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: 18 条 焦点: 8 条 快讯: 0 条

Executive Summary

今日AI行业核心进展集中在模型开源与性能优化方面。MiniMax 开源 M3 模型权重，该 428B 总参数模型采用 MSA（MiniMax Sparse Attention）架构，显著降低长上下文计算成本，是首个从预训练阶段就进行多模态交错混合训练的开源模型。Z Lab 与 SGLang 团队联合发布 DFlash 投机解码模型，采用块扩散+KV注入并行生成技术，在 Qwen 3.5 397B-A17B 模型上实现显著性能提升。月之暗面 推出 Kimi K2.7 Code 高速版，输出速度提升5-6倍达到180 Token/s，定价为普通版2倍。

行业结构变化体现为生态合作深化与资本布局加速。OpenAI 推出 Partner Network 并投资1.5亿美元支持全球合作伙伴，设立三级认证体系推动解决方案生态建设。Salesforce 以36亿美元收购 AI客服平台Fin，强化其企业级 Agentforce 平台能力。Nvidia 发行200亿美元债券加入AI债务融资热潮，为基础设施扩张提供资金支持。Meta 在Facebook上线"AI Mode"功能，基于公开信息合成答案，推动社交平台AI化转型。

后续观察变量包括模型开放节奏与商业化平衡，M3 模型的多模态训练范式能否引领开源趋势。算力成本结构变化值得关注，Flash-KMeans 等专门化算法的GPU优化效果及其对AI基础设施成本的影响。监管政策走向成为关键因素，白宫AI监管决策的透明度争议可能影响全球AI治理格局。企业AI应用落地速度将受 Salesforce 收购 Fin 等交易影响，AI客服等垂直领域商业化进程值得持续跟踪。

重点 今日核心进展

★ 1. MiniMax 开源 M3 模型权重及 MSA 技术论文

公众号：MiniMax（稀宇科技） · 9 小时前 · 模型与工具能力

MiniMax 上周五开源了 428B 总参数、23B 激活参数的 M3 模型权重，同步发布 MSA（MiniMax Sparse Attention）技术论文，该架构显著降低长上下文计算成本。M3 是首个从预训练阶段就进行文本、图像等多模态交错混合训练的开源模型。发布两周后，M3 在 Artificial Analysis 综合智能指数、GDPval-AA 排行榜均获开源模型第一，Code Ar

能力进展 基础设施 新发布

https://mp.weixin.qq.com/s/AW6L89QZkwN-jD27hQ84ww

★ 2. 下一代投机解码：DFlash 与 Spec V2

LMSYS：Blog（Chatbot Arena 团队） · 6 小时前 · 模型与工具能力

Z Lab、Modal 与 SGLang 团队联合发布 DFlash 投机解码模型和 SGLang 的默认 Spec V2 引擎。DFlash 采用块扩散+KV 注入并行生成整块 draft token，在 Qwen 3.5 397B-A17B（BF16）的 HumanEval 数据集上、并发 1 时吞吐量达到基线的 4.3

能力进展 新发布

https://www.lmsys.org/blog/2026-06-15-next-generation-speculative-decoding-dflash-v2

★ 3. Flash-KMeans：IO感知的精确K-Means，在GPU上比FAISS快200倍以上

MarkTechPost（RSS） · 15 小时前 · 应用与商业化

UC Berkeley与UT Austin团队开源Flash-KMeans（Apache 2.0，`pip install flash-kmeans`），精确实现标准Lloyd's k-Means，通过重构GPU数据流而非改变数学或近似来提速。在NVIDIA H200上，端到端速度比最佳基线快17.9×，比cuML快33×，比FAISS快200×以上。其FlashAssign核避免物化完整N×K距

能力进展 基础设施 新发布

https://www.marktechpost.com/2026/06/15/meet-flash-kmeans-an-io-aware-exact-k-means-that-runs-over-200x-faster-than-faiss-on-gpus

★ 4. OpenAI 推出合作伙伴网络 OpenAI Partner Network

OpenAI：官网动态（RSS · 排除企业/客户案例） · 昨天 01:00 · 产业与基础设施

OpenAI 宣布推出 OpenAI Partner Network，并投资 1.5 亿美元支持全球合作伙伴构建、销售和交付 AI 解决方案。该计划设立 Select、Advanced、Elite 三级合作伙伴层级，提供 Codex、网络安全、智能体等专业方向认证，并试点 Forward Deployed Experts 项目以支持复杂企业部署。目标在 2026 年底前培训认证 30 万名顾问。案

能力进展 监管/资本 新发布

https://openai.com/index/introducing-openai-partner-network

★ 5. 6倍速！Kimi K2.7 Code 高速版已上线

公众号：月之暗面（Kimi） · 13 小时前 · 应用与商业化

Kimi K2.7 Code 高速版上线，与普通版为同一模型，输出速度约 5-6 倍，常规编程场景约 180 Token/s，短上下文可达 260 Token/s。API 定价为普通版 2 倍，模型 ID：kimi-k2.7-code-highspeed。Kimi Code Plan 用户可通过「抢先体验计划」使用，用量消耗为普通版 3 倍。使用须开启思考模式，关闭会报错或回退至 K2.6。庆祝发

能力进展 新发布

https://mp.weixin.qq.com/s/p87ebkY1xqKtkGZ2N3DGSw

★ 6. Meta 在 Facebook 上线"AI Mode"，基于平台公开信息合成答案

TechCrunch：AI（RSS） · 5 小时前 · 应用与商业化

Meta 宣布在 Facebook 推出"AI Mode"搜索功能，利用 Meta AI 从公开帖子（含群组和 Reels）提取信息并合成答案，用户可用自然语言提问获得摘要。同时新增视频拼贴剪辑、过渡效果及 AI 照片预设（可更换服装、发型和配饰），体育迷可在 Stories 中点击"AI Edit"虚拟穿上队服。这些更新延续了此前动态头像、Marketplace 自动回复和创作者 AI 助手的部

能力进展 新发布

https://techcrunch.com/2026/06/15/metas-new-ai-mode-on-facebook-pulls-from-public-info-across-its-platforms

★ 7. Salesforce以36亿美元收购AI客服平台Fin

TechCrunch：AI（RSS） · 9 小时前 · 产业与基础设施

Salesforce宣布以36亿美元收购AI客服平台Fin（前身为Intercom）。Fin提供可跨实时聊天、WhatsApp、短信、电话、Slack等多渠道解决客户问题的AI智能体。Salesforce计划利用Fin的技术和团队增强其企业级Agentforce平台，该平台允许企业构建自定义AI智能体以自动化任务。交易预计在Salesforce 2027财年第四季度（即2027年初）完成。Fin联

能力进展 监管/资本

https://techcrunch.com/2026/06/15/salesforce-acquires-ai-customer-service-platform-fin-for-3-6b

★ 8. Grok Build 推出 Agent Dashboard 管理多个编码会话

xAI：News（网页） · 昨天 08:00 · 应用与商业化

xAI 为 Grok Build 推出 Agent Dashboard，提供单一屏幕管理多个编码会话。仪表板按状态分组（等待输入、工作中、空闲），每行显示状态标记、名称、分支、权限模式和当前操作。选中会话可打开 peek 面板查看最新输出并直接回复，等待输入的会话支持用箭头键或数字键选择选项。底部输入框用于分派新会话，支持设置模型、启动计划模式或自动批准编辑。通过 `grok dashboard`

能力进展 新发布

https://x.ai/news/agent-dashboard

能力 模型与工具能力

1. MiniMax 开源 M3 模型权重及 MSA 技术论文

公众号：MiniMax（稀宇科技） · 9 小时前

MiniMax 上周五开源了 428B 总参数、23B 激活参数的 M3 模型权重，同步发布 MSA（MiniMax Sparse Attention）技术论文，该架构显著降低长上下文计算成本。M3 是首个从预训练阶段就进行文本、图像等多模态交错混合训练的开源模型。发布两周后，M3 在 Artificial Analysis 综合智能指数、GDPval-AA 排行榜均获开源模型第一，Code Ar

能力进展 基础设施 新发布

https://mp.weixin.qq.com/s/AW6L89QZkwN-jD27hQ84ww

2. 下一代投机解码：DFlash 与 Spec V2

LMSYS：Blog（Chatbot Arena 团队） · 6 小时前

Z Lab、Modal 与 SGLang 团队联合发布 DFlash 投机解码模型和 SGLang 的默认 Spec V2 引擎。DFlash 采用块扩散+KV 注入并行生成整块 draft token，在 Qwen 3.5 397B-A17B（BF16）的 HumanEval 数据集上、并发 1 时吞吐量达到基线的 4.3

能力进展 新发布

https://www.lmsys.org/blog/2026-06-15-next-generation-speculative-decoding-dflash-v2

产业 产业与基础设施

1. OpenAI 推出合作伙伴网络 OpenAI Partner Network

OpenAI：官网动态（RSS · 排除企业/客户案例） · 昨天 01:00

OpenAI 宣布推出 OpenAI Partner Network，并投资 1.5 亿美元支持全球合作伙伴构建、销售和交付 AI 解决方案。该计划设立 Select、Advanced、Elite 三级合作伙伴层级，提供 Codex、网络安全、智能体等专业方向认证，并试点 Forward Deployed Experts 项目以支持复杂企业部署。目标在 2026 年底前培训认证 30 万名顾问。案

能力进展 监管/资本 新发布

https://openai.com/index/introducing-openai-partner-network

2. Salesforce以36亿美元收购AI客服平台Fin

TechCrunch：AI（RSS） · 9 小时前

Salesforce宣布以36亿美元收购AI客服平台Fin（前身为Intercom）。Fin提供可跨实时聊天、WhatsApp、短信、电话、Slack等多渠道解决客户问题的AI智能体。Salesforce计划利用Fin的技术和团队增强其企业级Agentforce平台，该平台允许企业构建自定义AI智能体以自动化任务。交易预计在Salesforce 2027财年第四季度（即2027年初）完成。Fin联

能力进展 监管/资本

https://techcrunch.com/2026/06/15/salesforce-acquires-ai-customer-service-platform-fin-for-3-6b

3. Nvidia 加入 AI 债务热潮，发行 200 亿美元债券

The Decoder：AI News（RSS） · 8 小时前

Nvidia 计划通过自 2021 年以来的首次债券发行筹集至少 200 亿美元，消息援引知情人士透露。此举标志着 Nvidia 加入 AI 领域的债务融资热潮。

基础设施 监管/资本

https://the-decoder.com/nvidia-joins-ai-debt-boom-with-20-billion-bond-sale

---

应用 应用与商业化

1. Flash-KMeans：IO感知的精确K-Means，在GPU上比FAISS快200倍以上

MarkTechPost（RSS） · 15 小时前

UC Berkeley与UT Austin团队开源Flash-KMeans（Apache 2.0，`pip install flash-kmeans`），精确实现标准Lloyd's k-Means，通过重构GPU数据流而非改变数学或近似来提速。在NVIDIA H200上，端到端速度比最佳基线快17.9×，比cuML快33×，比FAISS快200×以上。其FlashAssign核避免物化完整N×K距

能力进展 基础设施 新发布

https://www.marktechpost.com/2026/06/15/meet-flash-kmeans-an-io-aware-exact-k-means-that-runs-over-200x-faster-than-faiss-on-gpus

---

2. 6倍速！Kimi K2.7 Code 高速版已上线

公众号：月之暗面（Kimi） · 13 小时前

Kimi K2.7 Code 高速版上线，与普通版为同一模型，输出速度约 5-6 倍，常规编程场景约 180 Token/s，短上下文可达 260 Token/s。API 定价为普通版 2 倍，模型 ID：kimi-k2.7-code-highspeed。Kimi Code Plan 用户可通过「抢先体验计划」使用，用量消耗为普通版 3 倍。使用须开启思考模式，关闭会报错或回退至 K2.6。庆祝发

能力进展 新发布

https://mp.weixin.qq.com/s/p87ebkY1xqKtkGZ2N3DGSw

---

3. Meta 在 Facebook 上线"AI Mode"，基于平台公开信息合成答案

TechCrunch：AI（RSS） · 5 小时前

Meta 宣布在 Facebook 推出"AI Mode"搜索功能，利用 Meta AI 从公开帖子（含群组和 Reels）提取信息并合成答案，用户可用自然语言提问获得摘要。同时新增视频拼贴剪辑、过渡效果及 AI 照片预设（可更换服装、发型和配饰），体育迷可在 Stories 中点击"AI Edit"虚拟穿上队服。这些更新延续了此前动态头像、Marketplace 自动回复和创作者 AI 助手的部

能力进展 新发布

https://techcrunch.com/2026/06/15/metas-new-ai-mode-on-facebook-pulls-from-public-info-across-its-platforms

---

4. Grok Build 推出 Agent Dashboard 管理多个编码会话

xAI：News（网页） · 昨天 08:00

xAI 为 Grok Build 推出 Agent Dashboard，提供单一屏幕管理多个编码会话。仪表板按状态分组（等待输入、工作中、空闲），每行显示状态标记、名称、分支、权限模式和当前操作。选中会话可打开 peek 面板查看最新输出并直接回复，等待输入的会话支持用箭头键或数字键选择选项。底部输入框用于分派新会话，支持设置模型、启动计划模式或自动批准编辑。通过 `grok dashboard`

能力进展 新发布

https://x.ai/news/agent-dashboard

---

5. OpenRouter新增免费模型gpt-oss-20b和Gemma4 26B

X：OpenRouter (@OpenRouter) · 6 小时前

OpenRouter 上新增免费容量，由 @eigenlabs 的 Darkbloom 提供：gpt-oss-20b 和 Gemma 4 26B。 今天就开始使用这些模型吧 ↓

能力进展 新发布

https://x.com/OpenRouter/status/2066585705581797616

格局 观点、资本与监管

1. AI 应用黄金时代已至：Fable 被禁、Nadella 的护城河论点与 Salesforce 收购 Fin

Tomer Tunguz 博客（VC 分析） · 昨天 08:00

美国政府关闭 Anthropic 的 Fable 访问，开源和本地模型成必备；Satya Nadella 主张 AI 生态护城河应是人类专业知识和模型外围系统；Salesforce 以 36 亿美元收购 Fin（前 Intercom），Fin 利用开源模型实现性价比。这三件事标志 AI 应用进入黄金时代。构建 AI 应用的难点：在 Kimi K2.6、Qwen 3.6 27b、GLM 5.1 等不

能力进展 监管/资本 新发布

https://www.tomtunguz.com/golden-age-of-applications

---

2. AI裁员浪潮成为火药桶

TechCrunch：AI（RSS） · 16 小时前

今年科技公司已累计裁员约15万人，日均974人，速度比去年快44%；上月裁员近4万创两年新高，AI连续三个月被列为裁员首要原因。Block近半数员工被裁后，CEO Jack Dorsey否认AI是根源，Marc Andreessen则称AI只是"银弹借口"。Uber裁撤23%人事部门，但此前CTO透露AI编码预算四个月内耗尽。与此同时，AI芯片商Cerebras上市首日市值达670亿美元，Spac

基础设施 监管/资本 新发布

https://techcrunch.com/2026/06/15/the-ai-layoff-wave-is-becoming-a-powder-keg

---

3. 乔木小说创作 Skill 开源发布

X：Vista (@vista8) · 昨天 22:27

开源乔木小说创作 Skill，用户只需说"我想写一个小说"或指定风格，AI 自动生成剧情梗概、人物设定、钩子、经典桥段、人物欲望、冲突升级和结尾。与 AI 讨论确认后，可生成完整、低 AI 味的小说。安装命令：npx skills add joeseesun/qiaomu-novel-generator，Github 开源地址见评论区。

能力进展 新发布

https://x.com/vista8/status/2066165703443726749

4. 白宫AI监管决定被指偏袒OpenAI与亚马逊

Gary Marcus：The Road to AI We Can Trust（RSS） · 昨天 00:15

白宫周五做出的AI监管决定被指偏袒OpenAI、亚马逊等企业，同时对Anthropic施压不足24小时，缺乏透明度和事实依据。Gary Marcus、Dean W Ball及卡托研究所Kevin Frazier等专家指出，这种由少数人闭门快速决策的做法带有腐败嫌疑，可能促使其他国家加速发展"主权AI"甚至中国AI，并导致美国人才流失。Anthropic声明称政府应在法定程序中基于技术事实阻止不安全

监管/资本 新发布

https://garymarcus.substack.com/p/what-washington-must-do

5. Satya Nadella：没有生态的前沿不稳定

X：Satya Nadella (@satyanadella) · 昨天 23:33

微软CEO Satya Nadella认为，AI驱动的平台转变首次实现人与数字系统间的认知循环。企业需同时构建人力资本（知识、判断、关系）与token资本（自有的AI能力），且人力资本不会贬值，反而随token资本增长而增值。真正的机会在于建立人力资本与token资本复合增长的学习循环--企业应能替换通用模型而不丢失已内化的专家知识，通过私有评估和强化学习让模型从内部真实轨迹中持续提升。他警告，若

能力进展

https://x.com/satyanadella/status/2066182223213293753

6. 项目负责人揭秘为何苹果 AI 版 Siri 姗姗来迟：推倒重来，彻底重构

IT之家（RSS） · 1 小时前

苹果AI版Siri迟迟未上线，项目负责人迈克·罗克韦尔在WWDC技术分享会上透露，去年团队曾做出在原有Siri基础上小幅改良、新增工具调用的可运行版本，但因无法达到产品愿景，最终选择推倒重来，完整从零重构系统，依托全新大模型搭建。重构后的Siri拥有独立应用程序，原生支持多模态交互，隐私保护贯穿底层架构，并覆盖iPhone、iPad、Mac、Apple Watch、Vision Pro、CarPl

能力进展

https://www.ithome.com/0/964/620.htm

7. GitHub Copilot CLI 初学者指南：常用斜杠命令概览

GitHub Blog · 4 小时前

GitHub Copilot CLI 为初学者提供了常用斜杠命令的概述，帮助用户通过命令控制终端中的 AI 智能体。

能力进展

https://github.blog/ai-and-ml/github-copilot/github-copilot-cli-for-beginners-overview-of-common-slash-commands

8. Skydio CEO Adam Bry：硅谷不应为无人机使用画红线

The Verge：AI（RSS） · 10 小时前

Skydio是美国最大的无人机制造商，主攻公共安全、军事、能源、基建巡检等企业市场。CEO Adam Bry表示，特朗普政府去年底禁止中国产无人机后，廉价消费级无人机几乎消失，Skydio产品成为主要替代方案。公司认为无人机正从工具转向自主基础设施--通过机库、远程操控和软件整合实现规模化应用，AI在其中扮演关键角色。访谈还涉及Skydio与军方合作的态度，以及自主技术如何带动公司扩张。

监管/资本

https://www.theverge.com/podcast/949195/skydio-ceo-adam-bry-autonmous-drones-china-red-lines-military