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: 17 条 焦点: 8 条 快讯: 0 条

Executive Summary

今日AI行业迎来多项重要模型发布与开源进展。MiniMax发布M3开源权重模型，具备约428B总参数、23B激活参数，在编码与智能体能力方面表现突出，已上架HuggingFace。智谱的GLM-5.2全量开放，支持真正的1M上下文长度，被定位为最强国产Coding模型，计划下周开源并上线API。字节豆包推出"任务模式"，支持定时执行、零代码网页生成等功能，原"思考模式"升级为"专家模式"，强化深度推理能力。

行业监管与资本动态显著影响竞争格局。Anthropic面临美国政府对其模型的整治行动，可能源于亚马逊CEO与美国官员的会谈，该公司同时秘密提交IPO申请，估值高达9650亿美元。谷歌遭遇内部动荡，Android安全负责人因反对军事AI合作辞职，批评公司AI能耗问题。OpenAI正接受多州总检察长联合调查。基础设施层面，OpenRouter推出Fusion API，以半价实现Fable级别智能，Hermes Agent已处理超过17万亿tokens。

后续需关注模型开源节奏与监管政策走向。MiniMax M3和智谱GLM-5.2的开源实施情况将影响开源生态竞争，Anthropic的监管应对措施可能重塑企业AI合规标准。算力成本优化方面，OpenRouter的定价策略和Suno的音轨分离技术升级值得关注。企业AI转型实践如Meta的组织调整经验，将为行业提供重要参考。

重点 今日核心进展

★ 1. MiniMax M3 开源权重模型发布，已上架 HuggingFace

X：MiniMax (@MiniMax_AI) · 昨天 22:11 · 模型与工具能力

MiniMax 发布开源权重模型 M3，约 428B 总参数、23B 激活参数，已上传 HuggingFace。该模型融合三种前沿能力：编码与智能体方面达 59.0% SWE-Bench Pro、66.0% Terminal Bench 2.1、34.8% SWE-fficiency、28.8% KernelBench Hard、74.2% MCP Atlas；采用 MiniMax 稀疏注意力将上

能力进展 新发布

https://x.com/MiniMax_AI/status/2065436935188058208

★ 2. 智谱 GLM-5.2 全量开放，支持 1M 上下文且下周开源

公众号：智谱（GLM） · 18 小时前 · 模型与工具能力

GLM-5.2 是智谱迄今能力最强的开源模型，支持真正可用的 1M 上下文，在长程任务中继续保持领先，并被智谱称为最强的国产 Coding 模型。今晚 5：21 起面向 GLM Coding Plan 全量用户开放（覆盖 Lite、Pro、Max、团队版）。API 将于下周上线，模型下周正式开源，遵循 MIT 协议。

能力进展 新发布

https://mp.weixin.qq.com/s/LDrbtLM0wiCTJorvd5GY9w

★ 3. 谷歌Android安全负责人因反对军事AI合作辞职

IT之家（RSS） · 15 小时前 · 产业与基础设施

谷歌Android平台安全负责人René Mayrhofer辞职，他在5月18日内部告别信中指责公司"丧失道德指针"，批评谷歌悄悄放弃碳中和目标（因AI模型能耗），并与美国战争部签署允许AI用于"任何合法目的"的协议。今年4月下旬谷歌宣布向五角大楼提供AI用于机密工作，2025年2月更新AI原则时移除了不使用AI开发武器或监控工具的承诺。Mayrhofer担忧谷歌AI产品可能被用于针对公民的大规模

能力进展 监管/资本 新发布

https://www.ithome.com/0/963/888.htm

★ 4. 亚马逊首席执行官与美国官员会谈引发对 Anthropic 模型的整治

Hacker News 热门（buzzing.cc 中文翻译） · 5 小时前 · 产业与基础设施

亚马逊 CEO 与美国官员的会谈直接导致美国政府对 Anthropic 公司的 AI 模型采取整治行动。此次事件涉及对 Anthropic 旗下大语言模型的监管升级，具体措施及模型版本细节尚未披露。

能力进展 监管/资本 新发布

https://www.wsj.com/tech/ai/amazon-ceos-talks-with-u-s-officials-triggered-crackdown-on-anthropic-models-dcc90578

★ 5. olmo-eval：面向模型开发循环的评估工作台

Hugging Face：Blog（RSS） · 昨天 23:56 · 应用与商业化

olmo-eval 是基于 OLMES 标准构建的评估工作台，专为 LLM 持续开发中的反复评测场景设计。相比 OLMES，它减少了新增评测的实现工作量，支持 agentic 和多轮评测作为一等用例，并允许根据基准需求选择轻量直接运行或容器化隔离运行。采用模块化架构，模型、工具、容器环境、辅助模型均可独立替换。评测结果同时报告分数、标准误差和最小可检测效应。与 Harbor 侧重于发布不同，olm

能力进展 新发布

https://huggingface.co/blog/allenai/olmo-eval

★ 6. 字节豆包上线"任务模式"：支持定时执行与文件生成，"思考模式"升级为"专家模式"

IT之家（RSS） · 昨天 23:33 · 应用与商业化

6月12日，字节跳动旗下AI应用豆包大范围上线"任务模式"，支持定时执行、零代码网页生成、一键PPT生成、数据可视化分析等全链路Agent执行。原"思考模式"升级为"专家模式"，调用豆包大模型2.0 Pro版本，强化深度推理能力。App顶部模式切换改为"快速、专家、任务"。基础功能免费，高阶服务付费，专业版三档：标准版68元/月或688元/年，加强版200元/月或2048元/年，专业版500元/月

能力进展 新发布

https://www.ithome.com/0/963/725.htm

★ 7. Fusion API：半价达Fable级智能

X：OpenRouter (@OpenRouter) · 6 小时前 · 应用与商业化

推出Fusion API，市场上最智能的复合模型。 Fusion以一半的价格实现Fable级别的智能。 工作原理如下⊠

能力进展 新发布

https://x.com/OpenRouter/status/2065856853989270011

★ 8. Anthropic的安全警告可能适得其反--政府已撤回其最强大AI

TechCrunch：AI（RSS） · 21 小时前 · 产业与基础设施

Anthropic对政府撤回其最强大AI模型表达不满，称仅基于一个狭窄的潜在越狱发现就召回已部署给数亿用户的商业模型不合理。

能力进展 监管/资本

https://techcrunch.com/2026/06/12/anthropics-safety-warnings-may-have-just-backfired-the-government-has-pulled-the-plug-on-its-most-powerful-ai

能力 模型与工具能力

1. MiniMax M3 开源权重模型发布，已上架 HuggingFace

X：MiniMax (@MiniMax_AI) · 昨天 22:11

MiniMax 发布开源权重模型 M3，约 428B 总参数、23B 激活参数，已上传 HuggingFace。该模型融合三种前沿能力：编码与智能体方面达 59.0% SWE-Bench Pro、66.0% Terminal Bench 2.1、34.8% SWE-fficiency、28.8% KernelBench Hard、74.2% MCP Atlas；采用 MiniMax 稀疏注意力将上

能力进展 新发布

https://x.com/MiniMax_AI/status/2065436935188058208

2. 智谱 GLM-5.2 全量开放，支持 1M 上下文且下周开源

公众号：智谱（GLM） · 18 小时前

GLM-5.2 是智谱迄今能力最强的开源模型，支持真正可用的 1M 上下文，在长程任务中继续保持领先，并被智谱称为最强的国产 Coding 模型。今晚 5：21 起面向 GLM Coding Plan 全量用户开放（覆盖 Lite、Pro、Max、团队版）。API 将于下周上线，模型下周正式开源，遵循 MIT 协议。

能力进展 新发布

https://mp.weixin.qq.com/s/LDrbtLM0wiCTJorvd5GY9w

产业 产业与基础设施

1. 谷歌Android安全负责人因反对军事AI合作辞职

IT之家（RSS） · 15 小时前

谷歌Android平台安全负责人René Mayrhofer辞职，他在5月18日内部告别信中指责公司"丧失道德指针"，批评谷歌悄悄放弃碳中和目标（因AI模型能耗），并与美国战争部签署允许AI用于"任何合法目的"的协议。今年4月下旬谷歌宣布向五角大楼提供AI用于机密工作，2025年2月更新AI原则时移除了不使用AI开发武器或监控工具的承诺。Mayrhofer担忧谷歌AI产品可能被用于针对公民的大规模

能力进展 监管/资本 新发布

https://www.ithome.com/0/963/888.htm

2. 亚马逊首席执行官与美国官员会谈引发对 Anthropic 模型的整治

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

亚马逊 CEO 与美国官员的会谈直接导致美国政府对 Anthropic 公司的 AI 模型采取整治行动。此次事件涉及对 Anthropic 旗下大语言模型的监管升级，具体措施及模型版本细节尚未披露。

能力进展 监管/资本 新发布

https://www.wsj.com/tech/ai/amazon-ceos-talks-with-u-s-officials-triggered-crackdown-on-anthropic-models-dcc90578

3. Anthropic的安全警告可能适得其反--政府已撤回其最强大AI

TechCrunch：AI（RSS） · 21 小时前

Anthropic对政府撤回其最强大AI模型表达不满，称仅基于一个狭窄的潜在越狱发现就召回已部署给数亿用户的商业模型不合理。

能力进展 监管/资本

https://techcrunch.com/2026/06/12/anthropics-safety-warnings-may-have-just-backfired-the-government-has-pulled-the-plug-on-its-most-powerful-ai

4. OpenAI 遭多州总检察长联合调查

Bloomberg：Technology（RSS） · 23 小时前

OpenAI 正被一个由多州总检察长组成的联盟调查，该联盟已向这家人工智能公司索取涵盖广泛主题的信息。

新发布

https://www.bloomberg.com/news/articles/2026-06-13/openai-probed-by-coalition-of-state-attorneys-general

5. Anthropic 秘密申请上市，估值 9650 亿美元

Bloomberg：Technology（RSS） · 13 小时前

Anthropic，这家估值达 9650 亿美元的 AI 巨头、史上增长最快的初创公司之一，在秘密提交 IPO 申请后，再次投下重磅炸弹。

监管/资本

https://www.bloomberg.com/news/articles/2026-06-13/global-capitalism-bets-it-all-on-ai-future-that-alarms-voters

6. 扎克伯格承认 Meta AI 转型"脱轨"：裁员 10%、转岗 7000 人后组织调整过快

IT之家（RSS） · 17 小时前

路透社披露的 Meta 内部备忘录显示，CEO 扎克伯格承认公司在 AI 转型中组织调整节奏过快，带来员工安置、管理跨度等问题，预计未来"几乎肯定会犯更多错误"。Meta 今年 5 月已裁减全球 10% 员工，并将 7000 人转入 AI 相关新项目。为缓解协作问题，公司将增加团队建设预算，7 月举办黑客松。新成立的应用 AI 工程部门个人贡献者与管理者比例最高达 50：1。扎克伯格重申今年不再全

https://www.ithome.com/0/963/858.htm

应用 应用与商业化

1. olmo-eval：面向模型开发循环的评估工作台

Hugging Face：Blog（RSS） · 昨天 23:56

olmo-eval 是基于 OLMES 标准构建的评估工作台，专为 LLM 持续开发中的反复评测场景设计。相比 OLMES，它减少了新增评测的实现工作量，支持 agentic 和多轮评测作为一等用例，并允许根据基准需求选择轻量直接运行或容器化隔离运行。采用模块化架构，模型、工具、容器环境、辅助模型均可独立替换。评测结果同时报告分数、标准误差和最小可检测效应。与 Harbor 侧重于发布不同，olm

能力进展 新发布

https://huggingface.co/blog/allenai/olmo-eval

2. 字节豆包上线"任务模式"：支持定时执行与文件生成，"思考模式"升级为"专家模式"

IT之家（RSS） · 昨天 23:33

6月12日，字节跳动旗下AI应用豆包大范围上线"任务模式"，支持定时执行、零代码网页生成、一键PPT生成、数据可视化分析等全链路Agent执行。原"思考模式"升级为"专家模式"，调用豆包大模型2.0 Pro版本，强化深度推理能力。App顶部模式切换改为"快速、专家、任务"。基础功能免费，高阶服务付费，专业版三档：标准版68元/月或688元/年，加强版200元/月或2048元/年，专业版500元/月

能力进展 新发布

https://www.ithome.com/0/963/725.htm

3. Fusion API：半价达Fable级智能

X：OpenRouter (@OpenRouter) · 6 小时前

推出Fusion API，市场上最智能的复合模型。 Fusion以一半的价格实现Fable级别的智能。 工作原理如下⊠

能力进展 新发布

https://x.com/OpenRouter/status/2065856853989270011

4. Suno 音轨分离：从零生成更纯净

X：Suno (@suno) · 5 小时前

重大更新：Suno 的音轨分离刚刚大幅升级。⊠我们现在从零重新生成音轨，而非仅仅隔离频率。结果如何？纯净无伪影的音轨，可直接拖入你的 DAW。

新发布

https://x.com/suno/status/2065862499765821916

格局 观点、资本与监管

1. SemiAnalysis 洞察 Token 经济：200 美元 AI 订阅榨出 70 倍用量

IT之家（RSS） · 18 小时前

SemiAnalysis 购买了 Anthropic 和 OpenAI 的全部订阅方案，模拟高强度编码任务直至触及每周上限。月费 200 美元的 Claude Max 20x 方案，按 API 价格换算最高可消耗约值 8000 美元的 token；ChatGPT Pro 20x 方案对应最高约值 14000 美元的 token。用户通过订阅可获取 40 至 70 倍的 API 价值，该机构指出这种

能力进展 新发布

https://www.ithome.com/0/963/834.htm

2. Oran Ge 开源《人味儿写作心法.skill》解决AI写作缺人味

X：Oran Ge (@oran_ge) · 昨天 06:48

Oran Ge 让 Claude Fable 5 打磨文案三遍，发现改稿越来越讲究却缺"人味儿"。他与 AI 讨论后得出结论：人写的文字背后有"存在感"--作者在具体位置付出过具体代价，而 AI 无法复现。为此他制作了《人味儿写作心法.skill》，专用于自写文章或口述后让 AI 改稿的场景，旨在保留文字的人味。该技能已开源免费发布在 GitHub。

能力进展 新发布

https://x.com/oran_ge/status/2065566882774868125

3. Hermes Agent 在 OpenRouter 上的使用指南：设置、模型与路由

OpenRouter：Announcements（RSS） · 昨天 00:00

Hermes Agent 已通过 OpenRouter 处理超过 17 万亿 tokens。使用指南包括设置流程、选择支持 64K 上下文窗口的模型，以及调整路由策略以兼顾成本与可靠性。

能力进展 新发布

https://openrouter.ai/blog/tutorials/hermes-agent

---

4. 如何在OpenRouter上获得最低成本的LLM推理

OpenRouter：Announcements（RSS） · 昨天 00:00

在OpenRouter上追加`：floor`可获取最便宜提供商，通过`max_price`设定花费上限，并可免费使用20多个零成本模型。同时需注意避免计费陷阱。

能力进展 新发布

https://openrouter.ai/blog/tutorials/how-to-get-the-lowest-cost-llm-inference-on-openrouter

---

5. /architect：减少80%的Fable token，Fable负责协调/审核，Codex负责构建

Hacker News 热门（buzzing.cc 中文翻译） · 7 小时前

/architect项目将Fable token减少80%，由Fable进行协调和审核，Codex负责构建任务。

https://github.com/DanMcInerney/architect-loop
