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: 49 条 焦点: 8 条 快讯: 0 条

Executive Summary

Nemotron 3.5 Content Safety模型发布，基于Gemma 3 4B IT提供多模态安全评估能力，支持128K上下文窗口和自定义安全策略执行。NVIDIA推出Nemotron 3 Ultra专门针对长时间运行AI智能体优化推理效率。neolab发布Nex-N2-Pro，基于Qwen3.5-397B-A17B的MoE模型达到GPT-5.5水平，支持262K上下文。Google发布Magenta RealTime 2实时音乐模型，Miso One开源8B参数语音模型实现110ms低延迟。OpenAI在API中集成内容审核评分功能。

Anthropic扩展Claude Partner Network建立服务分级体系，OpenClaw和OpenShell分别增加Windows节点和Vertex AI支持，显示智能体框架生态持续扩展。Cloudflare Radar数据显示机器人流量首次超过人类占比达57.5%，反映AI驱动流量增长趋势。微软AI负责人表示因Anthropic模型成本过高正自研替代方案，DeepSeek连续四周占据token份额榜首。Stanford发布OpenJarvis设备端AI智能体框架，强调本地优先架构。

需关注Nemotron系列模型的企业部署情况和安全策略定制能力普及程度。Qwen3.5生态的商业化进展及与GPT-5.5性能对标验证。各大厂商在智能体长程推理优化方面的技术路径差异。语音识别和合成模型的实时性能指标在实际应用场景的表现。企业AI安全合规需求推动的内容审核工具标准化进程。

重点 今日核心进展

★ 1. Nemotron 3.5 Content Safety：面向全球企业AI的可定制多模态安全

Hugging Face：Blog（RSS） · 5 小时前 · 模型与工具能力

Nemotron 3.5 Content Safety基于Gemma 3 4B IT，提供128K上下文窗口，支持用户提示、可选图像与助手响应的统一多模态安全评估。新增自定义策略执行，允许企业用自然语言定义专属安全规则；THINK模式可输出可审计的逐步推理痕迹。显式训练覆盖12种语言，并借助基座模型零样本泛化至约140种语言。输出提供低延迟二分类、带分类标签、THINK推理痕迹三种模式。安全分类遵

能力进展 基础设施 监管/资本

https://huggingface.co/blog/nvidia/nemotron-3-5-content-safety

★ 2. NVIDIA Nemotron 3 Ultra 为长时间运行的智能体带来更快、更高效的推理能力

NVIDIA Technical Blog（开发者技术博客 · RSS） · 11 小时前 · 模型与工具能力

NVIDIA 发布 Nemotron 3 Ultra 模型，专为长时间运行的 AI 智能体设计。该模型能够在多轮对话中保持上下文、调用工具、调用子智能体，并高效处理复杂工作流。随着多智能体协作导致模型 token 数量快速增长，Nemotron 3 Ultra 通过优化推理流程显著提升速度并降低计算成本，使长期运行的智能体任务更加可行。

能力进展 基础设施 新发布

https://developer.nvidia.com/blog/nvidia-nemotron-3-ultra-powers-faster-more-efficient-reasoning-for-long-running-agents

★ 3. Nex-N2-Pro 发布：基于 Qwen3.5 的 397B MoE 推理模型，性能达 GPT-5.5 水平

X：硅基流动 SiliconFlow (@SiliconFlowAI) · 9 小时前 · 模型与工具能力

neolab 推出 Nex-N2-Pro，基于 Qwen3.5-397B-A17B，总参数 397B 的 MoE 推理模型，支持 262K 上下文与多模态（VLM），性能达到 GPT-5.5 和 Claude Opus 4.7 级别。模型可自动调节推理深度，减少 30-50% 思考 token 且无性能折损，在 Terminal Bench 2.1、GDPVal、SWE-Verified 上取得

能力进展 新发布

https://x.com/SiliconFlowAI/status/2062549952266723493

★ 4. 介绍 Claude Partner Network 的 Services Track 和 Partner Hub

Anthropic：Newsroom（网页） · 昨天 21:30 · 应用与商业化

Anthropic 扩展 Claude Partner Network，推出 Services Track 分级体系和 Partner Hub 门户。Services Track 设 Select、Preferred、Global Premier 三级，按认证人数、投产客户数及客户推荐信量化评定。Partner Hub 提供每日更新仪表盘和公开目录，方便合作伙伴查看进展、客户寻找供应商。该网络三月

能力进展 基础设施 新发布

https://www.anthropic.com/news/services-track-partner-hub

★ 5. Google Magenta RealTime 2 （MRT2） 实时音乐模型发布

X：Google AI for Developers (@googleaidevs) · 5 小时前 · 模型与工具能力

Google AI for Developers 宣布推出开放权重的实时音乐模型 Magenta RealTime 2 （MRT2）。该模型可通过 MIDI 键盘、实时文本提示甚至手势进行演奏。MRT2 在 MacBook 上原生运行，延迟低于 200ms，提供开放权重、开源推理引擎以及配套应用和插件套件。

能力进展 新发布

https://x.com/googleaidevs/status/2062603374789263646

★ 6. Miso One 开源语音模型：8B 参数、110ms 延迟、一次语音克隆

X：Kim (@kimmonismus) · 昨天 00:32 · 模型与工具能力

Miso One 正式发布，一个 8B 参数的开源权重语音模型（TTS），旨在模拟真实人类朗读的温暖与节奏。它支持一次语音克隆（只需短样本），推理延迟仅 110ms。模型权重已开源至 GitHub，无需 API 即可自托管，音频数据不离开本地。API 访问即将推出。演示已上线，可先试听再克隆仓库。

能力进展 新发布

https://x.com/kimmonismus/status/2062210845308780639

★ 7. OpenClaw 2026.6.1发布：新增Windows节点与技能工坊

X：OpenClaw (@openclaw) · 昨天 05:40 · 应用与商业化

OpenClaw 2026.6.1 已上线 ⌧⌧原生 Windows 节点主机 ⌧用于自主学习型智能体的技能工坊（Skill Workshop） ⌧工作板（Workboard）编排 ⌧支持 MiniMax M3 Windows 加入集群。无需企鹅服。 https：//github.com/openclaw/openclaw/releases/tag/v2026.6.1

能力进展 基础设施 新发布

https://x.com/openclaw/status/2062288421406785710

★ 8. OpenShell v0.0.55 发布：新增 Vertex AI 推理支持

X：NVIDIA AI (@NVIDIAAI) · 昨天 00:29 · 应用与商业化

OpenShell v0.0.55 ⌧Google Vertex AI 推理提供者 ⌧基于配置文件的策略可见性 ⌧网关中更好的 Podman 检测 ⌧恢复 GPU procfs 基准行为 ⌧CI 与文档修复 运行智能体对接 Vertex AI，同时拥有改进的策略可见性以及更可靠的 Podman 和 GPU 沙箱行为。 https：//github.com/NVIDIA/OpenS

能力进展 基础设施 新发布

https://x.com/NVIDIAAI/status/2062210034109677665

能力 模型与工具能力

1. Nemotron 3.5 Content Safety：面向全球企业AI的可定制多模态安全

Hugging Face：Blog（RSS） · 5 小时前

Nemotron 3.5 Content Safety基于Gemma 3 4B IT，提供128K上下文窗口，支持用户提示、可选图像与助手响应的统一多模态安全评估。新增自定义策略执行，允许企业用自然语言定义专属安全规则；THINK模式可输出可审计的逐步推理痕迹。显式训练覆盖12种语言，并借助基座模型零样本泛化至约140种语言。输出提供低延迟二分类、带分类标签、THINK推理痕迹三种模式。安全分类遵

能力进展 基础设施 监管/资本

https://huggingface.co/blog/nvidia/nemotron-3-5-content-safety

2. NVIDIA Nemotron 3 Ultra 为长时间运行的智能体带来更快、更高效的推理能力

NVIDIA Technical Blog（开发者技术博客 · RSS） · 11 小时前

NVIDIA 发布 Nemotron 3 Ultra 模型，专为长时间运行的 AI 智能体设计。该模型能够在多轮对话中保持上下文、调用工具、调用子智能体，并高效处理复杂工作流。随着多智能体协作导致模型 token 数量快速增长，Nemotron 3 Ultra 通过优化推理流程显著提升速度并降低计算成本，使长期运行的智能体任务更加可行。

能力进展 基础设施 新发布

https://developer.nvidia.com/blog/nvidia-nemotron-3-ultra-powers-faster-more-efficient-reasoning-for-long-running-agents

3. Nex-N2-Pro 发布：基于 Qwen3.5 的 397B MoE 推理模型，性能达 GPT-5.5 水平

X：硅基流动 SiliconFlow (@SiliconFlowAI) · 9 小时前

neolab 推出 Nex-N2-Pro，基于 Qwen3.5-397B-A17B，总参数 397B 的 MoE 推理模型，支持 262K 上下文与多模态（VLM），性能达到 GPT-5.5 和 Claude Opus 4.7 级别。模型可自动调节推理深度，减少 30-50% 思考 token 且无性能折损，在 Terminal Bench 2.1、GDPVal、SWE-Verified 上取得

能力进展 新发布

https://x.com/SiliconFlowAI/status/2062549952266723493

4. Google Magenta RealTime 2 （MRT2） 实时音乐模型发布

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

Google AI for Developers 宣布推出开放权重的实时音乐模型 Magenta RealTime 2 （MRT2）。该模型可通过 MIDI 键盘、实时文本提示甚至手势进行演奏。MRT2 在 MacBook 上原生运行，延迟低于 200ms，提供开放权重、开源推理引擎以及配套应用和插件套件。

能力进展 新发布

https://x.com/googleaidevs/status/2062603374789263646

5. Miso One 开源语音模型：8B 参数、110ms 延迟、一次语音克隆

X：Kim (@kimmonismus) · 昨天 00:32

Miso One 正式发布，一个 8B 参数的开源权重语音模型（TTS），旨在模拟真实人类朗读的温暖与节奏。它支持一次语音克隆（只需短样本），推理延迟仅 110ms。模型权重已开源至 GitHub，无需 API 即可自托管，音频数据不离开本地。API 访问即将推出。演示已上线，可先试听再克隆仓库。

能力进展 新发布

https://x.com/kimmonismus/status/2062210845308780639

6. Ideogram v4.0 发布：2K 分辨率和 JSON 提示支持

X：Krea AI (@krea_ai) · 昨天 01:40

介绍 Ideogram v4.0。 原生 2K 分辨率，出色的文字渲染，支持 JSON 提示词。 立即在 Krea 中体验。

新发布

https://x.com/krea_ai/status/2062227837130887567

产业 产业与基础设施

1. Cloudflare Radar：机器人流量首次超过人类占比57.5%

X：小互 (@xiaohu) · 21 小时前

Cloudflare Radar 实时统计显示，过去一周（5月28日至6月4日）全球所有 HTML 网页请求流量中，57.5% 来自机器人（爬虫、AI 抓取、自动化脚本），仅42.5%来自真人浏览器，机器人流量首次超过人类。按所有 HTTP 流量返回内容分类，JSON（API 机器通信）占33.1%居首，HTML 仅12%。互联网流量主体已从人类浏览网页转向机器间通信和机器人抓取。

能力进展 基础设施

https://x.com/xiaohu/status/2062367357868355622

2. DeepSeek连续四周登顶Token份额榜

X：OpenRouter (@OpenRouter) · 10 小时前

DeepSeek 现已连续四周在我们平台的 token 份额排行榜上位居第一： https：//openrouter.ai/rankings

能力进展 新发布

https://x.com/OpenRouter/status/2062538625225548118

3. 微软AI负责人：Anthropic模型太贵，正自研更便宜的替代模型

Bloomberg：Technology（RSS） · 13 小时前

微软AI部门负责人表示，Anthropic推出的模型成本过高，公司目前正在内部研发更廉价的替代模型，以降低成本。

能力进展 新发布

https://www.bloomberg.com/news/newsletters/2026-06-04/microsoft-says-anthropic-models-are-too-expensive

4. 联合国报告：2030年AI数据中心水电消耗将翻倍

IT之家（RSS） · 23 小时前

联合国大学水、环境与健康研究所报告指出，受AI需求驱动，去年全球数据中心耗电448太瓦时（AI占五分之一），耗水4.5万亿升，碳排放1.89亿吨。预计到2030年，年耗电量将翻倍至945太瓦时（AI占40%），耗水增至9.3万亿升，碳排放升至3.99亿吨，占地面积从6900平方公里扩展至14500平方公里。报告警告若忽视环境成本，AI落地还将加剧土地紧张与电子废弃物问题。

基础设施

https://www.ithome.com/0/959/607.htm

5. Nemotron Parakeet ASR 印尼语准确率达 97.7%

X：NVIDIA (@nvidia) · 2 小时前

当法律和监督依赖于转录内容时，70-80% 是不够的。 http：//Rafiqspace.ai 通过微调 Nemotron Parakeet ASR 达到了 97.7% 的印尼语准确率（2.3% WER）-- 优于全球工具，同时每小时成本降低高达 90%。⚖

基础设施

https://x.com/nvidia/status/2062647600424128773

6. 微软与OpenAI分道扬镳--如今双方准备正面交锋

The Verge：订阅版科技（RSS） · 昨天 22:04

微软与OpenAI的合作关系已彻底破裂，双方进入正面竞争态势。前DeepMind高管、现任微软AI主管Mustafa Suleyman明确表示，微软现在必须从头证明自己能独立完成所有必需的任务。这标志着两家科技巨头从紧密合作伙伴转变为直接竞争对手。

新发布

https://www.theverge.com/ai-artificial-intelligence/942242/microsoft-build-ai-agents-openai-competition

7. 台积电难以跟上AI需求："我们只能支持这么多"

The Verge：订阅版科技（RSS） · 10 小时前

全球最大芯片制造商台积电表示，通过美国本土生产满足客户需求可能需要"非常长的时间"，凸显AI需求带来的产能压力。

基础设施

https://www.theverge.com/tech/943066/tsmc-ai-demand-struggles

8. Suno完成4亿美元D轮融资

X：Suno (@suno) · 昨天 22:44

我们激动地宣布Suno的新篇章：4亿美元D轮融资，估值54亿美元！我们的使命一直很简单：让更多人能体验制作音乐的乐趣。非常感谢我们不可思议的社区和投资者与我们共同建设。 点击此处阅读Mikey的博客：https：//suno.com/blog/series-d-announcement

监管/资本

https://x.com/suno/status/2062183524887675243

应用 应用与商业化

1. 介绍 Claude Partner Network 的 Services Track 和 Partner Hub

Anthropic：Newsroom（网页） · 昨天 21:30

Anthropic 扩展 Claude Partner Network，推出 Services Track 分级体系和 Partner Hub 门户。Services Track 设 Select、Preferred、Global Premier 三级，按认证人数、投产客户数及客户推荐信量化评定。Partner Hub 提供每日更新仪表盘和公开目录，方便合作伙伴查看进展、客户寻找供应商。该网络三月

能力进展 基础设施 新发布

https://www.anthropic.com/news/services-track-partner-hub

2. OpenClaw 2026.6.1发布：新增Windows节点与技能工坊

X：OpenClaw (@openclaw) · 昨天 05:40

OpenClaw 2026.6.1 已上线 ⌧⌧原生 Windows 节点主机 ⌧用于自主学习型智能体的技能工坊（Skill Workshop） ⌧工作板（Workboard）编排 ⌧支持 MiniMax M3 Windows 加入集群。无需企鹅服。 https：//github.com/openclaw/openclaw/releases/tag/v2026.6.1

能力进展 基础设施 新发布

https://x.com/openclaw/status/2062288421406785710

3. OpenShell v0.0.55 发布：新增 Vertex AI 推理支持

X：NVIDIA AI (@NVIDIAAI) · 昨天 00:29

OpenShell v0.0.55 ⌧Google Vertex AI 推理提供者 ⌧基于配置文件的策略可见性 ⌧网关中更好的 Podman 检测 ⌧恢复 GPU procfs 基准行为 ⌧CI 与文档修复 运行智能体对接 Vertex AI，同时拥有改进的策略可见性以及更可靠的 Podman 和 GPU 沙箱行为。 https：//github.com/NVIDIA/OpenS

能力进展 基础设施 新发布

https://x.com/NVIDIAAI/status/2062210034109677665

4. Meet OpenJarvis：一个本地优先的设备端个人AI智能体框架，支持工具、记忆与学习

MarkTechPost（RSS） · 17 小时前

Stanford 研究人员发布 OpenJarvis，一个完全在设备端运行推理、智能体、记忆与学习的开源框架。它将个人 AI 系统分解为五个可组合原语：Intelligence、Engine、Agents、Tools & Memory 和 Learning。该框架与最佳云端模型的性能差距在 3.2 points 以内，边际 API 成本降低约 800 倍。

能力进展 基础设施 新发布

https://www.marktechpost.com/2026/06/03/meet-openjarvis-a-local-first-framework-for-on-device-personal-ai-agents-with-tools-memory-and-learning

5. Anthropic 开源 AI 驱动漏洞发现框架

Hacker News 热门（buzzing.cc 中文翻译） · 2 小时前

Anthropic 将其用于 AI 驱动漏洞发现的开源框架代码托管在 GitHub 上。该框架借助 AI 技术进行漏洞发现，旨在帮助识别软件中的安全缺陷。

能力进展 监管/资本 新发布

https://github.com/anthropics/defending-code-reference-harness

6. OpenAI API 新增内容审核评分

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

Moderation scores 现已在 Responses API 和 Completions API 中可用。 在与生成相同的请求流程中返回审核信号，然后由你的应用决定如何使用它们进行记录、路由、审核或拦截。 https：//developers.openai.com/api/docs/guides/moderation

能力进展 新发布

https://x.com/OpenAIDevs/status/2062619558440267801

7. 黄仁勋与纳德拉共议智能体AI时代

X：NVIDIA (@nvidia) · 昨天 01:44

智能体AI时代来了。 从台北，黄仁勋与@satyanadella共同出席#MSBuild，展示NVIDIA与@Microsoft如何携手构建，从Windows设备到规模化AI工厂。 ▶ 观看对话：https：//nvda.ws/4uefQbs

能力进展 基础设施

https://x.com/nvidia/status/2062228974273716457

8. Codex 推出 iOS 应用构建插件

X：OpenAI Developers (@OpenAIDevs) · 6 小时前

更多 iOS 应用循环，现已集成至 Codex。 Build iOS Apps 插件让 Codex 可在应用内浏览器查看和测试你的 iOS 应用，打开 SwiftUI 预览，并无需离开 Codex 即可热重载编辑。

能力进展 新发布

https://x.com/OpenAIDevs/status/2062599291479478275

9. NotebookLM 推出福尔摩斯游戏笔记本

X：NotebookLM (@NotebookLM) · 7 小时前

专业技巧：将笔记本游戏化 不要只是阅读笔记--去调查它们。我们全新的福尔摩斯笔记本将学习变成一款互动侦探游戏。推理事实，发现线索，证明即使是最复杂的问题也能迎刃而解。 ➡ https：//goo.gle/Sherlock

能力进展 新发布

https://x.com/NotebookLM/status/2062582348194197743

10. Dreaming： ChatGPT 推出更强的记忆系统，更好记住用户偏好

OpenAI：官网动态（RSS · 排除企业/客户案例） · 15 小时前

ChatGPT 推出名为 Dreaming 的新记忆系统，能够更有效地记住用户偏好，并在跨对话场景中保持上下文的新鲜感和相关性，从而提升助手的个性化体验。

能力进展 新发布

https://openai.com/index/chatgpt-memory-dreaming

---

11. Replit上线SEO Agent助应用被发现

X：Replit (@Replit) · 昨天 00:37

你发布了你的应用。然后呢？ 你的应用可能看起来很棒，但如果没人能找到它，它就依然不可见。 发布只是开始。 认识一下SEO Agent。它会为你运行一次扫描，并建议修复措施，帮助你的应用在网页搜索和AI搜索中被发现。

能力进展 新发布

https://x.com/Replit/status/2062211976995188871

---

12. Meta 面向 WhatsApp Business 的 AI 智能体现已全球上线

TechCrunch：AI（RSS） · 昨天 21:40

Meta 为 WhatsApp Business 打造的 AI 智能体面向全球商家开放，将按照模型 token 使用量向企业收费。

能力进展 新发布

https://techcrunch.com/2026/06/03/metas-ai-agent-for-whatsapp-business-is-now-available-globally

---

13. Grok模型登陆Cloudflare AI Gateway

X：xAI (@xai) · 昨天 06:03

在 @Cloudflare 的 AI Gateway 上尝试 Grok 模型！

能力进展 基础设施

https://x.com/xai/status/2062294202625696081

---

14. Hugging Face 为编码智能体重塑 hf CLI 输出格式

Hugging Face：Blog（RSS） · 昨天 08:00

Hugging Face 重新设计 hf CLI，使其同时服务人类用户和编码智能体（Claude Code、Codex 等）。CLI 通过环境变量自动检测智能体驱动，输出紧凑无截断的 TSV 格式，避免 ANSI 和交互提示，大幅降低 token 消耗。复杂多步任务中，不使用 CLI 的智能体 token 消耗最高达 hf CLI 的 6 倍。2026 年 4 月起，Hugging Face 追踪

能力进展

https://huggingface.co/blog/hf-cli-for-agents

---

15. Replit Agent 联手 Shopify 快速建店

X：Replit (@Replit) · 6 小时前

我们与 Shopify 合作，让你从想法到上线商店只需几分钟 只需告诉 Replit Agent 你想卖什么。它会： - 构建自定义店铺页面 - 创建你的 Shopify 商店 - 帮你添加商品 在 Shopify 中认领店铺，设置支付，即可开业。

能力进展

https://x.com/Replit/status/2062594881625940379

---

16. Gemini macOS 双击 Command 附加活动窗口

X：Gemini (@GeminiApp) · 2 小时前

使用适用于 macOS 的 Gemini 应用，获取针对屏幕内容的定制帮助。☒只需同时按下两个 Command ⌘ 键，即可将当前活动窗口无缝附加到聊天中，无需手动截图或切换标签页。

能力进展

https://x.com/GeminiApp/status/2062652523945836770

---

17. NotebookLM 来源归属功能上线

X：NotebookLM (@NotebookLM) · 2 小时前

今天，我们推出又一项呼声很高的功能：来源归属！☒无需再猜测。现在你可以看到每个创作物背后所用的确切公式（提示词 + 来源）。想要调整？只需轻点"迭代"，随心定制 ☒

新发布

https://x.com/NotebookLM/status/2062653124326863077

---

18. Perplexity Personal Computer 登陆 Windows

X：Perplexity (@perplexity_ai) · 昨天 23:05

Personal Computer 即将登陆 Windows。 面向 Windows 的 Personal Computer 在你的机器上运行，并协调你每天使用的应用和文件。 我们将首先向等候名单上的付费 Max 和 Enterprise Max 订阅用户推送。

https://x.com/perplexity_ai/status/2062189045728596080

研究 研究与开源进展

1. NVIDIA Research 在 CVPR 2026 发表三篇论文：规模化训练实现抓取、自动驾驶与智能体泛化

NVIDIA AI Blog · 昨天 23:00

NVIDIA Research 在 CVPR 2026 上发表三篇论文，展示规模化训练带来的泛化能力。GraspGen-X 是首个零样本抓取基础模型，基于 20 亿次模拟抓取训练，可为任意末端执行器生成抓取姿态。LCDrive 用紧凑潜在表示替代文本推理，让自动驾驶在嵌入式硬件上更快推理。NitroGen 基于 Isaac GR00T 架构，在大量虚拟环境中训练具身智能体。此外还发布了新的物理 A

能力进展 基础设施 新发布

https://blogs.nvidia.com/blog/cvpr-research-grasping-driving-agent-training

2. Google Research 发布被动心率监测系统 PHRM

Google Research：Blog（网页） · 3 小时前

Google Research 开发了一种被动心率监测系统（PHRM），利用智能手机前置摄像头在日常使用中（人脸解锁后数秒内）捕捉面部视频，通过深度学习估算心率，平均绝对百分比误差（MAPE）低于10%（对比心电图金标准），满足各肤色人群的行业精度标准。系统将全天心率测量整合为每日静息心率（RHR），平均绝对误差（MAE）低于5 bpm（对比可穿戴设备）。研究同时发布了迄今最大规模的公开智能手机视

能力进展 基础设施 新发布

https://research.google/blog/towards-passive-heart-health-monitoring-via-smartphone-camera

3. EVA-Bench Data 2.0 发布：覆盖三大领域、121 个工具、213 个场景

Hugging Face：Blog（RSS） · 11 小时前

EVA-Bench Data 2.0 将评估范围从单一企业领域扩展至航空公司客户服务管理（CSM）、企业 IT 服务管理（ITSM）和医疗 HR 服务交付（HRSD）三个领域，共涵盖 121 个工具、213 个场景，场景数较原始版本增长约 4 倍。每个场景均经 OpenAI GPT-5.4、Google Gemini 3.1 Pro 和 Anthropic Claude Opus 4.6 验证可解

能力进展 新发布

https://huggingface.co/blog/ServiceNow-AI/eva-bench-data

4. Nemotron 预训练的任务种子合成问答生成

Hugging Face：Blog（RSS） · 12 小时前

在 Nemotron-3 Nano 模型的 100B token 续训练实验中，任务种子合成数据生成（Task-Seeded SDG）使 MMLU-Pro 提升 1.8 分，平均代码提升 1.9 分，常识理解提升 1.6 分，GPQA 提升 11.1 分，数学成绩保持稳定。该流程利用 lm-eval-harness 中约 70 个公开任务（约 700 子任务）的训练集作为种子，生成新示例并补充推理

能力进展 基础设施

https://huggingface.co/blog/nvidia/task-seeded-sdg

5. NVIDIA PPISP：补偿光度变化提升3D重建

X：NVIDIA AI (@NVIDIAAI) · 22 小时前

辐射场的质量取决于其背后的图像。 PPISP 可帮助补偿不同拍摄之间的光度变化，使 3D 重建在光照和相机设置不完全一致时更加鲁棒。 ☒项目：https：//nvda.ws/43JeJpk

能力进展 基础设施

https://x.com/NVIDIAAI/status/2062358080222876041

6. 微软研究：装瓶厂AI从聊天到决策

X：Microsoft Research (@MSFTResearch) · 昨天 00:09

一份在中西部装瓶厂进行的三个月试点显示，当AI超越聊天进入决策领域时会发生什么--约束条件变化、风险真实、答案必须可靠。 https：//msft.it/6015vjYUN

https://x.com/MSFTResearch/status/2062204914223169635

格局 观点、资本与监管

1. Nemotron 3.5 ASR：为你的语言、领域或口音进行微调

Hugging Face：Blog（RSS） · 11 小时前

Nemotron 3.5 ASR 是一个 600M 参数的多语言流式语音识别模型，单个检查点覆盖 40 种语言-地区（含英、西、德、法、意、日、韩、中、阿拉伯等）。采用 Cache-Aware FastConformer 编码器与 RNNT 解码器，缓存内部状态避免重复计算，实现低延迟流式转录且不损失精度。模型原生输出带标点和大写的生产级文本，无需后处理。支持指定语言（target_lang=es

能力进展 基础设施 新发布

https://huggingface.co/blog/nvidia/fine-tuning-nemotron-35-asr

2. DharmaOCR 利用 DPO 将文本退化率降低 59.4%

Hugging Face：Blog（RSS） · 昨天 20:55

4月发布的DharmaOCR（结构化OCR模型）在巴西葡萄牙语文档提取任务中，使用直接偏好优化（DPO）作为监督微调（SFT）后的第二训练阶段。SFT无法直接惩罚文本退化（重复循环），而DPO以模型自身失败输出（退化循环）作为负样本进行偏好训练，使所有测试模型族的文本退化率平均降低59.4%，最高达87.6%（如Nanonets-OCR2-3B从1.61%降至0.20%）。传统DPO多用于聊天对齐

能力进展 基础设施 新发布

https://huggingface.co/blog/Dharma-AI/direct-preference-optimization-beyond-chatbots

3. Boson AI 与 LMSYS 发布基于 SGLang-Omni 的 Higgs Audio v3 TTS 端到端服务

LMSYS：Blog（Chatbot Arena 团队） · 7 小时前

Boson AI 与 LMSYS 联合推出基于 SGLang-Omni 推理框架的 Higgs Audio v3 TTS 端到端服务。该模型约 4B 参数，基于 Qwen3-4B 骨干，支持 100 种语言（内部评测覆盖 111 种），在 Seed-TTS、CV3、MiniMax-Multilingual 及 Higgs-Multilingual 零样本语音克隆任务中达到单字级 WER/CER。开

能力进展 新发布

https://www.lmsys.org/blog/2026-06-04-higgs-audio-v3-tts

---

4. OpenRouter 翻遍 11 款 LLM 找最快的决策模型：Claude vs. Grok 领衔

OpenRouter：Announcements（RSS） · 12 小时前

OpenRouter 用总价 482 美元的推理花费，让 11 款大语言模型在 30 轮实时决策的"大逃杀"挑战中正面竞争。实验结果表明，传统的静态 benchmark 排名无法反映模型在需要即时反应的智能体任务（如自主控制机器人）中的真实表现，Claude 和 Grok 系列模型在决策速度与任务成功率上表现突出，而多项高分模型的实时调度能力未达预期。

能力进展 新发布

https://openrouter.ai/announcements/royale-last-agent-standing

---

5. 世界模型的功能分类

X：Fei-Fei Li (@drfeifei, World Labs) · 昨天 02:57

World Labs团队与李飞飞发文，梳理"世界模型"这一被滥用的术语。对比语言模型学习文本统计，世界模型学习空间与时间统计（如光照、物理规律）。基于部分可观马尔可夫决策过程（POMDP）框架，智能体通过动作影响世界状态，观测是部分视图。当前被称为"世界模型"的不同系统本质上是同一循环的不同投影：第一类为渲染器，输出给人眼看的像素，以视觉保真度为核心。文章着重于概念分层，未给出具体模型名、参数或基

能力进展

https://x.com/drfeifei/status/2062247238143996275

---

6. OpenAI 发布《智能时代的生物防御》行动计划，以 AI 驱动生物韧性

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

OpenAI 提出一项名为"Biodefense in the Intelligence Age"的行动计划，旨在利用 AI 增强生物防御与韧性。该计划聚焦于构建 AI 驱动的生物安全能力，以应对未来可能出现的生物威胁。

监管/资本 新发布

https://openai.com/index/biodefense-in-the-intelligence-age

---

7. 马斯克谈SpaceX上市：正处大规模资本扩张期

X：cb_doge (@cb_doge) · 40 分钟前

马斯克在JPMorgan活动上回应SpaceX上市问题：他已被建议上市近10年，自2014-2015年起SpaceX就已实现正现金流并自筹资金，之前的私募轮次实际是面向投资者和员工的流动性/回购轮次。当前不同之处在于SpaceX正进入显著资本增长阶段，计划发射约10万颗通信卫星（可能超10万颗），AI和机器人对带宽需求巨大，还将在太空中建设AI数据中心，马斯克认为这将成为AI扩张的主要手段。

基础设施

https://x.com/cb_doge/status/2062681226633523250

---

8. OpenAI称AI递归自我改进迹象初现

X：Kim (@kimmonismus) · 11 小时前

OpenAI刚刚写道："我们也看到了当今系统中递归自我改进（RSI）的早期迹象：AI开发本身正被AI加速。 我们预计这将加剧开发者与国家之间的竞争压力，并带来现有机构无法应对的治理挑战。 随着RSI的出现，社会将需要找到塑造AI发展轨迹的方法，确保其服务于人类利益。" 气氛变了，有事正在发生。

新发布

https://x.com/kimmonismus/status/2062517474277675102

---

9. 共存与协同智能的终结

Ethan Mollick：One Useful Thing（RSS） · 3 小时前

Ethan Mollick 在 One Useful Thing 博客中，以"共存与协同智能的终结"为题，并附带介绍了如何向 AI 推销一本书。

https://www.oneusefulthing.org/p/co-existence-and-the-end-of-co-intelligence

---

10. Alex Imas 和 Phil Trammell：AGI 后什么仍然稀缺?

Dwarkesh Patel：Podcast & Blog（RSS） · 8 小时前

经济学家 Alex Imas 和 Phil Trammell 指出，AGI 时代机器人数量可以快速复制增长，但人类独特技能（以芭蕾舞演员为例）的数量保持不变，揭示了即使技术大幅进步，某些稀缺资源仍不可替代。

https://www.dwarkesh.com/p/alex-imas-phil-trammell

---

11. 优步每月 1，500 美元的 AI 使用上限为 AI 工具定价提供参考

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 03:00

优步将 AI 工具每月使用上限定为 1500 美元，这一做法为行业 AI 工具定价提供了有价值的参考信号。

https://simonwillison.net/2026/Jun/3/uber-caps-usage
