

AI前沿发展日报 | 2026-10-03 (Asia/Shanghai)

日期：2026-10-03；覆盖窗口：2026-10-01 00:00—2026-10-03 06:56 (Asia/Shanghai)；重点筛选上一期后新增核验的信息及2026-10-02生效事项；海外公告按原文日期标注，窗口外内容单列为近期补充。

今日要点

- 商家建站开始直接生成和修改可运行页面。Shopify Canvas 将 Sidekick 接入全店设计工作区，发布采用逐步开放方式。短期看可用范围，中期看商家从设计到上线所需的人力是否下降。Shopify 公告 (<https://www.shopify.com/news/introducing-canvas>)
- 实时视频交互出现值得复测的研究信号。Tavus Griffin-Lite 在厂商组织的一分钟通话实验中，让54名参与者中的26名认为对方是真人；这是有限实验结果，尚非面向客户的正式产品。Tavus 研究说明 (<https://www.tavus.io/griffin>)
- 算力投资新增一个具体方向：计算芯片与内存之间的光连接。Volantis 披露8,800万美元融资，计划利用已有激光器供应链改善内存连接距离；商业兑现仍取决于硬件交付。Reuters 采访 / WTVB 刊载 (<https://wtvbam.com/2026/10/01/volantis-raises-88-million-for-tech-to-connect-ai-memory-chips/>)

今日结论

1. 建站服务的竞争应转向上线验收与持续运营。当平台内代理能修改真实页面，服务商更有价值的工作是品牌一致性、交互检查和业务效果验证；仅出售页面初稿的议价空间可能收窄。这是基于 Canvas 能力范围的编辑判断。
2. 数字人项目应把交互自然度和任务完成率分开验收。一分钟内看起来像真人，并不能证明长时间服务可靠。企业试点应记录问题是否解决、打断是否正确处理，以及何时需要人工接手，而不是直接采用厂商的“通过图灵测试”表述。
3. 硬件投资尽调需要单独检验内存供给能力。Volantis 的路线说明，更多计算单元之外，数据搬运也值得投入；现阶段应关注样片、封装、功耗和交付节点，不把目标内存规模当作已经形成的收入或性能优势。

AI 产品与应用

Shopify Canvas：从聊天建议到可交互的全店设计

公告日期：2026-10-01。Canvas 将商店各页面放在同一可平移、缩放的工作区。商家既能直接点击修改，也能要求 Sidekick 编辑；预览渲染真实代码，可检查交互、动画，以及不同商品和屏幕尺寸下的效果。官方发布 (<https://www.shopify.com/news/introducing-canvas>)

为什么重要：代理生成的结果进入商家原有的经营平台，设计反馈可以直接落到页面修改，减少在独立生成工具与商店后台之间搬运内容的步骤。

商业影响与边界：适合先用新店或独立页面检验制作周期和返工量。官方明确表示仍在逐步开放、存在功能缺口，也尚未替代现有编辑器。其“二十分钟建店”属于产品负责人举例，不能当作普遍交付承诺。发布范围说明 (<https://www.shopify.com/news/introducing-canvas>)

模型与技术进展

Griffin-Lite：让感知、接话和视频生成同时进行

研究发布：2026-10-01。Tavus 将 Griffin-Lite 描述为全双工视频到视频模型：接收人的声音和视觉行为，同时决定何时回应并生成语音、视频。它试图改善“等用户说完，再开始处理”的交互节奏。技术说明 (<https://www.tavus.io/griffin>)

新增证据：在一分钟视频通话实验中，26 / 54 名 Griffin-Lite 参与者认为对方是真人；前一套系统为 1 / 41。参与者事前被告知将与另一名参与者通话，实验结束后才获知对方是 AI。两组人数不同，结果来自厂商研究，不能推定所有人群和通话时长都成立。

商业影响与边界：培训陪练、可视化客服值得跟踪，但当前只向受信任测试者提供研究预览，未面向客户开放，也未进入 Tavus 正式平台。拟人度是交互指标，不能替代业务准确率或客户采购可用性。评测方法与开放限制 (<https://www.tavus.io/griffin>)

投融资与商业动态

Volantis 获8,800万美元：用光连接拓宽内存布局

Reuters 报道日期：2026-10-01。Volantis 表示已筹集8,800万美元，由 Lachy Groom 与 Abstract Ventures 领投。公司拟用垂直腔面发射激光器 (VCSEL) 连接计算与内存芯片，以降低电连接距离带来的布局约束。Reuters 原始采访 / WTVB 刊载 (<https://wtvbam.com/2026/10/01/volantis-raises-88-million-for-tech-to-connect-ai-memory-chips/>)

为什么重要：VCSEL 已用于大量消费设备，公司试图复用已有器件供应链解决 AI 硬件问题，而不是让所有关键组件都从零开始量产。

商业影响与边界：公司提出让单个 GPU 周围容纳220颗内存芯片的目标，并希望在2027年交付芯片。以上属于路线与时间计划，尚不能写成已量产能力；报道也未提供足以独立验证的性能、良率与客户收入数据。相较上一期的供应商融资，本条新增变量是获得资本支持的物理连接路线。

政策、伦理与安全

加州向 OpenAI 送达调查传票，网络事件调查增加正式取证动作

官方公告日期：2026-10-01；送达日期按公告为2026-09-30。加州司法部表示，已在既有调查中向 OpenAI 送达调查传票，询问涉及公司及其模型的网络安全事件与风险；此前已就 Hugging Face 事件开展正式调查。加州司法部公告 (<https://www.oag.ca.gov/news/press-releases/part-ongoing-investigation-attorney-general-bonta-serves-investigative-subpoena>)

为什么重要：新事实是州执法机关采取了具体调查措施，且关注范围涵盖模型测试、开发及投入服务阶段。它不同于上一期关于模型权重访问记录的国会议员问询。

商业影响与边界：对模型企业及部署方，测试环境权限、事件记录和供应商配合调查能力将成为更具体的尽调问题。这是编辑推断；公告并未公布违法认定、处罚决定或传票全文，不能将调查写成定罪。

X 平台高信号

当期可直接核验的 X 原帖信号偏少。本栏保留一条已定位的官方 X 线索，并以一手文档核实事实；其余四条明确采用官方渠道补充，不冒充已验证的 X 发言。各条影响均为编辑分析。

1. Perplexity：判断型调用有了独立价格

信号标题：Decisions API 每百万输入 token 收费0.04美元，输出免费

类型：官方 X 线索 / API 定价与准入

来源或链接：开发者账号原帖入口 (<https://x.com/perplexitydevs/status/2105725598882832414>)
；官方 API 文档 (<https://docs.perplexity.ai/docs/decisions/quickstart>)

核心观点：2026-10-01发布线索所对应的 API 已有正式文档：任何 Perplexity API 密钥均可调用，每个组织每秒10次请求；返回选择、是非或评分的概率，而非生成文本。

为什么重要：分类、分流等高频任务可以单独核算调用成本。

影响：团队可先用真实工单校准阈值，再决定哪些任务交给更昂贵的生成模型；概率数值本身并不保证业务判断正确。

验证状态：价格、准入和限流已用一手文档核验；X 正文重试后仍不可读取，发布时间线索来自链接该原帖的报道，未引用原帖措辞，也未实测账单。

2. GitHub Copilot：模型退役日期已经到达

信号标题：四个模型的计划退役日为2026-10-02

类型：官方渠道补充 / 访问范围变更

来源或链接：GitHub 模型退役公告 (<https://github.blog/changelog/2026-09-03-upcoming-deprecation-of-selected-github-copilot-models/>)

核心观点：2026-09-03预告在本覆盖期到达执行日期：Gemini 3.5 Flash、Gemini 3.6 Flash、Kimi K2.7 Code、Claude Opus 4.7列入退役，推荐替代分别为 Gemini 3.8 Flash、Kimi K3、Claude Opus 5。

为什么重要：员工可选模型取决于平台支持及组织策略，不能只看模型厂商是否仍提供服务。

影响：企业管理员应检查替代模型是否获准使用，并对关键代码流程做回归抽样；日期到达是本期新增变量，并非重新发布旧预告。

验证状态：官方时间表已核验，未登录客户租户逐一确认实际下线状态；对应 X 原帖未核验。

3. arXiv：月度投稿额度按提交人计算

信号标题：每个自然月最多提交两篇，同时最多三篇处于处理中

类型：官方渠道补充 / 科研平台规则变更

来源或链接：arXiv 2026-10-01政策公告 (<https://blog.arxiv.org/2026/10/01/updated-rate-limit-policy/>)

核心观点：月度上限跨所有学科计算，被拒稿件也计入；由别人提交的合著论文不占用本人额度。三个活跃投稿上限为既有规则，本次新增的是统一月度限制。

为什么重要：自动化写作增加的论文供给开始直接影响研究成果的发布排期。

影响：实验室需协调提交责任人与会议截止日；该规则不是每位作者只能署名两篇，也不是禁止 AI 辅助研究。

验证状态：政策正文及问答已直接核验；对应 X 原帖未核验。

4. Cloudflare：KV 可限定持久存储区域

信号标题：Workers KV 区域限定正式可用，但不限制所有缓存位置

类型：官方渠道补充 / 数据部署条件变更

来源或链接：Cloudflare 2026-10-02变更记录 (<https://developers.cloudflare.com/changelog/post/2026-10-02-kv-jurisdictions-ga/>)

核心观点：创建命名空间时可选 eu、us 或 fedramp，创建后不能追加或修改；数据仍可能在该区域之外缓存，也可从全球访问。

为什么重要：使用 KV 保存代理状态或应用配置的团队，多了一个存储位置控制选项。

影响：与上一期 AI Search 计费变化不同，本次新增的是存储区域能力；不能据此宣称整条 AI 数据链都被限制在同一区域。

验证状态：官方当期公告已核验；未进行部署实测，对应 X 原帖未核验。

5. Ideogram 4.5：最低图片报价存在调用条件

信号标题：Runware 的0.008美元最低档不适用于无参考图请求

类型：官方渠道补充 / 图像服务价格与接入限制

来源或链接：Runware 供应商文档与价目表 (<https://runware.ai/docs/models/ideogram-4-5>)

核心观点：近期补充，模型页标注发布于2026-09-30。本期核验该渠道四档每图价格为0.008、0.03、0.06、0.20美元；文档同时规定，无参考图时只能选择 low、medium、high。

为什么重要：最低宣传价格与实际生成方式必须一起看，不能直接套入所有素材预算。

影响：纯文字生成按此渠道最低可用档应从0.03美元估算；连续修改还需累加各轮费用。该报价只代表 Runware，不代表所有供应商统一价格。

验证状态：服务提供商一手文档已核验，未实测扣费；此条为窗口外近期补充，对应 X 原帖未核验。

前沿研究速递

Google：将联邦学习计算移入可验证的服务器环境

本期新增：2026-10-02官方技术说明；配套论文为此前预印本。研究说明 (<https://research.google/blog/toward-provably-private-learning-from-federated-data/>)；论文 (<https://arxiv.org/abs/2609.31494>)

做了什么：设备加密上传训练样本，服务器在可信执行环境（TEE）内按预先授权的程序处理；访问策略进入公开透明日志，对外释放匿名化结果与具有差分隐私保护的模型权重。

新在哪里：将一部分原本受设备算力、在线时间限制的计算转移到服务器，同时让外部审查者检查获准执行的代码。官方披露已用于 Gboard 英语及日语的下一词预测模型。

潜在应用方向：对隐私敏感、又受终端训练资源限制的个性化模型；是否适用仍取决于可信硬件、审计能力和具体数据约束。

一句话判断：值得关注的是可核验的数据处理过程；数据加密上传意味着它并非“原始数据永不离开设备”的传统叙述，侧信道等问题也仍需研究。

RAC：科研代理在执行中重新分工，额外协调未必更好

提交日期：2026-10-01；预印本，单随机种子的探索性实验。 Can AI Scientists Coordinate at Runtime? (<https://arxiv.org/abs/2610.00980>)

做了什么：在 Agent Laboratory、EvoScientist 和 ARK 三套科研代理中，引入运行时选人、限定任务及基于产物的验证，并在保持各宿主模型、工具和权限的条件下比较表现。

新在哪里：将通信、运行时选择、任务约定和验证逐项拆开。三个宿主中，运行时选择取得各自最高观察均分；继续加入任务约定和验证，均分反而下降。

潜在应用方向：自动实验、资料分析和研发代理的任务调度；可以据此设计在固定预算内比较“新增协调步骤是否值得”的消融实验。

一句话判断：协调机制也消耗预算；论文提供了值得复测的反例，尚不足以推出验证步骤普遍无用。