

AI前沿发展日报 | 2026-08-17 (Asia/Shanghai)

日期：2026-08-17；覆盖窗口：2026-08-16 06:45 至 2026-08-17 06:45 (Asia/Shanghai)，以窗口内生效、临近截止或经官方页面出现实质更新的信息为主；周末公开信号偏少，未用旧闻补量。

今日要点

- DeepSeek 的 V4 峰谷价已于北京时间 8 月 17 日零时生效。同一模型、同一任务在不同时段可相差一倍，批量 agent 的执行时间从运维参数变成毛利参数。
- OpenAI 在截稿前更新 ChatGPT 广告说明：Free 与 Go 可能展示广告，Free 用户可用更低消息额度和更少工具换取无广告体验。免费层开始被明确拆成“注意力换算力”和“功能换清静”两种产品。
- Atlassian 新 AI 条款今日生效：一方面为符合条件的输出版权索赔提供有限赔偿，另一方面明确客户要对 agent 动作的适当性与监督负责。生成内容与执行动作的责任被合同化地区分开。
- 今日更密集的信号不是新模型发布，而是入口、模型端点、权限和计费规则集中切换。它们看似是维护事项，实际会直接改变成本、可用性和责任边界。

今日结论

1. AI 服务的商业竞争已进入“动态供给管理”：厂商开始用时段定价、广告补贴和功能限额分配算力。企业不能再用一个静态 token 单价或席位价估算全年成本。
2. agent 的合同责任正在脱离普通生成式 AI。供应商可以对特定输出提供有限知识产权保护，但执行动作仍由客户判断、授权和监督；采购合同与运行权限必须同步审查。
3. 模型基准越来越容易被接口传输和执行路径掩盖。对会发命令的 coding agent，真正可比的是端到端最终状态，而不是模型在理想接口上的单一分数。

AI 产品与应用

ChatGPT 把无广告选择改造成显式的功能交换

OpenAI 在截稿前更新的帮助页明确：广告可能出现在 Free 与 Go 计划，Plus、Pro、Business、Enterprise 和 Edu 保持无广告；Free 用户还可选择免费的 Ads-Free 模式，但代价是更低的消息额度，以及可能失去图像生成、深度研究等工具。关闭个性化并不会去除广告；即使关闭，

当前对话上下文仍可用于选择广告。OpenAI 同时称，不向广告主提供聊天、记忆或个人信息，只提供汇总后的浏览与点击数据。OpenAI 官方说明 (<https://help.openai.com/en/articles/20001047-ads-in-chatgpt>)

产品上的新增变量不是“聊天里有广告”本身，而是算力、工具和注意力被放进同一个交换关系。对面向消费者的 AI 产品，这会提高免费层的可持续性，也会让使用额度、工具可用性和广告负担共同成为转化漏斗。品牌方若进入此渠道，应把“对话相关”与“答案背书”严格分开，避免把赞助位误读为模型推荐。

模型与技术进展

QuoteBench 显示 coding agent 的接口路径可让成绩失真数十个百分点

8 月 13 日提交、在周末研究队列中仍具新增解释力的 QuoteBench，用 56 个一次性任务、14 类真实事故来源的命令问题，比较模型原始 Bash 输出与多经过一层未转义解析器后的最终状态。八个配置中，同一回复经额外解析后成功率下降 55.4—73.2 个百分点；披露这一边界后，六个配置能恢复 30.4—60.7 个百分点，另外两个几乎没有恢复。论文还展示，表面只有 -3.6 分差距的配置，内部可能同时藏着 -64.3 分的传输损伤与 +60.7 分的适配补偿。QuoteBench 论文 (<https://arxiv.org/abs/2608.13547>)

这说明 coding agent 的“模型能力”不能脱离序列化、shell 包装、转义、执行器和最终状态验证来讨论。采购或内部评测至少要固定并披露模型版本、生成约定、命令传输路径、运行环境和验收器；否则换一个 SDK 或 wrapper，就可能比换模型更影响成功率。

投融资与商业动态

DeepSeek 峰谷价正式生效，批量推理获得 2 倍时段价差

DeepSeek 新价格已于 2026 年 8 月 16 日 16:00 UTC、即北京时间 8 月 17 日零时生效。V4-Flash 的缓存命中输入、未命中输入和输出峰值价分别为每百万 token 0.014、0.44 和 1.32 美元；V4-Pro 分别为 0.044、1.32 和 3.96 美元。01:00—04:00 与 06:00—10:00 UTC 为峰时，其余时段半价；换算到上海时间，峰时为 09:00—12:00 与 14:00—18:00。DeepSeek 更新日志 (<https://api-docs.deepseek.com/updates/>)；实时价格页 (https://api-docs.deepseek.com/quick_start/pricing/)

相较 8 月 15 日报告中的“即将启用”，今日新增事实是价格已进入执行状态。对异步评测、批量内容、离线检索和非紧急代码任务，调度到夜间或午间窗口可直接把模型费减半；对客服、交易辅助等实时 workflow，则应按峰值费率做预算，不能把半价当作默认成本。值得继续跟踪的是

，其他低价 API 是否也会用时间差异替代统一折扣。

政策、伦理与安全

Atlassian 新 AI 条款把“输出”与“agent 动作”拆成两套责任

Atlassian AI Terms 于 8 月 17 日生效。条款把 agent 定义为可代表客户执行动作的 AI 产品，并明确这些动作不属于“输出”；客户负责判断动作是否适当并进行适当监督。对于按协议授权使用的 AI 输出，Atlassian 可为直接侵犯第三方版权的索赔提供辩护与赔偿，但排除客户修改、与第三方产品组合、明知可能侵权、第三方产品输出等情形；总责任上限为相关云产品前 12 个月已付金额与 100 万美元两者中的较低者。Atlassian 官方 AI 条款 (<https://www.atlassian.com/legal/ai-terms>)

这份条款的商业意义在于：生成结果的有限版权保护不能覆盖 agent 的实际执行风险。企业应分别登记内容输出、第三方模型或插件、可执行动作与审批人；若采购表只问“是否有 AI indemnity”，很可能漏掉真正会造成业务损失的动作责任。

X 平台高信号

1.

信号标题：Claude 旧版 Workbench 与三项实验性 prompt API 今日停止访问

类型：开发工具退役 / 接口访问变化

来源或链接：Anthropic 官方发布说明 (<https://platform.claude.com/docs/en/release-notes/overview>)

核心观点：旧版 Workbench 的访问在 8 月 17 日结束；新版不承接已保存的 prompts、变量和 evals。generate_prompt、improve_prompt、templatize_prompt 三个实验端点同步退役，后续请求将返回错误。

为什么重要：这不是界面换皮，而是保存资产与自动化端点同时失效；只在控制台里维护提示词的团队可能失去可复现记录。

影响：开发团队应把关键提示、变量和评测导出到版本库，并扫描生产代码是否仍调用三个实验端点。

验证状态：退役日期、数据不迁移范围和端点行为均由 Anthropic 官方发布说明确认。

2.

信号标题：Groq 免费与开发者层停止两款 Llama 模型，企业承诺消费客户暂不受影响

类型：模型下线 / 访问分层

来源或链接：Groq 官方模型退役页 (<https://console.groq.com/docs/deprecations>)

核心观点：llama-3.1-8b-instant 与 llama-3.3-70b-versatile 已于 8 月 16 日在免费和开发者层下线；官方分别建议迁移到 openai/gpt-oss-20b，以及 openai/gpt-oss-120b 或 qwen/qwen3.6-27b。有承诺消费合同的企业客户不受此次退役影响。

为什么重要：同一端点供应商开始按合同层级维持旧模型，模型生命周期本身成为企业套餐的一部分。

影响：依赖旧模型 ID 的免费与开发者应用会直接报错；迁移时必须重测结构化输出、工具调用、延迟和成本，不能只替换名称。

验证状态：下线范围、日期、例外客户和替代模型均由 Groq 官方文档确认。

3.

信号标题：Microsoft 消费版 Copilot Deep Research 明日退役，替代能力转向 Premium

类型：功能退役 / 付费访问变化

来源或链接：Microsoft 官方支持文档 (<https://support.microsoft.com/en-us/microsoft-copilot/deep-research-in-microsoft-copilot>)

核心观点：消费版 Copilot 的 Deep Research 将从 8 月 18 日起不可再用；Microsoft 365 Premium 用户可改用 Researcher。既有报告不会随此次变化删除，Personal 与 Family 用户仍可在聊天历史中访问，也可保存到 Word。

为什么重要：研究型能力正在从通用消费入口迁入更高价、与办公数据连接更深的计划，而历史内容的保留方式与新建权限被分开处理。

影响：依赖 Copilot 生成定期研究报告的个人用户需要在截止前确认替代计划；企业也应避免把消费版功能写入长期流程承诺。

验证状态：退役日期、替代计划和历史内容处理方式均由 Microsoft 官方支持文档确认。

4.

信号标题：Snowflake 启动 Cortex 模型权限从 allowlist 向角色授权迁移

类型：权限规则 / 企业数据平台

来源或链接：Snowflake 官方行为变更说明 (<https://docs.snowflake.com/en/release-notes/bcr-bundles/un-bundled/bcr-2378>)

核心观点：Snowflake 从 8 月 17 日至 9 月 4 日分批把既有 CORTEX_MODELS_ALLOWLIST 映射到模型 RBAC 应用角色，覆盖 AI_EMBED 与 Cortex Search；官方提醒自动回填是 best-effort，禁用 secondary roles、Native Apps 和受限调用者权限存储过程等执行环境可能无法继承 PUBLIC 授权。

为什么重要：模型可用性开始由数据库角色而非账户级名单控制，安全粒度提高，但隐含权限继承可能在迁移后暴露断点。

影响：数据团队应在非生产账户复测嵌入、搜索和存储过程，并把需要的模型角色显式授予真实执行角色，再移除旧 allowlist。

验证状态：迁移窗口、覆盖产品与已知例外均由 Snowflake 官方文档确认；具体账户何时完成迁移取决于分批进度。

5.

信号标题：Google Ads 今日改变预算受限广告的目标出价行为

类型：平台算法规则 / 广告投放

来源或链接：Google Ads 官方帮助 (<https://support.google.com/google-ads/answer/17061251>)

核心观点：从 8 月 17 日起，处于“Limited by budget”且使用目标 CPA 或目标 ROAS 的广告系列，会更贴近广告主填写的目标，即使预算变化。若历史实际 CPA 显著优于设定目标而不调整，新机制可能主动消耗这部分超额表现。

为什么重要：机器学习投放系统从“尽量优于目标”转向“更严格执行目标”，表面不改预算也可能改变流量、成本和波动。

影响：营销团队应把近期实际 CPA、ROAS 与系统目标重新对齐，并在变更后单独监控预算受限系列，避免把算法切换误判为素材或需求变化。

验证状态：生效日期、适用条件和官方建议均由 Google Ads 帮助页确认；实际幅度取决于各广告系列历史表现。

前沿研究速递

PlayWorld：让 agent 玩家评价世界模型的长程一致性

做了什么：论文 (<https://arxiv.org/abs/2608.13552>) 提出 171 个带目标的交互场景，让多模态 agent 自主操作九个世界模型，并从几何一致性、交互真实性、视野外演化和视野内演化四个维度检查长期表现。

新在哪里：它不再要求所有模型执行固定动作序列，而是让 agent 为相同目标选择各自路径，更贴近用户实际游玩与探索。结果显示，现有模型在空间一致性和状态持续演化上仍不可靠。

潜在应用方向：游戏原型、机器人仿真、交互视频、虚拟训练环境，以及需要跨长时间保持状态的世界模型验收。

一句话判断：世界模型的视频观感已经领先于它保持可交互因果状态的能力，长程目标测试比短片 demo 更接近产品风险。

Vero：首次把代码与机器证明的联合生成推到仓库级

做了什么：论文 (<https://arxiv.org/abs/2608.13522>) 构建 43 个来自真实 Python、Dafny、Verus 与 Coq 项目的多模块 Lean 4 实例，要求 coding agent 同时完成实现与形式化证明，并允许 agent 证明规范不可满足或参考代码错误。

新在哪里：评测对象从单函数证明扩展到跨模块 API、实现选择和证明的一致性。最强配置只完整解决 43 个实例中的 27 个，在最难仓库上没有闭合任何规范。

潜在应用方向：密码协议、分布式系统、关键基础软件、合规代码生成，以及高风险 agent 产出验收。

一句话判断：形式化验证能把“看起来能运行”提升为机器可检查，但当前 agent 距离仓库级可信交付仍有明显距离。

StreamTTT：把长时记忆写入快速权重，给实时视觉保留短上下文

做了什么：论文 (<https://arxiv.org/abs/2608.13416>) 让流式视觉语言模型把长期历史写进在线更新的 fast weights，同时用短滑动 KV cache 专注近期画面，并联合训练离线长视频问答与实时问答数据。

新在哪里：StreamTTT-4B 在 OVO-Bench 上相对 SimpleStream-4B 的实时感知提高 1.4 分、回溯提高 3.7 分，并在 StreamingBench 的 RTVU 子集上接近更大的 8B 对手。代码尚待公开，结论仍需复现。

潜在应用方向：可穿戴助手、机器人持续感知、直播理解、行车视频和需要边看边记的实时 agent。

一句话判断：实时多模态系统不必把全部历史塞回注意力上下文，外置可更新记忆可能是兼顾当前反应与长期回忆的更有效路径。