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: 37 条 焦点: 8 条 快讯: 0 条

Executive Summary

Nano Banana Pro和Nano Banana 2正式发布，通过Gemini API投入生产使用，标志着多模态模型能力进一步成熟。OpenAI推出实时翻译模型gpt-realtime-translate，支持70多种语言输入并翻译成13种输出语言，专门针对智能眼镜等设备优化。阶跃星辰发布开源多模态模型，体现小模型在特定场景下的竞争力。OpenRouter支持apply_patch功能，允许模型生成文件补丁，提升代码协作效率。

Anthropic在最新融资轮中估值达到9650亿美元，首次超越OpenAI，反映市场竞争格局变化。阿里云与Qwen成为UEFA全球AI合作伙伴，合作期覆盖2027年至2033年，显示大厂AI服务向体育赛事等垂直领域渗透。GitHub Copilot转向按token计费模式引发开发者争议，可能影响企业级AI工具采用策略。NVIDIA推出MCG Toolkit自动化模型文档生成，应对AB-2013和欧盟AI法案等监管要求，降低合规成本。

需关注Google新发布的Agents API生态建设情况，以及OpenAI实时翻译模型的商业化定价策略。ComfyUI集成OpenRouter模型调用后的市场接受度值得跟踪。Tiny-vLLM等轻量化推理引擎的性能表现将影响边缘部署成本结构。Anthropic估值超越OpenAI后的竞争态势变化，以及GitHub Copilot新计费模式对开发者生态的长期影响需要持续观察。

重点 今日核心进展

★ 1. Nano Banana Pro与Nano Banana 2正式发布

X：Google AI for Developers (@googleaidevs) · 12 小时前 · 模型与工具能力

ICYMI：Nano Banana Pro 【gemini-3-pro-image】 和 Nano Banana 2 【gemini-3.1-flash-image】 现已正式发布，可通过 Gemini API 投入生产使用。 查看这些优秀的社区示例，了解两个模型的实际能力 ☒

能力进展 新发布

https://x.com/googleaidevs/status/2060685345738375640

★ 2. OpenAI推出实时翻译模型，支持70+语言输入

X：Greg Brockman (@gdb) · 昨天 04:03 · 模型与工具能力

OpenAI 实时翻译功能--使用70多种输入语言说话，翻译成13种输出语言： gpt-realtime-translate 接收任意语言的语音输入，并输出目标语言的语音。 大语言模型很棒，但特定用例需要专用模型。 我们正在智能眼镜上运行此功能。

能力进展 新发布

https://x.com/gdb/status/2060452095279415725

★ 3. 小即是美：开源多模态模型发布

X：阶跃星辰 StepFun (@StepFun_ai) · 13 小时前 · 模型与工具能力

小即是美。😊。该条来自X：阶跃星辰 StepFun (@StepFun_ai)，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 新发布

https://x.com/StepFun_ai/status/2060678753030947226

★ 4. 如何使用 NVIDIA MCG Toolkit 实现 AI 模型文档自动化

NVIDIA Technical Blog（开发者技术博客 · RSS） · 昨天 00:00 · 观点、资本与监管

随着 AI 模型复杂性增加及加州 AB-2013、欧盟 AI 法案等监管框架加强，软件团队面临在模型发布前生成全面、可审计文档的挑战。NVIDIA MCG Toolkit 旨在自动化模型卡创建流程，该文档需涵盖模型工作原理、预期用途、许可证、训练数据与性能等关键信息。

能力进展 基础设施 监管/资本

https://developer.nvidia.com/blog/how-to-automate-ai-model-documentation-with-the-nvidia-mcg-toolkit

★ 5. 阿里云与Qwen成为UEFA多年全球AI合作伙伴

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 23 小时前 · 产业与基础设施

阿里云和Qwen成为UEFA官方独家AI、云计算与电子商务合作伙伴，合作期覆盖2027/2028赛季至2032/2033赛季的UEFA男子俱乐部赛事，以及UEFA EURO 2028。阿里巴巴集团主席蔡崇信表示，将投入云计算、全栈AI及全球电商平台能力，支持赛事运营。合作将利用Qwen大语言模型部署先进AI技术，增强球迷互动与媒体内容体验，并依托阿里云基础设施打造全球沉浸式观赛体验。

能力进展 基础设施

https://x.com/alibaba_cloud/status/2060520586489770167

★ 6. "开玩笑吧"：GitHub Copilot 新的基于 token 的计费模式引发开发者不满

TechCrunch：AI（RSS） · 7 小时前 · 产业与基础设施

微软旗下 GitHub Copilot 的黄金时代似乎正在终结。其新推出的计费模式改为按 token 计量，这一变化引发了开发者的广泛担忧与不满。

能力进展 新发布

https://techcrunch.com/2026/05/30/what-a-joke-github-copilots-new-token-based-billing-spurs-consternation-among-devs

★ 7. Codex现已支持Windows端计算机使用功能

X：OpenAI (@OpenAI) · 昨天 02:30 · 应用与商业化

Windows用户，这条消息是给你的。 计算机使用功能现已在Windows上可用，因此Codex可以在你的Windows电脑上执行操作。 通过ChatGPT移动应用中Codex的Windows支持，你可以在工作继续在Windows电脑上进行时，随时随地启动、审查和引导任务。 这是一项早期体验，但我们正在努力提供更多方式，让你的工作无论身在何处都能持续进行。

能力进展 新发布

https://x.com/OpenAI/status/2060428604727771421

★ 8. ComfyUI现已支持OpenRouter模型直接调用

X：OpenRouter (@OpenRouter) · 昨天 07:58 · 应用与商业化

现在你可以直接在ComfyUI工作流中使用你的OpenRouter模型了！ 【引用 @ComfyUI】：ComfyUI刚刚添加了@OpenRouter支持。 你不再局限于单一的大语言模型，现在可以直接在Comfy中访问20多个模型。 更多灵活性，更少摩擦，同样的工作流。 工作流链接在下方

能力进展 新发布

https://x.com/OpenRouter/status/2060511136932315259

能力 模型与工具能力

1. Nano Banana Pro与Nano Banana 2正式发布

X：Google AI for Developers (@googleaidevs) · 12 小时前

ICYMI：Nano Banana Pro 【gemini-3-pro-image】 和 Nano Banana 2 【gemini-3.1-flash-image】 现已正式发布，可通过 Gemini API 投入生产使用。 查看这些优秀的社区示例，了解两个模型的实际能力

能力进展 新发布

https://x.com/googleaidevs/status/2060685345738375640

2. OpenAI推出实时翻译模型，支持70+语言输入

X：Greg Brockman (@gdb) · 昨天 04:03

OpenAI 实时翻译功能--使用70多种输入语言说话，翻译成13种输出语言： gpt-realtime-translate 接收任意语言的语音输入，并输出目标语言的语音。 大语言模型很棒，但特定用例需要专用模型。 我们正在智能眼镜上运行此功能。

能力进展 新发布

https://x.com/gdb/status/2060452095279415725

3. 小即是美：开源多模态模型发布

X：阶跃星辰 StepFun (@StepFun_ai) · 13 小时前

小即是美。。该条来自X：阶跃星辰 StepFun (@StepFun_ai)，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 新发布

https://x.com/StepFun_ai/status/2060678753030947226

产业 产业与基础设施

1. 阿里云与Qwen成为UEFA多年全球AI合作伙伴

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 23 小时前

阿里云和Qwen成为UEFA官方独家AI、云计算与电子商务合作伙伴，合作期覆盖2027/2028赛季至2032/2033赛季的UEFA男子俱乐部赛事，以及UEFA EURO 2028。阿里巴巴集团主席蔡崇信表示，将投入云计算、全栈AI及全球电商平台能力，支持赛事运营。合作将利用Qwen大语言模型部署先进AI技术，增强球迷互动与媒体内容体验，并依托阿里云基础设施打造全球沉浸式观赛体验。

能力进展 基础设施

https://x.com/alibaba_cloud/status/2060520586489770167

2. "开玩笑吧"：GitHub Copilot 新的基于 token 的计费模式引发开发者不满

TechCrunch：AI（RSS） · 7 小时前

微软旗下 GitHub Copilot 的黄金时代似乎正在终结。其新推出的计费模式改为按 token 计量，这一变化引发了开发者的广泛担忧与不满。

能力进展 新发布

https://techcrunch.com/2026/05/30/what-a-joke-github-copilots-new-token-based-billing-spurs-consternation-among-devs

3. Anthropic估值9650亿美元超越OpenAI

Bloomberg：Technology（RSS） · 昨天 01:26

Anthropic在最新融资轮中筹集了650亿美元，融资后公司估值达到9650亿美元，首次超越竞争对手OpenAI的估值水平。

监管/资本 新发布

https://www.bloomberg.com/news/videos/2026-05-29/anthropic-valuation-of-965-billion-passes-openai-video

4. xAI放弃JAX GPU转向自研训练框架

X：SemiAnalysis (@SemiAnalysis_) · 20 小时前

突发新闻：JAX NVIDIA GPU & XLA：GPU最大客户刚刚宣布已放弃JAX GPU，宁愿用Grok Build"氛围编程"一个C训练框架。据报道xAI的JAX堆栈MFU低于10%。NVIDIA JAX团队过去两年全部主力996专注于支持xAI却失败了，我想他们现在可以休息并兑现期权了。

基础设施

https://x.com/SemiAnalysis_/status/2060571944575963482

5. 据报道：软银将在法国投资750亿欧元用于AI

Bloomberg：Technology（RSS） · 5 小时前

软银集团计划投资高达750亿欧元（约合870亿美元），用于在法国建设人工智能数据中心。该投资计划由《论坛报》与《金融时报》报道。

基础设施

https://www.bloomberg.com/news/articles/2026-05-30/softbank-to-invest-some-75-billion-in-ai-in-france-reports-say

6. AI 骗子正在创建虚假的黑人形象来销售 Shein 劣质商品

The Verge：AI（RSS） · 11 小时前

有卖家利用 AI 生成虚假的黑人形象，在 TikTok、Facebook 和 Instagram 上扮演手工制品创作者进行销售。例如一个名为 Aliyah 的 AI 生成形象，以带泪诉说的方式售卖所谓手工皮带扣，但该形象及其产品均为虚构。此类 AI 虚拟网红被用于推广通过代发货模式销售的批量生产品。

https://www.theverge.com/ai-artificial-intelligence/938844/ai-tiktok-shop-blackface-shein-dropshipping

7. 新加坡防务论坛：AI 风险超过核武器

Bloomberg：Technology（RSS） · 12 小时前

在新加坡举行的防务论坛上，专家警告AI风险已超越核武器。面板讨论指出，AI可能大幅压缩反应时间，导致决策者做出草率决定，对战略稳定构成威胁。

https://www.bloomberg.com/news/articles/2026-05-30/ai-dangers-eclipse-nuclear-weapons-at-singapore-defense-forum

应用 应用与商业化

1. Codex现已支持Windows端计算机使用功能

X：OpenAI (@OpenAI) · 昨天 02:30

Windows用户，这条消息是给你的。 计算机使用功能现已在Windows上可用，因此Codex可以在你的Windows电脑上执行操作。 通过ChatGPT移动应用中Codex的Windows支持，你可以在工作继续在Windows电脑上进行时，随时随地启动、审查和引导任务。 这是一项早期体验，但我们正在努力提供更多方式，让你的工作无论身在何处都能持续进行。

能力进展 新发布

https://x.com/OpenAI/status/2060428604727771421

2. ComfyUI现已支持OpenRouter模型直接调用

X：OpenRouter (@OpenRouter) · 昨天 07:58

现在你可以直接在ComfyUI工作流中使用你的OpenRouter模型了！ 【引用 @ComfyUI】：ComfyUI刚刚添加了@OpenRouter支持。 你不再局限于单一的大语言模型，现在可以直接在Comfy中访问20多个模型。 更多灵活性，更少摩擦，同样的工作流。 工作流链接在下方▫

能力进展 新发布

https://x.com/OpenRouter/status/2060511136932315259

3. OpenRouter支持模型生成文件补丁

X：OpenRouter (@OpenRouter) · 昨天 00:17

OpenRouter 现已支持 "apply_patch"，这是一个服务器工具，允许任何模型通过 Responses API 使用 V4A diffs 提出文件编辑建议。 模型生成一个补丁（创建、更新或删除文件）。OpenRouter 在服务器端验证 diff 语法。

能力进展 新发布

https://x.com/OpenRouter/status/2060395056196936054

4. Gemini 本月更新：全新界面与智能体助手

X：Gemini (@GeminiApp) · 昨天 23:55

从全新设计的 Gemini 界面，到 Gemini Spark 提供的全天候智能体辅助，以下是本月 Gemini 更新概览。▫

能力进展 新发布

https://x.com/GeminiApp/status/2060389565052096911

5. Show HN： Tiny-vLLM--基于 C 和 CUDA 的高性能大型语言模型推理引擎

Hacker News 热门（buzzing.cc 中文翻译） · 20 小时前

Tiny-vLLM 是一个用 C 和 CUDA 编写的高性能大语言模型推理引擎，项目代码已开源至 GitHub。

能力进展 新发布

https://github.com/jmaczan/tiny-vllm

6. ChatGPT对话目录功能现已上线

X：ChatGPT (@ChatGPTapp) · 昨天 05:03

对于每个始于"就问一件事"却演变成完整长篇的ChatGPT对话：目录功能现已推出。 适用于包含5条以上回复的对话。

能力进展 新发布

https://x.com/ChatGPTapp/status/2060467129066070182

7. Runway API持续扩展模型与端点支持

X：Runway (@runwayml) · 昨天 04:10

我们持续为 Runway API 添加新模型和端点，以便您能将最佳生成能力直接集成到应用、产品和平台中。通过 Runway API，您可以在一个地方获得所需的所有模型，包括 Seedance 2.0、GPT Image 2、HappyHorse 1.0、Nano Banana Pro、Magnific Precision Upscaler V2 等更多内容。请通过下方链接开始使用。

能力进展

https://x.com/runwayml/status/2060453805519765548

8. Luma Agents 自动生成宣传图，输入内容即可

X：Luma AI (@LumaLabsAI) · 昨天 04:40

博客文章完成了思考。现在让宣传来发挥作用。 输入内容。定义钩子。Luma Agents 从那里构建每一张宣传图。 投入使用 → http：//lumalabs.ai/app

能力进展

https://x.com/LumaLabsAI/status/2060461313713909783

9. Gemini Omni可将草图变为现实

X：Gemini (@GeminiApp) · 昨天 02:59

Gemini Omni甚至能将简单的草图变为新的现实。 在Gemini应用中亲自尝试。上传一段有人画圆的视频，然后输入这个提示词：当我画完这个圆时，它变成了___。

能力进展

https://x.com/GeminiApp/status/2060435981946503243

10. Codex可自主管理对话线程与并行任务

X：Greg Brockman (@gdb) · 昨天 06:19

Codex用于管理Codex界面： 【引用 @guinnesschen】：如果你厌倦了管理Codex对话线程，就让Codex自己管理自己吧！Codex现在可以创建对话线程、搜索它们、整理它们、固定重要的线程，并为并行任务启动工作树。

https://x.com/gdb/status/2060486309886443787

11. Canvas新功能与Clerk自定义登录介绍

X：Replit (@Replit) · 昨天 23:58

Canvas 新功能 + 使用 Clerk 自定义登录 https：//x.com/i/broadcasts/1pJdRRzreMRKW

https://x.com/Replit/status/2060390354910859401

研究 研究与开源进展

1. GPIC：大规模视觉生成基准数据集发布

X：Fei-Fei Li (@drfeifei, World Labs) · 昨天 00:56

我对这个适用于大规模生成模型新时代的视觉生成基准数据集感到非常兴奋！

能力进展 新发布

https://x.com/drfeifei/status/2060404846734512205

格局 观点、资本与监管

1. 如何使用 NVIDIA MCG Toolkit 实现 AI 模型文档自动化

NVIDIA Technical Blog（开发者技术博客 · RSS） · 昨天 00:00

随着 AI 模型复杂性增加及加州 AB-2013、欧盟 AI 法案等监管框架加强，软件团队面临在模型发布前生成全面、可审计文档的挑战。NVIDIA MCG Toolkit 旨在自动化模型卡创建流程，该文档需涵盖模型工作原理、预期用途、许可证、训练数据与性能等关键信息。

能力进展 基础设施 监管/资本

https://developer.nvidia.com/blog/how-to-automate-ai-model-documentation-with-the-nvidia-mcg-toolkit

2. 在浏览器中通过 Pyodide 和 Service Worker 运行 Python ASGI 应用

Simon Willison 博客 · 3 小时前

作者展示了如何在浏览器中通过 Pyodide 和 Service Worker 运行 Python ASGI 应用。此前的 Datasette Lite 使用 Web Workers，但无法执行 `<script>` 标签中的 JavaScript。新方案由 Claude Opus 4.8 协助完成开发，解决了这一问题。作者已展示了基础的 ASGI FastCGI 演示和运行 Datasette

能力进展 新发布

https://simonwillison.net/2026/May/30/pyodide-asgi-browser

3. 免费领取6个月ChatGPT Pro及AI工具思考

X：阿易 AI Notes (@AYi_AInotes) · 9 小时前

OpenAI为开源项目维护者提供福利，可免费领取6个月ChatGPT Pro（价值$1200），申请无硬性Star数要求，有项目链接即可。同时，文章引用讨论了AI工具的分类：一类是"agent型"（如Claude Code、Codex），可自主运行；另一类是"实习生型"（如Cursor），需人工决策，有助于使用者以术入道、培养判断力，但受限于需人在场。作者推荐了网易的UU远程工具，称其免费两年，支

能力进展 新发布

https://x.com/AYi_AInotes/status/2060740414273941874

---

4. 这个 skill 看着不错，可将文字、URL 或文章直接生成公众号首图、小红书图文卡、教程步骤卡等视觉物料，支持 28 种布局和 10 种主题。

X：洪明 (@hongming731) · 昨天 06:23

claude-design-card 是一款专为中文内容创作者设计的 Skill。它能将文字、URL 或文章直接转化为可发布的视觉卡片，如公众号首图、小红书图文卡、教程步骤卡等，支持 28 种布局与 10 种主题。其核心价值在于自动化了"写完文章"后最繁琐的流程：自动提炼重点、选择版式、生成 HTML 并截图成 PNG，替代了以往手动使用 Figma 或 Canva 等工具的步骤。该工具开源，适合

能力进展 新发布

https://x.com/hongming731/status/2060487110906527820

---

5. LlamaIndex 团队基于 Google Agents API 构建 LlamaParse/LiteParse 智能体模板

X：Google AI for Developers (@googleaidevs) · 昨天 03:15

LlamaIndex 团队基于 Google 新发布的 Agents API 构建了一个模板，使智能体能够访问 LlamaParse 和 LiteParse，从而自动处理非结构化文档。其工作流程为：配置数据与输出的 Git 仓库，将仓库克隆至智能体沙箱，安装 LiteParse CLI 与 LlamaParse SDK 及相关技能，最后通过提示词驱动智能体自主执行任务。该模板最终形成一个可直接使用

能力进展 新发布

https://x.com/googleaidevs/status/2060439904929382700

---

6. 亲测为实：难以置信的推理速度

X：Rohan Paul (@rohanpaul_ai) · 昨天 01:14

Kog团队在标准数据中心GPU上实现了极高的单用户推理速度，在8× AMD MI300X GPUs上达到3，000 tokens/s，在8× NVIDIA H200上达到2，100 tokens/s。相比常规推理速度（约100-300 tokens/s），实现了10-30倍提升。其核心思路是将LLM解码视为内存流问题，通过协同设计monokernel、重建同步机制、针对性内存访问映射及采用延迟张量

能力进展 基础设施

https://x.com/rohanpaul_ai/status/2060409504693645440

---

7. DynoSim：模拟帕累托前沿

NVIDIA Technical Blog（开发者技术博客 · RSS） · 昨天 06:31

现代大语言模型服务难以调优，因为每个部署都涉及模型后端、张量并行形状、预填充/解码分割、Worker数量、调度器设置、路由策略、KV缓存行为、自动扩展阈值和拓扑等相互关联的选择。这些选择在多个层级相互作用，局部优化可能导致瓶颈转移至其他环节。

能力进展 基础设施

https://developer.nvidia.com/blog/dynosim-simulating-the-pareto-frontier

---

8. Gemini架构师分享AI前沿探索幕后故事

X：Google AI (@GoogleAI) · 昨天 00:05

聆听Gemini的架构师们回顾他们持续推动AI前沿的旅程，本期Release Notes节目。 @JeffDean、@koraykv、@OriolVinyalsML和@NoamShazeer一同出镜，分享模型背后团队的幕后故事，以及他们如何见证愿景的实现。

能力进展 新发布

https://x.com/GoogleAI/status/2060392191508488493

---

9. NVIDIA 或将于六月发布整合 Blackwell GPU 与 AI 单元的 ARM 笔记本芯片 N1X

X：阿易 AI Notes (@AYi_AInotes) · 6 小时前

NVIDIA、微软与 Arm 同步发布指向台北音乐中心的坐标，暗示 6 月 1 日发布会将有重大动作。此举被认为是 NVIDIA 与联发科合作的 ARM 笔记本芯片 N1X 的预告。该芯片整合了 CPU、基于 Blackwell 架构的 GPU 及 AI 单元，目标是使轻薄本具备接近 RTX 4070 的图形性能。这标志着 NVIDIA 的战略转变：从显卡供应商，转型为定义整机核心方案的提供商，将

基础设施 新发布

https://x.com/AYi_AInotes/status/2060779431648547016

---

10. 参与我们的 I/O 2026 测验：该测验由 Google AI Studio 氛围编程生成

Google Blog：AI（RSS） · 昨天 03:00

Google 使用其开发工具 Google AI Studio，通过氛围编程（vibe coding）方式，创建了一个关于 Google I/O 2026 主要公告的在线测验。

能力进展

https://blog.google/innovation-and-ai/technology/ai/io-2026-vibe-coded-quiz

---

11. 当公司过于"AI上瘾"时会发生什么?

TechCrunch：AI（RSS） · 昨天 01:57

Box创始人Aaron Levie指出，决定用AI替代员工的人往往最不了解工作的实际内容，他将此称为"AI psychosis"。ClickUp近期为部署AI智能体裁员22%即是一例。2026年的科技行业裁员规模已接近2025年全年。

能力进展

https://techcrunch.com/video/what-happens-when-companies-become-too-ai-pilled

12. Cognition的Scott Wu表示：AI编程智能体不应取代人类

TechCrunch：AI（RSS） · 昨天 00:13

Cognition公司开发了Devvin，这是一个号称首个且最成功的AI编程智能体。其著名程序员创始人Scott Wu明确表示，该智能体并非旨在取代人类程序员。

能力进展

https://techcrunch.com/2026/05/29/cognitions-scott-wu-says-ai-coding-agents-shouldnt-replace-humans

13. 随着成本飙升，美国企业开始对人工智能实施配给

Hacker News 热门（buzzing.cc 中文翻译） · 8 小时前

由于运行和使用AI工具的成本持续飙升，美国企业正开始对人工智能的使用实施配给制。企业通过限制使用量、设置分层级审批流程等方式控制开支，以应对AI费用增长过快的问题。这种从广泛采用转向精细化管理的策略，标志着企业在AI应用上从追求速度转向注重成本效益。

https://www.wsj.com/tech/ai/corporate-america-is-starting-to-ration-ai-as-cost-skyrockets-1eb99d7a

14. Kling AI助力电影RAPHAEL创作全流程揭秘

X：可灵 Kling AI (@Kling_ai) · 昨天 23:00

Kling AI戛纳展示--RAPHAEL：AI工作流幕后 深入了解RAPHAEL，一部使用Kling AI创作的AI驱动故事片。看看创作者如何在整个电影制作流程中运用Kling AI，从创意构思到最终电影画面，简化制作并释放新的创作可能性。

https://x.com/Kling_ai/status/2060375625404432757

15. 最后一次技术面试

Hacker News 热门（buzzing.cc 中文翻译） · 16 小时前

Steve Yegge 在 Medium 发表观点文章《最后一次技术面试》，探讨 AI 时代传统技术面试的意义变化。文章在 Hacker News 社区获得 100 分关注。

https://steve-yegge.medium.com/the-last-technical-interview-bc13ddcf4564