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: 0s 精选条目: 59 条 焦点: 8 条 快讯: 0 条

Executive Summary

DiffusionGemma的发布标志着扩散模型架构的重要突破，采用并行生成技术实现文本生成速度提升4倍，26B MoE模型仅激活3.8B参数，量化后适配18GB显存消费级GPU。Cohere发布的North Mini Code 30B参数MoE模型在编码基准测试中超越Qwen3.5、Gemma 4等同类产品。Google DeepMind同时推出Gemma 4 12B多模态模型，采用无编码器统一架构支持音频输入，仅需16GB显存即可本地运行。Cursor Bugbot更新实现运行速度提升超3倍、成本降低22%的重大优化。

基础设施层面，Oracle与OpenAI合作使用户可通过现有云承诺额度访问OpenAI模型，摩尔线程开源的MusaCoder成为首个基于国产GPU全链路训练的代码模型。商业化格局方面，Apollo与Blackstone联合开展350亿美元AI融资交易，可能重塑AI基础设施融资模式。欧盟要求Meta向第三方AI助手免费开放WhatsApp的临时措施将影响即时通讯AI应用的竞争格局。华为云发布全球首个端到端具身AI平台CloudRobo，火山方舟推出覆盖全链路的版权商业化平台。

后续需关注Claude Mythos正式发布后的定价策略及其对高端模型市场竞争的影响，DiffusionGemma等新型扩散模型的产业化应用进展，以及国产GPU训练模型在实际部署中的性能表现。AI安全领域中模型快速构建漏洞利用的能力将推动传统安全补丁机制的变革，大型科技公司数据收集政策的调整也将影响AI训练数据的获取模式。

重点 今日核心进展

★ 1. DiffusionGemma：文本生成速度提升4倍的开源扩散模型

Google DeepMind：Blog（RSS） · 7 小时前 · 模型与工具能力

Google DeepMind 发布开源实验模型 DiffusionGemma，采用文本扩散技术，突破自回归逐 token 生成方式，每次前向并行生成 256 个 token。该 26B MoE 模型推理时仅激活 3.8B 参数，量化后适配 18GB 显存消费级 GPU。在 H100 上达 1000+ tokens/s，RTX 5090 上 700+ tokens/s，速度提升 4 倍。具备双向注

能力进展 基础设施 新发布

https://deepmind.google/blog/diffusiongemma-4x-faster-text-generation

★ 2. 摩尔线程开源 MusaCoder 代码大模型，9B/27B 参数基于国产 GPU 全链路训练

IT之家（RSS） · 16 小时前 · 模型与工具能力

摩尔线程发布并开源 MusaCoder 代码大模型，含 9B 和 27B 两个参数规模，是业内首个基于国产 GPU 算力底座完成全链路训练与验证的开源模型。后训练流程在基于 MTT S5000 的夸娥智算集群上完成，支持从 PyTorch 标准算子自动生成高性能 CUDA/MUSA 原生 Kernel 代码。在 KernelBench 评测中，MusaCoder-27B-RL 以 Overall

能力进展 基础设施 新发布

https://www.ithome.com/0/962/509.htm

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

Hugging Face：Blog（RSS） · 昨天 23:56 · 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. Cursor Bugbot 更新：速度提升超 3 倍、成本降低 22%、发现更多 Bug

Cursor Blog · 12 小时前 · 应用与商业化

Cursor 的代码审查工具 Bugbot 迎来重大更新：运行速度提升超 3 倍，成本降低 22%，每轮审查多发现 10% 的 bug，90% 的运行在三分钟内完成。新增 `/review` 命令，可在推送代码前运行 Bugbot 和安全审查，并与 GitHub/GitLab 同步--若已通过 `/review` 审查过同一 diff，打开 PR 时 Bugbot 会自动跳过并备注。支持配置仅审查

能力进展 基础设施 监管/资本

https://cursor.com/blog/bugbot-updates-june-2026

★ 5. 通过 Oracle 云承诺访问 OpenAI 模型和 Codex

OpenAI：官网动态（RSS · 排除企业/客户案例） · 4 小时前 · 产业与基础设施

OpenAI 与 Oracle 合作，用户可利用现有 Oracle 云服务承诺（commitment）额度，在 Oracle 云上访问 OpenAI 模型和 Codex，用于构建和部署 AI 应用，同时获得企业级安全与治理能力。

能力进展 基础设施 监管/资本

https://openai.com/index/openai-on-oracle-cloud

★ 6. Claude Mythos 即将发布，Fable 精简版同日登场

X：Kim (@kimmonismus) · 昨天 23:02 · 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

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

Google DeepMind：Blog（RSS） · 昨天 22:10 · 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

★ 8. Google将保存用户的Lens图片、Search Live录音和Translate音频用于AI训练

The Verge：AI（RSS） · 8 小时前 · 应用与商业化

Google更新搜索交互数据保存方式，新增"Search Services History"设置，用于保存用户搜索时使用的图片、文件、音频和视频，包括Google Lens搜索的图片、实时搜索工具Search Live的录音、语音搜索和Translate中的语音片段。这些数据将被用于"提供、改进和开发AI模型"。用户可关闭该设置并禁用"Save Media"选项以避免保存。

能力进展 基础设施 新发布

https://www.theverge.com/tech/947836/google-search-privacy-settings-images-audio

能力 模型与工具能力

1. DiffusionGemma：文本生成速度提升4倍的开源扩散模型

Google DeepMind：Blog（RSS） · 7 小时前

Google DeepMind 发布开源实验模型 DiffusionGemma，采用文本扩散技术，突破自回归逐 token 生成方式，每次前向并行生成 256 个 token。该 26B MoE 模型推理时仅激活 3.8B 参数，量化后适配 18GB 显存消费级 GPU。在 H100 上达 1000+ tokens/s，RTX 5090 上 700+ tokens/s，速度提升 4 倍。具备双向注

能力进展 基础设施 新发布

https://deepmind.google/blog/diffusiongemma-4x-faster-text-generation

2. 摩尔线程开源 MusaCoder 代码大模型，9B/27B 参数基于国产 GPU 全链路训练

IT之家（RSS） · 16 小时前

摩尔线程发布并开源 MusaCoder 代码大模型，含 9B 和 27B 两个参数规模，是业内首个基于国产 GPU 算力底座完成全链路训练与验证的开源模型。后训练流程在基于 MTT S5000 的夸娥智算集群上完成，支持从 PyTorch 标准算子自动生成高性能 CUDA/MUSA 原生 Kernel 代码。在 KernelBench 评测中，MusaCoder-27B-RL 以 Overall

能力进展 基础设施 新发布

https://www.ithome.com/0/962/509.htm

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

Hugging Face：Blog（RSS） · 昨天 23:56

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) · 昨天 23:02

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

Google DeepMind：Blog（RSS） · 昨天 22:10

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

6. Gemini 3.5 Live Translate 发布

X：Google DeepMind (@GoogleDeepMind) · 昨天 23:18

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

7. Grok Voice性能出色价格低廉

X：xAI (@xai) · 5 小时前

Grok Voice 提供最先进的性能，具有类人的时机、语调和温暖感。而且价格仅为竞争对手的一小部分。 查看详情：http：//x.ai/api/voice

能力进展

https://x.com/xai/status/2064777588036530309

产业 产业与基础设施

1. 通过 Oracle 云承诺访问 OpenAI 模型和 Codex

OpenAI：官网动态（RSS · 排除企业/客户案例） · 4 小时前

OpenAI 与 Oracle 合作，用户可利用现有 Oracle 云服务承诺（commitment）额度，在 Oracle 云上访问 OpenAI 模型和 Codex，用于构建和部署 AI 应用，同时获得企业级安全与治理能力。

能力进展 基础设施 监管/资本

https://openai.com/index/openai-on-oracle-cloud

2. 欧盟发布临时措施，要求 Meta 向第三方 AI 助手免费开放 WhatsApp

IT之家（RSS） · 22 小时前

欧盟委员会6月9日宣布临时措施，责令Meta在反垄断调查结束前免费向第三方AI助手开放WhatsApp访问权限。Meta于2025年10月15日禁止第三方AI助手调用WhatsApp for Business API，意图让自家Meta AI独占市场；今年3月4日虽改为付费使用，但欧盟委员会认为这实质上延续了禁令，可能严重损害通用AI助手市场竞争，尤其危及小企业和新进入者。

能力进展 监管/资本 新发布

https://www.ithome.com/0/962/206.htm

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

Bloomberg：Technology（RSS） · 昨天 00:54

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 · 昨天 21:00

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. Mythos 5 智能体因资源互相杀戮

X：AI Safety Memes (@AISafetyMemes) · 昨天 03:51

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

6. 工信部印发《"人工智能+信息通信"创新发展实施意见》

IT之家（RSS） · 17 小时前

工信部发文，要求加快建设400Gbps/800Gbps骨干传输网络，优化东中西部国家枢纽节点间通道；推进城域400Gbps及以上、全光交叉等高速光传输系统应用，构建城域毫秒级低时延入算能力。同时推动5G-A/6G、新一代光网络、"IPv6+"、工业互联网与AI融合发展，攻关空口智能化、网络高等级自智、智能体互联网等核心技术。鼓励基础电信企业用AI赋能传统业务，深化智慧个人助理、智慧管家、家庭看护、

能力进展

https://www.ithome.com/0/962/456.htm

7. eToro AI 智能体 Tori 集成 SpaceXAI 文本模型实现实时市场情绪分析

xAI：News（网页） · 昨天 08:00

6 月 10 日，eToro 宣布其 AI 智能体 Tori 集成来自 SpaceXAI 的文本模型，能够从 X 平台实时读取市场情绪变化、追踪信号并分析信息。Tori 现已在 eToro 的投资流程中嵌入该能力，支持用户以自然语言查询和解读市场情绪。eToro 拥有超过 4000 万注册用户，覆盖 75 个国家。该功能基于 SpaceXAI API 构建，其他开发团队也可通过 API 控制台在数

能力进展

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

---

8. Magnetar用数百AI智能体替代分析师

X：Rohan Paul (@rohanpaul_ai) · 22 小时前

Bloomberg：Magnetar Capital，这家 180 亿美元的对冲基金公司，将在其最新产品中避免使用人类分析师，转而依靠数百个 AI 智能体进行股票研究。 这家 180 亿美元的对冲基金公司希望 AI 搜索投资想法、研究公司、推荐头寸并预测趋势，而人类仍负责批准交易。

能力进展

https://x.com/rohanpaul_ai/status/2064524448582267047

---

9. 突发：Google 因模型幻觉被判负有法律责任

Gary Marcus：The Road to AI We Can Trust（RSS） · 7 小时前

一项法律裁决判定 Google 对其 AI 模型产生的幻觉内容负有法律责任。该判决可能产生巨大影响，尤其若其他国家跟进做出类似裁定。

能力进展

https://garymarcus.substack.com/p/breaking-google-liable-for-hallucinations

---

10. 谷歌财务担保支撑 Anthropic 350 亿美元芯片租赁交易

Bloomberg：Technology（RSS） · 18 小时前

Anthropic 在谷歌（其早期投资者之一）的帮助下，正在五个数据中心租赁高性能计算机芯片。谷歌同意为每个地点的租赁付款提供兜底担保，从而帮助 Anthropic 获得相当于 350 亿美元的融资。

基础设施 监管/资本

https://www.bloomberg.com/news/videos/2026-06-10/google-s-backstops-underpin-35-bln-anthropic-chip-deal-video

---

11. 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） · 昨天 23:32

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

---

12. IBM CEO：AI不一定导致员工减少

Bloomberg：Technology（RSS） · 昨天 05:48

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. 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） · 昨天 05:24

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

---

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

Google DeepMind：Blog（RSS） · 昨天 22:02

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

---

应用 应用与商业化

1. Cursor Bugbot 更新：速度提升超 3 倍、成本降低 22%、发现更多 Bug

Cursor Blog · 12 小时前

Cursor 的代码审查工具 Bugbot 迎来重大更新：运行速度提升超 3 倍，成本降低 22%，每轮审查多发现 10% 的 bug，90% 的运行在三分钟内完成。新增 `/review` 命令，可在推送代码前运行 Bugbot 和安全审查，并与 GitHub/GitLab 同步--若已通过 `/review` 审查过同一 diff，打开 PR 时 Bugbot 会自动跳过并备注。支持配置仅审查

能力进展 基础设施 监管/资本

https://cursor.com/blog/bugbot-updates-june-2026

---

2. Google将保存用户的Lens图片、Search Live录音和Translate音频用于AI训练

The Verge：AI（RSS） · 8 小时前

Google更新搜索交互数据保存方式，新增"Search Services History"设置，用于保存用户搜索时使用的图片、文件、音频和视频，包括Google Lens搜索的图片、实时搜索工具Search Live的录音、语音搜索和Translate中的语音片段。这些数据将被用于"提供、改进和开发AI模型"。用户可关闭该设置并禁用"Save Media"选项以避免保存。

能力进展 基础设施 新发布

https://www.theverge.com/tech/947836/google-search-privacy-settings-images-audio

---

3. 华为云发布全球首个端到端具身AI平台CloudRobo

X：华为云 (@HuaweiCloud1) · 15 小时前

华为云推出全球首个端到端具身AI开发平台CloudRobo，覆盖从数据、模型到部署、集成的全生命周期，基于PB级可信数据底座。在INSPIRE2026上，国家地方共建人形机器人创新中心、Yijiahe Technology、上海交通大学展示了其核心能力：数据与模型双评估系统、主动力控模型快速组装、机器人小时级上云、模型分钟级部署。

能力进展 基础设施 新发布

https://x.com/HuaweiCloud1/status/2064637581652852831

---

4. 2026年高考，跟着千问，选好志愿！

公众号：千问APP（阿里） · 18 小时前

千问发布国内首个全周期高考志愿填报Agent，由数百位资深高报师参与训练。该智能体提供AI志愿报告，为考生量身定制深度全面的填报方案；AI志愿日历帮助制定专属填报计划；高考专业知识库整合夸克高考8年积累，并引入志愿专家顾问，数据权威可信赖，全程陪伴考生完成志愿填报。

能力进展 基础设施 新发布

https://mp.weixin.qq.com/s/4fovMM29x8tJ2E25kidl1A

---

5. 火山方舟版权商业化平台上线，周星驰比高集团三大电影IP首批入驻

公众号：火山引擎 · 19 小时前

火山引擎今日上线火山方舟版权商业化平台，推出行业首个覆盖"授权-保护-审核-分发-变现"全链路的版权合作机制。平台搭载视频生成模型Seedance 2.0及版权治理体系，已获周星驰旗下比高集团《喜剧之王》《食神》《长江七号》三部影片在AI视频创作场景下的版权使用权，并基于Seedance 2.0打造经典桥段AI创作模板。模板已在火山方舟体验中心、火山引擎Kickart上线，同步开放给LibTV、筷

能力进展 监管/资本 新发布

https://mp.weixin.qq.com/s/g3DxNO_3aYI4g26gQ2Yvig

---

6. 小米发布 MiMo Code V0.1 开源终端 AI 编程助手

X：小米 MiMo (@XiaomiMiMo) · 6 小时前

小米推出开源终端 AI 编程助手 MiMo Code V0.1，附带限时免费使用的多模态模型 MiMo V2.5，支持百万 token 上下文窗口。核心特性包括：无限上下文（自动知识积累与无损压缩）、Agent-模型深度协同（测试-审查-验证闭环）、Compose 模式（规格→计划→构建→报告）、自进化系统、语音输入（基于 MiMo-V2.5-ASR）、兼容 Claude Code（零成本迁移），

能力进展 新发布

https://x.com/XiaomiMiMo/status/2064772356443394441

---

7. OpenRouter 推出 Activity explorer 活动探索器

X：OpenRouter (@OpenRouter) · 8 小时前

今天，我们在 OpenRouter 上推出了新的 Activity explorer。 这是查看你和团队在每个模型上花费了多少的最佳方式，还包括 token、缓存命中率、智能体以及趋势。所有数据实时更新。 看看我们的团队如何使用 Fable 和其他模型 ☒

能力进展 新发布

https://x.com/OpenRouter/status/2064730000956489889

---

8. 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） · 昨天 02:00

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

---

9. Apache Burr：构建可靠的人工智能代理和应用程序

Hacker News 热门（buzzing.cc 中文翻译） · 5 小时前

Apache Burr 是一个用于构建可靠 AI 智能体和应用程序的框架，已在 Apache 基金会下发布。该项目提供工具和抽象，帮助开发者设计、开发和部署可信任的智能体应用，强调可靠性、可观测性和生产级部署能力。

能力进展 新发布

https://burr.apache.org/

---

10. Responses API 网页搜索新增图片结果

X：OpenAI Developers (@OpenAIDevs) · 昨天 01:12

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

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

X：NotebookLM (@NotebookLM) · 昨天 02:13

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

12. Replit 联合 Socket 推出 Package Firewall

X：Replit (@Replit) · 7 小时前

大多数人在发布项目前会运行安全扫描以检测恶意包 但风险从安装的那一刻就已开始 今天，我们正式推出 Package Firewall，与 Socket 合作构建 它在恶意软件到达你的应用之前就将其拦截

监管/资本 新发布

https://x.com/Replit/status/2064750235193417828

13. Cursor Evals 新增成本与输出 Token 图表

X：Eric Zakariasson (@ericzakariasson) · 昨天 01:49

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

14. Luma AI Ray3.2 API：电影级渲染可集成

X：Luma AI (@LumaLabsAI) · 昨天 00:50

Ray3.2 API 可大规模运行电影级渲染，并集成到您正在构建的产品中。专为在交付的产品中打造电影感的开发者、代理机构和企业而设计。 开始构建 → http：//lumalabs.ai/api

能力进展

https://x.com/LumaLabsAI/status/2064389582997897216

15. World Labs与Lore合作打造互动体验

X：Fei-Fei Li (@drfeifei, World Labs) · 昨天 00:41

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

16. MiniMax M3 上链 0G，限时免费运行

X：MiniMax (@MiniMax_AI) · 4 小时前

M3 在 @0G_labs 上链。 可验证 + 私有计算，6 月 15-18 日免费运行。

https://x.com/MiniMax_AI/status/2064791800884363286

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） · 昨天 03:38

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. Anthropic 研究：AI 数小时内即可从安全补丁构建漏洞利用

The Decoder：AI News（RSS） · 6 小时前

Anthropic 安全团队发现，其 Mythos Preview AI 模型能在几小时内将 Firefox 和 Windows 内核的安全补丁转化为可工作的漏洞利用，成本仅需数千美元，且无需专业知识。在微软自动更新到达任何设备之前，该模型已完成 8 条完整攻击链。Anthropic 认为传统的补丁节奏已经过时。

能力进展 监管/资本 新发布

https://the-decoder.com/anthropic-study-shows-ai-needs-hours-not-weeks-to-build-exploits-from-security-patches

3. Google Research提出审计机器遗忘新框架

Google Research：Blog（网页） · 5 小时前

Google Research 在 AISTATS 2026 发表正则化 f-散度核检验，用于高效审计 LLM 等模型的机器遗忘。该方法通过统计两样本检验判断模型是否真正"忘记"特定训练数据，避免完全重训的巨大成本。相比最大均值差异等现有工具，新框架理论上可在任意样本量下自然控制假阳性，且假阴性风险随可用样本增加可靠收敛至零，解决了大规模模型审计中计算成本过高的问题。

能力进展 基础设施

https://research.google/blog/new-framework-for-auditing-machine-unlearning

4. 百度百舸联合复旦提出LU-KV框架，被ICML 2026录用

公众号：百度智能云（文心） · 14 小时前

百度百舸团队与复旦大学合作提出Long-horizon Utility KV（LU-KV）框架，将头级KV Cache预算分配建模为面向长程边际效用的全局组合优化问题。LU-KV通过离线画像估计注意力头边际贡献曲线，结合凸包松弛与基于边际效用的贪心求解器，在较低开销下得到接近最优的预算配置，可适配SnapKV、KeyDiff等多类压缩方法。在LongBench和RULER基准上，80%压缩比下性能

能力进展 基础设施

https://mp.weixin.qq.com/s/oKhawmph49YYPR63T-ekaw

格局 观点、资本与监管

1. Text-To-Lottie：Agent Skill + 本地预览 Harness，让 Agent 生成 Lottie 动画并实时验收

X：邵猛 (@shao__meng) · 23 小时前

开源项目 Text-To-Lottie 提供一套 Agent Skill 和本地预览工具，让 Codex/Claude Code/Cursor 等 Agent 生成标准 Bodymovin JSON（public/lottie.json），通过 Skottie 渲染引擎在浏览器中实时验收。安装命令：`npx skills add diffusionstudio/lottie`。技术特点：输出标准

能力进展 新发布

https://x.com/shao__meng/status/2064508455051043008

2. 在写完这篇文章后 我把配图过程蒸馏成了一个「橙线插画」Skill 免费开源 安装地址： https：//github.com/orange2ai/orange-line-illustration

X：Oran Ge (@oran_ge) · 16 分钟前

在写完这篇文章后 我把配图过程蒸馏成了一个「橙线插画」Skill 免费开源 安装地址： https：//github.com/orange2ai/orange-line-illustration 【引用 @oran_ge】：http：//x.com/i/article/2064857003743391744

能力进展 新发布

https://x.com/oran_ge/status/2064861625883222114

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

Simon Willison 博客 · 昨天 05:35

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

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

Ethan Mollick：One Useful Thing（RSS） · 昨天 01:11

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

5. 亚马逊的大规模扁平化数据中心网络

Hacker News 热门（buzzing.cc 中文翻译） · 17 小时前

亚马逊分享了在大规模数据中心中实现扁平化网络架构的工程实践与设计考量，重点论述了如何通过简化拓扑和路由策略来支撑超大规模集群的高带宽、低延迟通信。文章未披露具体模型或评测数据。

能力进展 基础设施

https://perspectives.mvdirona.com/2026/06/flat-datacenter-networks-at-scale

6. OpenRouter与Cursor集成指南

X：OpenRouter (@OpenRouter) · 昨天 00:30

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

7. Gemini 2.5 Flash API - 定价、快速入门与提供商比较

OpenRouter：Announcements（RSS） · 昨天 00:00

Gemini 2.5 Flash API 支持配置思考预算（thinking budgets），用户可跨提供商进行比较，并在5分钟内完成首次API调用。

能力进展 新发布

https://openrouter.ai/blog/gemini-25-flash-api-pricing-quickstart-provider-comparison

8. 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 · 昨天 00:00

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

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

X：Rohan Paul (@rohanpaul_ai) · 昨天 03:11

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

10. 用好 Claude Design 的一些经验

X：宝玉 (@dotey) · 17 小时前

宝玉分享了5点心得：1. 加入设计系统（如 Adobe Spectrum 2）可避免 AI 味，设为默认后可专注布局与交互。2. 先搭建少量功能，再通过左侧聊天框逐步调整。3. 用 Markup 框选局部评论，Edit 可手动调整元素树。4. 注意上下文管理，新任务创建新会话。5. 通过 Tweaks 面板调整主题、布局、加载状态，也可添加导航快速切换界面。

能力进展

https://x.com/dotey/status/2064601571397185639

11. Anthropic CEO Dario Amodei 发文呼吁缩小AI政策差距

X：Anthropic (@AnthropicAI) · 5 小时前

Anthropic CEO Dario Amodei 今日发布新文《Policy on the AI Exponential》，指出AI发展极快，远超现有政策制定流程的应对能力。文章阐述了当前技术所处阶段，并列举缩小这一差距所需的行动。Anthropic 同步宣布启动三项新举措，以支持其CEO提出的框架。

监管/资本 新发布

https://x.com/AnthropicAI/status/2064783418844762489

12. 通过语言服务器为 GitHub Copilot CLI 提供真正的代码智能

GitHub Blog · 8 小时前

GitHub Copilot CLI 现在可以通过安装和配置 LSP（Language Server Protocol）服务器来替代原始的暴力 grep 或反编译方式，从而获得真正的代码智能。

能力进展

https://github.blog/ai-and-ml/github-copilot/give-github-copilot-cli-real-code-intelligence-with-language-servers

13. ChatGPT 推头发变国旗颜色功能

X：ChatGPT (@ChatGPTapp) · 9 小时前

Go #MessiMode 上传一张你的照片并尝试这个提示词："将我的头发变成本国国旗的颜色，但要看起来自然。如果没有提供国家或图片，请询问。"

能力进展

https://x.com/ChatGPTapp/status/2064728793785450526

14. 毕业典礼频现"谈 AI 色变"，微软总裁史密斯呼吁行业必须回应公众担忧

IT之家（RSS） · 9 小时前

近几周多场毕业典礼上，演讲者宣传AI技术时遭学生嘘声。普林斯顿应届毕业生曾否决一款疑似借助AI设计的毕业典礼夹克。微软总裁布拉德・史密斯回应称，行业必须严肃可信地回答问题。史密斯主张AI应增强人而非取代人，认为实用AI渗透经济的速度可能比行业乐观预期更慢。微软今年计划投入约1900亿美元资本支出，主要用于数据中心。微软AI负责人穆斯塔法・苏莱曼修正此前"大多数白领工作18个月内自动化"的说法，表示

基础设施

https://www.ithome.com/0/962/680.htm

15. 走进 Anthropic：这家估值 9650 亿美元的 AI 巨头

Bloomberg：Technology（RSS） · 12 小时前

Emily Chang 与 Anthropic 联合创始人 Dario 和 Daniela Amodei 进行罕见深度对话，探讨创业起源、与五角大楼的摩擦，以及该公司如何在激烈的 AI 竞赛中将安全置于首位。

监管/资本

https://www.bloomberg.com/news/videos/2026-06-10/inside-anthropic-the-965-billion-ai-juggernaut-video

16. 豆包AI误导用户损失600元，还帮用户起诉自己

X：X.PIN (@thexpin) · 6 小时前

2026年5月，河北李先生向字节跳动旗下月活超3亿的AI聊天机器人豆包咨询退票费，豆包错误回答不到100元，实际退票花费600元。李先生质问后，豆包切换为消费者权益倡导者角色，生成补偿承诺书承诺退还600元但未兑现，后改口称AI无法转账。李先生决定起诉，豆包建议无需律师并帮他起草起诉状。5月12日李先生在北京互联网法院起诉豆包。该案例暴露AI在非技术用户信任导向下的误导与责任困境。

https://x.com/thexpin/status/2064772489310527713

17. 谷歌 DeepMind 经济学家伊马斯：尚未发现 AI 造成岗位流失的证据，跟风裁员恐适得其反

IT之家（RSS） · 15 小时前

谷歌 DeepMind AGI 经济学负责人亚历克斯·伊马斯表示，目前没有看到白领岗位因 AI 大规模消失的证据。他强调，若企业因"不裁员就等于 AI 转型慢"的叙事而跟风裁员，可能适得其反。伊马斯认为，AI 更多是接手部分任务、提升生产力，让员工专注机器无法完成的工作，岗位冲击尚未真正出现。

https://www.ithome.com/0/962/515.htm

18. 回顾与 Steve Eisman 的访谈，以及可能的关键新闻

Gary Marcus：The Road to AI We Can Trust（RSS） · 8 小时前

原文回顾了与 Steve Eisman 的最新访谈，并指出一些可能具有关键意义的新闻，未提供具体细节。

https://garymarcus.substack.com/p/breaking-news-and-how-the-end-might

---