AI 行业日报

模型与工具能力 · 产业基础设施 · 应用商业化 · 研究开源 · 资本监管 | Data Source: aihot.virxact.com

API耗时: 1s 精选条目: 43 条 焦点: 8 条 快讯: 0 条

Executive Summary

GLM-5.2 模型正式发布并开源，采用753B参数MoE架构，支持1M上下文窗口，专注于Coding与长程任务处理能力。Cloudflare 推出One stack智能体工具集，实现AI驱动的Zero Trust环境自动配置。OpenAI 一季度现金消耗达37亿美元，超过同期收入一半，同时秘密提交IPO申请，估值或达1万亿美元。NVIDIA GEAR 实验室发布ENPIRE系统，首次实现8个AI智能体自主控制机器人完成物理实验。Anthropic 5月企业AI订阅份额达41%，首次超越OpenAI的39.5%。

AI基础设施领域呈现显著分化趋势。Google 发布Agentic Resource Discovery开放规范，推动智能体工具标准化发现机制。Vercel 开源Eve智能体框架，采用文件系统优先设计理念。阿里云 发布HappyOyster 1.0开放式世界模型，支持实时交互数字世界生成。算力成本压力持续显现，OpenAI年度运营亏损达209.2亿美元，反映大模型商业化仍面临严峻成本挑战。Anthropic 与DeepMind CEO在G7会议呼吁组建AI联盟排除中国，地缘政治因素进一步影响全球AI产业格局。

后续需关注GLM-5.2在国产算力平台的实际性能表现及其开源生态发展情况。OpenAI IPO进程与财务状况将直接影响行业投资预期。Anthropic 市场份额增长势头能否持续值得关注。智能体标准化协议ARD的采纳程度将影响未来工具互操作性。物理世界AI应用的商业化路径与安全性保障机制需要持续跟踪。中美AI技术脱钩趋势对全球供应链的影响程度将是重要观察变量。

重点 今日核心进展

★ 1. GLM-5.2 上线并开源：专注 Coding 与长程任务

智谱：研究（网页内嵌数据） · 昨天 00:00 · 模型与工具能力

GLM-5.2 已发布并开源，采用 MIT 协议，支持 1M 上下文窗口。Coding 方面能承载项目级上下文，长程任务执行更稳定，遵循生产级工程规范，并支持客户端与移动端真机调试闭环。通过极致 Infra 优化，发布首日即可在国产算力平台运行。模型已开源至 GitHub、Hugging Face、ModelScope、BigModel 开放平台、Z.ai、智谱清言、AutoClaw 及 ZCod

能力进展 基础设施 新发布

https://www.zhipuai.cn/zh/research/161

★ 2. GLM-5.2：可能是最强大的纯文本开源权重大语言模型

Simon Willison 博客 · 23 分钟前 · 模型与工具能力

智谱（Z.ai）于6月13日向编码计划订阅者发布GLM-5.2，6月16日以MIT许可证开源完整权重。该模型为753B参数、1.51TB的MoE架构，40个活跃参数，纯文本输入，上下文窗口提升至100万token。在Artificial Analysis Intelligence Index v4.1上以51分领先，超越MiniMax-M3（44）、DeepSeek V4 Pro（max，44）和

能力进展 新发布

https://simonwillison.net/2026/Jun/17/glm-52

★ 3. MolmoMotion：语言引导的3D运动预测模型

Hugging Face：Blog（RSS） · 8 小时前 · 模型与工具能力

MolmoMotion基于Molmo 2骨干网络，输入视频帧、物体上的3D点标记及文字动作指令（如"移动并旋转桌上放水果的木碗"），预测未来数秒内这些点的3D轨迹。提供两个变体：自回归的MolmoMotion-AR逐步预测坐标，流匹配的MolmoMotion-FM通过连续空间变换处理多可能性运动。同时发布MolmoMotion-1M数据集（含116万视频的3D点轨迹及动作描述）和PointMoti

能力进展 新发布

https://huggingface.co/blog/allenai/molmomotion

★ 4. Cloudflare 发布 Cloudflare One stack：智能体驱动的部署工具集

Cloudflare Blog · 11 小时前 · 应用与商业化

6月17日，Cloudflare 推出 Cloudflare One stack，一组可直接赋予 AI 智能体的技能文件，用于自动配置、部署和管理 Zero Trust 环境。工具集包含两个轻量级 skill：`cloudflare-one` 负责通用产品指导（VPN 替换、网络连接、安全策略等），`cloudflare-one-migration` 提供从 Zscaler、Palo Alto N

能力进展 基础设施 监管/资本

https://blog.cloudflare.com/cloudflare-one-stack

★ 5. 消息称 OpenAI 今年一季度现金消耗达 37 亿美元，超同期收入的一半

IT之家（RSS） · 19 小时前 · 产业与基础设施

OpenAI 在 2026 年第一季度现金消耗达 37 亿美元，超过同期 57 亿美元收入的一半。数据来自一份向股东披露的文件，直观体现 AI 大模型研发与规模化落地的巨额成本。OpenAI 正筹备上市，已在美国保密递交 IPO 申请，最早或于 9 月完成，估值最高可达 1 万亿美元。头部 AI 企业持续重金投入算力、模型研发与人才招募以维持竞争优势。

能力进展 基础设施 监管/资本

https://www.ithome.com/0/965/335.htm

★ 6. 谷歌发布Agentic Resource Discovery（ARD）开放规范

Google Developers Blog（RSS） · 8 小时前 · 产业与基础设施

Agentic Resource Discovery（ARD）是一项开放规范，用于在Web上发布、发现和验证AI工具、技能与智能体。它基于两个原语：组织在其自有域名下托管catalog描述可用能力，registry作为搜索引擎索引catalog并响应发现请求。ARD支持加密验证，使客户端与端点连接前确认发布者身份，然后直接通过原生协议调用能力。Google Cloud的Gemini Enterpr

能力进展 基础设施 新发布

https://developers.googleblog.com/announcing-the-agentic-resource-discovery-specification

★ 7. NVIDIA GEAR实验室发布ENPIRE：8个Codex智能体自主控制机器人完成物理实验

X：Jim Fan (@DrJimFan) · 7 小时前 · 研究与开源进展

NVIDIA GEAR实验室推出ENPIRE系统，首次实现物理世界自主研究。系统让8个Codex智能体控制8台机器人，配备GPU和token预算。安全方面采用硬运动极限切断和扭矩受限夹爪两层硬件保障，支持通宵无人运行。奖励函数通过视觉分类器离线固定并冻结，防止智能体作弊。实时监测机器人利用率（MRU）、token利用率（MTU）和GPU利用率，以Tokens-to-Success和Time-to-

能力进展 基础设施 监管/资本

https://x.com/DrJimFan/status/2067283904986517866

★ 8. Wolfram 语言和 Mathematica 15 版发布：内置 AI 助手、符号音乐等新功能

Hacker News 热门（buzzing.cc 中文翻译） · 20 小时前 · 应用与商业化

在 Mathematica 诞生近 38 年后，Wolfram 语言与 Mathematica 发布 Version 15。每个笔记本内置 AI 助手，支持从 AI 环境中直接调用 Wolfram 技术。新增符号音乐系统、大规模时间序列与事件序列处理、分类数据计算、模型拟合超函数 ModelFit。笔记本支持千兆字节级大小与实时查找，首次引入侧边栏、视觉主题及弃用功能样式。强化了表格连接、多点可视

能力进展 基础设施 新发布

https://writings.stephenwolfram.com/2026/06/launching-version-15-of-wolfram-language-mathematica-built-in-useful-ai-lots-of-new-core-functionality

能力 模型与工具能力

1. GLM-5.2 上线并开源：专注 Coding 与长程任务

智谱：研究（网页内嵌数据） · 昨天 00:00

GLM-5.2 已发布并开源，采用 MIT 协议，支持 1M 上下文窗口。Coding 方面能承载项目级上下文，长程任务执行更稳定，遵循生产级工程规范，并支持客户端与移动端真机调试闭环。通过极致 Infra 优化，发布首日即可在国产算力平台运行。模型已开源至 GitHub、Hugging Face、ModelScope、BigModel 开放平台、Z.ai、智谱清言、AutoClaw 及 ZCod

能力进展 基础设施 新发布

https://www.zhipuai.cn/zh/research/161

2. GLM-5.2：可能是最强大的纯文本开源权重大语言模型

Simon Willison 博客 · 23 分钟前

智谱（Z.ai）于6月13日向编码计划订阅者发布GLM-5.2，6月16日以MIT许可证开源完整权重。该模型为753B参数、1.51TB的MoE架构，40个活跃参数，纯文本输入，上下文窗口提升至100万token。在Artificial Analysis Intelligence Index v4.1上以51分领先，超越MiniMax-M3（44）、DeepSeek V4 Pro（max，44）和

能力进展 新发布

https://simonwillison.net/2026/Jun/17/glm-52

3. MolmoMotion：语言引导的3D运动预测模型

Hugging Face：Blog（RSS） · 8 小时前

MolmoMotion基于Molmo 2骨干网络，输入视频帧、物体上的3D点标记及文字动作指令（如"移动并旋转桌上放水果的木碗"），预测未来数秒内这些点的3D轨迹。提供两个变体：自回归的MolmoMotion-AR逐步预测坐标，流匹配的MolmoMotion-FM通过连续空间变换处理多可能性运动。同时发布MolmoMotion-1M数据集（含116万视频的3D点轨迹及动作描述）和PointMoti

能力进展 新发布

https://huggingface.co/blog/allenai/molmomotion

4. Grok 4.3 在 Amazon Bedrock 正式可用

xAI：News（网页） · 昨天 08:00

6 月 17 日，xAI 宣布 Grok 4.3 在 Amazon Bedrock 上全面可用。该模型在前沿模型中达成最低幻觉率，支持 100 万 token 上下文窗口，并提供可配置推理努力（none/low/medium/high）。在 Artificial Analysis Omniscience 基准排名第一，在 Tau2 Telecom 基准评估客服智能体真实工具调用性能排名第一，在 V

能力进展

https://x.ai/news/grok-amazon-bedrock

产业 产业与基础设施

1. 消息称 OpenAI 今年一季度现金消耗达 37 亿美元，超同期收入的一半

IT之家（RSS） · 19 小时前

OpenAI 在 2026 年第一季度现金消耗达 37 亿美元，超过同期 57 亿美元收入的一半。数据来自一份向股东披露的文件，直观体现 AI 大模型研发与规模化落地的巨额成本。OpenAI 正筹备上市，已在美国保密递交 IPO 申请，最早或于 9 月完成，估值最高可达 1 万亿美元。头部 AI 企业持续重金投入算力、模型研发与人才招募以维持竞争优势。

能力进展 基础设施 监管/资本

https://www.ithome.com/0/965/335.htm

2. 谷歌发布Agentic Resource Discovery（ARD）开放规范

Google Developers Blog（RSS） · 8 小时前

Agentic Resource Discovery（ARD）是一项开放规范，用于在Web上发布、发现和验证AI工具、技能与智能体。它基于两个原语：组织在其自有域名下托管catalog描述可用能力，registry作为搜索引擎索引catalog并响应发现请求。ARD支持加密验证，使客户端与端点连接前确认发布者身份，然后直接通过原生协议调用能力。Google Cloud的Gemini Enterpr

能力进展 基础设施 新发布

https://developers.googleblog.com/announcing-the-agentic-resource-discovery-specification

3. Anthropic 5月企业AI订阅份额首超OpenAI，特朗普政府禁令反促采用量创新高

TechCrunch：AI（RSS） · 昨天 06:34

Anthropic 5月企业AI订阅市场份额达41%，首次超越OpenAI（39.5%）。公司刚完成650亿美元融资、估值9650亿美元，并因首次盈利季度秘密提交IPO。特朗普政府以出口管制为由要求Anthropic禁止非美国人访问最新模型Mythos 5及Fable 5，导致两款模型下架。Ramp首席经济学家指出，类似争议（如3月被国防部列为供应链风险）反而推动Anthropic企业采用量创纪录

能力进展 监管/资本 新发布

https://techcrunch.com/2026/06/16/anthropics-latest-feud-with-the-trump-admin-may-actually-help-it-sales-data-suggests

4. 泄露文件显示OpenAI年营收130亿但亏损远超收入

Hacker News 热门（buzzing.cc 中文翻译） · 1 小时前

OpenAI 2025年营收130.7亿美元（2024年37亿），但研发成本达191.8亿（含向微软支付105.9亿），收入成本（推理计算）75亿，销售营销成本57.3亿，运营亏损209.2亿。2025年净亏损约390亿，扣除约300亿一次性会计费用后约80亿。2025年3月获1220亿融资（估值8520亿）。ChatGPT周活超9亿，付费约5000万。为控制成本已关闭Sora视频模型并削减非核心

能力进展 监管/资本 新发布

https://arstechnica.com/ai/2026/06/leaked-financial-docs-show-openai-is-losing-billions-of-dollars-a-year

5. Anthropic与DeepMind CEO呼吁G7组建AI联盟排除中国

X：Kim (@kimmonismus) · 6 小时前

Dario Amodei（Anthropic）与Demis Hassabis（Google DeepMind）在G7闭门会议上呼吁组建美国主导的联盟，为人工智能制定全球规则和标准。Amodei指出，该联盟应以前沿模型和硬件（包括芯片及其他关键组件）的访问权限为手段，将中国排除在外。这一主张被评论为高技术新冷战的开端，竞争方将从根本上被剥夺参与权。

能力进展 基础设施

https://x.com/kimmonismus/status/2067310431669223425

6. 微软考虑为 Copilot Cowork 集成 DeepSeek V4

X：Kim (@kimmonismus) · 昨天 02:08

微软正考虑为 Copilot Cowork 提供微软托管的 DeepSeek V4 版本，作为更便宜的模型选项。Copilot Cowork 将放弃无限定价，转向按使用量计费，原因是成本过高（用户每周执行数百项任务导致费用激增）。若采用 DeepSeek，该模型将是可选的、经过微调与安全防护，并完全托管于 Azure。Axios 报道称微软已微调了一个可用模型，最终决定待定。

能力进展 监管/资本

https://x.com/kimmonismus/status/2066946013026263110

7. 美国司法部援引国家安全为xAI未经许可的燃气轮机辩护

The Decoder：AI News（RSS） · 昨天 21:23

美国司法部在一份驳回诉讼的动议中称，xAI的聊天机器人Grok对军事行动至关重要，以此为其在密西西比州Southaven的Colossus 2设施运行未经许可的燃气轮机辩护。NAACP已提起诉讼，指控xAI的燃气轮机数量从4月的27台增至57台，导致氮氧化物排放飙升111%。国防部首席数字与人工智能官Cameron Stanley表示，Grok是支持机密和绝密网络军事任务的四款AI模型之一，包括近

能力进展 监管/资本

https://the-decoder.com/doj-invokes-national-security-to-defend-xais-unpermitted-gas-turbines-in-naacp-lawsuit

8. 中国加紧筹建世界人工智能合作组织

IT之家（RSS） · 21 小时前

中国正加紧筹建世界人工智能合作组织，欢迎各方加入。2025年7月26日，中国政府倡议成立该组织，作为践行多边主义、推动共商共建共享全球治理的举措，旨在弥合数字和智能鸿沟、促进人工智能向善普惠发展。初步考虑总部设在上海。同日，2025世界人工智能大会发表《人工智能全球治理行动计划》，呼吁各方遵循向善为民、尊重主权、发展导向、安全可控、公平普惠、开放合作的原则，协力推进全球人工智能发展与治理。

监管/资本 新发布

https://www.ithome.com/0/965/248.htm

9. Fable 遭美国政府封禁，TechCrunch 质疑真正原因并非模型越狱

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 00:45

美国政府对 Anthropic 的模型 Fable 实施封禁，但 TechCrunch 发文质疑，实际原因可能并非此前认为的"模型越狱"问题。该文章在 Hacker News 引发讨论，获得 103 个点赞。

能力进展

https://techcrunch.com/2026/06/15/the-us-governments-anthropic-models-ban-was-never-about-an-ai-jailbreak

10. 库克：AI 浪潮引发存储芯片价格暴涨，iPhone 等苹果产品涨价已"不可避免"

IT之家（RSS） · 1 小时前

苹果CEO库克确认，AI热潮导致存储芯片严重短缺和价格暴涨，苹果产品涨价已"不可避免"。库克未透露涨价具体细节。华尔街日报指出，全球AI巨头大幅增加资本开支，高带宽内存需求激增，挤压消费电子芯片供应。自2024年以来内存和存储芯片价格已翻四倍，涨势预计延续至2027年。研究机构估算，下一代iPhone 18 Pro售价或需增加约270美元。苹果已在上月提高Mac Mini起售价。摩根士丹利预测，今

基础设施

https://www.ithome.com/0/965/694.htm

11. 马斯克：AI将达Stockfish级编码

X：Elon Musk (@elonmusk, xAI) · 昨天 21:47

AI 将实现 Stockfish 级别的编码和通用计算机使用

https://x.com/elonmusk/status/2066880262668247091

应用 应用与商业化

1. Cloudflare 发布 Cloudflare One stack：智能体驱动的部署工具集

Cloudflare Blog · 11 小时前

6月17日，Cloudflare 推出 Cloudflare One stack，一组可直接赋予 AI 智能体的技能文件，用于自动配置、部署和管理 Zero Trust 环境。工具集包含两个轻量级 skill：`cloudflare-one` 负责通用产品指导（VPN 替换、网络连接、安全策略等），`cloudflare-one-migration` 提供从 Zscaler、Palo Alto N

能力进展 基础设施 监管/资本

https://blog.cloudflare.com/cloudflare-one-stack

2. Wolfram 语言和 Mathematica 15 版发布：内置 AI 助手、符号音乐等新功能

Hacker News 热门（buzzing.cc 中文翻译） · 20 小时前

在 Mathematica 诞生近 38 年后，Wolfram 语言与 Mathematica 发布 Version 15。每个笔记本内置 AI 助手，支持从 AI 环境中直接调用 Wolfram 技术。新增符号音乐系统、大规模时间序列与事件序列处理、分类数据计算、模型拟合超函数 ModelFit。笔记本支持千兆字节级大小与实时查找，首次引入侧边栏、视觉主题及弃用功能样式。强化了表格连接、多点可视

能力进展 基础设施 新发布

https://writings.stephenwolfram.com/2026/06/launching-version-15-of-wolfram-language-mathematica-built-in-useful-ai-lots-of-new-core-functionality

3. 小米 MiMo Claw 正式版发布：旗舰模型+金山办公，全新订阅服务上线

公众号：小米 MiMo · 昨天 21:32

小米推出云端轻量化 Claw 类产品 MiMo Claw 正式版，搭载与 OpenClaw 框架深度适配的 MiMo-V2.5-Pro 旗舰模型。该模型原生兼容 MCP 工具调用协议，内置百万级超长上下文，支持单会话千次以上连续工具调用；依托 MTP 三层解码架构，在 OpenClaw 标准 Agent 工作流中吞吐效率提升约 3 倍。ClawEval 测试中任务达标率（Pass3）达 63.8%

能力进展 基础设施 新发布

https://mp.weixin.qq.com/s/N5ac768a8LkhEjOVpkR1dQ

4. Strands Robots SDK：用单一智能体打通 Hugging Face Hub 到物理机器人

Hugging Face：Blog（RSS） · 14 小时前

AWS（Apache 2.0）开源的 Strands Robots SDK 将 LeRobot 栈封装为 AgentTools，构建统一智能体。默认用 MuJoCo 模拟（无需硬件），mode="real" 切换至真实机器人。可记录演示数据为 LeRobotDataset 并推送 Hugging Face Hub，运行 GR00T 或 LerobotLocal 策略推理，经 Zenoh mesh

能力进展 基础设施 新发布

https://huggingface.co/blog/amazon/strands-lerobot-hub-to-hardware

5. Vercel 发布开源 AI 智能体框架 Eve：每个智能体就是一个文件目录

MarkTechPost（RSS） · 6 小时前

Vercel 发布开源 AI 智能体框架 Eve（npm 包，Apache-2.0 许可）。Eve 采用文件系统优先设计：每个智能体对应一个磁盘目录，目录结构直接映射模型、指令、工具、技能、连接、子智能体等能力，无需额外注册代码。内置六大生产级能力：持久执行（每步检查点，崩溃后可恢复）、沙箱计算、人机审批、安全连接（支持 MCP 和 OpenAPI）、多通道（Slack、Discord、Teams

能力进展 监管/资本 新发布

https://www.marktechpost.com/2026/06/17/vercel-releases-eve

---

6. 阿里云发布HappyOyster 1.0：一句话生成可实时交互的数字世界

IT之家（RSS） · 11 小时前

6月17日，阿里云发布开放式世界模型HappyOyster 1.0（快乐生蚝）。该产品基于原生多模态架构，支持多模态输入与音视频联合生成，可在生成过程中持续接收用户指令并实时响应画面。它深度学习物理世界状态转移规律，保持人物和环境长程一致性。官网开放"实时导演"与"世界探索"两种玩法：前者可随时叫停改写故事、与虚拟男友实时互动等；后者支持自由漫游、滑板冲刺、翼装滑翔、骑马奔驰、攻击打怪等交互。该产

能力进展 基础设施 新发布

https://www.ithome.com/0/965/652.htm

---

7. Kickart 3.0发布，让广告视频创作更精准高效

公众号：火山引擎 · 14 小时前

火山引擎一站式营销创作平台Kickart 3.0（原"创作Agent"）正式上线，升级为对话式视频生成模式，用户可通过多轮对话调整商品图、故事板等，用自然语言生成营销视频。新增"爆款裂变"能力，上传视频链接后自动拆解爆款逻辑并重构至新商品视频，支持抖音电商内容合规与质量预审核。平台开放SaaS、API及Skill等多种交付方式，并已接入Seedance 2.0 mini，助力降低广告营销成本。

能力进展 监管/资本 新发布

https://mp.weixin.qq.com/s/e6SOMeRbdMB_zATsNcYmUQ

---

8. Claude Code v2.1.181 发布

Claude Code：GitHub Releases（RSS） · 2 小时前

Claude Code v2.1.181 发布，新增 `/config key=value` 语法允许在提示中直接设置任意配置项，新增 `sandbox.allowAppleEvents` 选项使沙盒命令支持 Apple Events，新增 `CLAUDE_CLIENT_PRESENCE_FILE` 环境变量用于抑制移动端推送通知。内置 Bun 运行时升级至 1.4，改进了长段落流式输出（逐行显示

能力进展 新发布

https://github.com/anthropics/claude-code/releases/tag/v2.1.181

---

9. Claude Design 更新：跨项目保持品牌一致，与Claude Code协同

Claude：Blog（网页） · 3 小时前

6月17日，Claude Design 更新，支持跨项目使用统一设计系统，并与Claude Code同步工作流。用户可直接拖拽、对齐和缩放画布元素，编辑器稳定性大幅提升。设计系统可从GitHub、设计文件或原始上传导入，团队管理员可锁定标准系统防止篡改。新增桌面端侧边栏入口及独立网页端claude.ai/design。使用限制与聊天、Claude Cowork、Claude Code共享，每次任务

能力进展 新发布

https://claude.com/blog/claude-design-stays-on-brand-for-daily-work

---

10. Google发布99美元Gemini智能音箱

TechCrunch：AI（RSS） · 7 小时前

Google推出首款专为Gemini打造的智能音箱Google Home Speaker，售价99.99美元。支持自然语言请求和多步指令，可在说话中途纠正，并具备连续对话功能。内置10种新声音。高级AI功能需订阅Google Home Premium（月费10美元或年费100美元），包括Gemini Live自由对话、Nest摄像头活动摘要等。即日起预售，本月发货。

能力进展 新发布

https://techcrunch.com/2026/06/17/google-bets-on-gemini-to-reinvent-the-smart-home-speaker

---

11. Omnigent开源：AI智能体团队元框架

X：Yuchen Jin (@Yuchenj_UW) · 8 小时前

编程的未来不是单一智能体，而是一个完整的AI团队。 Omnigent让你在一个实时会话中运行一个智能体团队：Claude Code、Codex、Cursor、Pi，以及你自己的智能体。 它是一个面向AI智能体的元框架，基于我们内部的Databricks开发工具构建，现已开源给所有人。 由传奇人物@matei_zaharia和Databricks AI团队打造。没错，Matei仍然编写大量代码

能力进展 新发布

https://x.com/Yuchenj_UW/status/2067273020352380950

---

12. 借助 Workload Identity Federation 安全访问 Claude Platform

Claude：Blog（网页） · 3 小时前

Workload Identity Federation （WIF） 已在 Claude Platform 上全面可用。WIF 兼容任何 OIDC 身份提供者，覆盖所有 Claude API 端点（包括第一方 SDK 和 Claude Code）。WIF 用短生命期凭证替代静态 API 密钥，并引入服务账户，每个工作负载拥有独立身份、角色和审计日志。Claude Console 提供引导设置流程，

能力进展 监管/资本

https://claude.com/blog/workload-identity-federation

---

13. GitHub 发布 CC0-1.0 开源多语言仓库级数据集，覆盖 README、Issue 和 PR

GitHub Blog · 4 小时前

GitHub 推出一个新的仓库级数据集，采用 CC0-1.0 许可证，旨在帮助研究人员和开发者发现跨 README、Issue 和 Pull Request 的多语言开发者内容，加速多语言 AI 开发。

能力进展 新发布

https://github.blog/ai-and-ml/github-copilot/getting-more-from-each-token-how-copilot-improves-context-handling-and-model-routing

14. Copilot Cowork 全球正式可用，支持多模型

X：Satya Nadella (@satyanadella) · 昨天 23:50

Copilot Cowork 现已全球正式可用，并支持多模型！ 每个组织都可以让长期运行的智能体处理复杂的多步骤任务，基于你组织的独特知识和专有技术。 https：//www.microsoft.com/en-us/microsoft-365/blog/2026/06/16/copilot-cowork-is-now-generally-available/？v=15

能力进展

https://x.com/satyanadella/status/2066911399494963335

15. Claude Design与Replit联动，设计变应用

X：Replit (@Replit) · 4 小时前

在Claude中设计。在Replit中构建。 你现在可以将Claude Design中的设计发送到Replit，将其变成一个可工作的应用。

能力进展

https://x.com/Replit/status/2067328501003497684

16. Midjourney V8.1 推出 Draft mode 草稿模式与新功能预览

Midjourney：Updates（RSS） · 昨天 06:04

Midjourney V8.1 的 Draft mode 草稿模式每次生成24张低分辨率低质量图片。用户可对任意图片点击 "Vary"，将其渲染为全质量、全分辨率版本。草稿任务消耗的快速小时数减半。

新发布

https://updates.midjourney.com/draft-mode-for-v8-1-and-new-feature-previews

研究 研究与开源进展

1. NVIDIA GEAR实验室发布ENPIRE：8个Codex智能体自主控制机器人完成物理实验

X：Jim Fan (@DrJimFan) · 7 小时前

NVIDIA GEAR实验室推出ENPIRE系统，首次实现物理世界自主研究。系统让8个Codex智能体控制8台机器人，配备GPU和token预算。安全方面采用硬运动极限切断和扭矩受限夹爪两层硬件保障，支持通宵无人运行。奖励函数通过视觉分类器离线固定并冻结，防止智能体作弊。实时监测机器人利用率（MRU）、token利用率（MTU）和GPU利用率，以Tokens-to-Success和Time-to-

能力进展 基础设施 监管/资本

https://x.com/DrJimFan/status/2067283904986517866

2. 用SGLang-JAX在TPU上优化Ling-2.6-1T：一个Pallas核将MoE数据移动隐藏在计算中

LMSYS：Blog（Chatbot Arena 团队） · 6 小时前

SGLang-JAX现已支持inclusionAI的Ling-2.6-1T（1T稀疏MoE，63B激活参数，256路由专家，top-8路由加共享专家）在TPU v7x上高效推理。团队开发了Fused MoE V2--一个融合scatter、专家FFN和gather的Pallas核，通过将MoE数据移动隐藏在计算中，使MoE预填充延迟从5.16ms降至2.42ms（降幅53%），解码核延迟从0.24

能力进展 基础设施

https://www.lmsys.org/blog/2026-06-17-ling-2-6-tpu

3. LifeSciBench 发布

OpenAI：官网动态（RSS · 排除企业/客户案例） · 昨天 08:00

2026 年 6 月，OpenAI 联合 173 位博士级生命科学家发布 LifeSciBench 评测基准，涵盖 750 个真实研究任务，覆盖证据处理、分析、设计优化等七个工作流及七个生物领域。每项任务配有约 25 条细化评分标准（共 19，020 条），评估模型的科学正确性与实用价值。79% 的任务需多步推理，53% 要求解读图表、PDF 等附件数据，旨在衡量 AI 在复杂、不确定的研究任务中

能力进展 新发布

https://openai.com/index/introducing-life-sci-bench

4. 公开聊天数据能否预测真实世界AI失调?

OpenAI：Alignment 研究博客（RSS） · 昨天 02:00

OpenAI利用WildChat公开数据集（2023年4月至2024年5月收集的100万条对话）模拟模型部署，预测GPT-5.1、GPT-5.2、GPT-5.4在真实生产环境中的不良行为率。与私有生产数据对比发现，WildChat模拟的平均预测误差约3倍；但对技术性和智能体型失调的预测精度下降。研究验证了公开数据集作为外部审计工具的可行性。

能力进展 新发布

https://alignment.openai.com/validating-public-evals

5. Google 医学推理 AI 系统 AMIE 新研究：从诊断迈向长期疾病管理

Google Blog：AI（RSS） · 9 小时前

今日发表在《自然》杂志上的研究展示了 Google 的医学推理 AI 系统 AMIE（Articulate Medical Intelligence Explorer）从单次诊断对话演进到长期疾病管理的能力。AMIE 利用 Gemini 模型的长上下文能力，整合共情对话智能体和深度思考管理推理智能体，可交叉引用数百页临床指南。在盲测中，AMIE 与 21 名初级保健医生相比，在整体管理推理上匹配临

能力进展

https://blog.google/innovation-and-ai/models-and-research/google-research/amie-for-disease-management-in-nature

格局 观点、资本与监管

1. Matt Pocock 开源 skills v1：将技能描述 Token 成本降低 63%

X：阿易 AI Notes (@AYi_AInotes) · 4 小时前

Matt Pocock（Total TypeScript 作者）开源了 skills v1，将技能描述的 Token 成本降低 63%。该工具包将技能分为模型可调用和用户可调用，新增 /codebase-design、/domain-modeling、/grilling 三项技能；重写 /writing-great-skills；将 /diagnose 更新为 /diagnosing-bugs 并

能力进展 新发布

https://x.com/AYi_AInotes/status/2067327021005656135

2. baoyu-design 本地动画视频导出功能更新

X：宝玉 (@dotey) · 昨天 08:21

baoyu-design（本地运行 Claude Design 的 Skill）新增动画视频导出功能。其声明式动画引擎基于 f（t） 设计：任意时间点 t 可绝对确定画面状态。导出采用无头 Chromium 逐帧截图 + ffmpeg 编码，每帧等待两帧 requestAnimationFrame 确保渲染完成。截图以 2 倍 DPR（3840×2160）再缩回 1080p，保证细节清晰。95 秒

能力进展 新发布

https://x.com/dotey/status/2067039941960327204

3. Google 分享 A2UI 与 MCP Apps 三种集成架构模式

Google Developers Blog（RSS） · 3 小时前

Google 分享了三种集成 A2UI 与 MCP Apps 的架构模式，旨在结合两者优势。A2UI 采用声明式框架，通过 JSON payload 定义 UI，由宿主原生渲染，确保一致性与安全性，但受限于预定义组件库。MCP Apps 在 iframe 中使用标准 Web 技术提供自定义界面，但存在设计碎片化、性能与安全挑战。三种模式包括：通过 MCP 服务器提供 A2UI，利用 MCP Res

能力进展 监管/资本

https://developers.googleblog.com/a2ui-and-mcp-apps

4. WorkBuddy日活飙升至行业第二的3-4倍，非技术用户涌入

公众号：数字生命卡兹克 · 昨天 20:42

从3月至今，WorkBuddy日活用户数已达行业第二名的3-4倍，用户不再限于开发者，大量HR、运营、行政等非技术岗位也在使用。其企业版和项目功能进一步扩展了Agent办公场景。同期，Trae Work、QoderWork、Kimi Work等产品纷纷改名或出新，争夺市场。腾讯云认为这可能是十年一遇的机遇。

能力进展 基础设施

https://mp.weixin.qq.com/s/iAgbUzPrsbpqvu4s2XiH0g

5. 人工智能是否已经让自助类非虚构书籍销声匿迹了？

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 07:09

2026年Q1美国成人非虚构书籍销量同比下降9%，自助类下跌26.3%，仅手工艺/爱好/古董/游戏和宗教两个子类别增长。一位出版了多本《纽约时报》畅销书的作者透露，其五本书的印刷版年销量从2022年基准连续下滑：2023年-5%，2024年-13%，2025年-46%，2026年年化跌幅达-57% vs 2025年；若持续，2026年销量将比2022年减少约80%。所有格式在2025年下半年环比上

能力进展

https://tim.blog/2026/06/12/has-ai-already-killed-nonfiction

6. OpenAI 的领先优势正在快速缩小

Gary Marcus：The Road to AI We Can Trust（RSS） · 昨天 05:54

评论认为 OpenAI 正面临多重危机：缺乏护城河导致市场领先地位下滑；最大投资者微软持续疏远，近期甚至公开考虑将主要产品外包给中国；亏损速度远超预期，年亏损额以 8 倍增长。华盛顿方面可能打压 Anthropic，但也可能反而帮助其崛起，而 Elon Musk 成为另一个潜在的竞标者。

新发布

https://garymarcus.substack.com/p/openais-lead-is-dwindling-fast

7. Meta 解散工程部门引发热议

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 02:37

6月16日，一篇标题为"Why is Meta destroying its engineering organization? "的博客文章出现在 Hacker News，获得110个点赞。文章指出 Meta 正在解散其工程组织，引发业界广泛讨论。具体原因和后续影响尚未明确。

https://newsletter.pragmaticengineer.com/p/why-is-meta-destroying-its-engineering