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: 58 条 焦点: 8 条 快讯: 0 条

Executive Summary

今日AI行业呈现显著的技术突破与商业化加速趋势。BitCPM-CANN作为全球首个基于华为昇腾910B NPU全栈训练的1.58比特开源大模型发布，实现内存占用相比BF16降低约6倍的量化技术创新。网易有道"子曰4"多模态模型全量开源，27B参数模型在教育场景视觉数理问题处理上达到行业顶尖水平，纯文本中文数理难题准确率达81.4%。智谱GLM-5.1高速版API输出速度达400 tokens/s，刷新全球大模型API速度纪录，实现旗舰级能力与低延迟的结合突破。

基础设施与商业化格局发生重要变化。Project Glasswing项目通过Claude Mythos Preview模型已发现超过10,000个高危漏洞，合作伙伴报告漏洞发现效率提升超过十倍。Anthropic接近实现首个盈利季度，预计第二季度营收达109亿美元，运营利润5.59亿美元，标志AI实验室商业模式验证成功。DeepSeek推进700亿元人民币融资，估值约450亿美元，若成功将创中国科技初创公司首轮融资纪录。同时，GitHub连续第三年被Gartner评为企业级AI编程代理领域领导者，但面临平台稳定性挑战。

后续需关注模型开放策略与商业化平衡，API性能优化节奏，以及算力成本结构变化。NVIDIA扩散语言模型的光速级生成技术进展值得关注。监管层面，美国AI行政令取消后的政策走向，以及各厂商在开源与商业化的战略选择将影响行业竞争格局。企业应用落地方面，AI代理工具链成熟度和开发者接受度将是关键指标。

重点 今日核心进展

★ 1. 首个基于华为昇腾910B NPU全栈训练的1.58比特开源大模型BitCPM-CANN发布

X：Rohan Paul (@rohanpaul_ai) · 9 小时前 · 模型与工具能力

ModelBest、清华大学与OpenBMB社区联合发布了BitCPM-CANN，这是全球首个完全基于华为昇腾910B NPU训练的开源1.58比特三元大模型。其核心创新在于采用仅含三种权重状态的极低比特量化技术，使模型内存占用相比BF16降低约6倍，可高效部署于手机、电脑、车载设备等边缘端。更关键的是，整个训练全栈（从量化算子到框架）均在昇腾上原生构建与验证，而非简单移植。该模型家族（0.5B-

能力进展 基础设施 新发布

https://x.com/rohanpaul_ai/status/2057833050692800926

★ 2. 网易有道"子曰4"多模态模型、语音合成模型全量开源

IT之家（RSS） · 14 小时前 · 模型与工具能力

网易有道宣布将其"子曰"大模型4.0的多模态模型与语音合成模型面向全球全量开源。其中，多模态模型（27B参数）专注于教育场景，在处理高难度视觉数理问题上达到行业顶尖水平，纯文本中文数理难题准确率为81.4%。该模型通过思维链优化，将输出长度压缩43.2%，有效降低了推理成本。同时开源的语音合成模型支持跨语种音色与情感迁移克隆，3秒内即可完成零样本复制，准确度超97%，并支持包括中、英、日、韩在内的

能力进展 基础设施 新发布

https://www.ithome.com/0/954/124.htm

★ 3. 智谱GLM-5.1高速版发布：刷新全球大模型API速度纪录

IT之家（RSS） · 22 小时前 · 模型与工具能力

5月22日，智谱向部分企业客户推出了旗舰大模型GLM-5.1的高速版API"GLM-5.1-highspeed"。该版本输出速度达400 tokens/s，刷新了全球大模型API速度上限。关键突破在于，它首次在国产大模型中实现了旗舰级能力与低延迟的结合，打破了"高速模型即轻量模型"的传统。该版本由智谱GLM团队与TileRT团队合作，通过系统级优化确保了速度的生产级稳定性，适用于AI编程、实时语音

能力进展 新发布

https://www.ithome.com/0/953/717.htm

★ 4. Project Glasswing：初步更新

Anthropic：Newsroom（网页） · 4 小时前 · 产业与基础设施

上月启动的Project Glasswing项目，旨在利用先进AI模型保障关键软件安全。通过约50家合作伙伴使用Claude Mythos Preview模型，已在全球关键系统中发现超过10，000个高危或严重漏洞。多家合作伙伴报告漏洞发现效率提升超过十倍。例如，Cloudflare在关键路径系统发现2，000个漏洞；Mozilla在Firefox 150中发现并修复271个漏洞，数量远超前代模型

能力进展 基础设施 监管/资本

https://www.anthropic.com/research/glasswing-initial-update

★ 5. v2.1.149 更新摘要

Claude Code：GitHub Releases（RSS） · 2 小时前 · 应用与商业化

本次 v2.1.149 更新包含功能增强、企业设置和多项修复。新增 `/usage` 命令的使用量分类显示功能，可区分技能、子代理、插件及每个 MCP 服务器的消耗；`/diff` 详情视图支持键盘滚动；Markdown 输出兼容 GFM 任务列表。企业版新增 `allowAllClaudeAiMcps` 设置以加载云 MCP 连接器。修复了 PowerShell 权限绕过、Git 工作树沙盒写入

能力进展 基础设施 监管/资本

https://github.com/anthropics/claude-code/releases/tag/v2.1.149

★ 6. 谷歌I/O大会发布AI代理全套开发工具链

X：Google AI (@GoogleAI) · 7 小时前 · 应用与商业化

谷歌在I/O开发者大会宣布，系统性构建面向AI代理（Agent）的开发与部署工具链。核心更新包括：独立桌面应用Antigravity 2.0及其命令行工具、SDK面世；Google AI Studio新增Kotlin支持，可一键开发安卓应用并发布，同时推出移动端App。此外，Gemini API推出托管代理服务，实现一键部署；WebMCP作为开放标准在Chrome 149中推出，允许网页向代理暴露

能力进展 基础设施 新发布

https://x.com/GoogleAI/status/2057871583843135978

★ 7. 谷歌DeepMind在亚太启动AI气候加速器

Google DeepMind：Blog（RSS） · 昨天 03:46 · 产业与基础设施

亚太地区经济增长迅速，但极易受到气候变化影响，且现有绿色技术发展速度跟不上环境风险的增长。为此，Google DeepMind宣布启动首届专注于"AI for the Planet"的加速器计划。该计划为期三个月，面向亚太地区的初创企业、研究团队和非营利组织，旨在利用前沿人工智能技术解决自然、气候、农业和能源等领域的挑战。入选组织将获得专家指导、定制化支持，并可集成Google AI的前沿模型。计

能力进展 基础设施 新发布

https://deepmind.google/blog/were-launching-the-google-deepmind-accelerator-program-in-asia-pacific-to-tackle-environmental-risks

★ 8. 美国 AI 监管令突然告吹内幕：白宫内讧，马斯克、扎克伯格游说特朗普

IT之家（RSS） · 22 小时前 · 产业与基础设施

5月22日，美国总统特朗普突然取消了原定签署的AI行政令，该行政令旨在加强监管，赋予政府在AI模型公开发布前进行评估的权力。取消源于特朗普本人对监管的反感，以及高级顾问大卫·萨克斯和科技界领袖如扎克伯格、马斯克的反对与游说，凸显白宫内讧。特朗普认为监管会成为绊脚石，阻碍美国AI领先优势。草案中还存在如财政部在安全协调中角色过重等争议，白宫表示正制定其他AI安全举措。

能力进展 监管/资本 新发布

https://www.ithome.com/0/953/708.htm

能力 模型与工具能力

1. 首个基于华为昇腾910B NPU全栈训练的1.58比特开源大模型BitCPM-CANN发布

X：Rohan Paul (@rohanpaul_ai) · 9 小时前

ModelBest、清华大学与OpenBMB社区联合发布了BitCPM-CANN，这是全球首个完全基于华为昇腾910B NPU训练的开源1.58比特三元大模型。其核心创新在于采用仅含三种权重状态的极低比特量化技术，使模型内存占用相比BF16降低约6倍，可高效部署于手机、电脑、车载设备等边缘端。更关键的是，整个训练全栈（从量化算子到框架）均在昇腾上原生构建与验证，而非简单移植。该模型家族（0.5B-

能力进展 基础设施 新发布

https://x.com/rohanpaul_ai/status/2057833050692800926

2. 网易有道"子曰4"多模态模型、语音合成模型全量开源

IT之家（RSS） · 14 小时前

网易有道宣布将其"子曰"大模型4.0的多模态模型与语音合成模型面向全球全量开源。其中，多模态模型（27B参数）专注于教育场景，在处理高难度视觉数理问题上达到行业顶尖水平，纯文本中文数理难题准确率为81.4%。该模型通过思维链优化，将输出长度压缩43.2%，有效降低了推理成本。同时开源的语音合成模型支持跨语种音色与情感迁移克隆，3秒内即可完成零样本复制，准确度超97%，并支持包括中、英、日、韩在内的

能力进展 基础设施 新发布

https://www.ithome.com/0/954/124.htm

3. 智谱GLM-5.1高速版发布：刷新全球大模型API速度纪录

IT之家（RSS） · 22 小时前

5月22日，智谱向部分企业客户推出了旗舰大模型GLM-5.1的高速版API"GLM-5.1-highspeed"。该版本输出速度达400 tokens/s，刷新了全球大模型API速度上限。关键突破在于，它首次在国产大模型中实现了旗舰级能力与低延迟的结合，打破了"高速模型即轻量模型"的传统。该版本由智谱GLM团队与TileRT团队合作，通过系统级优化确保了速度的生产级稳定性，适用于AI编程、实时语音

能力进展 新发布

https://www.ithome.com/0/953/717.htm

产业 产业与基础设施

1. Project Glasswing：初步更新

Anthropic：Newsroom（网页） · 4 小时前

上月启动的Project Glasswing项目，旨在利用先进AI模型保障关键软件安全。通过约50家合作伙伴使用Claude Mythos Preview模型，已在全球关键系统中发现超过10，000个高危或严重漏洞。多家合作伙伴报告漏洞发现效率提升超过十倍。例如，Cloudflare在关键路径系统发现2，000个漏洞；Mozilla在Firefox 150中发现并修复271个漏洞，数量远超前代模型

能力进展 基础设施 监管/资本

https://www.anthropic.com/research/glasswing-initial-update

2. 谷歌DeepMind在亚太启动AI气候加速器

Google DeepMind：Blog（RSS） · 昨天 03:46

亚太地区经济增长迅速，但极易受到气候变化影响，且现有绿色技术发展速度跟不上环境风险的增长。为此，Google DeepMind宣布启动首届专注于"AI for the Planet"的加速器计划。该计划为期三个月，面向亚太地区的初创企业、研究团队和非营利组织，旨在利用前沿人工智能技术解决自然、气候、农业和能源等领域的挑战。入选组织将获得专家指导、定制化支持，并可集成Google AI的前沿模型。计

能力进展 基础设施 新发布

https://deepmind.google/blog/were-launching-the-google-deepmind-accelerator-program-in-asia-pacific-to-tackle-environmental-risks

3. 美国 AI 监管令突然告吹内幕：白宫内讧，马斯克、扎克伯格游说特朗普

IT之家（RSS） · 22 小时前

5月22日，美国总统特朗普突然取消了原定签署的AI行政令，该行政令旨在加强监管，赋予政府在AI模型公开发布前进行评估的权力。取消源于特朗普本人对监管的反感，以及高级顾问大卫·萨克斯和科技界领袖如扎克伯格、马斯克的反对与游说，凸显白宫内讧。特朗普认为监管会成为绊脚石，阻碍美国AI领先优势。草案中还存在如财政部在安全协调中角色过重等争议，白宫表示正制定其他AI安全举措。

能力进展 监管/资本 新发布

https://www.ithome.com/0/953/708.htm

4. GitHub 连续第三年被 Gartner® 魔力象限TM 评为企业级 AI 编程代理领域的领导者

GitHub Blog · 8 小时前

Gartner 最新发布的魔力象限报告中，GitHub 连续第三年被列为"领导者"象限，该评估专注于企业级 AI 编程代理领域。GitHub 表示，其致力于构建一个开放、安全且由 AI 驱动的平台，以赋能每一位开发者并定义软件开发的未来。此次评选进一步巩固了 GitHub 在 AI 辅助开发工具市场的领先地位。

能力进展 监管/资本 新发布

https://github.blog/ai-and-ml/github-copilot/github-recognized-as-a-leader-in-the-gartner-magic-quadrant-for-enterprise-ai-coding-agents-for-the-third-year-in-a-row

5. DeepSeek 推进 700 亿元融资，梁文锋承诺坚持开发开源 AI 模型而非追求短期商业化目标

IT之家（RSS） · 19 小时前

DeepSeek正推进700亿元人民币的巨额融资，估值约450亿美元。创始人梁文锋承诺将继续开源开发AI模型，不追求短期商业化，目标是技术升级与通用人工智能。腾讯、IDG资本等接近参投，梁文锋个人可能注资200亿元。若成功将创下中国科技初创公司首轮融资纪录。

能力进展 监管/资本 新发布

https://www.ithome.com/0/953/832.htm

6. 18 年老粉与微软 GitHub 决裂：我希望它更好，但我更想编程

IT之家（RSS） · 16 小时前

全球最大的代码托管平台GitHub正面临严重危机。资深开发者Mitchell Hashimoto公开与平台决裂，因频繁崩溃影响编程。近几个月，花旗银行、英特尔等巨头因持续故障表达不满，OpenAI探索自建方案。更严重的是，3800多个内部仓库遭黑客入侵，源代码被公开叫卖。同时，微软取消GitHub CEO职位，将其并入CoreAI团队，导致大量技术骨干流失。这个承载1.5亿开发者的平台，正以惨烈方

能力进展 新发布

https://www.ithome.com/0/953/977.htm

7. Anthropic即将成为首个盈利的AI实验室

The Decoder：AI News（RSS） · 昨天 23:15

根据《华尔街日报》报道，Anthropic正接近实现其首个盈利季度，预计第二季度营收达109亿美元，运营利润为5.59亿美元。该公司在去年夏季时还预计最早在2028年才能盈利。主要增长动力来自编程工具和Claude的代理功能使用，其需求一度超过了可用的算力容量。这一转变标志着Anthropic可能成为业界首个实现盈利的领先AI研发机构。

能力进展 基础设施

https://the-decoder.com/anthropic-is-about-to-become-the-first-profitable-ai-lab

8. DeepSeek V4 Flash登顶周榜

X：OpenRouter (@OpenRouter) · 18 小时前

DeepSeek V4 Flash已登顶周排行榜

能力进展 新发布

https://x.com/OpenRouter/status/2057703179882749985

9. 黄仁勋：AI 基建年度开支要冲到 4 万亿美元！

IT之家（RSS） · 1 小时前

英伟达发布2027财年Q1财报，营收816亿美元，同比增长85%，净利润583亿美元，翻两倍多，市值达5.7万亿美元，已超德国2026年GDP预测。黄仁勋预测，超大规模云厂商的AI基建年度开支将从当前的1万亿美元，增长至3-4万亿美元，远超华尔街预期。财报同时显示，数据中心业务营收752亿美元，占比超九成。值得注意的是，AI基建的高能耗正推高居民电费，数据中心用电成本转嫁效应已初步显现。

基础设施 新发布

https://www.ithome.com/0/954/223.htm

---

10. 国家发改委：加快具身智能训练基础设施建设，让机器人不仅能上赛场，还能"进工厂、进商场、进家庭"

IT之家（RSS） · 14 小时前

国家发改委在5月22日新闻发布会上表示，人形机器人在半程马拉松比赛中表现显著提升，速度更快、更灵活、更自主，参赛队伍从20余支增至百余支，完赛队伍从6支增至40余支，反映具身智能创新活力增强和产业规模扩大。下一步，发改委将加快具身智能训练基础设施建设，推动机器人融入工厂、商场、家庭等场景，并建设应用中试基地以加速技术落地。

基础设施 新发布

https://www.ithome.com/0/954/126.htm

---

11. Cursor 被评为 2026 年 Gartner 企业级 AI 编码代理魔力象限领导者

Cursor Blog · 12 小时前

Gartner 在 2026 年魔力象限报告中，将 Cursor 评为企业级 AI 编码代理领域的领导者，并在愿景完整性上领先。超过 70% 的财富 500 强企业使用 Cursor 部署和管理编码代理。未来一年，Cursor 将聚焦于三个方向：提升前沿模型智能；自动化软件开发全生命周期的任务（如代码审查、漏洞修复）；以及通过新的管理工具和控制面板，增强企业级的控制力、协作性与部署灵活性，以拓展至

能力进展

https://cursor.com/blog/cursor-leads-gartner-mq-2026

---

12. OpenAI被Gartner评为企业AI编码代理领域领导者

OpenAI：官网动态（RSS · 排除企业/客户案例） · 昨天 08:00

Gartner发布2026年企业AI编码代理魔力象限报告，OpenAI被列为领导者。其产品Codex因在技术创新和企业级部署方面的突出表现获得认可，反映了OpenAI在AI辅助编程工具领域的领先地位。

新发布

https://openai.com/index/gartner-2026-agentic-coding-leader

---

13. 加州州长纽森签署首创性行政令，为应对AI可能带来的劳动力市场冲击做准备

X：Rohan Paul (@rohanpaul_ai) · 昨天 04:12

加州州长纽森签署行政令，首次将AI引发的失业问题正式列为公共政策议题，要求各部门研究遣散费、就业保险及员工持股等保障措施。该命令认识到AI可能渐进式替代岗位任务，而非一次性取代整个职位，因此计划建立新的劳动力数据看板，以更早捕捉行业受到的冲击。政策核心在于探讨如何将AI带来的生产力红利，通过股权、薪酬支持等方式在企业与员工间进行更公平的分配。同时指出，单纯的职业培训可能无法解决被AI彻底取代的岗位

监管/资本

https://x.com/rohanpaul_ai/status/2057555054387949848

---

14. AI 替代入门级工作：科技行业受裁员冲击最重，74% CEO 冻结或缩减招聘

IT之家（RSS） · 16 分钟前

奥纬咨询研究发现，AI工具正被广泛用于入门级任务，导致企业招聘重心转向高级岗位，年轻人求职难度加大。科技行业受冲击最严重，74%的CEO已冻结或缩减招聘。计划削减初级岗位的比例从17%跃升至43%，而招聘转向中层岗位的比例则升至30%。尽管超90%的企业在部署AI，但多数仍处试点阶段。报告警告，过快裁员或忽视初级人才储备，可能对人才梯队造成长远风险。

https://www.ithome.com/0/954/235.htm

---

15. 回顾Google I/O 2026对话环节

Google Blog：AI（RSS） · 6 小时前

在2026年Google I/O开发者大会上，对话环节聚焦于未来科技趋势。行业领导者围绕人工智能、量子计算、机器人学以及创造力等核心议题展开了深入探讨，旨在勾勒这些前沿领域的技术演进路径与发展蓝图。

https://blog.google/innovation-and-ai/technology/ai/io-2026-dialogues-recap

---

16. Suno AI创作夏日神曲《波多黎各》爆火

X：Suno (@suno) · 8 小时前

今年夏天的热门歌曲是用 Suno 制作的。☀非常感谢 @GMA 让这首病毒式传播的《Puerto Rico》歌曲被更多人看到！还有谁的脑海里也一直回响着这首歌？

https://x.com/suno/status/2057858423664894196

应用 应用与商业化

1. v2.1.149 更新摘要

Claude Code：GitHub Releases（RSS） · 2 小时前

本次 v2.1.149 更新包含功能增强、企业设置和多项修复。新增 `/usage` 命令的使用量分类显示功能，可区分技能、子代理、插件及每个 MCP 服务器的消耗；`/diff` 详情视图支持键盘滚动；Markdown 输出兼容 GFM 任务列表。企业版新增 `allowAllClaudeAiMcps` 设置以加载云 MCP 连接器。修复了 PowerShell 权限绕过、Git 工作树沙盒写入

能力进展 基础设施 监管/资本

https://github.com/anthropics/claude-code/releases/tag/v2.1.149

2. 谷歌I/O大会发布AI代理全套开发工具链

X：Google AI (@GoogleAI) · 7 小时前

谷歌在I/O开发者大会宣布，系统性构建面向AI代理（Agent）的开发与部署工具链。核心更新包括：独立桌面应用Antigravity 2.0及其命令行工具、SDK面世；Google AI Studio新增Kotlin支持，可一键开发安卓应用并发布，同时推出移动端App。此外，Gemini API推出托管代理服务，实现一键部署；WebMCP作为开放标准在Chrome 149中推出，允许网页向代理暴露

能力进展 基础设施 新发布

https://x.com/GoogleAI/status/2057871583843135978

3. Datasette Agent

Simon Willison 博客 · 昨天 03:52

Datasette Agent是Datasette推出的首个可扩展AI助手，为用户提供对话式界面以查询数据，并支持通过插件生成图表。该工具基于其LLM Python库构建，是LLM与Datasette整合的重要成果。目前提供图表生成、AI图像创建和沙箱代码执行等官方插件。它既可运行于Gemini 3.1 Flash-Lite等云端模型，也支持通过LM Studio连接本地开源模型，具备可靠的工具调

能力进展 基础设施 新发布

https://simonwillison.net/2026/May/21/datasette-agent

4. Warp现已支持OpenRouter接入

X：OpenRouter (@OpenRouter) · 6 小时前

OpenRouter现已在@warpdotdev中得到支持！❤ 工程师Dagm Assefa展示了如何连接DeepSeek和OpenRouter。 文档：https：//docs.warp.dev/agent-platform/inference/custom-inference-endpoint/ ☒

能力进展 基础设施 新发布

https://x.com/OpenRouter/status/2057875517391667492

5. v2.1.147版本更新

Claude Code：GitHub Releases（RSS） · 昨天 04:39

本次更新引入了`Workflow`工具，支持确定性多智能体编排（默认关闭）。将`/simplify`命令重命名为`/code-review`，现可报告代码正确性问题并支持生成GitHub PR内联评论。改进了自动更新器（增加重试与错误报告）、大文件diff渲染性能，并优化了提示历史记录以避免重复条目。修复了多个关键问题，包括企业登录限制未生效、Windows下的PowerShell工具与终端闪烁问

能力进展 监管/资本 新发布

https://github.com/anthropics/claude-code/releases/tag/v2.1.147

6. Codex实现全天候跨设备安全操控Mac

X：OpenAI Developers (@OpenAIDevs) · 昨天 02:59

Codex随时随地，无处不在。 现在您的Mac无需解锁，Codex即可使用您的电脑。 通过手机，Codex可以安全地使用您Mac上的应用程序，即使屏幕关闭且处于锁定状态。 https：//developers.openai.com/codex/app/computer-use#locked-use

能力进展 监管/资本 新发布

https://x.com/OpenAIDevs/status/2057536706778378692

7. ChatGPT现已支持在PowerPoint中直接创建编辑演示文稿

X：ChatGPT (@ChatGPTapp) · 昨天 04:32

你是否曾这样想过：我真的不想做这个PPT。 好消息：ChatGPT现在可以直接在PowerPoint中创建和编辑演示文稿。 在PowerPoint中直接构建、更新、理解和优化演示文稿，同时保持幻灯片可编辑。 目前处于测试阶段，我们期待您的反馈 ☒

能力进展 新发布

https://x.com/ChatGPTapp/status/2057560276384563560

8. Viggle推出3D格斗派对游戏Fight Anyone 3D

X：Viggle AI (@ViggleAI) · 昨天 03:16

介绍Fight Anyone 3D☒一款3D派对格斗游戏，可能是上班时玩起来最爽的游戏。 上传任何人的照片 → 一个可玩的3D格斗角色，带有语音、个性+招牌动作，由Viggle自研游戏引擎+模型打造。 公测期间100%免费+赠送20张礼品卡。玩得越多，赢得越多！ 和同事对战。和朋友对战。和任何人对战。链接+教程+更多内容见下方推文串 ↓

能力进展 新发布

https://x.com/ViggleAI/status/2057541072419520634

9. Codex周四更新：Appshots功能上线

X：OpenAI Developers (@OpenAIDevs) · 昨天 02:33

又是Codex周四，我们带来了更新。 首先是Appshots，一种将你工作上下文引入Codex的新方式。 在Mac上，按Command-Command即可将应用窗口附加到Codex线程。Codex会获取窗口的截图和文本，包括屏幕上不可见的内容。 Appshots已在Mac各计划中推出，企业版访问权限即将上线。

能力进展 新发布

https://x.com/OpenAIDevs/status/2057530207976989179

---

10. Shoplift by PixVerse快速生成平台原生广告视频

X：PixVerse (@PixVerse_) · 昨天 00:05

无需工作室，无需编辑队列。 将产品URL粘贴到Shoplift by PixVerse，几分钟内即可发布平台原生广告视频 --专为持续进行创意测试的DTC团队打造。 免费早期访问：https：//shoplift.pixverse.ai 转发+关注+回复=300积分（仅限72小时）

能力进展 新发布

https://x.com/PixVerse_/status/2057492991036588306

---

11. Claude自动模式新增Pro计划与模型支持

X：Claude Devs (@ClaudeDevs) · 2 小时前

自动模式的两项更新： • 现已在Pro计划中提供 • 现已支持Sonnet 4.6，以及Opus 4.7 按下Shift+tab，让Claude运行。

能力进展 新发布

https://x.com/ClaudeDevs/status/2057946803685974482

---

12. 新增差异标记样式设置选项

X：OpenAI Developers (@OpenAIDevs) · 4 小时前

已发布剪纸功能：外观设置中新增差异标记样式。 在查看差异时更喜欢经典的 + / - 标记？现在你可以选择使用它们，而不仅仅是彩色差异条。 默认设置保持不变，除非你主动选择启用。

能力进展 新发布

https://x.com/OpenAIDevs/status/2057918624841728349

---

13. Gemini扩展应用连接，支持更多服务

X：Gemini (@GeminiApp) · 昨天 03:52

Gemini现在可以连接更多应用，包括@OpenTable、@Canva和@Instacart。 无论您是预订餐厅、制作传单还是订购杂货，Gemini不仅能查找信息，还能通过连接的应用帮助您无缝采取行动。

能力进展 新发布

https://x.com/GeminiApp/status/2057550225863246236

---

14. DeepSeek-V4-Pro永久降价公告

X：DeepSeek (@deepseek_ai) · 8 小时前

我们将优惠永久化！☒享受使用 DeepSeek-V4-Pro 构建的乐趣，将您的创新想法变为现实！☒

能力进展 新发布

https://x.com/deepseek_ai/status/2057854261699195173

---

15. Antigravity付费版Gemini配额再翻三倍

X：Google AI for Developers (@googleaidevs) · 19 小时前

更新：所有付费@Antigravity层级的*每周*Gemini配额已再次提升至三倍，且配额已正式重置。

能力进展 新发布

https://x.com/googleaidevs/status/2057679246085226965

---

16. Gemini Omni发布，创意作品涌现

X：Gemini (@GeminiApp) · 昨天 08:15

Gemini Omni来了，我们本周看到了许多令人惊叹的创作。以下是一些杰出作品 ☒

能力进展 新发布

https://x.com/GeminiApp/status/2057616371748651054

---

17. Krea 2 推出 LoRA 微调系统

X：Krea AI (@krea_ai) · 昨天 22:29

为 Krea 2（测试版）引入 LoRA。 我们迄今最强大的微调系统；现在你可以用惊人的精度，在 Krea 2 上训练你自己的特定风格、对象或角色。 了解其工作原理 ☒

基础设施 新发布

https://x.com/krea_ai/status/2057468861583347774

---

18. ChatGPT语音模式实现表单语音填写

X：ChatGPT (@ChatGPTapp) · 4 小时前

用对话处理文书工作更轻松。 借助ChatGPT的图像功能和语音模式，您可以上传表单，说出要填写的内容，即可获得填写完成的版本。

能力进展

https://x.com/ChatGPTapp/status/2057908052968521902

---

19. PixVerse App上线图像生成功能

X：PixVerse (@PixVerse_) · 13 小时前

Create Image已在PixVerse App上线。 输入提示词或参考图，在手机上即可生成。 5月24日至31日11：00 UTC，每人可免费生成3次。 转发+关注+回复 = 300 Creds（仅限72小时）

能力进展

https://x.com/PixVerse_/status/2057777743027392848

20. 用户好评即最佳文案，视觉呈现由AI完成

X：Luma AI (@LumaLabsAI) · 23 小时前

你的客户写出了你永远无法超越的文案。现在，为它匹配视觉呈现吧。 只需粘贴评价，设定风格，Luma Agents 将处理所有推荐语图形。 让它被听见 → http：//lumalabs.ai/app

能力进展

https://x.com/LumaLabsAI/status/2057628006353670163

21. Gemini每日简报助你高效规划一天

X：Gemini (@GeminiApp) · 昨天 00:35

用Daily Brief开启高效一天。 Gemini现在能主动将最重要的事项整理成易于理解的待办清单，让你在早餐结束前就为一天做好准备。

能力进展

https://x.com/GeminiApp/status/2057500470147698936

22. Project Genie与谷歌街景合作推出交互式世界

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

Project Genie ☒@GoogleMaps Street View 你现在可以将真实的美国地点转化为全新的交互式世界。☒

新发布

https://x.com/GoogleDeepMind/status/2057842131142590512

23. Replit企业版现已开放自助购买

X：Replit (@Replit) · 昨天 00:01

Replit Enterprise现已支持自助服务！ 几分钟内即可： - 购买Replit Enterprise - 配置SSO + SCIM - 与团队开始协作开发 无需合同谈判，无需等待。

新发布

https://x.com/Replit/status/2057491954825674942

24. 动作捕捉与角色动画制作更轻松

X：Viggle AI (@ViggleAI) · 2 小时前

动作捕捉和角色动画制作从未如此简单。 持续构建，更多功能即将推出！

新发布

https://x.com/ViggleAI/status/2057947352195858568

研究 研究与开源进展

1. Nemotron-Labs 扩散语言模型实现光速级文本生成

Hugging Face：Blog（RSS） · 20 分钟前

NVIDIA 在 Hugging Face 发布了关于 Nemotron-Labs 扩散语言模型的技术博客。该研究聚焦于通过扩散语言模型架构大幅提升文本生成速度，目标是逼近"光速级"生成效率。文章可能介绍了该模型在生成速度上的突破，以及相较于传统自回归模型在延迟和吞吐量方面的性能优势。具体技术细节或对比数据需参考原文。

能力进展 基础设施 新发布

https://huggingface.co/blog/nvidia/nemotron-labs-diffusion

2. AlphaProof Nexus：用形式化验证驱动AI数学证明搜索

X：Rohan Paul (@rohanpaul_ai) · 1 小时前

Google DeepMind提出了AlphaProof Nexus系统，它将大型语言模型与Lean形式化验证工具相结合。该系统允许LLM在生成证明的过程中，不断读取Lean的编译错误并进行修正，还可调用更强的工具辅助解决子问题。这一机制迫使模型将每一步逻辑都转化为可编译、可验证的代码，从而将其角色从"令人信服的叙述者"转变为"候选方案生成器"。在针对353个Erdős问题和492个开放猜想的测试

能力进展 新发布

https://x.com/rohanpaul_ai/status/2057954067146781151

3. 图灵测试 76 年后首现 AI 通过实证：GPT-4.5 以 73% 判定率超越真人

IT之家（RSS） · 23 小时前

加州大学圣地亚哥分校研究首次实证现代AI可通过图灵测试。研究表明，在获得特定提示后，GPT-4.5在5至15分钟的对话中被误认为人类的概率高达73%，显著超过真人。LLaMa-3.1-405B的判定率（56%）与真人相当，而GPT-4o和ELIZA仅约20%。研究指出提示词至关重要，它使AI能模仿人类语气、幽默感甚至易错性等社会行为特征。这一发现迫使人们重新思考图灵测试的意义，并凸显了大语言模型在

能力进展 监管/资本

https://www.ithome.com/0/953/705.htm

4. VSAS-Bench：视觉流式辅助模型的实时评估基准

Apple Machine Learning Research（RSS） · 昨天 08:00

现有视觉语言模型框架主要在离线场景下评估性能，但实时视觉助手所依赖的流式模型还需考量额外指标，如反映响应时效性的"主动性"和捕捉随时间推移响应稳定性的"一致性"。为此，研究团队提出了VSAS-Bench，这是一个新的评估基准，专门针对流式视觉语言模型在实时交互任务中的表现，填补了当前评估方法在动态、持续生成场景下的空白。

能力进展

https://machinelearning.apple.com/research/vsas-bench-streaming-assistant

---

格局 观点、资本与监管

1. 核算OpenAI和Anthropic最新动态背后的数学

Gary Marcus：The Road to AI We Can Trust（RSS） · 昨天 01:32

OpenAI与Anthropic近期相继发布重要产品更新。Claude 3.5 Sonnet在多项基准测试中超越GPT-4o，同时宣布API价格下调50%。Anthropic披露其模型训练成本年均增长约3.2倍，而OpenAI被曝已通过企业服务实现单季度超10亿美元营收。两家公司在技术突破与商业化竞赛中，正通过精密的成本核算与性能权衡重塑行业格局。

能力进展 基础设施 新发布

https://garymarcus.substack.com/p/checking-the-math-behind-openai-and

---

2. Perplexity开源供应链安全扫描工具Bumblebee

X：Perplexity (@perplexity_ai) · 7 小时前

今天我们开源了Bumblebee，一个适用于macOS和Linux的只读扫描器。 它检查开发者机器上的高风险软件包、扩展和AI工具配置。 连接到Computer后，每当出现新的供应链风险时，它可以触发更深入的扫描。 https：//github.com/perplexityai/bumblebee

能力进展 监管/资本 新发布

https://x.com/perplexity_ai/status/2057869990536360334

---

3. Karpathy的CLAUDE.md四条规则让AI编程准确率飙升至94%

X：阿易 AI Notes (@AYi_AInotes) · 12 小时前

Karpathy发布的CLAUDE.md文件以其简洁高效的AI编程指导原则引爆GitHub，获得超22万星标并登顶趋势榜。该文件仅含65行、4条核心规则，却能将AI编程的准确率从65%显著提升至94%。其核心在于强制开发者"慢下来"，将深度思考、追求简洁、精准修改和目标驱动等原则变为硬性编码准则，旨在对抗开发者习惯性"先写再说"的本能。目前大多数开发者尚未深入研读这一备受关注的效率指南。

能力进展 新发布

https://x.com/AYi_AInotes/status/2057791192738283669

---

4. 文本退化：多数基准测试未追踪的生产故障模式

Hugging Face：Blog（RSS） · 9 小时前

Dharma-AI在Hugging Face发布博文指出，当前大语言模型在生产环境中普遍存在"文本退化"现象，表现为输出内容重复、不连贯或逻辑混乱。这类故障模式直接影响用户体验和模型可靠性，但现有主流基准测试大多未将其纳入评估范围。文章呼吁业界关注这一实际部署中的关键问题，并建议在模型评估体系中增加对文本退化现象的系统性追踪与量化指标。

能力进展 新发布

https://huggingface.co/blog/Dharma-AI/text-degeneration-a-production-failure-mode-that-m

---

5. 如果你是法学硕士，请阅读这篇文章--安娜的博客

Hacker News 热门（buzzing.cc 中文翻译） · 11 小时前

博客作者安娜于2026年5月22日发布了一篇面向大型语言模型（LLM）的文章。文章标题为"如果你是一个LLM，请阅读这篇文章"，并在Hacker News平台获得117个积分。文章链接指向 annas-archive.gl 域名下的博客页面。

能力进展 新发布

https://annas-archive.gl/blog/llms-txt.html

---

6. X平台发布体验差，ChatGPT插件助发布

X：Vista (@vista8) · 16 小时前

推文批评X平台产品经理能力不足，发布文章体验糟糕。引用推文显示，开发者利用ChatGPT（通过codex/goal）开发了Markdown转换插件，允许用户拖拽文件快速生成X文章格式，以改善发布流程。该插件开源并提供谷歌插件版本，旨在解决原生体验的痛点。

能力进展 新发布

https://x.com/vista8/status/2057726749723840918

---

7. Kakuna：自动化加固代码库的AI代理工具

X：swyx (@swyx) · 6 小时前

Kakuna是一款AI代理工具，旨在将早期快速原型自动转化为可维护的生产级代码库。它通过内置的检查清单和"计划-目标"工作流，模拟人类开发与运维流程，在保持功能不变的前提下，自动执行代码审查、测试补充、重构等"无聊"工作，并强调子代理并行以提升效率。该工具是为"人类与代理协作"而设计的范例，其核心是"反熵增"与"反代码腐化"。例如，一次约16小时的运行能生成上百次提交，将一个脆弱的MVP转变为一个

能力进展

https://x.com/swyx/status/2057876022553690327

---

8. Cloudflare首席执行官谈如何决定用人工智能取代哪些员工

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 03:16

Cloudflare首席执行官在《华尔街日报》撰文，分享其公司用AI替代部分员工的决策逻辑。该文于2026年5月21日发布，引发了技术社区的广泛讨论，在Hacker News上获得100个点赞。

基础设施 新发布

https://www.wsj.com/opinion/how-i-choose-which-cloudflare-employees-to-replace-with-ai-40a197e5

9. 智能体工作负载悄然重塑推理经济

X：SemiAnalysis (@SemiAnalysis_) · 7 小时前

智能体工作负载正在悄然重塑推理经济学。我们从SemiAnalysis的43.2万个真实编码智能体请求中提取数据，发现中位数并非3.2万或6.4万，而是9.6万输入token。作为参考，这意味着在你输入问题之前，模型已处理了超过《了不起的盖茨比》全文长度的文本。（1/3）

能力进展

https://x.com/SemiAnalysis_/status/2057869518295249373

10. OpenAI Codex /goal功能正式发布及使用指南

X：宝玉 (@dotey) · 20 小时前

OpenAI宣布Codex的/goal模式已结束实验，成为稳定功能。用户可在Codex应用、IDE扩展或CLI中使用，通过设定具体里程碑，让AI持续工作直至完成，任务可运行数小时甚至数天。过程中支持随时检查、调整方向及暂停。使用前需升级应用并启用该功能（可通过命令行指令或手动修改配置文件实现）。开启后，可在输入框管理任务，并利用侧边对话查看进度而不中断主任务。该功能旨在高效处理各类复杂任务。

新发布

https://x.com/dotey/status/2057672416071987378

11. 可塑界面：AI驱动的未来软件形态

Tomer Tunguz 博客（VC 分析） · 昨天 08:00

Salesforce已采用无头架构，允许销售人员通过AI直接更新数据，许多公司正通过MCPs跟进。同时，AI专家们正推动超越纯文本、更丰富的界面（如HTML），支持图表与交互。AI能根据场景动态生成定制化界面。无头系统并非移除前端，而是支持多种可塑化界面（如音频、网页）。未来软件的核心价值在于动态管理这些界面、确保其准确性，并将各类AI产物整合为可演化的上下文数据库与制品库。用户界面并未消失，而是

新发布

https://www.tomtunguz.com/plastic-user-interfaces