

OpenAI自研芯片首秀基准、中国开源模型成西方底座、Meta 167亿美元和解

头部企业动态

• OpenAI 8月25日公布与博通合作的首款自研推理芯片 Jalapeño 首批基准：峰值吞吐下每瓦 AI 工作量较在售最优系统高1.5–1.9倍、端到端延迟低1.7–3.6倍，测试模型为 GPT-OSS 120B、DeepSeek R1、Kimi K2.5，2026年底小批量部署、2027年放量。 **B+ • 91% 多源一致**

• Meta 于8月26日在奥克兰联邦庭审中与29州达成和解，最高支付166.8亿美元（含或有部分口径约171亿、分10年支付），并须在全美为青少年账号上每日时长上限、夜间时段限制、强化年龄核验与家长控制。

B+ • 75% 多源一致 • 链接探活失败

• 路透8月26日独家：月之暗面正与微软、亚马逊、谷歌洽谈由三大云托管 Kimi K3，寻求 K3 相关服务收入最高30%的分成；谈判仍处早期，分账方式、数据访问权限与 token 用量审计等关键问题未决。 **B- • 68% 同源引述 • 链接探活失败**

• 亚马逊8月25–26日宣布9月30日关停运营21年的众包平台 Mechanical Turk，同日一并下线 SageMaker Ground Truth 与 Augmented AI，峰值曾超50万名标注工作者，等同退出人工数据采集基础设施业务。

B+ • 73% 多源一致 • 链接探活失败

模型发布

• Anthropic 8月25日打通 Claude 记忆：普通对话与 Cowork 云端任务共享同一份记忆，内容以文件/主题列表可查可删；健康、种族、宗教、政治等敏感话题默认不入库，需用户手动开关，Free/Pro/Max 全量覆盖 Web/桌面/移动端。

B+ • 92% 多源一致

• 法律 AI 独角兽 Harvey（OpenAI 投资）8月21日发布首个自研模型 Tenet，底座弃用 GPT，改为在月之暗面开源权重模型 Kimi K3 上做法律领域后训练（与 Fireworks 合作），面向长周期法律工作。 **B+ • 90% 多源一致**

• 商汤8月21日正式开源轻量原生统一多模态模型 SenseNova U1.5 Lite（8B），单卡可部署，原生支持4K图像直出、精准局部编辑与超长指令跟随，权重与代码同步上架 GitHub、Hugging Face 与魔搭社区（8月3日先出 Preview 版）。

B- • 81% 同源引述

• IBM 发布 Granite 4.2 系列开源模型（3B/8B/30B 三档），128K 上下文窗口、Apache 2.0 许可，主打本地化企业推理与智能体场景。 **B • 72% 待复核**

热点与趋势

• 中国开源权重模型正快速变成西方产品的底座层：Harvey 用 Kimi K3 做法律专用模型，同期月之暗面与美三大云谈 K3 托管分成，开源—闭源的商业分工、成本优势与地缘合规风险被同时推到台前。 **B+ • 88% 多源一致**

• 自研推理芯片开始动摇英伟达的定价叙事：Jalapeño 基准由 SemiAnalysis 的公开 InferenceX 出具并现场复核部分结果；分析指出更公平的对手是同用 HBM4 的 Vera Rubin，Jalapeño 每兆瓦输出 token 更高，但每 token 总成本大致相当。 **B+ • 88% 多源一致**

• 8月26日阿拉巴马州总检察长就 Hugging Face 数据泄露事件向 OpenAI 发出传票，事件由一起平台安全事故升级为州级调查，成为本轮 AI 训练数据合规执法的新样本。 **B • 74% 待复核**

• Anthropic 8月25日在官网发布《资助更好地评估 AI 对福祉的影响》，是其8月20日以来唯一一条官方新闻，延续把外部评估、安全治理做成公共品的路线（此前8月还上线了全球输出水印以对齐欧盟 AI 法案第50(2)条）。 **B • 76% 待复核**

主动剔除(未核实,不外发)

• Anthropic 潜在 IPO 估值约 2 万亿美元、可寻址市场超 30 万亿美元——仅见于聚合稿转述《华尔街日报》，未打开原文、无公司一手确认，属未证实估值说法。

- DeepSeek 以约 740 亿美元估值募资约 70-80 亿美元、并筹备 2027 年科创板 IPO——彭博等多家均为『知情人士』口径且时间跨度从7月延续至今，无公司公告或监管文件佐证。
- SSI (Ilya 团队) 首个 TTT 架构模型『本周发布』、英伟达为其下一代算力投资 50 亿美元——仅见于中文自媒体日报单源，无任何一手或权威媒体印证。
- 软银计划发行 100-200 亿美元债券为其对 OpenAI 约 650 亿美元投资再融资——仅见于聚合稿转述，未核到一手发行文件或彭博/路透原文。
- 英伟达『8月推出面向智能体的次世代旗舰机柜 Vera Rubin、真实 Agent 编码测试每兆瓦吞吐较前代提升最高 30 倍』——核查发现 Vera Rubin 实为 2026年1月 CES 发布，官方口径为在 DeepSeek-R1 上较 GB NVL72 约 10 倍，时点与倍数均对不上，判为串稿。
- Harvey Tenet 在其内部基准 LAB 上全通过率从约 11% 升至 19.7%、用约 150 张 B300 训练两个月、推理成本不到闭源前沿模型四分之一——这些数字仅出现在搜索摘要中，SCMP 全文未见，未获一手确认。
- 微软与沙特 HUMAIN 将 ALLAM 阿拉伯语模型接入 Foundry 与 Copilot——仅见于单一地区媒体转述且明确标注为『计划中能力、上线日期待定』。

信息源清单(可审计)

条目	等级	可信度	探活	来源
OpenAI 8月25日公布与博通合作的首款自研	B+	91%	✓活	techcrunch.com/2026/08/25/openais-jalapeno-chi...
Meta 于8月26日在奥克兰联邦庭审中与29州	B+	75%	✗死	cnbc.com/2026/08/26/meta-social-media-trial-se...
路透8月26日独家：月之暗面正与微软、亚马逊、谷	B-	68%	✗死	marketscreener.com/news/moonshot-ai-negotiatin...
亚马逊8月25-26日宣布9月30日关停运营21	B+	73%	✗死	cnbc.com/2026/08/25/amazon-service-that-jeff-b...
Anthropic 8月25日打通 Claude	B+	92%	✓活	engadget.com/2243753/claude-memory-now-works-a...
法律 AI 独角兽 Harvey (OpenAI	B+	90%	✓活	scmp.com/tech/tech-trends/article/3364827/open...
商汤8月21日正式开源轻量原生统一多模态模型 S	B-	81%	✓活	tech.ifeng.com/c/8vlpDDeeY7F...
IBM 发布 Granite 4.2 系列开源模	B	72%	✓活	techstartups.com/2026/08/26/top-tech-news-toda...
中国开源权重模型正快速变成西方产品的底座层：Ha	B+	88%	✓活	scmp.com/tech/tech-trends/article/3364827/open...
自研推理芯片开始动摇英伟达的定价叙事：Jalap	B+	88%	✓活	the-decoder.com/openais-first-custom-chip-jala...
8月26日阿拉巴马州总检察长就 Hugging	B	74%	✓活	theverge.com/ai-artificial-intelligence/984239...
Anthropic 8月25日在官网发布《资助更	B	76%	✓活	anthropic.com/news...

评分:A 官方一手抓原文 / B+ ≥2独立源 / B 单家权威或未复核 / B- 同源引述 / C 二手(不采)。本期 WebSearch=17 WebFetch=11 成本≈\$1.5700009999999998。自动生成,源已分级核验。