AI早报：Agent工作流前移，治理与安全同步升温

中文内部早报 · 5分钟版

今日总览

今天的AI新闻主线集中在智能体工作流、企业部署和治理安全。头条里最重要的几条分别是Claude Code v2.1.196 发布、从任何地点构建——Cursor for iOS 公测版发布、为 Amazon Bedrock 和 Google Cloud 推出的 Claude apps gateway，整体呈现出从模型能力展示转向真实工作流和企业环境落地的趋势。

智能体、编程与工作流：最重要的是从任何地点构建——Cursor for iOS 公测版发布。这类新闻的共同点是，AI 工具正在进入任务拆解、跨系统协作、版本控制和长期执行等环节，不再只是停留在代码补全或单轮问答。 同板块还包括美军用AI选目标却误炸伊朗学校，Anthropic Claude嵌入Palantir系统首日建议约1000目标、OpenAI 报告：绘制欧洲 AI 劳动力机遇版图。

前沿模型与多模态：最重要的是Runway API 推出广告本地化 Recipe。模型能力的亮点更多体现在产品嵌入和场景化调用上，视频编辑、旅行对话、模型聚合入口等消息都指向同一个方向：能力需要被放进具体工作流，读者更应关注实际可用性和成本。

大厂与产业链：最重要的是为 Amazon Bedrock 和 Google Cloud 推出的 Claude apps gateway。企业部署、云上交付、内容合作和基础设施投入仍是产业侧重点；算力、水电、版权和企业采购会继续影响 AI 服务成本与可落地范围。

监管、安全与研究：监管安全侧最重要的是Claude Code 打开 GitHub 仓库即执行隐藏恶意代码，攻击者可获完全控制。今天的信号集中在漏洞修复、治理控制面和数据驻留等主题，说明企业在采用模型时需要同步处理审计、权限和供应商边界。研究评测方面，EverOS：开源Markdown优先智能体记忆运行时，支持混合检索与自进化技能值得放入方法论更新。开发者工具方面，Claude Code v2.1.196 发布可作为近期试用入口。

头条 今日头条

1. Claude Code v2.1.196 发布

Claude Code：GitHub Releases · 06/30 08:00 · 开源项目与开发者工具

新增组织默认模型支持，未选模型时显示“Org default”。聊天会话可读默认名称，文件附件支持 Cmd/Ctrl-click 定位。安全方面，`claude mcp list/get` 不再启动通过 `.claude/settings.json` 自批准的不安全服务器；不受信任工作区显示“Pending approval”。`/code-review` 合并五个清理查找器，token 用量减少约 25%。终端 UI 跳过空子树遍历减少渲染。流式空闲看门狗默认开启，5 分钟无事件自动中止重试。修复背景对话误删、远程会话自动恢复、MCP OAuth 作用域冲突、Agents 侧边栏焦点丢失等多项问题。

2. 从任何地点构建——Cursor for iOS 公测版发布

Cursor Blog · 06/30 08:00 · 智能体、编程与工作流

Cursor 推出 iOS 原生公测版，所有付费计划可用。开发者可在手机上启动始终在线的云端智能体，或远程操控电脑端智能体。支持语音输入、斜杠命令和选择前沿模型。智能体运行后，锁屏 Live Activities 和推送通知实时更新状态，完成或需要输入时提醒。云端智能体在隔离虚拟机中运行，可自动迭代生成合并就绪的 PR，并输出演示、截图和日志。本地与云端智能体支持双向切换。移动端 Composer 2.5 享受 75% 折扣，优惠至 2026 年 7 月 5 日。

3. 为 Amazon Bedrock 和 Google Cloud 推出的 Claude apps gateway

Claude 博客 · 06/30 08:00 · 大厂与产业链

Anthropic 今日推出 Claude apps gateway，一个自托管控制平面，让企业能在 Amazon Bedrock 和 Google Cloud 上运行 Claude Code。它作为单个无状态容器部署于 Linux，后端使用 PostgreSQL，提供企业级 SSO 登录（通过 OIDC 对接 Google Workspace、Microsoft Entra ID、Okta 等）、集中策略管理、角色权限、路由（支持故障转移）以及按日/周/月、按组织/群组/用户的消费上限。遥测数据通过 OTLP 发送至用户配置的收集器。gateway 不会向 Anthropic 发送推理流量或使用数据（除非配置使用 Claude A

4. Claude Code 打开 GitHub 仓库即执行隐藏恶意代码，攻击者可获完全控制

The Decoder：AI News · 06/30 08:00 · 监管、安全与版权

安全研究人员在 Mozilla 的 GenAI 漏洞赏金平台 0DIN 发现新攻击向量。一个看似正常的 GitHub 仓库包含 setup 脚本，该脚本运行时从 DNS 条目拉取命令并执行，恶意代码从未存在于仓库中，对扫描器、代码审查和 AI 智能体不可见。开发者使用 Claude Code 等 AI 编码工具打开该仓库时，Claude Code 在设置过程中遇到常规错误消息后自动运行该脚本，打开反向 shell，攻击者可窃取 API 密钥和登录凭据并维持持久访问。研究人员建议 AI 智能体应在运行前显示 setup 脚本内容，开发者应将第三方仓库的 setup 说明视为不受信任代码。

5. 三星和SK海力士计划投资5900亿美元扩产芯片，AI需求推高内存价格

The Decoder：AI News · 06/30 08:00 · 大厂与产业链

在韩国政府支持下，三星和SK海力士计划投入5900亿美元扩大芯片产能，包括800万亿韩元新建四座工厂、81万亿韩元建封装中心，以及未来15年30万亿韩元用于研发下一代芯片。AI数据中心需求是主要驱动力。Jefferies预测，2026年Q3内存价格将上涨40%至50%，Q4再涨30%至40%，2027年继续上涨40%至45%，到2028年新产能仅上线15%至20%才可能缓解。两家公司合计控制全球近80%的高带宽内存芯片市场。内存涨价已推高消费电子产品成本，苹果已上调Mac和MacBook售价。

模型 前沿模型与多模态

1. Runway API 推出广告本地化 Recipe

X：Runway (@runwayml) · 06/27 21:02

广告本地化现在可通过 Runway API 以 Recipe 形式使用。 现在您可以通过单次 API 调用翻译静态广告和图形资产。

2. Ask an AI expert: What exactly is the full stack?

Google AI Blog · 06/30 00:00

Ask an AI expert: What exactly is the full stack?。

3. 小红书 RedKnot 推理引擎：将 KV Cache 按注意力头拆解实现长文本加速

公众号：小红书技术（dots.llm） · 06/30 08:00

RedKnot 将 KV Cache 沿注意力头维度拆解，通过头分类稀疏（局部头占 83.4%–96.8%）、稀疏 FFN 和 SegPagedAttention 三个机制统一算法与存储粒度。在 8 卡 H800 上，TTFT 最高加速 1.6–3.54×，单卡并发提升 4.7–7.8×，预填充 FLOPs 削减 67%–79.5%。DeepSeek-V4-Flash 上 128K 上下文 TTFT 加速达 5.16×，KV 传输最多省 6.3×。精度通常不低于稠密 F1 的 95%。

4. Amazon engineers are reportedly distilling Anthropic models to cut costs before new token-based pricing kicks in

The Decoder · 06/30 02:05

Amazon engineers are reportedly distilling Anthropic models to cut costs before new token-based pricing kicks in。

智能体 智能体、编程与工作流

1. 美军用AI选目标却误炸伊朗学校，Anthropic Claude嵌入Palantir系统首日建议约1000目标

The Decoder：AI News · 06/30 08:00

美军在打击伊朗时首次大规模使用AI选择目标（Anthropic的Claude模型嵌入Palantir的Maven Smart System，首日建议约1000个目标），但对一所学校的导弹袭击导致约120名儿童死亡。调查发现，情报分析师早在2019年就通过数字工具标记该地点已变为小学，但该工具未连接军方官方目标数据库MIDB，信息从未送达指挥官。MIDB建于1980年代，依赖手动输入，替代系统MARS多年延迟。五角大楼事后宣布推出agentic AI initiative。Project Maven创建人Jack Shanahan批评目标验证不力不可原谅。

2. OpenAI 报告：绘制欧洲 AI 劳动力机遇版图

OpenAI · 06/30 08:00

OpenAI 发布新报告，分析 AI 对欧盟就业的影响，划定哪些职业面临自动化、增长或工作流程变化。

3. Grok 4.5 私测于 SpaceX 和 Tesla，性能接近 Opus

X：Elon Musk (@elonmusk, xAI) · 06/28 18:50

Grok 4.5，基于我们的1.5T V9基础模型，并在补充训练中加入Cursor数据，现已在SpaceX和Tesla进入私测。初步评估显示其性能接近，或许超越Opus。 强化学习仍在持续显著改进模型，Grok Build工具链也在日益完善。 所有参与者的出色工作！ 今年，@SpaceX 将每月发布完全从头训练的新模型。

4. 美团LongCat Owl Alpha：OpenRouter最流行模型，1.6万亿MoE，国产ASIC训练

X：Emad Mostaque (@EMostaque) · 06/30 08:00

美团LongCat的1.6万亿参数MoE模型Owl Alpha成为OpenRouter上最流行模型，累计消耗10万亿tokens，性能达Gemini/Opus 4.6级别。该模型使用35万亿tokens训练，完全在5万块国产ASIC上完成。据官方推文，Owl Alpha上线后每日调用量全球Top3，在Hermes Agent排名#1，Claude Code排名#2，OpenClaw排名#3。该模型即将退役，后续版本待公布。

产业 大厂与产业链

1. Artifacts 22：Zyphra、Cohere 和 Poolside 正在扩展生态系统广度

Nathan Lambert：Interconnects · 06/29 01:03

开源模型生态正变得更多元，参与者从少数中国公司扩展到全球各类组织。纯模型制造商包括 DeepSeek、智谱、MiniMax、Poolside、Arcee、Zyphra 及主权 AI 玩家 Cohere、Sovereign、Mistral、Trillion Labs；科技巨头如阿里 Qwen、Google Gemma 和 NVIDIA 各有不同动机；产品公司如 JetBrains、Zed、Krea、Photoroom 则训练高度专业的小模型。NVIDIA 发布 Nemotron-3-Ultra-550B-A55B-BF16，采用 LatentMoE 架构并改用 OpenMDW 许可证。Cohere 以 Apache 2.0 开源其旗舰模型 Command A+（05-2026-bf16），这是一款 218B-A25B MoE 模型，具备多模态、多语言和智能体能力。

2. DiScoFormer：一个跨分布同时估计密度与分数的单一Transformer模型

Hugging Face 博客 · 06/30 08:00

DiScoFormer（Density and Score Transformer）是一个无需重新训练即可从数据点估计分布密度和分数的单一模型。它利用Transformer的交叉注意力机制，在单次前向传播中输出密度和分数，并通过一致性损失实现分布外自适应。在100维空间中，DiScoFormer比最优调参的核密度估计（KDE）降低分数误差约6.5倍、密度误差超过37倍，且随样本量增加持续提升，而KDE内存耗尽。模型基于高斯混合模型训练，可泛化至非高斯分布（如Laplace、Student-t）及未见过的多模态混合。

3. HP Inc. launches Frontier strategic partnership with OpenAI

OpenAI 新闻 · 06/29 01:00

HP Inc. launches Frontier strategic partnership with OpenAI。

4. Anthropic：当AI成本超过工程师薪酬

Tomer Tunguz 博客（VC 分析） · 06/30 08:00

Anthropic在算力上的支出达到每位工程师每年51.5万美元，是其完全薪资（22.4万美元）的2.3倍。相比之下，顶尖1%软件公司的算力支出为8.9万美元，中位数仅为1.37万美元。三个2029年情景预测了这一差距的缩小路径。

开源 开源项目与开发者工具

1. eli-labz/Godcoder

GitHub · 06/27 08:56

eli-labz/Godcoder。

观点 KOL 与巨头言论

1. Qwen 3.6 27B 是本地开发的理想选择

Hacker News · 06/30 08:00

Qwen 3.6 27B 是一款密集参数本地大语言模型，原生支持 256k 上下文。在 Macbook Max M5 上运行 llama.cpp Q8_0 量化版（含多 token 预测）可达 30 tokens/s；用户反馈在 RTX 5090 上 Q6_K 量化可达 50 tokens/s。它可通过单个提示完成创意诗歌、用 pnpm 生成六边形扫雷游戏等任务，作者称其为首个真正具备通用智能的本地模型。另有一个 MoE 变体 35B A3B，但作者推荐 27B 版本。

监管 监管、安全与版权

1. Claude 在 Microsoft Foundry 正式可用

Claude 博客 · 06/30 08:00

从今天起，Claude 模型在 Microsoft Foundry 上正式可用，托管于 Azure 环境，运行在 NVIDIA GB300 GPU 上。首批提供 Claude Opus 4.8 和 Claude Haiku 4.5，通过 Messages API 调用，支持提示缓存和扩展思考。用户可选择推理处理位置，包括美国数据区域，由 Anthropic 负责推理运营。Azure 用户可使用现有身份验证、计费与治理控制，并获得统一账单；符合条件的 Enterprise Agreement 客户可将 Claude 用量计入 Azure 承诺消费。

2. 一次失败的（民族国家？）攻击的剖析

Hacker News · 06/27 23:39

作者收到伪装成新加坡VC Lua Ventures的虚假面试邮件，要求完成一个TypeScript仓库的"测试"。作者将仓库交给Claude扫描，在`typescript+5.9.2.patch`中发现base64混淆载荷，该载荷在`patch-package`安装时触发，向`~/.cache-`等目录写入`payload.js`和`mutex.js`，构成后门（命名PinpinRAT）。攻击者使用虚构身份和空洞LinkedIn资料，目标是作者在crates.io上的Rust包。相关信息已报告加拿大CCCS等机构。

研究 研究、评测与方法论

1. EverOS：开源Markdown优先智能体记忆运行时，支持混合检索与自进化技能

MarkTechPost · 06/29 18:42

EverMind 推出开源智能体记忆运行时 EverOS（Apache 2.0 许可）。它以可编辑的 Markdown 文件为记忆主体，经 SQLite 管理状态、LanceDB 实现混合检索（BM25 关键词 + 向量搜索 + 标量过滤）。每个完成的任务记录为 Case，离线提炼为可复用的 Skill，使记忆随使用自我进化。v1.1.0 新增 Knowledge APIs（支持分类与话题搜索的 Markdown 页面）和 Reflection（跨会话优化 Profile 和 Skill）。据 EverMind 报告，LoCoMo 得分 93.05%，LongMemEval 83.00%，HaluMem 93.04%，p95 检索延迟低于 500ms。运行时可本地优先部署，也提供 EverOS Cloud 托管选项，兼容 OpenAI 协议端点。

2. 仅有三个AI模型在500天创业测试中盈利超过起始资本

The Decoder：AI News · 06/28 18:16

普林斯顿大学推出CEO-Bench基准测试，让AI智能体在模拟环境中运营订阅软件公司NovaMind 500天，起始资金100万美元。14个测试模型中，仅Claude Fable 5（最佳轮次盈利4715万美元）、Claude Opus 4.8（2780万美元）和GPT-5.5（2130万美元）在最佳运行中超过起始资本。一个不调用语言模型的简单规则启发式方法通过固定定价、配额和针对性开发达到1576万美元，超越除上述三款外的所有模型。多数模型无法保持连贯策略，在模拟结束前破产。该测试旨在衡量AI的长期战略决策能力。

3. Meta发布Brain2Qwerty v2：非侵入式实时句子解码

X：AI at Meta (@AIatMeta) · 06/30 08:00

Meta公布Brain2Qwerty v2，这是非侵入式脑电信号解码研究的最新里程碑。基于当天发表在《Nature》的v1，v2是性能最高的端到端管道，能从原始脑信号实时解码句子。其从字符级性能提升至解码单词和语义，提高整体沟通准确性。该研究有望帮助数百万因脑损伤或疾病无法沟通的人群。