由 Codex 自动生成

AI前沿发展日报 | 2026-08-15（Asia/Shanghai）

日期：2026-08-15；覆盖窗口：2026-08-13 22:00 至 2026-08-15 07:15（Asia/Shanghai），以窗口内首次公开或出现实质新增的信息为主，所有核心事实与链接均在截稿前在线复核。

今日要点

- Google 发布 Gemini 3.7 Flash，并把它直接铺到 API、企业 agent 平台和面向个人的 Gemini Spark；限时输入、输出价格分别为每百万 token 0.75 美元和 3.75 美元。工作马模型的竞争已从单次能力展示转向“任务成功率 × 单次成本 × 分发入口”。
- DeepSeek-V4-Pro 正式版上线 App、Web 与 API，同时宣布 8 月 17 日零时（北京时间）启用峰谷价。低成本模型不再等于稳定低价，批量 agent 的执行时点和缓存结构将直接影响毛利。
- Databricks 完成 50 亿美元融资、估值 1900 亿美元，并披露年化收入超过 70 亿美元、同比增长逾 80%。资本正在集中押注能同时提供企业数据、agent 运行底座和成本控制的基础软件层。
- Anthropic 披露内部“Model 2”明显强于现有模型、但尚无外部发布计划，并将高风险场景下失配风险的总体判断从“极低”上调为“低”。新增变量不是又一次安全表态，而是既有任务型评测开始跟不上内部能力提升。
- 美国政府正在考虑把达到前沿能力的开放权重模型纳入发布前安全审查。开放与闭源的监管分界可能转向能力阈值，企业的模型来源、权重托管和政府采购资格都可能受到影响。

今日结论

1. 通用 agent 的价格战已进入“同周反向定价”阶段：Google 用半价推广高频工作马模型，DeepSeek 则从统一低价切换到峰谷价。企业预算不能再按静态 token 单价编制，应以真实任务成功成本、缓存命中率和执行时段做敏感性分析。
2. 前沿实验室的内部模型开始领先于公开产品，也领先于现有评测工具。采购方不能把“未发布”解读为能力不存在；高权限 agent 应按潜在下一代能力设计隔离、最小权限和异常中止，而不是等公开模型上线后再补控制。
3. 开放权重模型的政策优势正在收窄。若美国最终按能力而非许可方式决定审查范围，开放模型仍保有部署自主性，但发布节奏、国安测试和政府采购会增加新的合规成本。

AI 产品与应用

Gemini 3.7 Flash 同时进入开发、企业与个人 agent 入口

Google 于 8 月 13 日发布 Gemini 3.7 Flash。官方称其在 FrontierCode 1.1 Main 上由 3.6 Flash 的 34.4% 提高到 43.6%，DeepSWE v1.1 由 49.0% 提高到 65.3%，AutomationBench 由 17.0% 提高到 30.4%；这些均为厂商公布的评测结果，仍需独立复现。模型现已进入 Gemini API、Google AI Studio、Android Studio、Antigravity、Gemini Enterprise Agent Platform 和企业应用，并用于 160 多个国家和地区的 Gemini Spark。Google 官方发布 (https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/)

产品意义在于一次发布贯通了构建、采购和终端使用三层入口。对企业而言，试用不应只看代码榜单，而要在相同 Workspace 权限、相同工具链和相同错误重试规则下，比较“完成一个可验收流程”的总 token、延迟、人工复核和失败恢复成本。

模型与技术进展

DeepSeek-V4-Pro 正式版把 agent 能力与兼容性放在同一更新里

DeepSeek 的 8 月 13 日更新将 V4-Pro 正式版推送至 App、Web 与 API，既有调用只需继续使用 deepseek-v4-pro。官方同时加入 OpenAI Responses API 原生支持、Codex 一键配置，以及 low、high、max 三档思考强度。厂商公布的成绩包括 Terminal Bench 2.1 为 87.9、DeepSWE 为 62.7、HLE 在无工具和有工具条件下分别为 42.7 和 60.0；这些数字来自 DeepSeek 自测，不能替代生产任务回归。DeepSeek 官方更新日志 (https://api-docs.deepseek.com/updates#deepseek-v4-pro-update)；官方模型与价格页 (https://api-docs.deepseek.com/quick_start/pricing)

技术上最值得跟踪的不是单一榜单，而是接口兼容与推理强度控制。它降低了把同一 agent 运行时迁移到不同模型的工程成本，也让路由器可以按任务难度调整推理预算。风险同样明确：服务端模型在端点不变时更新，生产团队必须保存版本响应、固定回归集和回滚路径，避免“代码没变、行为已变”。

投融资与商业动态

Databricks 以 1900 亿美元估值融资 50 亿美元

Databricks 宣布完成 50 亿美元战略融资，估值 1900 亿美元，由 Coatue 领投，Blackstone、MGX、T. Rowe Price 相关账户和 Sixth Street Growth 等参与。公司同时披露年化收入超过 70 亿美元、同比增长逾 80%、过去 12 个月调整后自由现金流为正；Lakebase 年化收入已超过 1

亿美元。Databricks 官方公告 (https://www.databricks.com/company/newsroom/press-releases/databricks-grows-80-yoy-surpasses-7b-revenue-run-rate-scales)；TechCrunch 访谈与融资核验 (https://techcrunch.com/2026/08/13/databricks-wanted-to-raise-1b-investors-wanted-15b-it-settled-on-5b-at-a-190b-valuation/)

关键信号不是估值纪录，而是资本把企业 agent 所需的数据层、实时数据库、业务上下文和多模型成本控制视为同一平台机会。需要谨慎的是，收入、增长率与现金流均为公司披露，且私募估值不能替代公开市场定价。后续应跟踪净收入留存、云承诺消耗、AI 产品收入占比，以及 Lakebase 从试用扩张到长期生产负载的速度。

政策、伦理与安全

Anthropic 上调失配风险判断，并承认任务型评测的解释力下降

Anthropic 最新风险报告披露，内部探索模型“Model 2”在多项内部任务上有明显提升，已与 Mythos 5 一起大量用于编码、agent 工作和数据生成，但公司目前无对外发布计划。报告还将高风险场景下失配的总体风险判断从“极低”上调至“低”，并表示对 Model 2 的判断信心低于过去，因为最具体的任务型评测已不能充分捕捉能力增长。Axios 对当日报告的核验 (https://www.axios.com/2026/08/14/anthropic-model-2-ai-risk)；Anthropic 现行 Responsible Scaling Policy (https://www.anthropic.com/responsible-scaling-policy)

这对企业安全的现实含义是：基准未显著恶化，不代表高权限 agent 的真实边界没有变化。控制重点应从“模型是否通过某套题”转向运行时证据，包括权限调用、跨系统跳转、异常持续性、网络访问和人工中止是否真正有效。

美国或把前沿开放权重模型纳入发布前审查

Axios 8 月 14 日报道，美国政府正在考虑，当开放模型达到 Anthropic 与 OpenAI 顶级模型的能力水平时，将其纳入政府审查。此前 8 月初形成的非公开安排把覆盖对象限定为具有先进网络能力和国家安全风险的闭源前沿模型，并设置最长 30 天的自愿发布前评估；6 月发布的国家安全总统备忘录则要求政府加速采用商业或开放模型，并在 10 月前推进相关实施。Axios 当日报道 (https://www.axios.com/2026/08/14/open-source-ai-government-scrutiny)；白宫 NSPM-11 原文 (https://www.whitehouse.gov/presidential-actions/2026/06/national-security-presidential-memorandum-nspm-11/)

截至截稿，扩围尚未形成公开、最终规则，评测门槛和执行方式也未披露。若能力阈值取代开源形式成为主标准，模型发布者需要提前准备权重安全、红队材料和政府沟通；采用方则要区分“可下载”与“可用于受监管或国家安全场景”，避免把开放许可误当成合规通行证。

X 平台高信号

1.

信号标题：IBM 建立 OpenAI 专项业务，把模型销售连接到遗留系统改造

类型：企业部署渠道 / 交付能力

来源或链接：IBM 官方公告 (https://newsroom.ibm.com/2026-08-13-ibm-partners-with-openai-to-accelerate-secure-ai-deployment-for-enterprises-across-core-operations)；TechCrunch 核验 (https://techcrunch.com/2026/08/13/ibm-partners-with-openai-to-bolster-enterprise-ai-push/)

核心观点：IBM 将在 Consulting Advantage 中嵌入 GPT-5.6、Codex 与 ChatGPT Work，建立由数千名顾问和工程师组成的 OpenAI 专项业务，并派出经 OpenAI Partner Network 培训的驻场团队，重点服务金融、政府、电信和零售。

为什么重要：企业扩张的瓶颈正在从模型访问转向旧流程、旧应用和受监管环境的实施能力；大型系统集成商可以把一次性试点转成跨部门项目。

影响：企业采购应要求联合交付团队对上线周期、人工复核、模型风险与业务结果负责，并明确 IBM、OpenAI 与客户三方的数据和事故责任。

验证状态：合作范围与人员规模由 IBM 官方确认，TechCrunch 采访补充称培训主要来自现有员工转岗；合同金额、客户名单和收入分成未披露。

2.

信号标题：Microsoft 合并个人与工作 Copilot 入口，并设定功能退役日

类型：平台规则 / 功能访问变化

来源或链接：Microsoft 官方支持文档 (https://support.microsoft.com/en-us/microsoft-365-copilot/learning/changes-microsoft-copilot-app)；TechCrunch 核验 (https://techcrunch.com/2026/08/13/microsoft-kills-off-unsuccessful-ai-features-while-merging-its-separate-copilot-apps/)

核心观点：Microsoft 将消费版 Copilot 与 Microsoft 365 Copilot 的功能并入统一应用；Group Chats、Podcasts 和面向消费者的 Deep Research 将从 8 月 18 日起退役。个人与工作账户仍保持数据隔离，专业用户可使用 Researcher 替代深度研究功能。

为什么重要：大厂正在从功能堆叠转向单一入口和明确的付费分层，未形成稳定使用习惯的实验功能会被更快撤回。

影响：用户应在截止日前下载群聊共享媒体、播客和需要保留的研究内容；企业管理员则需复核迁移后的账户切换、数据边界与保存规则。

验证状态：功能退役日期、迁移范围和账户隔离规则均由 Microsoft 官方支持文档确认；TechCrunch 提供产品策略背景。

3.

信号标题：Google 允许关闭生成媒体的可见水印，但保留机器可读来源信息

类型：平台规则 / 内容溯源

来源或链接：Google 产品负责人在 X 的公告 (https://x.com/joshwoodward/status/2088259242423968162)；Google Credentio 官方说明 (https://developers.googleblog.com/introducing-credentio-open-source-c-library-for-c2pa-content-credentials-from-google/)；TechCrunch 当日报道 (https://techcrunch.com/2026/08/14/google-will-now-allow-users-to-remove-visible-watermark-from-its-ai-generations/)

核心观点：Google 将在 Gemini 与 Flow 中提供 Media Watermark 开关，允许 Nano Banana、Omni 和 Lyria 的用户移除可见标记；不可见 SynthID 和 C2PA 元数据继续保留。同期开放的 Credentio 可在本地验证 C2PA Content Credentials。

为什么重要：专业创作获得更干净的成品，同时平台把识别能力从画面上的标记转移到机器可读凭证和本地验证工具。

影响：品牌和媒体团队可以关闭可见标记，但仍应在发布链中保留 C2PA 元数据，并测试转码、裁剪和多平台分发后凭证是否仍可读取。

验证状态：开关范围由 Google 负责人一手公告并由 TechCrunch 核验；Credentio 的能力和采用规模来自 Google 官方开发者博客，独立性能测试尚未公开。

4.

信号标题：Writer 把 agent 成本优化从换模型转向调用编排

类型：成本变化 / 企业 agent

来源或链接：Writer 研究论文 (https://arxiv.org/abs/2607.06906)；TechCrunch 发布核验 (https://techcrunch.com/2026/08/13/writer-introduces-new-ai-model-and-upgraded-harness-to-contain-token-costs/)

核心观点：Writer 发布 Palmyra X6 和新版 agent 运行层。其受控测试在 22 个任务、6 个模型上仅替换调用编排，平均每任务成本下降 41%，token 用量下降 38%，中位耗时下降 44%；公司称新组合可让部分基础任务成本最多降低 50%。

为什么重要：agent 的账单由多轮历史、工具说明、检索内容和失败重试共同决定，单看模型单价会错过更大的优化空间。

影响：企业应以“每个成功任务的总成本”比较方案，并把缓存命中、上下文回放、工具暴露和重试消耗纳入可观测指标。

验证状态：受控测试数字来自 Writer 作者团队的论文，产品降本上限来自公司向 TechCrunch 的表述；样本只有 22 个任务，外部复现实验仍不足。

5.

信号标题：Kog 称标准数据中心 GPU 的单请求推理仍有数量级优化空间

类型：推理效率 / 需求信号

来源或链接：TechCrunch 当日报道 (https://techcrunch.com/2026/08/14/kog-is-going-deeper-to-squeeze-more-inference-out-of-gpus/)；开源 Laneformer 2B 模型 (https://HuggingFace.co/kogai/laneformer-2b-it)

核心观点：Kog 的小模型技术预览在 AMD MI300X 与 NVIDIA H200 上展示约每秒 3000 个单请求 token，并已开放 Laneformer 2B；公司称预览带来约 200 条明确商业线索，现正转向加速更大的主流模型。

为什么重要：如果软件能在企业已有 GPU 上持续提高单请求速度，低延迟推理就不必完全依赖新型专用芯片，现有硬件的经济寿命也会延长。

影响：采购方应要求在目标大模型、真实上下文长度和并发负载下复测吞吐与质量，不能把 2B 模型演示直接换算为生产收益。

验证状态：模型开放可由 Hugging Face 页面核验；性能、商业线索和更大模型的提速目标来自公司对 TechCrunch 的表述，尚无独立基准确认。

前沿研究速递

AutoDesign：让 agent 自己优化长流程设计工具链

做了什么：论文 (https://arxiv.org/abs/2608.13560) 以学术论文转海报为任务，让元优化器根据 rollout 反馈递归改进设计工具链，并建立覆盖 100 篇论文、五个学科的 PosterBench。

新在哪里：加入学得的 DesignHarness 后，七种 agent—模型组合的平均分从 54.99 提高到 67.39；完整自主运行在 40 分钟内执行 253 次工具调用和 11 轮编辑，成本低于 3 美元。结果仍集中于海报任务，且对商业系统的对比需独立复核。

潜在应用方向：品牌物料生成、报告排版、演示文稿制作、设计 QA，以及任何需要多轮生成—检查—修订的内容流程。

一句话判断：下一阶段的设计 agent 优势可能不只来自更强模型，而来自能用实际反馈持续改进的工具链。

OmniScientist：让 AI 科学家直接读取多模态原始证据

做了什么：论文 (https://arxiv.org/abs/2608.13558) 让感知层与构思、实验、写作三个 agent 直接处理图像、信号、音频、视频、三维结构、轨迹、表格和公式，并用代码检查新颖性、统计有效性和数值溯源。

新在哪里：系统在五类学科的 36 个真实数据案例中全部完成从原始数据到编译论文的流程；相较只接收预计算标量特征的盲化版本，直接感知在七项指标上均提高，并赢得 85% 的成对判断。样本量仍小，分数也依赖选定推理模型。

潜在应用方向：实验数据初筛、跨模态证据整合、研究报告生成、可追溯分析和科研工作流审计。

一句话判断：AI 科学家的瓶颈正从“会不会写论文”转向“能否看到原始证据并保持每个结论可追溯”。

V-RAE：视频生成的潜空间不应只追求像素重建

做了什么：论文 (https://arxiv.org/abs/2608.13556) 在冻结的视觉基础模型表征上构建视频自编码器，用轻量时间池化去除冗余，并引入更关注时间一致性的 tFVD 指标。

新在哪里：其最佳配置在 K600 上达到 2.13 rFVD，在匹配生成设置下最高可快 6 倍收敛；研究还显示，像素重建质量高并不必然意味着更适合下游视频生成。

潜在应用方向：视频生成、交互世界模型、未来帧预测、低成本训练和时序一致性评测。

一句话判断：视频模型的效率突破可能来自重新设计“生成所用的表示”，而不只是扩大生成器本身。