、服务系统和产品负载一起优化，谁就更可能在 token 价格、延迟、容量和毛利率上形成差异。

商业启发： 企业采购 AI 时不能只看模型榜单，还要看供应商的容量稳定性、单位 token 成本和长期价格曲线。对应用公司而言，模型能力接近后，推理成本会直接决定能否把 AI 放进高频工作流。

2. Qualcomm 将收购 Modular，补齐 AI 软件栈与开发者生态

发生了什么： Qualcomm 宣布已达成协议收购 Modular，称 Modular 的 AI-form 将加强其从 edge 到 data center 的生成式 AI 与 agentic AI Business 报道称交易约为 39 亿美元股票交易，预计 2026 年下半年完成，仍需监批准。来源：Qualcomm 官方发布 （https://www.qualcomm.com/new/qualcomm-to-acquire-modular）、WSJ （https://www.wscquire-ai-software-firm-modular-in-3-9-billion-stess （https://aibusiness.com/generative-ai/qualcomoper-modular）

关键信息： Modular 由 Chris Lattner 与 Tim Davis 创立，核心价值负载跨不同硬件更高效运行，并通过 Mojo / MAX 等技术降低开发者对单一硬件软件栈的依赖。Qualcomm 将其定位为“industry-friendly open software

为什么重要： AI 芯片竞争不只是谁的硅片更快，还包括开发者是否愿意迁移、工具链是否成熟、模型能否在多类设备和数据中心环境中稳定运行。Modular 正好补 Qualcomm 在 AI 开发者软件层的短板。

商业启发： 未来企业 AI 架构会更重视“可移植推理”。如果软件层足够抽象，企业可以在云端、边缘、PC、移动设备之间更灵活地调度模型，减少被单一 GPU 生态和单一云平台锁定。

3. 中国 LineShine 登顶 TOP500，AI 算力竞争扩展到国家级 H

发生了什么： TOP500 在 2026-06-23 发布第 67 版榜单，中国 LineShir排名第一，取代美国 El Capitan。LineShine 部署在深圳国家超级计算中心，由深圳云算中心建设，HPL 成绩为 2.198 Exaflop/s，使用 13,789,440 个核心。来新闻稿 （https://top500.org/news/lineshine-debuts-no-l-exascale-era/）、TOP500 2026-06 榜单 （https://top56/）

关键信息： TOP500 披露 LineShine 基于自定义 LingKun 平台、LX2 30
ngQi 互联和 Kylin OS，成为 2017 年 Sunway TaihuLight 之后首
国系统。其 HPL 性能较第二名 El Capitan 高出 20% 以上。

为什么重要： 虽然 HPL 不是 AI 训练或推理的完整指标，但 LineShine 证明中国在高
能计算自主体系上仍具备公开可验证的进展。AI 基础设施竞争正在同时发生在 GPU、CPU
、互联、操作系统、液冷、电力和国家级供应链上。

商业启发： 中国企业做 AI 长周期规划时，需要把“模型能力”和“算力可得性”分开看
。国产算力栈在科学计算和部分 AI 工作负载上可能先形成可用替代，但与主流 AI GPU
生态、框架、工具链的衔接仍是决定商业化速度的关键。

4. ByteDance Seed2.1 发布，强调从模型能力走向专业生产力

发生了什么： ByteDance Seed 团队在 2026-06-23 发布 Seed2.1，官
专业工作，支持信息分析、方案设计、内容规划和结果整合，可帮助用户推进过去依赖外部
顾问或专门服务团队完成的工作。来源：ByteDance Seed 官方博客 （https://se
dance.com/en/blog/seed2-1-officially-released-adv
teDance Seed 博客列表 （https://seed.bytedance.com/blo

关键信息： 官方材料把 Seed2.1 的价值放在复杂任务交付，而不是单轮问答。Seed 还在
2026-06-19 将 Seed-2.1-Pro-Preview 放到 Arena AI 的 C
两周内进入飞书 Spark 和 Coze。来源：Seed-2.1-Preview on Arena
tedance.com/en/blog/seed-2-1-preview-model-releas

为什么重要： 中国大模型竞争正在从“参数与榜单”转向“是否能进入企业办公、内容生
产、低代码 agent 和平台生态”。飞书、Coze 与 Seed 模型如果形成闭环，会更接近企
业真实场景。

商业启发： 对中国品牌、内容服务和中小企业来说，值得关注的不是单个模型是否领先全
球，而是模型能否直接嵌入素材规划、脚本生成、营销复盘、客服流程和内部知识工作。平
台分发能力会放大模型能力。

5. Thomson Reuters 报告：专业服务 AI 已从效率工具变成客户与

发生了什么： Thomson Reuters 发布 2026 Future of Professi
800 名专业人士的全球调查，称美国专业服务市场有高达 1,430 亿美元客户收入面临风险
；74% 的专业人士每周使用 AI，但 91% 认为组织没有兑现 AI 应有价值。来源：Thoms
Reuters 官方发布 （https://www.thomsonreuters.com/en/p
ai-is-ready-but-firms-are-not-how-falling-behind-
ting-clients-and-talent）

关键信息： 报告称三分之一的律师、会计和合规专业人士在使用未经组织批准的 AI；78%

的客户认为 ＡＩ 带来的质量提升已是必要条件，但只有 ６％ 认为多数服务商已经交付。另有四分之一专业人士表示，如果看不到预期 ＡＩ 价值，两年内会考虑离开。

为什么重要： 这说明企业 ＡＩ 落地的风险不再只是“投入浪费”。如果组织没有提供合规、可信、可解释的 ＡＩ 工作系统，员工会转向 ｓｈａｄｏｗ ＡＩ，客户会重新评估供应商，人才也会把 ＡＩ 能力视作工作环境的一部分。

商业启发： 律所、会计师事务所、咨询、合规和内容服务公司需要把 ＡＩ 纳入交付标准，而不是内部效率实验。可验证内容、保密边界、审计记录、人工复核和客户可解释性会成为服务溢价的一部分。

商业与应用解读

大模型公司： ＯｐｅｎＡＩ 今天释放的不是单纯硬件新闻，而是“模型公司要控制推理经济性”的信号。未来 ｆｒｏｎｔｉｅｒ ｌａｂ 的竞争会更像云厂商与芯片公司的混合体：模型能力吸引需求，推理芯片和服务系统决定供给成本。ＢｙｔｅＤａｎｃｅ Ｓｅｅｄ２．１ 则代表另一条路线：通过办公、内容和 ａｇｅｎｔ 平台把模型能力直接导入应用生态。

ａｇｅｎｔ / ｃｏｄｉｎｇ / ｗｏｒｋｆｌｏｗ： Ｑｕａｌｃｏｍｍ 收购 Ｍｏｄｕｌａｒ 与从 ｄｅｍｏ 到生产的摩擦。前者解决跨硬件部署和开发者工具链，后者解决复杂工作交付和平台入口。对企业来说，２０２６ 年下半年的 ａｇｅｎｔ 评估不应只看“能否完成任务”，还要看是否具备可移植运行、权限控制、审计日志、人工接管和成本测量。

中国企业与内容服务场景： ＬｉｎｅＳｈｉｎｅ 和 Ｓｅｅｄ２．１ 是两类不同但互补的中国信号：一个在算力自主栈，一个在应用生产力栈。内容服务公司、ＭＣＮ、品牌营销团队和客服中心更应优先试验 Ｓｅｅｄ / Ｃｏｚｅ / 飞书一类靠近业务流程的能力，而不是等待最强通用模型。关键指标应是内容周转时间、复用率、人工审核成本和客户可交付质量。

基础设施与成本： Ｊａｌａｐｅñｏ、Ｍｏｄｕｌａｒ 和 ＬｉｎｅＳｈｉｎｅ 合在一起说明，ＡＩ 成只来自模型蒸馏。硬件专用化、软件可移植、长上下文缓存压缩、调度和互联都会进入企业ＡＩ 成本模型。ＣＩＯ 需要建立 ｔｏｋｅｎ、延迟、ＧＰＵ / ＮＰＵ 占用、缓存命中率和供应商替的统一看板。

风险与治理： Ｔｈｏｍｓｏｎ Ｒｅｕｔｅｒｓ 的数据把 ｓｈａｄｏｗ ＡＩ 问题量化了。员工已经在组织不提供可信工具并不会降低风险，只会让风险不可见。专业服务、金融、医疗和法务团队应优先建立“批准工具清单 + 数据分级 + 结果引用 + 审批记录”的基本制度。

Ｘ 平台高信号观点

１． 信号标题： ＯｐｅｎＡＩ 在 Ｘ 把 Ｊａｌａｐｅñｏ 明确定义为服务 ＣｈａｔＧＰＴ、Ｃ第一款自研推理芯片。
来源或链接： ＯｐｅｎＡＩ ｏｎ Ｘ （ｈｔｔｐｓ：／／ｘ．ｃｏｍ／ＯｐｅｎＡＩ／ｓｔａｔｕｓ／２

penAI 官方发布 （https://openai.com/index/openai-broad
hip/）

核心观点： 这条公开信号不是单纯秀硬件，而是 OpenAI 首次把“模型服务负载需要什么芯片”直接说成公司级产品战略。

为什么重要： 当模型公司开始自己定义推理芯片，未来 token 定价、延迟、容量上限与产品稳定性都会更多由垂直整合决定，而不是只由 GPU 市场供给决定。

影响： 这会迫使企业客户在比较模型时，把供应稳定性、推理成本曲线和长期服务能力放进同一个采购表，而不再只看榜单成绩。

2. 信号标题： Qualcomm 宣布收购 Modular，公开把开发者软件层提升到并购级优先事。

来源或链接： Qualcomm 官方发布 （https://www.qualcomm.com/ne6/qualcomm-to-acquire-modular）、AI Business （https e-ai/qualcomm-acquire-ai-platform-developer-modul

核心观点： Qualcomm 看重的不是一套点状工具，而是让 AI 工作负载能跨 edge、PC、bile 和 data center 迁移的默认软件层。

为什么重要： 这说明 AI 基础设施竞争已经进入“谁掌握开发者迁移路径”的阶段；硬件性能不再足够，软件可移植性本身已成为战略资产。

影响： 对企业架构团队来说，未来更应评估模型与推理栈能否跨多类硬件运行，否则成本、供应和平台锁定风险会同时放大。

3. 信号标题： ByteDance Seed2.1 把价值主张从“模型更强”推进到“能接管高价值业工作”。

来源或链接： ByteDance Seed 官方博客 （https://seed.bytedanc 2-1-officially-released-advancing-ai-productivity （https://seed.bytedance.com/en/blog/seed-2-1-prev

核心观点： Seed 团队这次强调的是信息分析、方案设计、内容规划与结果整合，并同步给出进入飞书 Spark 与 Coze 的落地路径。

为什么重要： 这类产品信号比单一 benchmark 更有商业意义，因为它直接回答了“模型是否会进入企业真实流程和内容生产入口”。

影响： 中国企业和内容团队可以据此优先评估 AI 在脚本规划、营销复盘、客服协作和知识工作中的可交付价值，而不是继续停留在通用聊天试用。

4. 信号标题： Thomson Reuters 用客户收入、人才流失和 shadow AI 数据风险从抽象讨论变成经营指标。

来源或链接： Thomson Reuters 官方发布 （https://www.thomsonr s-releases/2026/june/ai-is-ready-but-firms-are-no implementation-is-costing-clients-and-talent）

核心观点： 74% 的专业人士每周使用 AI、91% 认为组织尚未兑现 AI 价值，且大量员工仍在使用未经批准的工具，说明风险已从“试验失败”变成“交付与治理失配”。

为什么重要： 这不是泛泛的 adoption 口号，而是直接把 AI 落地不足与收入风险、人才流失和客户预期落差挂钩。

影响： 律所、咨询、财税和内容服务公司需要尽快把 AI 纳入正式交付标准、审计记录与权限制度，否则既会丢客户，也会丢一线人才。

5. 信号标题： ChatGPT release notes 把默认模型优化重点放到决策、计划、研物等高频任务。

来源或链接： ChatGPT Release Notes（https://help.openai53-chatgpt-release-notes）

核心观点： OpenAI 释放的产品方向不是“再讲一次更强推理”，而是把默认体验进一步贴近日常工作和消费决策场景。

为什么重要： 默认模型的调优方向，往往比一次性新品发布更能反映平台真正押注的增长入口，因为它会直接改变用户日常使用分布。

影响： 对 AI 应用团队而言，这意味着下一阶段竞争点会更多落在默认入口、任务完成率和工作流嵌入深度，而不是只靠少量专家用户的复杂任务演示。

前沿研究速递

1. SAFARI：用主动调查解决长程 agent 失败归因

做了什么： SAFARI 提出一个面向长程 agent 轨迹的故障归因框架，用工具增强的诊断循环读取、搜索轨迹片段，并用短期记忆支持跨轮推理，避免把完整轨迹一次性塞进上下文。

来源：arXiv:2606.24626（https://arxiv.org/abs/2606.2

新在哪里： 论文指出复杂 multi-step / multi-agent 任务轨迹会超过上下文窗做法会出现 attention dilution。SAFARI 在 Who&When 和 TRAI结果，并能处理目标故障位于原生上下文窗口 5 倍之外的场景。

潜在应用方向： 企业 agent 运维、自动化工作流审计、代码 agent 事故复盘、客服机器人质量分析、多 agent 协作监控。

一句话判断： agent 规模化后，最稀缺的不是“会执行”，而是失败后能定位责任链。

2. GUI vs. CLI：计算机使用 agent 的执行瓶颈比较

做了什么： 论文构建 440 个桌面任务、18 个应用和 12 类工作流的匹配基准，比较 screen-only GUI agents 与 skill-mediated CLI agents 在表现。来源：arXiv:2606.24551（https://arxiv.org/abs/260

新在哪里： 最强 GUI agent 达到 59.1% full pass rate，强于原始 s的 48.2%；但加入 verifier-guided skill augmentation 后，。这说明 CLI 的短板很多来自 skill coverage，而不是模型能力本身。

潜在应用方向： 办公自动化 agent、RPA 替代、软件测试、企业桌面操作、内部工具编排。

一句话判断： 生产级 computer-use agent 不能只押 GUI 或 CLI，关键是把 all 接口做全。

3. CompressKV：长上下文推理的 KV cache 压缩

做了什么： CompressKV 面向 GQA-based LLM，识别 Semantic Retrieval Heads，保留语义重要 token，并按层分配 cache budget，以降低长上下文推理的内存和解码成本。来源：arXiv:2606.24467（https://arxiv.org/abs/2606.24

新在哪里： 论文称在 LongBench 问答任务中，CompressKV 用 3% KV cache 达到 97% full-cache performance；在 Needle-in-a-Haystack 上到 90% accuracy。

潜在应用方向： 长文档问答、企业知识库、代码库分析、客服历史上下文、低成本 agent memory。

一句话判断： 长上下文的商业化不只靠模型窗口变大，更靠把记忆成本压下来。

明日追踪清单

- 跟踪 OpenAI Jalapeño 是否释放更多制程、内存架构、互联与量产节奏细节，判断其对 API 定价和容量的真实影响。
- 跟踪 Qualcomm 收购 Modular 后对 Mojo、MAX 与异构推理部署路线的整合是否形成对 CUDA 生态的实质替代路径。
- 跟踪 Seed2.1 进入飞书 Spark 与 Coze 的具体时间表和公开案例，验证其是否真入企业生产流而不是停留在产品宣传。