

A I 前沿发展日报 | 2 0 2 6 - 0 7 - 3 1 (A s i a)

日期：2 0 2 6 - 0 7 - 3 1；覆盖窗口：2 0 2 6 - 0 7 - 3 0 0 7 : 0 0 至 2 0 2 6 - 0 7 - 3 1 a i)，仅纳入窗口内新发布、正式生效或获得可靠公开确认的事实，以及本期新进入公开研究筛选的论文。

今日要点

- 亚马逊把 A I 需求转成了可核收入：A W S Q 2 收入同比增长 3 7 % 至 4 2 2 亿美元，业务与芯片业务的年化收入均超过 2 5 0 亿美元。
- 增长仍以现金流换取。亚马逊过去 1 2 个月自由现金流转为净流出 7 6 亿美元，财报明确把主要原因指向 A I 相关资产投入；这为昨日微软、M e t a 的资本开支分化增加了第三个可比样本。
- 产品侧出现两种相反但互补的信号：A m a z o n A d s A g e n t 已带来更低的展示和获客成本，而 G i t H u b M o d e l s、A m a z o n Q B u s i n e s s 与旧版 B e d r o c k A g e n t 已停止接纳新客户。
- 自动驾驶商业化跨过新的监管门槛。美国监管机构允许 Z o o x 在取德州和地方许可后对无方向盘、无人工控制装置的专用 r o b o t a x i 收费。

今日结论

- 1 . A I 云需求已经足以让三大云厂商同时给出硬增长数据，但“收入增长”和“现金回收”正在分成两张成绩单；未来估值差异会更多来自资产效率，而不只是算力规模。
- 2 . 企业 a g e n t 正从功能展示进入结果计价期。降低获客成本、提高购物转化、缩短部署周期等业务指标，将比“可自主执行多步任务”更能决定预算续签。
- 3 . 平台选择必须把生命周期风险计入总成本。模型入口、知识助手和第一代 a g e n t 服务可以在需求仍存在时被主动收拢，迁移能力与数据可携带性已成为采购条件。

A I 产品与应用

A m a z o n 用转化数据证明 a g e n t 开始影响交易

亚马逊披露，使用 A d s A g e n t 的广告主平均展示成本低 8 %、获客成本低 6 %；由 R u f u s 与 A l e x a + 合并而来的 A l e x a f o r S h o p p i n g 在 Q 2 的活跃用户接近翻倍，增长超过 5 倍。美国用户使用 A l e x a f o r S h o p p i n g 后，平均每单消费额比未使用者高 1 0 % 以上；试用 A l e x a + 的客户加入 P r i m e 的概率高近 2 5 %。A m a z o n Q 2 业绩公告 (https://s2.q4cdn.com/299287126/files/doc_earnings/2026/026-Earnings-Release.pdf)

这里的新增变量不是“购物 agent 可以做什么”，而是它是否改变单位经济性。广告、会员和交易额三条指标开始共同指向可量化回报；但这些数字来自公司整体样本，尚不能证明因果关系，也未披露客户分层与长期留存。

模型与技术进展

AWS 自研芯片同时获得外部采用和规模证据

亚马逊称，AWS 芯片业务年化收入已超过 250 亿美元且同比增速为三位数；Anthropic 与 OpenAI 已对 Trainium 作出多年、多吉瓦承诺。Graviton5 已正式可用，相比 4 最高提升 25% 计算性能；Graviton 系列覆盖 AWS 前 1,000 大 EC2 客户收入承诺环比接近三倍。Amazon Q2 业绩公告 (https://s2.q4cdn.com/299287126/files/doc_earnings/2026/q2/earnings-result/AMZN-Q2-2026-Earnings-Release.pdf)

这说明云厂商自研芯片不再只是压低内部成本的替代品，而在形成独立收入与长期容量合约。对模型公司，算力采购会更像多供应商组合；对企业客户，评估重点应转向模型在特定芯片上的实际吞吐、迁移成本和可用区域，而不是只比较芯片峰值参数。

投融资与商业动态

AWS 加速至 37%，但 AI 投入使过去 12 个月自由现金流转负

亚马逊 Q2 总收入为 2,006 亿美元，同比增长 20%；AWS 收入 422 亿美元，同比增长 37%，为 18 个季度以来最快，AWS 营业利润升至 166 亿美元。与此同时，公司过去 12 个月自由现金流从上年同期净流入 182 亿美元转为净流出 76 亿美元，物业和设备净购置额同比增加 661 亿美元，主要用于 AI。Q2 净利润中的 534 亿美元税前非经营收益主要来自对 Anthropic 的投资重估，不能视为当期经营利润。Amazon Q2 业绩公告 (https://s2.q4cdn.com/299287126/files/doc_earnings/2026/q2/2026-Earnings-Release.pdf)

与昨日微软“高投入但利润同步增长”、Meta“资本开支几乎吞掉单季经营现金流”的观察相比，今天的新增事实是 AWS 已显著加速，却仍未阻止亚马逊整体自由现金流转负。商业判断必须同时看云增速、基础设施利用率和投资重估收益的会计影响。

政策、伦理与安全

Zoox 获准为无人工控制装置的专用 robotaxi 收费

美国国家公路交通安全管理局批准 Zoox 的临时豁免，使其在取德州和地方许可后可以对专用、无方向盘 robotaxi 的乘车服务收费。AP 报道称，这是此类专用车辆首次获得美国商业收费通道；安全倡议者同时指出，Zoox 未公开足够细节来证明其车辆在公共道路上的安全性。AP 报道 (<https://apnews.com/article/1bdb3bb8ecc>)

此前的联邦申请记录 (<https://www.govinfo.gov/app/details/FR730>)

监管意义在于，规则开始从“允许演示”推进到“允许收费”，但联邦豁免不等于全国通行证。商业扩张仍受州、地方许可、事故报告义务和真实道路安全数据约束；对具身 AI 企业，合规能力正在与模型和硬件能力一起决定上市速度。

X 平台高信号

1 .

信号标题：AWS AI 业务年化收入首次超过 250 亿美元

类型：硬采用与收入数据

来源或链接：Amazon Q2 业绩公告 (<https://s2.q4cdn.com/2992871ings/2026/q2/earnings-result/AMZN-Q2-2026-Earnings>)

核心观点：AWS 首次披露其 AI 业务年化收入超过 250 亿美元，且同比增长为三位数；Bedrock 在 Q2 的客户支出高于此前所有季度之和。

为什么重要：它把生成式 AI 云需求从合同储备和客户案例推进到可量化收入规模。

影响：企业模型市场正在快速集中到少数可同时提供模型目录、芯片、agent 运行时和全控制的云平台，但集中度上升也会放大迁移风险。

验证状态：收入运行率、增速与 Bedrock 支出均由亚马逊 7 月 30 日官方业绩材料验证，未披露精确收入额与客户集中度。

2 .

信号标题：GitHub Models 于 7 月 30 日全面停止服务

类型：平台访问与产品退役

来源或链接：GitHub 官方变更公告 (<https://github.blog/changelog/github-models-is-being-fully-retired-on-july-30-2026>)

核心观点：GitHub Models 的 playground、模型目录、推理 API 与 BYOC 用户关闭；新增变化不是限制新注册，而是现有活跃用户也失去访问。

为什么重要：开发者曾可直接在代码平台比较和调用模型，如今被引导到 Microsoft Foundry 或 GitHub Copilot 两个更明确的商业入口。

影响：依赖 GitHub Models API 的团队必须切换端点、重新验证模型行为与成本，并避免把实验入口当成长期生产依赖。

验证状态：退役范围与生效日由 GitHub 官方公告验证。

3 .

信号标题：AWS 同日停止向新客户开放第一代知识助手与 agent 服务

类型：产品整合与访问变化

来源或链接：AWS 服务可用性更新 (<https://aws.amazon.com/about-aws>)

26/06/aws-service-availability/) Amazon Q Business .com/q/business/)

核心观点：Amazon Q Business、Amazon Kendra 和更名为 Bedrock 旧版 agent 服务自 7 月 30 日起不再向新客户开放；Q Business 用户被引导迁移到 Amazon Quick。

为什么重要：云厂商正在主动收拢早期生成式 AI 产品线，把知识问答、跨应用操作和持续运行的 agent 合并到更少入口。

影响：现有客户短期仍可使用旧服务，但新增功能、人才投入和生态集成更可能流向 Quick 与新版 AgentCore；采购方应提前验证索引、权限和工作流的迁移路径。

验证状态：生效日、受影响服务与推荐迁移方向均由 AWS 官方页面验证。

4 .

信号标题：Zoox 获得美国首个专用无人工控制 robotaxi 收费通道

类型：监管准入变化

来源或链接：AP 报道 (<https://apnews.com/article/1bdb3bb8e0ce>) 联邦申请记录 (<https://www.govinfo.gov/app/details/30>)

核心观点：7 月 30 日的新事实是 Zoox 从免费示范服务跨到可收费运营，而非仅获得上路测试许可；实际收费仍须取德州与地方批准。

为什么重要：专为无人驾驶设计、没有传统人工控制装置的车辆首次获得明确商业化路径。

影响：监管竞争将从“谁能测试”转向“谁能满足多层许可、持续报告和真实道路安全要求后规模化收费”。

验证状态：商业收费准入由 AP 当日报道并由亚马逊业绩材料确认；具体城市上线时间仍待地方许可。

5 .

信号标题：Zendesk 员工服务 AI agent 开始分批开放且早期免费

类型：企业软件访问与定价

来源或链接：Zendesk 官方支持公告 (<https://support.zendesk.com/es/10957525099290-Announcing-the-Employee-Service>)

核心观点：Zendesk 从 7 月 30 日开始向早期访问用户推出 Employee Services，启用后企业帮助中心的传统搜索栏会转为连接内部知识的对话入口；早期访问阶段不收费。

为什么重要：这不是会展预告，而是客服平台将 AI 从可选插件变为员工帮助中心默认交互的具体上线动作。

影响：免费期降低试用门槛，但企业应在正式定价前测量解决率、转人工率、权限泄露和知识更新延迟，避免只凭对话量决定续费。

验证状态：开放时间、界面变化和早期免费政策均由 Zendesk 官方公告验证；正式价格尚未公布。

前沿研究速递

Shadow evaluations：让原论文作者评审 AI 的开放式研究

做了什么：研究 (<https://arxiv.org/abs/2607.27191>) 把两篇未公开投稿的核心问题交给前沿 agent，每次给予六天和数千美元算力，再由原作者按论文标准评审结果。

新在哪里：它不只测可自动判分的小任务，也不依赖随机性较高的匿名同行评审。agent 独立完成了全部工程工作，却未在两项研究问题上取得实质进展，两份成果都被原作者明确拒绝；第二种模型与执行系统复现了类似失败。

潜在应用方向：AI 研发自动化评估、研究 agent 采购、实验室人机分工，以及对“自动化科学进展”预测的校准。

一句话判断：当前 agent 已能承担大量研究工程，但选题判断、失败回退和识别可发表贡献仍是明显瓶颈；样本只有两项，结论应谨慎外推。

APEX - Accounting：用真实会计系统而非问答题测专业工作

做了什么：APEX - Accounting (<https://arxiv.org/abs/2607.27179>) 编写的任务，覆盖十个包含会计系统、表格和 PDF 的工作环境，要求模型完成对账、计提、记账和报告。

新在哪里：九个前沿模型中，最高 Mean Criteria@3 为 56.4%；没有模型的 Pass@2 只有 2.6%，最高 Pass@8 也只有 21.5%。把单次预算从 1 美元提高到 50 美元总体提升，但同一预算条件下，花费更多 token 的任务反而得分更低。

潜在应用方向：财务 agent 验收、共享服务中心自动化、审计抽样、企业软件中的任务级权限与复核设计。

一句话判断：专业 agent 的核心差距不是会不会解释会计，而是能否在多文件、多系统环境中连续八次可靠完成整项工作；该基准的私有测试集限制了外部复核。

AI teammate 实验：发言更多不等于团队贡献更高

做了什么：随机对照研究 (<https://arxiv.org/abs/2607.27179>) 比较“加 AI”团队与 17 个三人团队完成高风险道德决策任务时的交流模式。

新在哪里：AI 在每个实验团队中都是发言最多、内部连贯性最高的成员，但新增信息最少、信息密度最低；有人机队友的组中，人类彼此回应和社会影响更弱，并报告更低的归属感与地位感。

潜在应用方向：会议助手、协作式决策产品、AI 发言节流、团队体验指标与人机协作界面

设计。

一句话判断：企业不能只优化 AI 的参与率，还要测它是否挤压人类之间的有效交流；样本量小且任务单一，仍需在真实组织和长期场景复现。