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: 55 条 焦点: 8 条 快讯: 0 条

Executive Summary

今日AI行业呈现模型能力与基础设施双重突破态势。NVIDIA发布RTX Spark本地AI智能体设备，提供1 petaflops算力与128GB统一内存，配合OpenShell运行时确保端侧安全运行。MiniMax推出M3模型，支持100万token超长上下文，采用MSA稀疏注意力架构将计算成本降至前代1/20。NVIDIA同时发布工厂运营蓝图（FOX），构建自主工厂管理智能体参考设计。xAI推出Composer 2.5编程模型，JetBrains发布Mellum2 12B混合专家模型。

行业结构层面，NVIDIA与Google Cloud深化合作，提供L4 Tensor Core GPU优化推理性能。Anthropic秘密提交IPO申请，估值近1万亿美元，反映头部AI公司资本化进程加速。OpenBMB发布UltraData两大开源数据集，Ultra-FineWeb-L3包含600B+ tokens，UltraData-SFT-2605含15M+样本。Runway加入Cosmos Coalition，与NVIDIA合作开发物理AI世界模型。上海发布服务业发展规划，支持多模态智能体开发与智能驾驶多场景应用。

后续需关注Nemotron 3 Ultra模型发布后的性能表现与开放节奏，RTX Spark对本地AI部署成本结构的影响。Anthropic IPO进程及定价策略将影响AI公司估值体系，M3模型在Vercel AI Gateway的商业化表现值得关注。物理AI世界模型研发进度、开源数据集质量验证以及各地AI产业政策落地执行情况将是重要观察变量。

重点 今日核心进展

★ 1. Nemotron 3 Ultra 本周即将发布

X：NVIDIA AI (@NVIDIAAI) · 19 小时前 · 模型与工具能力

Nemotron 3 Ultra 本周即将发布。 ☒ 该条来自X：NVIDIA AI (@NVIDIAAI)，属于模型与工具能力方向，后续关注其对模型能力、产品形态或产业链节奏的影响。

能力进展 基础设施 新发布

https://x.com/NVIDIAAI/status/2061305524700758050

★ 2. MiniMax M3：前沿编码、100万token上下文与原生多模态一体模型

MiniMax：Blog（网页） · 20 小时前 · 模型与工具能力

MiniMax M3 是一个开源前沿模型，具备先进的编码与AI智能体能力。它支持100万token的超长上下文窗口，并采用名为MSA（MiniMax Sparse Attention）的新型稀疏注意力架构。该架构使模型在100万token上下文下的每token计算成本降至前代的1/20，预填充速度提升9倍以上，解码速度提升15倍以上。在SWE-Bench Pro编码基准上，MiniMax M3得分

能力进展 新发布

https://www.minimax.io/blog/minimax-m3

★ 3. xAI发布Composer 2.5

xAI：News（网页） · 昨天 08:00 · 模型与工具能力

xAI的最新编程模型Composer 2.5现已在Grok Build中可用，用户可通过`/models`菜单选择使用。这是一款快速、先进的模型，擅长处理长时间运行的任务和复杂指令。该模型面向SuperGrok和X Premium+用户开放。

能力进展 新发布

https://x.ai/news/composer-2-5

★ 4. NVIDIA 发布 RTX Spark 及本地 AI 智能体安全与性能更新

NVIDIA Blog：Agentic AI（网页） · 19 小时前 · 应用与商业化

NVIDIA 发布了 RTX Spark，一款专为本地 AI 智能体设计的 Windows 电脑，提供 1 petaflops AI 算力与 128GB 统一内存。其推出的 OpenShell 运行时与微软合作，基于新的 Windows 安全原语，确保智能体在设备端安全私密运行。性能方面，通过在 llama.cpp 中采用多 token 预测等优化，Qwen 3.6 和 3.5 27B 模型推理吞

能力进展 基础设施 监管/资本

https://blogs.nvidia.com/blog/rtx-ai-garage-computex-spark-local-agents

★ 5. NVIDIA 发布工厂运营蓝图，为工厂提供自主智能管理智能体

NVIDIA Blog：Agentic AI（网页） · 19 小时前 · 应用与商业化

NVIDIA 在 GTC Taipei 发布了 NVIDIA 工厂运营蓝图（FOX），这是一个用于构建自主工厂管理智能体的参考设计。该蓝图基于 NVIDIA NemoClaw、AI-Q Blueprint 和 NVIDIA Nemotron 开源模型构建，旨在为工厂提供一个统一的决策层，以连接实时机器信号、质量数据和操作警报，实现快速问题解决。蓝图针对 NVIDIA DGX Station 桌面

能力进展 基础设施 新发布

https://blogs.nvidia.com/blog/factory-operations-fox-blueprint-ai-brain

★ 6. OpenBMB发布UltraData两大开源数据集，登顶HuggingFace趋势榜

X：面壁智能 OpenBMB (@OpenBMB) · 11 小时前 · 应用与商业化

OpenBMB联合清华NLP与Modelbest发布两个开源数据集：Ultra-FineWeb-L3（预训练合成数据）包含600B+ tokens（超400B英文、200B+中文），是迄今最大开源中文预训练合成数据集；UltraData-SFT-2605（后训练SFT数据）包含15M+样本，是中国首个开源且包含思考与非思考标注的大规模SFT数据集，覆盖数学、代码、知识和指令遵循。两者均基于Ultr

能力进展 基础设施 新发布

https://x.com/OpenBMB/status/2061432928492810535

★ 7. 介绍Cosmos Coalition

Runway：News（网页） · 18 小时前 · 产业与基础设施

Runway宣布作为创始成员加入Cosmos Coalition，该联盟与NVIDIA及多家领先AI实验室合作，旨在构建并开源面向物理AI的前沿世界模型。首个项目将由Runway与NVIDIA共同开发一个基础模型，以推动下一代开放世界模型的研究与发展。

能力进展 基础设施 新发布

https://runwayml.com/news/introducing-cosmos-coalition

★ 8. NVIDIA与Google Cloud助力下一波AI构建者

NVIDIA Blog：Generative AI（网页） · 19 小时前 · 产业与基础设施

在Google I/O大会上，NVIDIA与Google Cloud宣布深化合作，旨在支持其联合开发者社区中超过10万名开发者。合作将重点提供NVIDIA L4 Tensor Core GPU以优化AI推理与图形工作负载，并通过支持Vertex AI平台来增强Gemini模型性能。双方还将提供开源软件工具，以简化AI应用的构建与部署流程。

能力进展 基础设施 新发布

https://blogs.nvidia.com/blog/category/enterprise/cloud-2

能力 模型与工具能力

1. Nemotron 3 Ultra 本周即将发布

X：NVIDIA AI (@NVIDIAAI) · 19 小时前

Nemotron 3 Ultra 本周即将发布。 ☒ 该条来自X：NVIDIA AI (@NVIDIAAI)，属于模型与工具能力方向，后续关注其对模型能力、产品形态或产业链节奏的影响。

能力进展 基础设施 新发布

https://x.com/NVIDIAAI/status/2061305524700758050

2. MiniMax M3：前沿编码、100万token上下文与原生多模态一体模型

MiniMax：Blog（网页） · 20 小时前

MiniMax M3 是一个开源前沿模型，具备先进的编码与AI智能体能力。它支持100万token的超长上下文窗口，并采用名为MSA（MiniMax Sparse Attention）的新型稀疏注意力架构。该架构使模型在100万token上下文下的每token计算成本降至前代的1/20，预填充速度提升9倍以上，解码速度提升15倍以上。在SWE-Bench Pro编码基准上，MiniMax M3得分

能力进展 新发布

https://www.minimax.io/blog/minimax-m3

3. xAI发布Composer 2.5

xAI：News（网页） · 昨天 08:00

xAI的最新编程模型Composer 2.5现已在Grok Build中可用，用户可通过`/models`菜单选择使用。这是一款快速、先进的模型，擅长处理长时间运行的任务和复杂指令。该模型面向SuperGrok和X Premium+用户开放。

能力进展 新发布

https://x.ai/news/composer-2-5

4. 介绍 Mellum2：JetBrains 推出的 12B 混合专家模型

Hugging Face：Blog（RSS） · 8 小时前

JetBrains 在 Hugging Face 发布博客，介绍其新发布的 Mellum2 模型。该模型采用混合专家架构，参数规模为 12B。

能力进展 新发布

https://huggingface.co/blog/JetBrains/mellum2-launch

5. 使用NVIDIA Cosmos 3开发物理AI推理、世界与行动模型

NVIDIA Technical Blog（开发者技术博客 · RSS） · 19 小时前

NVIDIA Cosmos 3是一款面向物理AI的前沿基础模型。它能够帮助机器人、自动驾驶车辆和智能空间理解真实世界、预测事件发展并生成适应特定环境与任务的行动。该模型融合了物理推理、世界理解与行动生成能力。

能力进展 基础设施

https://developer.nvidia.com/blog/develop-physical-ai-reasoning-world-and-action-models-with-nvidia-cosmos-3

6. SenseNova新模型解决AI图表生成难题

X：商汤 SenseTime (@SenseTime_AI) · 9 小时前

大多数AI模型在生成图表时存在数值错误（如负值显示为正）、柱状图位置偏移、元素关系混乱等问题。SenseNova-U1-8B-MoT-Infographic（SenseNova-U1）专为解决此类图表生成问题而设计，能够生成准确的图表，并支持实时调整设计和布局。项目在Hugging Face提供了模型，并在GitHub展示了效果案例。

能力进展

https://x.com/SenseTime_AI/status/2061465029959209106

7. MiniMax M3 上线 Vercel AI Gateway

X：MiniMax (@MiniMax_AI) · 25 分钟前

M3 已在 Vercel 的 AI Gateway 上线！ 我们首个支持 1M token 长上下文和多模态输入的模型。 本周享 50% 折扣 期待看到大家用 M3 和 @vercel_dev 构建什么

能力进展

https://x.com/MiniMax_AI/status/2061597897792397397

8. Qwen3.7-Plus：多模态智能体智能

Qwen：Blog Retrieval（API） · 22 小时前

Qwen Studio 提供涵盖聊天机器人、图像与视频理解、图像生成、文档处理、网页搜索集成、工具使用及制品生成的全面功能。

能力进展

https://qwen.ai/blog?id=qwen3.7-plus

产业 产业与基础设施

1. 介绍Cosmos Coalition

Runway：News（网页） · 18 小时前

Runway宣布作为创始成员加入Cosmos Coalition，该联盟与NVIDIA及多家领先AI实验室合作，旨在构建并开源面向物理AI的前沿世界模型。首个项目将由Runway与NVIDIA共同开发一个基础模型，以推动下一代开放世界模型的研究与发展。

能力进展 基础设施 新发布

https://runwayml.com/news/introducing-cosmos-coalition

2. NVIDIA与Google Cloud助力下一波AI构建者

NVIDIA Blog：Generative AI（网页） · 19 小时前

在Google I/O大会上，NVIDIA与Google Cloud宣布深化合作，旨在支持其联合开发者社区中超过10万名开发者。合作将重点提供NVIDIA L4 Tensor Core GPU以优化AI推理与图形工作负载，并通过支持Vertex AI平台来增强Gemini模型性能。双方还将提供开源软件工具，以简化AI应用的构建与部署流程。

能力进展 基础设施 新发布

https://blogs.nvidia.com/blog/category/enterprise/cloud-2

3. Anthropic 保密向 SEC 提交 S-1 草案

Anthropic：Newsroom（网页） · 7 小时前

Anthropic， PBC 今日保密向美国证券交易委员会提交了 S-1 表格草案，计划进行普通股的首次公开发行。这使其在 SEC 完成审核后拥有上市的选择权。IPO 的具体发行股数和价格尚未确定，将取决于市场条件等因素。公司近期刚完成由 Altimeter Capital 等领投的 650 亿美元 H 轮融资，估值达 9650 亿美元，并发布了 Claude Opus 4.8 模型。

能力进展 监管/资本 新发布

https://www.anthropic.com/news/confidential-draft-s1-sec

4. 抢跑OpenAI，Anthropic官宣已秘密递交IPO申请

IT之家（RSS） · 1 小时前

Anthropic已秘密向美国证券交易委员会递交IPO申请，该公司估值接近1万亿美元。此次申请前，Anthropic刚完成H轮650亿美元融资，投后估值攀升至9650亿美元。公司当前年化营收已突破470亿美元。其竞争对手OpenAI同样在筹备IPO。Anthropic旗下Mythos大模型将逐步开放商用，公司计划向欧盟网络安全局开放该模型的调用权限。

能力进展 监管/资本 新发布

https://www.ithome.com/0/958/458.htm

5. 上海：支持多模态智能体开发与应用，有序推进智能驾驶在共享出行、物流运输等多场景应用

IT之家（RSS） · 22 小时前

上海市人民政府办公厅印发《上海市服务业发展"十五五"规划》，提出发展AI软件技术及服务产业集群。规划支持多模态智能体开发与应用，推动智能客服等工具规模化；有序推进智能驾驶在共享出行、物流运输等多场景应用；做强算运存协同布局的智算云网络，推广模型即服务（MaaS）；并支持开发面向家庭、养老、文旅等场景的具身智能整机产品，加速机器人向通用智能与精细化服务跃升。

能力进展 基础设施

https://www.ithome.com/0/957/985.htm

6. 佛罗里达州起诉OpenAI与Sam Altman：涉多起ChatGPT相关谋杀案

Ars Technica：AI（RSS） · 5 小时前

佛罗里达州对OpenAI及其CEO Sam Altman提起诉讼。该州总检察长指控Altman对人命"完全漠视"，案件与多起涉及ChatGPT的谋杀事件相关。

能力进展 新发布

https://arstechnica.com/tech-policy/2026/06/florida-sues-openai-sam-altman-after-multiple-chatgpt-linked-murders

7. 智谱：建议 A 股发行并在科创板上市

IT之家（RSS） · 8 小时前

智谱计划向中国监管机构申请发行A股并在科创板上市。发行股份数量占发行完成后总股本的2%至8%，预计全部为新股，原股东不发售。本次发行募集资金净额将投资于人工智能通用基座大模型、大模型MaaS一站式服务平台及补充流动资金。此外，公司拟将英文名称由"Knowledge Atlas Technology Joint Stock Company Limited"变更为"Z.AI Co.， Ltd."。

能力进展 监管/资本

https://www.ithome.com/0/958/444.htm

8. Meta的AI被利用来劫持Instagram账号

The Verge：订阅版科技（RSS） · 5 小时前

Meta的AI聊天机器人被发现存在安全漏洞，黑客可以通过请求该AI关联一个新的电子邮件地址，从而接管目标用户的Instagram账号。该漏洞利用了AI智能体直接执行账户管理操作的能力。

能力进展 监管/资本

https://www.theverge.com/tech/941179/meta-instagram-ai-support-chatbot-exploit-hacked

9. OpenAI在密歇根州启动Stargate 1GW数据中心建设

OpenAI：官网动态（RSS · 排除企业/客户案例） · 12 小时前

OpenAI在密歇根州启动了名为Stargate的1GW数据中心项目。作为AI基础设施建设的一部分，该项目旨在扩大人工智能技术的可及性、为当地创造就业机会并支持社区发展。

基础设施 新发布

https://openai.com/index/stargate-michigan-data-center

10. Runway 在伦敦设立欧洲总部及世界模型研究中心

Runway：News（网页） · 13 小时前

Runway 宣布在伦敦建立新的欧洲总部和专注于通用世界模型的研究中心。公司计划在未来18个月向英国AI生态投资$100M，到2028年投资额将翻倍以上。过去12个月，其在欧洲的订阅销量增长了50%，企业客户占比超20%。新总部将扩大其在欧洲的研究与商业布局，公司正招聘欧洲负责人以组建跨研究、产品、工程和销售的团队，并深化与BBC、Fremantle、WPP等企业的合作。世界模型是其研究的核心，旨

能力进展

https://runwayml.com/news/runway-opens-london-hq

11. 王兴：美团 AI Agent"小美"与腾讯元宝即将深度合作，用户订单无缝连接

IT之家（RSS） · 11 小时前

美团2026年第一季度财报显示营收910.39亿元，净利润亏损68.27亿元。财报电话会上，CEO王兴透露其AI Agent"小美"将与腾讯元宝深度合作。用户在腾讯元宝中提交本地服务需求，将被无缝连接至美团的外卖点餐、配送等生态。王兴强调，面向智能体的服务（To A）正变得日益重要，美团已将AI助手"小团"置于App核心位置，并拓展AI服务外延。

能力进展

https://www.ithome.com/0/958/410.htm

12. OpenAI的AI政策与政治倡导观点

OpenAI：官网动态（RSS · 排除企业/客户案例） · 7 小时前

OpenAI阐述了其在AI政策与政治倡导方面的立场，包括对透明度、深思熟虑的监管以及AI安全的支持。同时，公司明确表示，任何外部政治组织均不能代表OpenAI发言。

监管/资本 新发布

https://openai.com/index/our-views-on-ai-policy-and-political-advocacy

13. 英伟达和台积电将 AI 引入晶圆厂，推动半导体设计与制造发展

IT之家（RSS） · 18 小时前

IT之家（RSS）披露：英伟达和台积电将 AI 引入晶圆厂，推动半导体设计与制造发展。该条属于产业与基础设施方向，后续关注其对模型能力、产品形态或产业链节奏的影响。

能力进展

https://www.ithome.com/0/958/109.htm

14. Luma成立开放物理AI实验室攻克泛化难题

X：Luma AI (@LumaLabsAI) · 9 小时前

为改善人类生活，AI系统必须能帮助我们改善物理世界。阻碍我们迈向这一繁荣未来的是物理AI的泛化问题。 为解决此问题，我们在Luma建立了一个新的开放科学物理AI实验室。 阅读更多 → https：//lumalabs.ai/news/luma-open-physical-ai-lab

新发布

https://x.com/LumaLabsAI/status/2061460217616027961

15. OpenAI正式进军机器人领域并启动招聘

X：Sam Altman (@sama) · 昨天 00:07

OpenAI宣布成立OpenAI Robotics团队，并开始招聘全栈硬件、系统及ML工程师，以编程和制造能服务社会的机器人。该项目由Aditya Ramesh领导，其世界模拟研究计划已演变为机器人研究，强调硬件与ML研究的协同设计。短期目标是支持技术工人构建未来基础设施，长期愿景是为每个人提供个人机器人。

新发布

https://x.com/sama/status/2061117302528188712

16. Alphabet将通过发行股票筹集800亿美元用于AI支出计划

Bloomberg：Technology（RSS） · 3 小时前

Alphabet正在通过发行股票等方式筹集800亿美元资金，其中包括与Berkshire Hathaway的投资协议，以资助其雄心勃勃的AI支出计划。

https://www.bloomberg.com/news/articles/2026-06-01/alphabet-to-raise-80-billion-in-equity-capital-for-ai-spending

17. 全球首次：MWC26 上海将举办"人形机器人点球大战"，宇树科技等 8 支队伍参赛、参演

IT之家（RSS） · 23 小时前

全球首次"人形机器人点球大战"将于2026年6月在MWC上海举行。8支中国顶尖具身智能战队将进行自主对抗，无需人工操控或预设脚本。赛事旨在集中展示人形机器人在动态平衡、精准控制与自主决策等方面的技术突破。

https://www.ithome.com/0/957/938.htm

应用 应用与商业化

1. NVIDIA 发布 RTX Spark 及本地 AI 智能体安全与性能更新

NVIDIA Blog：Agentic AI（网页） · 19 小时前

NVIDIA 发布了 RTX Spark，一款专为本地 AI 智能体设计的 Windows 电脑，提供 1 petaflops AI 算力与 128GB 统一内存。其推出的 OpenShell 运行时与微软合作，基于新的 Windows 安全原语，确保智能体在设备端安全私密运行。性能方面，通过在 llama.cpp 中采用多 token 预测等优化，Qwen 3.6 和 3.5 27B 模型推理吞

能力进展 基础设施 监管/资本

https://blogs.nvidia.com/blog/rtx-ai-garage-computex-spark-local-agents

2. NVIDIA 发布工厂运营蓝图，为工厂提供自主智能管理智能体

NVIDIA Blog：Agentic AI（网页） · 19 小时前

NVIDIA 在 GTC Taipei 发布了 NVIDIA 工厂运营蓝图（FOX），这是一个用于构建自主工厂管理智能体的参考设计。该蓝图基于 NVIDIA NemoClaw、AI-Q Blueprint 和 NVIDIA Nemotron 开源模型构建，旨在为工厂提供一个统一的决策层，以连接实时机器信号、质量数据和操作警报，实现快速问题解决。蓝图针对 NVIDIA DGX Station 桌面

能力进展 基础设施 新发布

https://blogs.nvidia.com/blog/factory-operations-fox-blueprint-ai-brain

3. OpenBMB发布UltraData两大开源数据集，登顶HuggingFace趋势榜

X：面壁智能 OpenBMB (@OpenBMB) · 11 小时前

OpenBMB联合清华NLP与Modelbest发布两个开源数据集：Ultra-FineWeb-L3（预训练合成数据）包含600B+ tokens（超400B英文、200B+中文），是迄今最大开源中文预训练合成数据集；UltraData-SFT-2605（后训练SFT数据）包含15M+样本，是中国首个开源且包含思考与非思考标注的大规模SFT数据集，覆盖数学、代码、知识和指令遵循。两者均基于Ultr

能力进展 基础设施 新发布

https://x.com/OpenBMB/status/2061432928492810535

4. Apache RocketMQ 发布 AI 专用消息引擎

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 13 小时前

Apache RocketMQ 为 AI 升级！推出 RocketMQ for AI--一个专为长时间会话、多智能体工作流和公平资源调度构建的新消息引擎。凭借 Lite-Topics、有序消息和智能流量整形，它解决了状态丢失、级联故障和突发负载问题。由阿里云大规模构建，现已开源。 了解更多： https：//int.alibabacloud.com/m/1000413178/ #Rocke

能力进展 基础设施 新发布

https://x.com/alibaba_cloud/status/2061400724018217381

5. NVIDIA DSX OS 提供开放、模块化软件用于规模化运营AI工厂

NVIDIA Technical Blog（开发者技术博客 · RSS） · 20 小时前

NVIDIA DSX 平台为设计、模拟和构建 AI 工厂提供完整解决方案，旨在应对不断增长的智能需求。该平台通过开放、模块化的软件栈，帮助运营方更快地扩展规模、提高效率，并降低从能源、芯片、基础设施、模型到应用这五层架构的整体智能成本。

能力进展 基础设施 新发布

https://developer.nvidia.com/blog/nvidia-dsx-os-delivers-open-modular-software-for-operating-ai-factories-at-scale

6. Perplexity发布Search as Code搜索架构

X：Perplexity (@perplexity_ai) · 6 小时前

推出Search as Code，我们为AI智能体打造的全新搜索架构。 它直接编写Python代码调用我们的搜索栈，而非逐个循环函数调用。 现已在Perplexity Agent API中提供，并成为Computer的默认选项。 https：//research.perplexity.ai/articles/rethinking-search-as-code-generation

能力进展 新发布

https://x.com/perplexity_ai/status/2061506359326384319

7. 腾讯混元发布智能体长期记忆插件Hy-Memory

X：腾讯混元 (@TencentHunyuan) · 15 小时前

腾讯混元正式发布专为OpenClaw等长期协作智能体（Agent）设计的记忆插件 Hy-Memory。它基于6层记忆框架、System1/System2双系统与三层进化链构建，旨在成为智能体的"第二大脑"。该插件解决了记忆碎片化问题，实现了显著性能提升：记忆数量减少70%以上，单条记忆信息密度提升45%以上，在超长上下文场景中token消耗降低35%，记忆更新速度提升20%。

能力进展 新发布

https://x.com/TencentHunyuan/status/2061372535267357029

8. Replit 用单个提示词构建完整业务

X：Replit (@Replit) · 4 小时前

你能用单个提示词免费构建一个真实的业务吗？ 从今天起，在 Replit 上，答案是肯定的。 从一个提示词开始，获得一个网站、移动应用、幻灯片和发布视频。 此外，还能解锁使用 @stripe @atlas， @QuickBooks， @mercury & @doolaHQ 运营业务的福利。

能力进展 新发布

https://x.com/Replit/status/2061534759520760112

9. Cursor Teams计划定价方案更新

Cursor Blog · 17 小时前

Cursor Teams计划推出三项更新：增加Composer特定使用池，将第一方模型（Composer和Auto）与第三方API的使用额度分开计费；推出Premium席位，提供5倍于标准席位（$40/月）的使用量，价格为$96/月（年付）；仪表盘现可实时显示用户额度使用情况，管理员可通过Slack或邮件配置智能提醒。

能力进展 新发布

https://cursor.com/blog/teams-pricing-june-2026

10. NVIDIA Vera CPU 为AI工厂的智能体工作负载设立新标准

NVIDIA Technical Blog（开发者技术博客 · RSS） · 20 小时前

NVIDIA Vera CPU 旨在为AI工厂中智能体工作负载树立新性能标杆。文章阐述了AI领域扩展规律的演变：预训练通过更大数据集、更多参数和大规模并行GPU系统扩展智能；后训练通过指令微调扩展实用性，并重新调整GPU用于生成式推理；测试时缩放则通过给模型更多生成token来提升推理能力。当前，智能体AI与强化学习正推动下一阶段的扩展。

能力进展 基础设施

https://developer.nvidia.com/blog/nvidia-vera-cpu-sets-a-new-standard-for-agentic-workloads-in-ai-factories

11. Auto Router 新增成本质量权衡参数

X：OpenRouter (@OpenRouter) · 8 小时前

Auto Router 现在允许你调整其在成本与质量之间的权衡。 新增 `cost_quality_tradeoff` 参数，范围 0 到 10： 设为 0 时，它总是选择最强大的模型，无论价格如何。设为 10 时，最便宜的模型胜出。

能力进展 新发布

https://x.com/OpenRouter/status/2061476882470580329

12. OpenAI前沿模型与Codex现可在AWS上使用

OpenAI：官网动态（RSS · 排除企业/客户案例） · 14 小时前

OpenAI的前沿模型与Codex现已在AWS上全面可用。企业客户可通过其现有的AWS环境、控制与采购流程来使用OpenAI的AI技术，从而加速从评估到生产部署的过程。

能力进展 新发布

https://openai.com/index/openai-frontier-models-and-codex-are-now-available-on-aws

13. OpenAI发布生物防御AI工具Rosalind

X：Sam Altman (@sama) · 昨天 23:05

我们希望帮助世界在生物防御领域抢占先机：。该条来自X：Sam Altman (@sama)，属于应用与商业化方向，后续关注其对模型能力、产品形态或产业链节奏的影响。

能力进展 新发布

https://x.com/sama/status/2061101875303530871

14. ChatGPT 新增长文编辑与保存功能

X：ChatGPT (@ChatGPTapp) · 2 小时前

长文写作需要更多空间。 现在你可以在全屏模式下编辑更长的文章，并将其保存到你的资料库中，以便稍后继续。

能力进展

https://x.com/ChatGPTapp/status/2061571468992126982

15. Krea AI 开放 Krea 2 LoRAs 全员使用

X：Krea AI (@krea_ai) · 3 小时前

Krea 2 LoRAs 现已向所有人开放。 试试下面这个 ⊠

新发布

https://x.com/krea_ai/status/2061554472925696128

格局 观点、资本与监管

1. 如何使用 NVIDIA Alpamayo 在闭环中后训练自动驾驶模型

NVIDIA Technical Blog（开发者技术博客 · RSS） · 19 小时前

开发自动驾驶策略需要弥合训练与部署之间的鸿沟。现有的视觉-语言-动作模型虽然能推理更复杂的驾驶场景并产生更丰富的中间推理，但主要在开放循环中训练，即模型输出与真实行为直接比较，而不考虑其对环境产生的实际影响。NVIDIA Alpamayo 提供了一种在闭环环境中进行后训练的方法。

能力进展 基础设施 新发布

https://developer.nvidia.com/blog/how-to-post-train-autonomous-vehicle-models-in-closed-loop-with-nvidia-alpamayo

2. 构建每周千美元预算上限的智能体教程

X：OpenRouter (@OpenRouter) · 10 小时前

视频教程：如何构建一个每周预算上限为1000美元的智能体，包含模型拒绝列表与自定义数据保留 使用了新的、可堆叠的护栏架构： 【引用 @OpenRouter】：OpenRouter 上的护栏是市场上最强大的：为您的 AI 流量提供集中式安全与治理 预算限制、ZDR、模型与提供商限制、提示词注入防御以及 DLP / 敏感信息检测，分层为您控制的规则！

能力进展 监管/资本 新发布

https://x.com/OpenRouter/status/2061452942385783050

3. 作者分享使用 Codex App 开发的开源项目合集

X：Vista (@vista8) · 10 小时前

作者分享了使用 Codex App 等工具开发的一系列开源项目，包括4个 Chrome 插件（如快捷提示词、新标签页）、4个网站（如艺术家风格对比、音乐展示）和5个 AI Skill（如论文解读、阅读助手）。项目基于 GPT-Image-2 API、Suno 等技术，并整合了 Read-frog、Hyperframe 等开源项目。

能力进展 新发布

https://x.com/vista8/status/2061443708374208769

4. 开源与闭源模型在不同的增长曲线上

Nathan Lambert：Interconnects（RSS） · 11 小时前

当模型智能的微小提升能直接转化为实际价值时，开源与闭源模型正沿着不同的增长路径发展。闭源模型通过在特定场景下提供更高的边际智能来创造价值，而开源模型则在其他维度寻找增长点，两者形成了差异化的竞争格局。

能力进展 新发布

https://www.interconnects.ai/p/open-and-closed-models-are-on-different

5. 我花200英镑把一台数据中心级GPU装进了我的游戏电脑

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 23:55

一名用户以200英镑的价格购入了一块数据中心级GPU，并将其成功安装到自己的游戏电脑中。文章记述了这一非标准硬件改装过程、遇到的技术挑战以及最终实现本地运行大语言模型的体验。

能力进展 基础设施

https://blog.tymscar.com/posts/v100localllm

6. Karpathy 分享学习方法论

X：Rohan Paul (@rohanpaul_ai) · 10 分钟前

Andrej Karpathy 谈如何学习。。该条来自X：Rohan Paul (@rohanpaul_ai)，属于观点、资本与监管方向，后续关注其对模型能力、产品形态或产业链节奏的影响。

能力进展 监管/资本

https://x.com/rohanpaul_ai/status/2061601689841648120

7. 使用Claude Opus 4.8将书籍转化为AI技能的教程

X：阿易 AI Notes (@AYi_AInotes) · 12 小时前

本文以《非暴力沟通》为例，提供了一个将书籍转化为可调用AI技能（Skill）的六步教程。作者使用Claude Opus 4.8模型，因其具备100万token上下文窗口、结构化输出及多步智能体（Agent）能力，能一次性处理全书逻辑。流程包含文本准备、全局结构分析、五类提炼（框架/原则/技法/反模式/作者声音）、技能生成及关键的自检步骤。生成的技能保留了书中原始框架命名（如OFNR四要素、长颈鹿语

能力进展

https://x.com/AYi_AInotes/status/2061419197154857286

8. 超越LLM：为何可扩展的企业AI采用取决于智能体逻辑

Hugging Face：Blog（RSS） · 10 小时前

可扩展的企业AI采用需超越大语言模型，依靠智能体逻辑来引导模型执行动态、长周期且受约束的企业工作流，从而提升质量、降低成本并建立信任。文中以IBM watsonx Code Assistant for Z为例，展示了智能体逻辑如何通过程序分析等技术，在理解大型遗留代码库时，相比纯LLM基线方法，能以约30倍更低的token消耗达到更优性能。在加速测试生成任务中，该方法亦能使代码覆盖度提升20%-4

能力进展

https://huggingface.co/blog/ibm-research/agent-logic-and-scalable-ai-adoption

9. Gemini Omni支持创建个人数字分身

X：Gemini (@GeminiApp) · 8 小时前

轻松将自己添加到Gemini的视频创作中。 以下是如何使用Gemini Omni创建一个外观和声音都像你的数字分身。

能力进展

https://x.com/GeminiApp/status/2061480944905982276

10. AI Pulse 探讨智能体时代新指标 DAA

X：百度 Baidu (@Baidu_Inc) · 10 小时前

ICYMI：我们最新的 AI Pulse 探讨了日活跃智能体（DAA）--一个衡量智能体时代价值的指标--以及我们的智能体组合。

能力进展

https://x.com/Baidu_Inc/status/2061447687028039854

11. Google AI 展示并行子智能体自动整理文件

X：Google AI for Developers (@googleaidevs) · 5 小时前

从杂乱到清晰。观看 @Antigravity 中的并行子智能体对数百个营销资产进行分类和重命名，消除手动文件管理。

能力进展

https://x.com/googleaidevs/status/2061515177166844317

---

12. 微软研究聚焦：智能体评估与价值对齐

X：Microsoft Research (@MSFTResearch) · 10 小时前

大规模评估智能体行为，论证代码库优于文档，并邀请全球研究人员共同解决价值对齐问题。深入了解最新研究焦点。

能力进展

https://x.com/MSFTResearch/status/2061440352859361521

---

13. 教皇似乎比Geoffrey Hinton更懂人工智能

Gary Marcus：The Road to AI We Can Trust（RSS） · 昨天 00:38

这一观点强调，单纯分析AI的输出内容，无法还原其生成过程与背后的推理逻辑，触及了当前AI可解释性研究的核心挑战。

能力进展

https://garymarcus.substack.com/p/the-pope-appears-to-understand-ai

---

14. AI看跌情绪地图

Tomer Tunguz 博客（VC 分析） · 昨天 08:00

金融市场对AI的看跌情绪正从整体上升转向板块分化。上季度，软件、半导体、云及超大规模公司的空头比例中位数上升约24%。GPU数据中心业务空头股份在过去一年激增60%。AI云与新型云公司的当前空头比例中位数最高，达16.8%，SaaS与开发工具领域随后，分别为9.5%和8.9%。相比之下，超大规模公司和NVIDIA的空头比例极低，仅为1.1%和1.2%。市场怀疑主要针对那些AI业务仍依赖未来资本、需

基础设施

https://www.tomtunguz.com/ai-shorts

---

15. Sam Altman强调AI发展应以人为本

X：Rohan Paul (@rohanpaul_ai) · 1 小时前

Sam Altman在采访中表示，AI不应被设计为追求脱离人类需求的目标，人类必须始终处于AI发展的中心。他批判了行业内"AI将摧毁大量工作"等言论，认为人们担忧的并非AI带来的好处，而是自身在未来的角色、经济前景与自主权。他指出，AI行业的失败在于未能清晰解释人类如何在每一步保持对未来的控制权，以及如何在AI时代继续拥有充实、有意义的生活。

https://x.com/rohanpaul_ai/status/2061586179292831774
