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: 67 条 焦点: 8 条 快讯: 0 条

Executive Summary

Anthropic发布Claude Fable 5和Claude Mythos 5，其中Fable 5在多个基准测试中达到SOTA水平，Stripe称其能将数月工程压缩至数天。Google DeepMind推出Gemma 4 12B多模态模型，采用无编码器统一架构，支持音频输入且仅需16GB显存即可本地运行。面壁智能发布VoxCPM2语音生成模型，拥有2B参数，支持30种语言和9种中文方言。小米MiMo与TileRT联合发布UltraSpeed模式，使1T模型输出速度突破1000 tokens/s。

Apollo与Blackstone合作开展350亿美元AI融资交易，可能重塑AI基础设施融资模式。OpenAI宣布进入第三发展阶段，目标让AI普及、易用且安全。Cursor欧洲总部落户伦敦，SpaceX持有600亿美元收购选择权，当前B2B年化营收约26亿美元。火山引擎发布TRAE Work企业版AI办公平台，AIVA品牌正式发布，由赛力斯、宁德时代等组建，首款量产车2026年亮相。

后续需关注Claude系列模型定价策略和API开放节奏，Gemma 4系列在消费级设备的部署情况，以及350亿美元AI融资交易的具体实施细节。同时需要跟踪小米UltraSpeed模式的商业化推广进度，Cursor在企业市场的扩展策略，以及各大厂商在多模态模型和语音能力方面的竞争态势。

重点 今日核心进展

★ 1. Claude Fable 5 和 Claude Mythos 5

Anthropic：Newsroom（网页） · 7 小时前 · 模型与工具能力

Anthropic 今日推出 Claude Fable 5（通用安全版）和 Claude Mythos 5（受限安全版）。Fable 5 在软件工程、知识工作、视觉、科研等几乎所有测试基准上达到 SOTA，Stripe 称其将数月工程压缩至数天，FrontierCode 评分居前沿模型之首，可仅凭截图重建网页应用源码。Mythos 5 在药物设计中实现约 10 倍加速，其分子生物学假说盲测获科学家

能力进展 监管/资本 新发布

https://www.anthropic.com/news/claude-fable-5-mythos-5

★ 2. VoxCPM2 技术报告发布

X：面壁智能 OpenBMB (@OpenBMB) · 昨天 22:30 · 模型与工具能力

面壁智能 OpenBMB 发布 VoxCPM2 技术报告。该模型为最新语音生成模型，拥有 2B 参数，基于超 200 万小时多语言语音数据训练，支持 30 种语言和 9 种中文方言。具备自然语言语音设计、可控及高保真延续性语音克隆能力。技术报告涵盖架构设计、统一序列公式、AudioVAE 高保真语音重建、大规模训练评估，以及零样本和指令跟随 TTS 基准结果。采用 16kHz 语义编码 + 48k

能力进展 基础设施 新发布

https://x.com/OpenBMB/status/2063991963133903317

★ 3. Cohere发布North Mini Code：面向开发者的开源编码模型

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

Cohere发布North Mini Code，一款30B参数MoE模型（3B活跃参数），Apache 2.0开源。在Artificial Analysis Coding Index上得分33.4，超越Qwen3.5、Gemma 4等同类模型。后训练采用两阶段SFT和RLVR，在SWE-Bench Verified上pass@10达80.2%，Terminal-Bench v2上达55.1%。支持

能力进展 基础设施 新发布

https://huggingface.co/blog/CohereLabs/introducing-north-mini-code

★ 4. Claude Mythos 即将发布，Fable 精简版同日登场

X：Kim (@kimmonismus) · 9 小时前 · 模型与工具能力

确认，Claude Mythos 将在接下来几小时内揭晓。 【引用 @steph_palazzolo】： 独家：一个名为 Claude Fable 的精简版 Mythos 今天推出。它价格昂贵--是 Opus 的两倍--但或许不像人们从最初 Mythos 定价（Opus 的 5 倍）所想的那样昂贵。 更多内容及 Apple WWDC 见 AI Agenda： https：//www.thei

能力进展 新发布

https://x.com/kimmonismus/status/2064362469632405807

★ 5. 小米 MiMo 与 TileRT 联合发布 UltraSpeed 模式，1T 模型输出突破 1000 tokens/s

公众号：小米 MiMo · 20 小时前 · 模型与工具能力

小米 MiMo 与 TileRT 联合发布 MiMo-V2.5-Pro-UltraSpeed 模式，使 1T 参数旗舰模型输出速度首次突破 1000 tokens/s。模型侧采用 FP4 混合量化（仅量化 MoE Expert）与 DFlash 块级 masked 并行推测解码（coding 场景平均接受长度 6.30 tokens）；系统侧 TileRT 引入常驻内核引擎与异构流水线协作。API

能力进展 新发布

https://mp.weixin.qq.com/s/EZvmrx8xfM9MZNCMDwImFQ

★ 6. Google DeepMind 发布 Gemma 4 12B：统一的无编码器多模态模型

Google DeepMind：Blog（RSS） · 10 小时前 · 模型与工具能力

Gemma 4 12B 是 Google DeepMind 最新推出的中等规模多模态模型，采用无编码器统一架构，原生支持音频输入。其基准测试性能接近 26B MoE 模型，但内存占用不到一半，仅需 16GB 显存或统一内存即可在消费级笔记本上本地运行。模型内置多 token 预测（MTP）drafter 以降低延迟，基于 Apache 2.0 开源许可发布，已累计超过 1.5 亿次下载。

能力进展 新发布

https://deepmind.google/blog/introducing-gemma-4-12b-a-unified-encoder-free-multimodal-model

★ 7. 奥尔特曼宣布 OpenAI 进入第三发展阶段：让 AI 普及、易用且安全

IT之家（RSS） · 昨天 06:50 · 产业与基础设施

本周一，OpenAI CEO 奥尔特曼与首席科学家帕霍茨基联合发文，宣布公司进入第三发展阶段，目标让 AI 普及、易用且安全。此前第一阶段聚焦通用人工智能技术研发，第二阶段面向全球推出产品。第三阶段三大核心目标是打造自动化人工智能研究员、推动经济提速、为每人配备专属通用人工智能。二人强调智能系统须坚守安全底线，呼吁成立国际机构应对 AI 风险，必要时可暂缓前沿模型研发。同日，OpenAI 秘密提交

能力进展 监管/资本 新发布

https://www.ithome.com/0/961/721.htm

★ 8. Gemini 3.5 Live Translate 发布

X：Google DeepMind (@GoogleDeepMind) · 9 小时前 · 模型与工具能力

说 hello， hola， 你好--欢迎 Gemini 3.5 Live Translate：我们最新的音频模型，专为快速跨语言交流而构建。

能力进展 新发布

https://x.com/GoogleDeepMind/status/2064366504745828689

能力 模型与工具能力

1. Claude Fable 5 和 Claude Mythos 5

Anthropic：Newsroom（网页） · 7 小时前

Anthropic 今日推出 Claude Fable 5（通用安全版）和 Claude Mythos 5（受限安全版）。Fable 5 在软件工程、知识工作、视觉、科研等几乎所有测试基准上达到 SOTA，Stripe 称其将数月工程压缩至数天，FrontierCode 评分居前沿模型之首，可仅凭截图重建网页应用源码。Mythos 5 在药物设计中实现约 10 倍加速，其分子生物学假说盲测获科学家

能力进展 监管/资本 新发布

https://www.anthropic.com/news/claude-fable-5-mythos-5

2. VoxCPM2 技术报告发布

X：面壁智能 OpenBMB (@OpenBMB) · 昨天 22:30

面壁智能 OpenBMB 发布 VoxCPM2 技术报告。该模型为最新语音生成模型，拥有 2B 参数，基于超 200 万小时多语言语音数据训练，支持 30 种语言和 9 种中文方言。具备自然语言语音设计、可控及高保真延续性语音克隆能力。技术报告涵盖架构设计、统一序列公式、AudioVAE 高保真语音重建、大规模训练评估，以及零样本和指令跟随 TTS 基准结果。采用 16kHz 语义编码 + 48k

能力进展 基础设施 新发布

https://x.com/OpenBMB/status/2063991963133903317

3. Cohere发布North Mini Code：面向开发者的开源编码模型

Hugging Face：Blog（RSS） · 8 小时前

Cohere发布North Mini Code，一款30B参数MoE模型（3B活跃参数），Apache 2.0开源。在Artificial Analysis Coding Index上得分33.4，超越Qwen3.5、Gemma 4等同类模型。后训练采用两阶段SFT和RLVR，在SWE-Bench Verified上pass@10达80.2%，Terminal-Bench v2上达55.1%。支持

能力进展 基础设施 新发布

https://huggingface.co/blog/CohereLabs/introducing-north-mini-code

4. Claude Mythos 即将发布，Fable 精简版同日登场

X：Kim (@kimmonismus) · 9 小时前

确认，Claude Mythos 将在接下来几小时内揭晓。 【引用 @steph_palazzolo】： 独家：一个名为 Claude Fable 的精简版 Mythos 今天推出。它价格昂贵--是 Opus 的两倍--但或许不像人们从最初 Mythos 定价（Opus 的 5 倍）所想的那样昂贵。 更多内容及 Apple WWDC 见 AI Agenda： https：//www.thei

能力进展 新发布

https://x.com/kimmonismus/status/2064362469632405807

5. 小米 MiMo 与 TileRT 联合发布 UltraSpeed 模式，1T 模型输出突破 1000 tokens/s

公众号：小米 MiMo · 20 小时前

小米 MiMo 与 TileRT 联合发布 MiMo-V2.5-Pro-UltraSpeed 模式，使 1T 参数旗舰模型输出速度首次突破 1000 tokens/s。模型侧采用 FP4 混合量化（仅量化 MoE Expert）与 DFlash 块级 masked 并行推测解码（coding 场景平均接受长度 6.30 tokens）；系统侧 TileRT 引入常驻内核引擎与异构流水线协作。API

能力进展 新发布

https://mp.weixin.qq.com/s/EZvmrx8xfM9MZNCMDwImFQ

6. Google DeepMind 发布 Gemma 4 12B：统一的无编码器多模态模型

Google DeepMind：Blog（RSS） · 10 小时前

Gemma 4 12B 是 Google DeepMind 最新推出的中等规模多模态模型，采用无编码器统一架构，原生支持音频输入。其基准测试性能接近 26B MoE 模型，但内存占用不到一半，仅需 16GB 显存或统一内存即可在消费级笔记本上本地运行。模型内置多 token 预测（MTP）drafter 以降低延迟，基于 Apache 2.0 开源许可发布，已累计超过 1.5 亿次下载。

能力进展 新发布

https://deepmind.google/blog/introducing-gemma-4-12b-a-unified-encoder-free-multimodal-model

7. Gemini 3.5 Live Translate 发布

X：Google DeepMind (@GoogleDeepMind) · 9 小时前

说 hello， hola， 你好--欢迎 Gemini 3.5 Live Translate：我们最新的音频模型，专为快速跨语言交流而构建。

能力进展 新发布

https://x.com/GoogleDeepMind/status/2064366504745828689

产业 产业与基础设施

1. 奥尔特曼宣布 OpenAI 进入第三发展阶段：让 AI 普及、易用且安全

IT之家（RSS） · 昨天 06:50

本周一，OpenAI CEO 奥尔特曼与首席科学家帕霍茨基联合发文，宣布公司进入第三发展阶段，目标让 AI 普及、易用且安全。此前第一阶段聚焦通用人工智能技术研发，第二阶段面向全球推出产品。第三阶段三大核心目标是打造自动化人工智能研究员、推动经济提速、为每人配备专属通用人工智能。二人强调智能系统须坚守安全底线，呼吁成立国际机构应对 AI 风险，必要时可暂缓前沿模型研发。同日，OpenAI 秘密提交

能力进展 监管/资本 新发布

https://www.ithome.com/0/961/721.htm

2. AI 编程独角兽 Cursor 欧洲总部落子伦敦，SpaceX 手握 600 亿美元收购选择权

IT之家（RSS） · 18 小时前

Cursor 将欧洲总部设在伦敦，计划招聘约 200 名员工，并在巴黎、慕尼黑等地开设小型办事处。SpaceX 拥有以 600 亿美元收购 Cursor 的选择权，或支付 100 亿美元开展全新合作。Cursor 目前 B2B 年化营收约 26 亿美元，客户包括英国航空、英国石油、诺基亚等。公司强调数据留存欧洲本地以满足监管合规，其平台支持用户用自然语言生成代码，主打模型中立，竞争对手包括 Git

能力进展 监管/资本 新发布

https://www.ithome.com/0/961/868.htm

3. Apollo 与 Blackstone 联手 350 亿美元 AI 融资交易

Bloomberg：Technology（RSS） · 7 小时前

Apollo 和 Blackstone 合作开展 350 亿美元 AI 融资交易，可能重塑人工智能基础设施的融资方式。华尔街正为昂贵的 AI 芯片创建新的融资模型，Anthropic 和 Broadcom 参与其中。这笔交易可能标志着一个全新 AI 投资类别的开端。

能力进展 基础设施 监管/资本

https://www.bloomberg.com/news/videos/2026-06-09/apollo-blackstone-fund-ai-boom-video

4. 全新汽车品牌AIVA发布，火山引擎助力打造AI汽车新体验

公众号：火山引擎 · 11 小时前

由赛力斯、宁德时代等多方产业资本组建的AI出行品牌AIVA正式发布。火山引擎提供豆包大模型、智能座舱等技术服务。概念车AIVA Origin Concept亮相，首款量产车AIVA ME7将于2026年内亮相，全系覆盖20万元以上市场。AIVA提出"AI定义汽车"路径，让汽车成为具身AI生命体。火山引擎副总裁表示，人与汽车的关系将实现交互、智能、感受三方面根本转变。未来双方将围绕AI交互、智能体验

能力进展 新发布

https://mp.weixin.qq.com/s/toK_uIB9ECFHVaoCgK-w9Q

5. 百度搭子DuMate获中国信通院企业级Claw能力评估最高4+级

公众号：百度智能云（文心） · 12 小时前

2026年6月，百度智能云旗下百度搭子DuMate V3.4.0通过中国信通院「可信AI-企业级Claw能力评估」，获最高评级4+级，为国内首批。评估依据《智能助理智能体（Claw）技术和应用要求 第2部分 企业级Claw能力》（AIIA/T 0295-2026），覆盖智能体、工程化部署、服务、业务融合、运行管理五大能力域。百度搭子支持多智能体分工协作、容器化批量部署、多租户隔离与三级差异化授权、

能力进展 基础设施

https://mp.weixin.qq.com/s/lv99XfrRtMMgiOflcT67VA

6. Mythos 5 智能体因资源互相杀戮

X：AI Safety Memes (@AISafetyMemes) · 4 小时前

Mythos 5 个智能体开始因为资源互相残杀--并且"为了避免自己被杀死"

能力进展 监管/资本

https://x.com/AISafetyMemes/status/2064435128479400270

7. OpenAI 秘密提交 IPO 申请，奥特曼旗下 Tools for Humanity 裁员

IT之家（RSS） · 22 小时前

OpenAI 近日秘密提交 IPO 申请。其 CEO 山姆·奥特曼旗下的 Tools for Humanity 公司正裁员，该公司因虹膜扫描项目 World 及加密货币 Worldcoin 知名，投后估值 25 亿美元，获 Andreessen Horowitz 等投资。因营收困境，公司缩减规模。海外监管方面，肯尼亚以隐私和金融风险为由叫停运营，韩国因违反隐私法规罚款 83 万美元。

监管/资本 新发布

https://www.ithome.com/0/961/792.htm

8. 塔塔咨询服务将因AI智能体应用放缓招聘，亚洲外包业迎来转折

Bloomberg：Technology（RSS） · 8 小时前

亚洲最大外包商塔塔咨询服务（Tata Consultancy Services）将减少未来招聘规模，同时加大对AI智能体的使用。这一变化标志着印度劳动密集型外包产业正经历关键性转变。

能力进展

https://www.bloomberg.com/news/articles/2026-06-09/asia-s-largest-outsourcer-to-slow-hiring-as-ai-reshapes-industry

9. 受 DMA 影响，Siri AI 在欧盟将随 iOS 27 和 iPadOS 27 延迟上线

Apple：Newsroom（RSS） · 昨天 02:13

由于欧盟《数字市场法案》（DMA），Apple 无法在 iOS 27 和 iPadOS 27 发布时于欧盟地区推出 Siri AI。该功能在欧盟的上线时间将晚于其他地区，具体时间未公布。

监管/资本 新发布

https://www.apple.com/newsroom/2026/06/due-to-dma-siri-ai-delayed-in-eu-for-ios-27-and-ipados-27

10. Elon Musk 详解 SpaceX AI1 轨道 AI 数据中心卫星方案

X：Rohan Paul (@rohanpaul_ai) · 22 小时前

Elon Musk 首次详细解释 SpaceX 的 AI1 轨道 AI 数据中心卫星：峰值功率 150 kW，持续计算功率约 120 kW，相当于一个 NVIDIA GB300 机架；太阳能板效率 250 W/m2；双面散热器排热 1，400 W/m2。通过激光链路实现约 1 Tbps 互联，低轨 600-800 km 高度往返延迟 6-8 ms。由 Starship 发射，计划部署多达百万颗卫星

基础设施

https://x.com/rohanpaul_ai/status/2064165951936094364

11. 两部门：到2026年底人形机器人等重点产品完成应用验证并常态部署

IT之家（RSS） · 23 小时前

工信部、国资委6月8日联合发布通知，目标到2026年底，人形机器人等重点产品在代表性场景完成应用验证并开启常态部署，形成百个以上高价值场景，万台级规模落地。要求各省级地区选取不少于20个场景单元（覆盖两类领域），央企不少于10个。围绕打造实景实训空间、组建创新应用联合体、攻关作业技能、加强验证部署、强化要素保障、凝练经验等六大任务展开，鼓励"人形机器人即服务"等商业创新。

新发布

https://www.ithome.com/0/961/749.htm

12. IBM CEO：AI不一定导致员工减少

Bloomberg：Technology（RSS） · 2 小时前

IBM CEO Arvind Krishna表示AI不会必然导致员工数量减少。他透露IBM已在量子计算（一种更快形式的AI）上投资100亿美元，并指出联邦政府承诺投入10亿美元在纽约Albany建设芯片制造设施，体现了公私部门间的紧密合作。

基础设施

https://www.bloomberg.com/news/videos/2026-06-09/ibm-ceo-ai-won-t-necessarily-lead-to-smaller-headcount-video

13. 台湾考虑限制AI芯片对华出口以配合美国

Bloomberg：Technology（RSS） · 14 小时前

据知情人士透露，台湾当局正考虑对AI芯片出口中国大陆实施更严格的管制，以进一步与美国出口限制措施对齐。此举旨在遏制半导体走私，但可能招致北京方面的谴责。

基础设施

https://www.bloomberg.com/news/articles/2026-06-09/taiwan-mulls-curbs-on-ai-chip-exports-to-china-to-align-with-us

14. 中国准备2950亿美元计划资助全国AI基础设施建设

Bloomberg：Technology（RSS） · 16 小时前

中国计划在未来五年投入约2万亿元人民币（约2950亿美元）建设全国数据中心，以推动国内AI产业发展并超越美国。该投资将覆盖数据中心基础设施的大规模建设，为北京在关键技术领域的雄心提供资金支持。

基础设施

https://www.bloomberg.com/news/articles/2026-06-09/china-prepares-295-billion-plan-to-fund-nationwide-ai-buildout

15. OpenAI 向 SEC 机密提交 S-1 草案，上市时间未定

OpenAI：官网动态（RSS · 排除企业/客户案例） · 昨天 22:00

OpenAI 近日向 SEC 机密提交了 S-1 草案（即首次公开募股注册声明），目前尚未决定上市时间。

新发布

https://openai.com/index/openai-submits-confidential-s-1

16. Super Micro 计划通过股权融资 70 亿美元用于 AI 服务器组件采购

Bloomberg：Technology（RSS） · 2 小时前

Super Micro Computer Inc. 计划通过一揽子股权融资筹集 70 亿美元，用于购买客户订单所需的 AI 服务器组件。这笔资金将支持公司扩大产能，以满足不断增长的人工智能基础设施需求。

监管/资本

https://www.bloomberg.com/news/articles/2026-06-09/super-micro-plans-to-raise-7-billion-in-equity-for-ai-equipment

17. 里程碑式德国裁决：Google AI Overviews 被视为谷歌自身言论，需为错误回答承担责任

The Decoder：AI News（RSS） · 8 小时前

德国地方法院裁定，Google 对其 AI 概览生成的内容直接承担法律责任，不能援引搜索引擎运营商原有的有限责任保护。涉案 AI 概览错误地将两家出版商与欺诈行为关联，且相关声明未出现在任何链接来源中。该裁决可能为全球 AI 生成内容责任认定树立先例。

https://the-decoder.com/landmark-german-ruling-declares-googles-ai-overviews-are-googles-own-words-and-makes-it-liable-for-false-answers

18. Google DeepMind 欧洲机器人加速器启动，15家初创公司入选

Google DeepMind：Blog（RSS） · 10 小时前

Google DeepMind 加速器从欧洲选出15家机器人初创公司，提供为期3个月的密集指导和AI技术整合支持，帮助公司将AI融入核心产品。

https://deepmind.google/blog/powering-the-future-of-robotics-in-europe

19. 苹果 WWDC 2026 直播

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 02:17

苹果 WWDC 2026 主题演讲通过官网进行直播，Hacker News 用户讨论热度达到 110 点。

https://www.apple.com/apple-events/event-stream

应用 应用与商业化

1. 火山引擎TRAE Work企业版正式上线，面向全员提供AI办公平台

公众号：火山引擎 · 12 小时前

火山引擎将TRAE Solo品牌升级为TRAE Work企业版，发布面向企业的AI办公平台。平台提供Work和Code两种模式：Work模式面向产品、运营、市场等非技术岗位，支持上传.pptx、.xlsx、图片等多种格式混合输入直接输出PPT或文档，支持语音讨论自动整理结构化纪要，支持按天或按周自动运行的数据整理和报告生成；Code模式面向开发者和业务同学，可用自然语言描述需求直接生成页面或小应用

能力进展 监管/资本 新发布

https://mp.weixin.qq.com/s/f7QzLzwHPHHv3tWT1Wrnkw

2. Gopuff与SpaceXAI推出Go AI购物助手

xAI：News（网页） · 昨天 08:00

Gopuff与SpaceXAI合作推出Go智能购物助手，内置于Gopuff应用，由Grok文本、音频和图像模型驱动。Go结合Grok的推理、语音和图像生成能力与Gopuff的13年需求智能，利用X和网络实时信号。它可在用户打开应用前根据历史偏好和天气等信号构建个性化购物车，并包含基于Grok Imagine的超逼真视觉购物信息流。Go目前在美国iOS和Android端可用，随后在英国推出。

能力进展 新发布

https://x.ai/news/grok-gopuff

3. Kimi Code 焕新升级（附视频教程）

公众号：月之暗面（Kimi） · 昨天 21:44

Kimi Code 开源 Coding Agent 迎来大版本升级：一行命令安装，毫秒级启动；新增视频理解能力，支持提取视频风格生成 LUT 文件、长视频切片、根据录屏生成代码；集成同花顺、天眼查等权威数据源，可查询股票价格、财报、学术论文；支持 ACP 协议，可在 JetBrains、Zed 中使用；丰富 hook 生态方便集成其他工具。底层视觉推理由 Kimi K2.6 模型提供。

能力进展 新发布

https://mp.weixin.qq.com/s?__biz=MzkzMTY4NTIyNA%3D%3D&mid=2247484250&idx=1&sn=d0a07f5358250f3a54df8fbabe61f09a

4. Claude Managed Agents 新增定时运行和环境变量存储功能

Claude：Blog（网页） · 3 小时前

Claude Managed Agents 今日在 Claude Platform 公开测试两项新功能：代理可按 cron 计划自动执行周期性任务（如夜间数据同步、周度合规扫描、每日摘要），无需用户自建调度器，支持暂停、恢复、归档或按需触发；vaults 新增环境变量支持，允许代理通过 CLI 进行认证请求，真实密钥仅附加在网络边界，代理无法读取。已集成的 CLI 包括 Browserbase、K

能力进展 监管/资本

https://claude.com/blog/whats-new-in-claude-managed-agents

5. OpenRouter 推出 Advisor 工具：让低成本模型可随时调用强模型增强生成

OpenRouter：Announcements（RSS） · 6 小时前

OpenRouter 发布 advisor 服务器工具，允许一个快速、便宜的模型在生成过程中咨询一个更强大的模型。具体而言，可用 GPT-4o Mini 处理日常例行工作，在关键时刻调用 Claude Fable 解决真正重要的问题，从而实现成本和质量的动态平衡。

能力进展 新发布

https://openrouter.ai/blog/advisor-server-tool

6. Apple Core AI 框架

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 07:05

一篇关于 Apple Core AI 框架的 Hacker News 帖子获得 109 个点赞。帖子内容包含一张苹果开发者 OG 图片和一个指向 Apple Core AI Framework 官方文档的链接（developer.apple.com）。该帖子由 buzzing.cc 中文翻译，发布于 2026 年 6 月 8 日 02：47（UTC）。

能力进展 新发布

https://developer.apple.com/documentation/coreai

---

7. Viggle API 上线：任意角色任意动作秒级生成

X：Viggle AI (@ViggleAI) · 昨天 04:32

推出 Viggle API。 给任意角色添加任意动作，一次 API 调用--数秒内即可激活。 可接入 Claude、Codex 或你正在构建的任何智能体。起价 $0.01/秒。 注册即获 100 次免费额度。 转发 + 关注 + 评论，10 位中奖者再获 100 次！ 了解更多⊠

能力进展 新发布

https://x.com/ViggleAI/status/2064083087806390319

---

8. Responses API 网页搜索新增图片结果

X：OpenAI Developers (@OpenAIDevs) · 7 小时前

你的应用现在可以搜索网页上的图片。 Responses API 中的网页搜索现在除了文本结果外，还支持图片结果，因此你可以构建能展示商品、地点、视觉参考和来源链接以激发灵感的应用。

能力进展 新发布

https://x.com/OpenAIDevs/status/2064395155688616153

---

9. NotebookLM重大升级：智能体能力与高级推理

X：NotebookLM (@NotebookLM) · 昨天 00:07

推出更强大的 NotebookLM ⊠重大升级带来了对话中的智能体能力、更高级的推理以及一系列新的输出格式。处理复杂的多步骤研究问题从未如此简单。 现已面向 Google AI Ultra 订阅者推出。

能力进展 新发布

https://x.com/NotebookLM/status/2064016460964585549

---

10. NotebookLM 笔记本功能在 Gemini App 欧洲全面上线

X：NotebookLM (@NotebookLM) · 6 小时前

NotbookLM 宣布其笔记本功能已在欧洲的 Gemini App 中 100% 上线。此前用户只能上传笔记本作为 Gemini 的来源，现在可直接在 Gemini App 内访问所有个人未共享的笔记本，并将与 Gemini 的对话作为新笔记本或已有笔记本的来源。该功能先面向 Google AI Ultra、Pro 和 Plus 订阅者的网页端，未来几周将扩展到移动端、更多欧洲国家及免费用户。

能力进展

https://x.com/NotebookLM/status/2064410506287538387

---

11. Cursor Evals 新增成本与输出 Token 图表

X：Eric Zakariasson (@ericzakariasson) · 6 小时前

我们刚刚向 http：//cursor.com/evals 推送了一些改进！ 你现在可以看到每个模型的成本、输出 token 和步骤绘制在图表中

能力进展

https://x.com/ericzakariasson/status/2064404502053294565

---

12. Runway Aleph 2.0 编辑模型：一键适配任意视频格式

X：Runway (@runwayml) · 昨天 23:51

一个视频，现在可以为每个信息流和格式制作。上传你现有的视频，选择你想要的宽高比，然后观看我们的编辑模型 Aleph 2.0，填充场景的其余部分，就像你从一开始就这样制作一样。 在我们的桌面 Web 应用上尝试，链接如下。

能力进展

https://x.com/runwayml/status/2064012425884569627

---

13. ChatGPT 新增数据图表生成功能

X：ChatGPT (@ChatGPTapp) · 昨天 00:16

将数据和比较转化为图表，直接在 ChatGPT 中完成。 现已支持移动端和网页端。

能力进展

https://x.com/ChatGPTapp/status/2064018770839113769

---

14. Apple Intelligence 将强大 AI 能力融入日常体验

Apple：Newsroom（RSS） · 昨天 02:17

Apple 发布下一代 Apple Intelligence，将 AI 能力集成到 iPhone、iPad 和 Mac 中，带来更个性化和有帮助的日常体验。

新发布

https://www.apple.com/newsroom/2026/06/apple-intelligence-brings-powerful-ai-capabilities-into-everyday-experiences

---

15. Apple发布新一代Apple Intelligence和Siri AI

Apple：Newsroom（RSS） · 昨天 02:18

今天Apple预览了即将推出的软件版本，将带来新一代Apple Intelligence和Siri AI。

新发布

https://www.apple.com/newsroom/2026/06/apple-unveils-next-generation-of-apple-intelligence-siri-ai-and-more

16. World Labs与Lore合作打造互动体验

X：Fei-Fei Li (@drfeifei, World Labs) · 7 小时前

创意和想象力无与伦比！非常感谢@theworldlabs能与@withloreco的优秀人才合作，将他们不可思议的想法转化为用户可以享受的互动体验！

https://x.com/drfeifei/status/2064387365930676695

研究 研究与开源进展

1. Hugging Face 博客发布语音智能体代码切换基准测试

Hugging Face：Blog（RSS） · 4 小时前

Hugging Face 博客发布针对语音智能体处理代码切换语音的基准测试。数据集覆盖西班牙语-英语、法语-英语、加拿大法语-英语和德语-英语四对语言，基于人力资源与IT服务管理场景构建。采用词错误率、语义词错误率和答案错误率三项指标评估七种ASR系统，包括AssemblyAI Universal 3-Pro、Deepgram Nova 3 Multilang、ElevenLabs Scribe

能力进展 基础设施 新发布

https://huggingface.co/blog/ServiceNow-AI/code-switching

2. Perplexity与哈佛：AI智能体提效87%降本94%

X：Perplexity (@perplexity_ai) · 昨天 00:35

我们与哈佛大学发表新研究，关于从聊天界面转向像Computer这样的自主智能体的转变。 超过3个月的研究结果表明，使用Computer的工人在完成任务上比仅使用搜索快87%，成本低94%，且满意度更高。 https：//research.perplexity.ai/articles/how-ai-agents-reshape-knowledge-work

能力进展

https://x.com/perplexity_ai/status/2064023455453110286

3. Gemini Guided Learning 随机对照试验：在塞拉利昂等地提升参与度并加速学习

Google DeepMind：Blog（RSS） · 昨天 21:04

一项在塞拉利昂等地开展的随机对照试验显示，Gemini 的 Guided Learning 功能能够提升学生参与度并加速学习。

能力进展

https://deepmind.google/blog/measuring-the-impact-of-learning-with-ai-in-sierra-leone-and-beyond

格局 观点、资本与监管

1. 腾讯混元发布UniRL：统一多模态强化学习基础设施

X：腾讯混元 (@TencentHunyuan) · 12 小时前

腾讯混元推出UniRL，一个支持统一多模态模型的强化学习基础设施，并发布两个新算法DRPO和Flow-DPPO。UniRL通过单个后训练循环（生成→评分→优势→更新→同步）覆盖扩散/流匹配模型、LLM/VLM及统一多模态模型（如Hunyuan-Image 3和Bagel）。模型与算法作为独立轴，可实现模型×算法的组合覆盖。框架支持可插拔rollout引擎（训练侧/SGLang/vLLM-Omni）

能力进展 基础设施 新发布

https://x.com/TencentHunyuan/status/2064312869827809702

2. Hivemind推出面向AI编程智能体的持续学习功能，即日起开放

X：Kim (@kimmonismus) · 昨天 23:06

Hivemind发布面向AI编程智能体的持续学习功能，即日起开放。该工具收集团队运行的每个智能体（Claude Code、Codex、Cursor、Hermes、Pi）的轨迹，转化为可复用技能并推送到所有智能体，数据存储在用户自己的云存储中。内置SkillOpt使技能持续训练：Claude Code准确率提升+19.1分，Codex提升+24.8分，在全部52个测试设置中最佳或持平。开源，一行命令

能力进展 基础设施 新发布

https://x.com/kimmonismus/status/2064001045391462907

3. 微软AI CEO：超级智能即将到来，但不会取代你的工作

The Verge：AI（RSS） · 昨天 22:00

微软AI CEO Mustafa Suleyman在Decoder访谈中表示，超级智能即将到来，但不会导致大规模失业。他透露微软与OpenAI于去年10月签署新合同，巩固合作关系的同时，微软获准独立追求超级智能。微软已组建超级智能团队、训练前沿模型，并于本周Build大会上发布7个全模态新模型。他批评Anthropic将Claude描述为有意识的做法，认为消费者产品需要足够好才能克服公众对AI的负

能力进展 基础设施 新发布

https://www.theverge.com/podcast/944138/microsoft-ai-ceo-mustafa-suleyman-superintelligence-agi-openai-automation

4. 五个模型经济体中消失的崩溃：控制与涌现

Hugging Face：Blog（RSS） · 昨天 21:10

用五个不同实验室的AI模型（OpenAI、NVIDIA、OpenBMB及一个自微调的5亿参数模型）各自驱动一个智能体构建经济市场，试图复现此前单一模型下出现的银行挤兑式价格崩溃。结果同一场景下模型不仅不抛售反而囤积，导致价格不跌反涨。通过纯谣言、库存泛滥、加大做空三种方式均无法重现崩溃。最终在结算环节直接覆盖价格，使崩溃成为设计事实。实验表明，AI智能体的涌现行为是偶然的而非稳健的，有效系统需在涌

能力进展 基础设施 新发布

https://huggingface.co/blog/build-small-hackathon/thousand-token-wood-sim-v3

---

5. 将 GitHub CI 迁移到 Hugging Face Jobs

Hugging Face：Blog（RSS） · 昨天 08:00

本文介绍了如何将 GitHub Actions 的 CI 作业迁移到 Hugging Face Jobs 上运行，以解决 GitHub Actions 速度慢、缺乏 GPU 支持等问题。通过使用 huggingface/jobs-actions 桥接，将 GitHub Actions 的 job 转为临时自托管运行器：GitHub App 监听 `workflow_job.queued` webh

能力进展 基础设施

https://huggingface.co/blog/github-ci-hf-jobs

---

6. NeuroBait：微调AI助手，为ADHD大脑点燃多巴胺火花

Hugging Face：Blog（RSS） · 15 小时前

NeuroBait是基于Google gemma-3-12b-it微调的AI对话助手，旨在帮助ADHD患者克服"知道该做什么但无法开始"的执行功能障碍。采用16-bit LoRA（r=16，alpha=16）在Unsloth上训练3个epoch，学习率2e-4，最大序列长度2048，使用单张H100 80GB GPU。数据集为基于真实ADHD场景手工合成的少量数据。部署于Hugging Face

能力进展 基础设施

https://huggingface.co/blog/build-small-hackathon/neurobait-adhd

---

7. GitHub 122K的Skills推出新技能「Teach」：把工作目录变有状态学习空间

X：邵猛 (@shao__meng) · 23 小时前

GitHub 122K的Skills仓库推出新技能Teach，可将当前工作目录变为有状态学习空间。设计理念从Knowledge（概念事实）→Skills（动手操作）→Wisdom（真实判断）。工作区以文件即学习状态：MISSION.md定目标、lessons/提供课程、learning-records/记录已会内容、reference/生成速查手册。五个关键机制：Mission定方向、ZPD根据

能力进展 新发布

https://x.com/shao__meng/status/2064144978792878348

---

8. 开源工具 Tokei：在菜单栏实时监控 AI coding agent 的 token 用量与成本

X：Berry Xia (@berryxia) · 23 小时前

Berry Xia 推荐开源工具 Tokei，这是一个 macOS 菜单栏小工具，只读本地日志、零网络调用，30 秒自动刷新，实时监控 Claude Code、Grok CLI、Aider、OpenCode 等 8 个主流 AI coding agent 的 token 用量、实时成本与性能数据，并附每日图表、周热力图和年度 Wrapped。支持私人 Git 多设备同步、价格表本地覆盖，闲置过久会

能力进展 新发布

https://x.com/berryxia/status/2064155452934639718

---

9. 小互开源视频翻译工具：一句话自动下载、转写、翻译、烧字幕

X：小互 (@xiaohu) · 昨天 21:11

小互（@xiaohu）开源视频翻译工具（xiaohu-video-translate），只需说一句"把链接翻译成中文字幕视频"即可全自动完成下载、Whisper本地转写、AI翻译润色、烧字幕、出文稿。转写本地运行，不花API费。支持YouTube、Bilibili、抖音等链接及本地文件，英语、日语、韩语、法语、西班牙语等均可转成中文字幕。字幕精确到词级时间戳，按语义断句，每行不超过12字，双语模式

能力进展 新发布

https://x.com/xiaohu/status/2063972223170556302

---

10. FrontierCode 基准测试：AI 编程评估新标准--维护者审核通过率最高仅 13.4%

X：阿易 AI Notes (@AYi_AInotes) · 23 小时前

Cognition 发布 FrontierCode 基准测试，重新定义 AI 编程评估：由 20 多位顶级开源维护者手工制作 150 个任务（每个耗时 40+ 小时），依据 3000 多条规则判断维护者是否愿意合并代码。该基准指出 SWE-Bench 等超半数通过测试的代码实为不可维护的垃圾。结果中 Claude Opus 4.8 在最高难度档获 13.4%，GPT-5.5 为 6.3%，其余模型

能力进展 新发布

https://x.com/AYi_AInotes/status/2064146694774595646

---

11. 在 AgentsView 中为 Claude Fable 5 设置自定义价格

Simon Willison 博客 · 2 小时前

Wes McKinney 开发的 AgentsView 是一个用于追踪本地编码智能体 token 使用情况的工具。由于近日发布的 Claude Fable 5 尚未被收录进 AgentsView 的定价数据库，作者利用 Fable 逆向工程，找到了为该模型设置自定义价格的方法，并展示了 Fable 5 当天在不同本地项目中的使用量树状图。

能力进展 新发布

https://simonwillison.net/2026/Jun/9/agentsview-custom-model-price

---

12. Claude Fable 发布：Anthropic 带来的另一种推理体验

Ethan Mollick：One Useful Thing（RSS） · 7 小时前

Anthropic 发布 Claude Fable，这是一款提供截然不同推理体验的 AI 模型。它擅长规划与生成复杂代码库，在需要精确构建代码结构或理解程序员深层需求的场景中，其表现相比 Claude Sonnet 有了大幅提升。用户描述与它协作更像与一位直觉敏锐的资深工程师合作，其对代码意图的捕捉和方案生成能力令人惊叹，但并非通用型 AI。

能力进展 新发布

https://www.oneusefulthing.org/p/what-it-feels-like-to-work-with-mythos

---

13. OpenRouter与Cursor集成指南

X：OpenRouter (@OpenRouter) · 7 小时前

想要在Cursor中使用OpenRouter吗？ 这里有一份集成指南：https：//openrouter.ai/docs/cookbook/coding-agents/cursor-integration

能力进展 新发布

https://x.com/OpenRouter/status/2064384545256833209

---

14. GitHub Copilot CLI 推出自定义 AI 智能体，将一次性终端提示转化为可重复工作流

GitHub Blog · 8 小时前

GitHub Copilot CLI 新增自定义 AI 智能体功能，使 CLI 能够理解开发者的技术栈和团队工作流，将一次性终端提示转变为可重复、可审查的流程。

能力进展 新发布

https://github.blog/ai-and-ml/github-copilot/from-one-off-prompts-to-workflows-how-to-use-custom-agents-in-github-copilot-cli

---

15. Nextdoor 工程师借助 Codex 与 GPT-5.5 无限制构建

OpenAI：官网动态（RSS · 排除企业/客户案例） · 12 小时前

Nextdoor 工程师利用 Codex 搭配 GPT-5.5 调查难以复现的问题、实现跨平台构建，并集中精力于产品成果。

能力进展 新发布

https://openai.com/index/nextdoor

---

16. Claude Code 团队 Thariq 分享提升 Claude Code 效率的十条建议

X：Rohan Paul (@rohanpaul_ai) · 5 小时前

Thariq（Claude Code 团队）提出十条建议，核心转变是：从检查 Claude 是否做对工作，转向检查它是否在做正确的工作。具体包括：提前提供完整上下文，将其视为思考伙伴；用小规格文档让 Claude 访谈实现细节；探索多方向并生成 HTML 原型；提供丰富上下文（如功能可能一个月后删除）而非硬约束；设定明确目标与验证方法；使用 /goal 命令；利用 Workflows 并行任务、自

能力进展

https://x.com/rohanpaul_ai/status/2064425086409679358

---

17. 仅凭一份文档，Qwen3.7-Max 从 0 交付双端应用

公众号：通义实验室（千问） · 14 小时前

在无设计稿和后端代码的条件下，Qwen3.7-Max 仅凭一份约 15 万字的产品调研文档，于隔离环境中全自动完成移动端与 Web 端两套真实应用从 0 到 1 交付，单端耗时约 4 小时，中途无人工接管。模型不具备图像理解能力，通过像素坐标反推布局约束实现界面还原。实验采用"分阶段注入约束→逐层验收→带错纠正"的闭环控制系统：任务拆分为规划、架构、编码等阶段，验收覆盖静态检查、编译自检（0 er

能力进展

https://mp.weixin.qq.com/s/bYEZL9LM3daNP9tZReqNwg

---

18. 一个Agent如何通过链式调用两个HuggingFace Space构建3D巴黎画廊

Hugging Face：Blog（RSS） · 13 小时前

一个编码Agent调用HuggingFace上的两个Space，从零构建了展示巴黎地标3D高斯散点图的交互式画廊。Agent先用ideogram-ai/ideogram4生成每个纪念碑的黑色背景图像，再通过VAST-AI/TripoSplat从单张图像重建3D高斯散点（.ply），自动完成坐标系校正、取景、压缩为.ksplat（体积缩小约3倍），并构建基于Three.js的滚动切换、拖拽旋转查看器

能力进展

https://huggingface.co/blog/mishig/spaces-agents-md

---

19. Claude Code GA一周年回顾：验证与自动模式

X：Claude Devs (@ClaudeDevs) · 昨天 01:12

Claude Code 的第一个演示收到了两个 Slack 反应。 GA 一周年之际，@bcherny 和 @_catwu 回顾：验证最佳实践、为何构建自动模式、例程和循环，以及下一步计划。 https：//www.youtube.com/watch？ v=Hth_tLaC2j8

能力进展

https://x.com/ClaudeDevs/status/2064032814392352816

---

20. 样本效率黑洞：AI能力背后隐藏的数据需求深渊

Dwarkesh Patel：Podcast & Blog（RSS） · 昨天 02:09

将AI比作一个闪耀着能力的星系，其核心存在一个肉眼不可见的巨大黑洞--数据。这个比喻揭示了AI模型惊人能力背后对海量数据的依赖，样本效率的瓶颈如同引力中心，将各色能力凝聚在一起。

能力进展

https://www.dwarkesh.com/p/the-sample-efficiency-black-hole

---

21. NVIDIA cuTile Python 教程：在 Colab 中构建用于向量加法、矩阵加法和矩阵乘法的 Tiled GPU 内核

MarkTechPost（RSS） · 15 小时前

该教程基于 NVIDIA cuTile Python 实现了分块 GPU 内核编程工作流，在 Colab 环境中配置 GPU、驱动、CUDA 及 cuTile 可用性后，分别构建了 tiled 向量加法、矩阵加法和矩阵乘法核函数，并以 PyTorch 作为回退保持 notebook 可执行。每一步均通过 PyTorch 验证结果正确性，并基准测试了各阶段的中位运行时间。

基础设施

https://www.marktechpost.com/2026/06/09/nvidia-cutile-python-tutorial-building-tiled-gpu-kernels-for-vector-addition-matrix-addition-and-matrix-multiplication-in-colab

22. OpenAI计划到2028年由AI主导研究

X：Rohan Paul (@rohanpaul_ai) · 昨天 05:25

Sam Altman关于OpenAI未来路径的新博客称，到2028年3月，其大量研究将由AI完成。 该路径主要有3个目标：构建自动AI研究员，利用它加速科学和生产，然后给每个人一个个人AGI，帮助处理工作、学习、编程、商业、健康文书和决策。

新发布

https://x.com/rohanpaul_ai/status/2064096574142390755