

A I 前沿发展日报 | 2 0 2 6 - 0 8 - 0 5 (A s i a)

日期：2 0 2 6 - 0 8 - 0 5；覆盖窗口：2 0 2 6 - 0 8 - 0 4 0 7：0 0 至 2 0 2 6 - 0 8 - 0 5 a i)，以窗口内公开发布、实施或获得一手材料确认的信息为准。

今日要点

- 今天最硬的两组数据来自 A I 软件与算力两端：P a l a n t i r 季度营收同比增长 9 3 %，A M 数据中心收入同比增长 1 0 7 %。前者证明高价值企业 A I 已能快速兑现合同，后者说明推理、a g e n t 与服务器扩容仍在持续拉动非英伟达算力。
- Q w e n 把 Q w e n 3 . 8 - M a x 从既有预览访问推进到“下周开放权重”的明确承诺，并同步预告 2 7 B 开放权重版本。真正可验证的变化是交付方式与时间表，性能排名仍主要来自厂商口径。
- 美国先进模型预发布评测的当天新增变量不是“已经完成”，而是适用范围：据参加白宫会议的人士披露，开放权重模型被排除，覆盖对象被限定为具有前沿能力和国家安全风险的闭源模型。
- 当天没有出现经官方确认、重要性高于上述事项的新大型 A I 融资。资本市场的有效信号转而来自收入、合同与基础设施利用率，而非新一轮估值叙事。

今日结论

- 企业 A I 已出现可审计的商业化分层：能把模型接入核心运营、形成大额合同的供应商正在加速增长；只提供通用助手的产品仍需证明续费、扩席与单位任务价值。
- 开放模型竞争的关键变量正从“是否开放”变成“旗舰权重何时交付、能否低成本部署、评测能否复现”。Q w e n 的下一步应按实际权重、许可证和第三方测试验收，而不是按参数规模采购。
- 美国的预发布评测安排正在制造闭源与开放权重模型之间的监管不对称。模型供应商的发布形态将同时影响合规路径、政府可见性和企业客户的版本选择。

A I 产品与应用

P a l a n t i r 的美国商业收入增长 1 4 9 %，企业 A I 从试点叙事进入合同兑现

P a l a n t i r 8 月 3 日美股盘后公布、并在本覆盖窗口持续被市场消化的二季度结果显示：总营收达到 1 9 . 3 5 亿美元，同比增长 9 3 %；美国商业收入为 7 . 6 4 亿美元，同比增长 1 4 %、环比增长 2 8 %；美国商业合同总价值为 2 1 . 3 2 亿美元，同比增长 1 5 3 %。公司将 2 0 2 年全年营收指引上调至 8 1 . 5 0 亿—8 1 . 5 8 亿美元，并预计美国商业收入至少达到 3 4 . 2 4 亿美元。P a l a n t i r 向 S E C 提交的业绩公告 (<https://www.sec.gov/A>

ta / 1321655 / 000132165526000039 / a2026q2ex991pressre
复核 (https : // apnews . com / article / stocks - markets - dol
e6128d618509e33d45a493c2615b1)

这组数字不能代表整个企业 AI 市场，但它给出了比席位数更硬的验证：客户正在为嵌入运营与决策流程的系统签订更大合同。采购方仍应区分合同总价值与当期收入——前者包含未来期限与可能不被行使的选项——并要求供应商披露上线周期、活跃 workflows、人工接管率和已实现财务收益。

模型与技术进展

Qwen 3 . 8 - Max 承诺下周开放权重，旗舰模型的交付形态发生变化

Qwen 团队在 8 月 3 日宣布，下周将开放 Qwen 3 . 8 - Max 权重，并同时推出开放权重 Qwen 3 . 8 - 27 B。当前可调用的是 qwen 3 . 8 - max - preview : QwenCloud 文档，支持推理，支持 low、medium、xhigh 三档推理强度，最高映射到 262,144 tokens 预算；阿里称 Max 版本总参数量为 2.4 万亿，面向 coding 与 cowork 场景。Qwen 发布入口 (https : // qwen . ai / blog) ; QwenCloud API 文档 (https : // docs . qwen . ai / api - reference / chat / openai - chat) ; AP 对模型发布状态的独立追踪 (https : // www . com / article / 4c66a2e0f557ce79d3cc2d769c9a6226)

当天最重要的新增事实是开放权重时间表，而不是厂商给出的性能名次。权重、模型卡、许可证和可复现实测尚未同时落地，因此企业现在适合准备评测集与部署预算，不适合预先把它视为已验证的本地旗舰方案。若 27 B 版本也按期交付，实际影响可能更偏向单机或小集群部署，而非 2.4 万亿参数本身。

投融资与商业动态

AMD 数据中心收入翻倍，第三季度收入指引升至 130 亿美元

AMD 于 8 月 4 日美股盘后公布二季度营收 115 亿美元，同比增长 50%；其中数据中心收入 67 亿美元，同比增长 107%，公司将增长归因于 EPYC 处理器与 Instinct GPU 需求。AMD 预计第三季度营收约 130 亿美元、上下浮动 3 亿美元，对应中值同比增长约 41%、环比增长约 13%。AMD 官方业绩公告 (https : // ir . amd . com / news - releases / detail / 1295 / amd - reports - second - quarter - 2024)

这说明 AI 基础设施需求正在同时拉动 CPU 与加速器，且竞争已从单颗 GPU 延伸到服务器与机架级组合。对云厂商和大型企业，AMD 的增长扩大了第二供应源的现实性；但采购决策仍要把 ROCm 软件兼容、网络互连、供货周期和实际推理吞吐一起核算，不能把收入增速直接等同于可替代性。

当天未发现达到同等可信度与重要性的新大型 AI 融资。相比融资额，AMD 的出货收入与

下季指引更能说明当前算力需求是否真正兑现。

政策、伦理与安全

美国预发布评测排除开放权重模型，闭源模型成为主要覆盖对象

Axios 8 月 4 日在白宫与行业举行 staff-level 会议后披露：适用对象被定义为兼
沿能力与国家安全风险的闭源模型；开放权重模型被排除，且文件明确表示不应被解释为限
制已经发布的开放模型。最长 30 天的政府预发布访问期内，模型需存放在高安全环境并
记录访问者，评测由多个政府部门参与；“前沿”与“国家安全风险”的具体门槛仍未公开
。Axios 当日报道 (<https://www.axios.com/2026/08/04/tru-del-s>)；白宫 6 月 2 日行政命令 (<https://www.whitehouse.gov/s/2026/06/promoting-advanced-artificial-intelligence/>)

相较前一日“安排已完成”的状态，这次新增的是适用边界与操作要求。直接后果是，同等
能力若采用闭源 API 或开放权重发布，可能进入不同的政府评测路径。企业客户应要求闭
源供应商披露预发布审查是否影响版本日期、区域可用性与回退方案，同时对开放权重模型
加强自行红队测试；排除政府预评测不等于风险更低。

X 平台高信号

1. SpaceX 首份上市后财报单列 AI 收入

信号标题：SpaceX 二季度 AI 收入达到 25.6 亿美元

类型：硬收入数据 / AI 商业化

来源或链接：Axios 对 SpaceX 二季度结果的当日报道 (<https://www.axios.com/2026/08/04/spacex-earnings-elon-musk>)；SpaceX 官方财报活动页 (<https://updates/releases-details/2026/SpaceX-to-Post-Second-Host-Webcast-on-August-4-2026-2026-g8layJlbFm/>)

核心观点：SpaceX 在上市后首份季度报告中把 xAI 与 Grok 订阅计入 AI 分部；Axios
报道该分部收入为 25.6 亿美元，同比增长 247%。

为什么重要：合并后的 SpaceX AI 首次给出可单独观察的收入基线，外界可以开始判断 AI
业务是否足以承担快速扩张的算力投入。

影响：后续应持续跟踪 AI 分部毛利、资本开支与内部交易口径；只有收入增长而没有现
金消耗拆分，仍不足以判断商业质量。

验证状态：财报发布日期由 SpaceX 一手活动页确认；具体 AI 分部收入与增速来自 Axios
当日报道，官方可抓取页面尚未呈现完整表格，故按一级媒体二次验证处理。

2. AI 数据中心需求外溢到燃气轮机与工程机械

信号标题：Caterpillar 报告数据中心发电设备订单增加

类型：基础设施需求 / 硬件订单

来源或链接：AP 对 Caterpillar 二季度结果的报道 (<https://apnews.com/stocks-markets-dollar-yen-trump-iran-war-fbbe6128d61eerpillar> 官方财报入口 (<https://www.caterpillar.com/en/releases/h/2Q26-earnings-media-advisory.html>))

核心观点：Caterpillar 二季度销售与收入首次超过 200 亿美元；管理层称主要业务订单率强、积压增加，AP 进一步确认数据中心用燃气轮机订单是受益来源之一。

为什么重要：AI 基础设施的约束正从芯片扩展到现场发电、施工和交付周期，传统工业供应商成为算力扩张的领先指标。

影响：数据中心项目评估应把发电设备排产、燃料供应与建设周期放进容量计划，而不是只按 GPU 到货日倒排。

验证状态：财报节点由 Caterpillar 一手页面确认；超过 200 亿美元、订单与数据中心轮机需求由 AP 当日报道复核，未取得公司逐项订单明细。

3. 欧盟 AI 透明度义务进入实际适用期

信号标题：AI 交互告知与生成内容机器可读标记自 8 月 2 日起适用

类型：合规生效 / 内容标识

来源或链接：欧盟委员会官方说明 (<https://digital-strategy.ec.europa.eu/en/commission-publishes-guidelines-transparency-obligations-certain-ai-systems>)

核心观点：面向欧盟用户的 AI 系统需要告知用户正在与 AI 交互，提供者还需为生成或操纵内容加入机器可读标记；部署者对深度伪造、无人工编辑的公共利益内容、情绪识别与生物特征分类负有额外告知义务。

为什么重要：透明度已从产品承诺变成上线要求，影响聊天界面、内容导出、元数据链路和下游分发。

影响：产品团队应逐项核查界面提示、导出文件标记、内容再编码后的标记保留，以及公共利益内容是否有人类编辑控制。

验证状态：适用日期、主体范围和主要义务均由欧盟委员会一手指南确认。

4. 明尼苏达州把“提供生成裸照工具”本身列为违法

信号标题：美国首个禁用 nudification 工具的州法已生效

类型：平台规则变化 / 合成内容安全

来源或链接：明尼苏达州议会官方法律摘要 (<https://www.house.mn.gov/New/2026/5741>)；AP 对 xAI 诉讼的报道 (<https://apnews.com/article/de093b567b12c9e16>)

核心观点：自 8 月 1 日起，网站、应用或服务不得让用户访问、下载或使用将真实人物图像“裸化”的工具，也不得为此类服务做广告；每次违法最高可处 50 万美元民事罚款。

。

为什么重要：责任点从传播非自愿私密图像前移到工具访问和平台提供，直接改变图像生成

产品的地域访问控制要求。

影响：面向美国市场的生成图像服务需要识别地域、拦截相关功能与广告，并保留执行证据；xAI 的诉讼意味着法律边界仍可能变化。

验证状态：生效日期、禁止行为、私人诉权与罚款上限由州议会一手摘要确认；xAI 的宪法挑战由 AP 核验，诉讼尚未终结。

5. GitHub Actions 对可疑 workflow 增加自动人工闸门

信号标题：公共仓库的潜在恶意 workflow 需获写权限协作者批准

类型：平台规则变化 / 软件供应链

来源或链接：GitHub 官方变更说明 (<https://github.blog/changelog/2024-09-18-github-actions-holds-unproven-workflows-for-approval/>)

核心观点：GitHub 现在会自动暂缓被识别为潜在恶意的 Actions workflow；只有具备写权限的仓库协作者通过已认证网页会话批准后，workflow 才会执行，无需仓库自行配置。

为什么重要：coding agent 与自动化工具能更快地产生和提交 workflow，平台侧默认闸门可以降低被盗凭据直接触发密钥窃取脚本的风险。

影响：公共仓库维护者需要明确批准责任人和响应时限，并把“被暂缓”与普通 CI 失败分开监控；当前保护不覆盖 GitHub Enterprise Server。

验证状态：适用对象、批准权限、默认启用方式与不覆盖范围均由 GitHub 一手变更日志确认。

前沿研究速递

ParEvalLayer：判断 agent 评测何时可以提前结束

做了什么：ParEvalLayer (<https://arxiv.org/abs/2608.02474>) 成对评测增加预先设定的决策层，在部分任务完成后输出“达到优势门槛”“未达到”“需要更多证据”或“拒绝判断”，并用已完成的公开基准回放验证早停结果。

新在哪里：它不把部分得分直接当结论，而是同时约束优势门槛、任务观察比例与未决比较。论文称在三个公开基准上，只观察 15%—25% 的任务结果即可得到与完整评测相同的判断；其他基准则需要更多样本。

潜在应用方向：模型路由 A/B 测试、coding agent 回归测试、持续集成中的评测预算控制，以及供应商版本更新验收。

一句话判断：这项工作把“跑到多少算够”变成可审计决策，但可靠性依赖预先锁定规则，不能在看到中途结果后再选择有利门槛。

EchoCache：用音频能量决定视频生成缓存何时刷新

做了什么：EchoCache (<https://arxiv.org/abs/2608.02474>)

型，用音频时频能量定位更重要的时刻，指导潜变量缓存更新，并结合动态时间步缓存与量化缓存管理降低计算和显存开销。

新在哪里：现有方法主要利用视觉时间冗余，它把“声音在哪些时刻真正驱动画面变化”纳入缓存策略。作者报告在 W a n 2 . 2 - S 2 V 与 E M T D 基准上实现 2 . 4 6 倍加速，同时维持质量和音画一致性；代码已公开。

潜在应用方向：数字人、口播视频、音乐可视化、实时虚拟主播和批量短视频生产。

一句话判断：跨模态显著性比固定刷新率更合理的缓存依据，但 2 . 4 6 倍结果仍需在不同音频类型、分辨率和硬件上复测。

C M u o n ：拆开融合权重后再做正交化，缩短扩散模型训练

做了什么：C M u o n (<https://arxiv.org/abs/2608.02502>) 针对批 a L N、Q K V 等融合张量，把功能不同的子矩阵先拆分，再分别执行动量正交化，减少标准 M u o n 优化器造成的子空间耦合。

新在哪里：作者称 6 . 7 5 亿参数 D i T 在 I m a g e N e t 2 5 6 上训练 2 0 0 个 e p F I D，相比 A d a m W 获得超过 2 倍训练加速，并缓解标准 M u o n 的后期收敛停滞。

潜在应用方向：图像与视频扩散模型预训练、需要降低优化器成本的生成模型实验，以及融合线性层较多的 T r a n s f o r m e r 训练。

一句话判断：改动简单且潜在节省可观，但当前证据集中在单一规模和视觉任务，扩展到更大模型前仍需验证稳定性与收益是否保持。