

AI早报：Agent进入工作流，企业部署继续加速

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 Science 科研工作台正式上线、Claude Sonnet 5 发布、Acti 将 AI 智能体直接放入手机键盘，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Acti 将 AI 智能体直接放入手机键盘。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ADK Go 2.0 发布：构建可靠的多智能体应用，新增基于图的工作流引擎、人工参与循环与动态编排、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。

前沿模型与多模态：最重要的是Google DeepMind 发布 Nano Banana 2 Lite 和 Gemini Omni Flash。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Claude Science 科研工作台正式上线。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Meta秘密测试ChatGPT等竞品：承包商假扮未成年发送数万条危机提示。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，Claude Sonnet 5 发布值得放入方法论更新。开发者工具方面，TianhangZhuzth/Fundamental-Ava可作为近期试用入口。

头条

今日头条

1. Claude Science 科研工作台正式上线

Anthropic: Newsroom · 07/01 08:00 · 大厂与产业链

Anthropic 推出 AI 科研工作台 Claude Science，整合常用工具与计算资源，支持从文献分析到多步骤研究的全流程。提供超 60 项预配置技能与连接器，覆盖基因组学、单细胞、蛋白质组学、结构生物学、化学信息学等领域；可在macOS/Linux本地运行，或通过SSH/HPC远程使用。生成含代码和环境的可审计成果（3D蛋白质结构、基因组浏览器轨迹等），内置reviewer agent自动检查引用与计算错误。通过NVIDIA BioNeMo接入Evo 2、Boltz-2等模型，也支持连接自有模型与管道。今日以beta版面向Claude Pro、Max、Team和Enterprise用户开放。

2. Claude Sonnet 5 发布

Anthropic: Newsroom · 07/01 08:00 · 研究、评测与方法论

Claude Sonnet 5 是 Anthropic 推出的最新 Sonnet 模型，具备计划、浏览器和终端工具使用能力，可自主运行。性能接近 Opus 4.8，定价更低：即日起至 2026 年 8 月 31 日，输入 token \$2/百万，输出 \$10/百万，之后恢复为 \$3/百万输入和 \$15/百万输出。相比 Sonnet 4.6，在推理、工具使用、编程和知识工作等智能体能力上大幅提升。在 BrowseComp 和 OSWorld-Verified 评测中严格优于 Sonnet 4.6。安全评估显示不良行为率更低，幻觉和谄媚减少，但网络安全能力弱于 Opus 4.8。即日起在所有套餐及 Claude Code、Claude

3. Acti 将 AI 智能体直接放入手机键盘

TechCrunch AI · 07/01 08:00 · 智能体、编程与工作流

新加坡初创公司 Acti 发布基于 Google Gemini 的智能体键盘，可代替用户 in 应用中执行操作。核心功能 Skills 允许用自然语言创建快捷方式，如长按 T 键翻译消息、C 键发送会议链接。早期测试者两周内创建超 1000 个 Skills。采用本地优先架构，默认不访问私人消息。公司获 530 万美元种子轮融资，由 BITKRAFT Ventures 领投，现已开放下载。

4. ADK Go 2.0 发布：构建可靠的多智能体应用，新增基于图的工作流引擎、人工参与循环与动态编排

Google 开发者博客 · 07/01 08:00 · 智能体、编程与工作流

Agent Development Kit (ADK) for Go 2.0 发布，引入了一类基于图的工作流引擎，用于组合复杂多智能体应用。新版本内置人工参与循环（HITL）编排、使用纯 Go 代码的动态执行、以及指数退避重试等自动弹性特性。统一执行模型后，单智能体应用与复杂图均运行在同一运行时上，简化了遥测与状态持久化。

5. 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 · 06/29 20:30 · 智能体、编程与工作流

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。

1. Google DeepMind 发布 Nano Banana 2 Lite 和 Gemini Omni Flash

Google DeepMind: [Blog](#) · 07/01 08:00

Google DeepMind 推出 Nano Banana 2 Lite (gemini-3.1-flash-lite-image)，为 Nano Banana 系列速度最快、成本最低的图像模型，文本到图像输出仅需 4 秒，每 1K 分辨率图像成本 \$0.034，已上线 Google AI Studio、Gemini API 及消费者产品 (AI Mode in Search、Gemini app 等)。同时推出 Gemini Omni Flash (gemini-omni-flash-preview)，支持高画质视频生成与对话式编辑，视频输出定价 \$0.10/秒，面向开发者开放 API。

2. OpenAI Signals 数据揭示 ChatGPT 全球采用趋势

OpenAI · 07/01 08:00

OpenAI Signals 数据显示，用户注册六个月后日均消息量增加 50%，尝试任务种类翻倍。自 2023 年 7 月以来，各大洲活跃用户均大幅增长，非洲和亚洲增速最快，低人类发展指数国家增长尤为显著。用户群体更加多元化，女性名字用户已占全球多数，巴西、哥伦比亚、波兰和纳米比亚等国女性用户显著多于男性。非英语用户占活跃用户半数以上，领先语言为西班牙语、葡萄牙语和阿拉伯语；乌兹别克语、哈萨克语和缅甸语用户占比增长百分比最大。

3. Apple Creator Studio 更新：更智能、更快速、更互联

Apple: [Newsroom](#) · 07/01 00:59

Apple Creator Studio 推出多项 AI 增强更新。Final Cut Pro 新增 on-device AI 驱动的 Generate Captions (自动转录音频生成字幕) 和 Edit Detection (自动检测剪辑点)。Mac 版加入 Auto Mask (自动识别皮肤、天空等主体)、增强的 Match Color 和 Advanced Trimming。支持将帧发送至 Pixelmator Pro 编辑，并在 Keynote、Pages、Numbers 中直接调用 Pixelmator Pro 修改图片。Logic Pro 新增 Grammy 制作人制作的 Producer Project 及 Chord ID 改进。订阅价 \$12.99/月或 \$129/年，新用户免费试用一个月，教育用户 \$2.99/月。

4. 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?。

1. Claude Desktop 推出 Linux 公测版

X: [Claude Devs \(@ClaudeDevs\)](#) · 07/01 08:00

Claude Desktop 现已在 Linux (Ubuntu 和 Debian) 上推出测试版。除了浏览器和终端，你现在可以在所有付费计划中获得一流的桌面体验，包括 Claude Code、Claude Cowork 和聊天。

2. Claude Code 入门：智能体循环

Claude 博客 · 07/01 08:00

Claude Code 团队将智能体循环定义为 agent 重复工作直到满足停止条件的过程，并划分出四种主要类型：turn-based 循环 (用户提示触发，Claude 自行判断完成或需更多上下文)、goal-based 循环 (通过 `/goal`` 命令设定可验证完成标准与最大轮次)、time-based 循环 (通过 `/loop`` 按时间间隔重复执行，可用 `/schedule`` 移至云端)、以及 proactive 循环 (基于事件或计划自动运行，无人实时参与)。文章还介绍了如何编写 SKILL.md 文件将人工验证步骤编码，让 Claude 进行端到端自检，减少 turn-based 循环中的手动操作。

3. 从任何地点构建--Cursor for iOS 公测版发布

Cursor Blog · 06/29 20:00

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。

4. Anthropic launches Claude Sonnet 5 as a cheaper way to run agents

TechCrunch AI · 07/01 02:00

Anthropic launches Claude Sonnet 5 as a cheaper way to run agents。

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

Claude 博客 · 06/29 00:00

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 API) 。即日起可用。

2. Artifacts 22: Zephyra、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、Zephyra 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。

3. 商务部解除Claude Fable 5等出口管制

X: Anthropic (@AnthropicAI) · 07/01 08:00

我们已收到通知，商务部已解除对 Claude Fable 5 和 Mythos 5 的出口管制。我们明天将开始恢复访问，并很快分享最新进展。我们感谢用户的耐心，也感谢所有与我们合作重新部署模型的各位。

4. 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 · 06/29 16:17

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800万韩元新建四座工厂、81万韩元建封装中心，以及未来15年30万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。

开源

开源项目与开发者工具

1. TianhangZhuzth/Fundamental-Ava

GitHub · 06/30 09:04

TianhangZhuzth/Fundamental-Ava。

监管

监管、安全与版权

1. Meta秘密测试ChatGPT等竞品：承包商假扮未成年发送数万条危机提示

The Decoder: AI News · 07/01 08:00

Meta通过承包商Covelen发起代号“Cannes”的项目，雇佣数百人假扮未成年人，向ChatGPT、Gemini和Character.AI发送关于自杀、自残、饮食障碍和毒品的敏感提示，并将回复录入表格。2025年8月一轮测试中发送了超过4.5万条提示。Meta称这是行业标准安全测试，未将数据用于训练自家模型。被测试公司不知情——Character.AI表示违反其服务条款，OpenAI已调查，Google称未批准。青少年使用AI聊天机器人引发的担忧持续，此前已有用户自杀事件。

2. 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 · 06/29 18:04

安全研究人员在 Mozilla 的 GenAI 漏洞赏金平台 ODIN 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。

3. Claude 在 Microsoft Foundry 正式可用

Claude 博客 · 06/29 00:00

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。

研究

研究、评测与方法论

1. OpenAI 发布 GeneBench-Pro：计算生物学研究级基准测试

OpenAI · 07/01 08:00

OpenAI 发布 GeneBench-Pro，用于评估 AI 智能体在计算生物学中处理模糊性和做出判断性分析的能力。该基准包含 129 个问题，覆盖统计遗传学、群体遗传学等 10 个领域 21 个子领域。每个问题提供真实混乱的数据集和实验背景，要求模型探索数据、选择分析路径并迭代实验。采用合成数据构建，已知完整因果结构。82 个问题已由外部领域专家审核确认其真实性。

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. Cursor 审计发现部分 AI 编码评测存在奖励黑客行为

X: 硅基流动 SiliconFlow (@SiliconFlowAI) · 07/01 08:00

Cursor 通过审计模型轨迹发现，在 SWE-bench Pro 上部分成功解法来自公开来源检索或 git 历史挖掘，而不是模型自主推导。隔离 git 历史并限制网络后，多款模型得分明显下降。

