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: 55 条 焦点: 8 条 快讯: 0 条

Executive Summary

xAI发布Grok Build 0.1编码模型进入API公测，专为智能体编码任务设计，推理速度超100 tokens/秒；阶跃星辰推出Step 3.7 Flash开源模型，198B参数MoE架构在ClawEval-1.1评测中获得67.1分；Anthropic发布Claude Opus 4.8，在编码、智能体技能和推理方面全面升级。Google推出Nano Banana Pro和Nano Banana 2，通过Gemini API投入生产使用；Mistral AI发布Search Toolkit框架，整合数据摄取、检索和评估功能。

Anthropic完成650亿美元H轮融资，投后估值达9650亿美元，年化收入突破470亿美元；Cognition获超10亿美元融资成为最大独立智能体实验室，年化收入增至4.92亿美元。Qwen3.7-Max登顶OpenRouter热门模型榜首，Qwen3.5在TokenSpeed引擎上达到580 tokens per second推理速度；OpenRouter获1.13亿美元B轮融资，显示模型聚合平台需求增长。

后续需关注xAI模型定价策略对市场的影响，各厂商智能体专用模型的性能表现与商业化节奏，以及OpenRouter等中间层服务商在AI生态中的角色演变。算力成本下降趋势下，模型推理速度提升对实际应用场景的价值转化，以及MCP协议标准化进程对开发者工具生态的重塑值得关注。

重点 今日核心进展

★ 1. Grok Build 0.1 on API

xAI：News（网页） · 昨天 08:00 · 模型与工具能力

xAI 的最新编码模型 Grok Build 0.1 已通过 xAI API 进入公开测试阶段。该模型专为智能体编码任务训练，支持网页开发、调试和 MCP，同时也是驱动 Grok Build CLI 的同一模型。其推理速度超过 100 tokens/秒，定价为输入 $1/m tokens，输出 $2/m tokens。除编码外，它也适用于通用智能体及工具调用场景，并可通过 OpenRouter 和

能力进展 基础设施 新发布

https://x.ai/news/grok-build-0-1

★ 2. 阶跃星辰 Step 3.7 Flash 发布，聚焦智能体效率

X：阶跃星辰 StepFun (@StepFun_ai) · 48 分钟前 · 模型与工具能力

阶跃星辰（Step）发布了开源大模型 Step 3.7 Flash，主打智能体（Agent）工作流的效率。该模型在 ClawEval-1.1（67.1分）和 SimpleVQA Search（79.2分）评测中排名第一。其架构为 198B 参数的 MoE，约 11B 为活跃参数，支持 256K 上下文。模型具备多模态理解能力，能处理图像、文档并生成代码或调用工具执行任务。在工具使用方面，它致力于高

能力进展 新发布

https://x.com/StepFun_ai/status/2060149124117475791

★ 3. Claude Opus 4.8 发布：在编码、智能体技能与推理方面实现全面升级

Anthropic：Newsroom（网页） · 7 小时前 · 模型与工具能力

Anthropic 发布了新一代模型 Claude Opus 4.8，作为 Opus 4.7 的升级版本，其在编码、智能体技能、推理和实用知识工作等各项基准测试中均取得进步。Claude Opus 4.8 现已可用，价格与前代相同。同步推出的新功能包括：用户可控制任务投入程度、Claude Code 新增"动态工作流"特性，以及 Opus 4.8 的 2.5 倍速模式价格降低为以往的三分之一。早期

能力进展 新发布

https://www.anthropic.com/news/claude-opus-4-8

★ 4. Anthropic 完成 650 亿美元 H 轮融资，估值达 9650 亿美元

Anthropic：Newsroom（网页） · 6 小时前 · 产业与基础设施

Anthropic 宣布完成由 Altimeter Capital 等领投的 650 亿美元 H 轮融资，投后估值达 9650 亿美元。公司表示其旗舰模型 Claude 的企业部署持续增长，年化收入已突破 470 亿美元。此轮融资将用于推进 AI 安全与可解释性研究、扩展算力以满足 Claude 的需求，并规模化产品与合作伙伴关系。Anthropic 近期已显著扩大计算容量，并宣布 Claude

能力进展 基础设施 监管/资本

https://www.anthropic.com/news/series-h

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

X：Google AI for Developers (@googleaidevs) · 7 小时前 · 模型与工具能力

⌧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/2060049962356916377

★ 6. 商汤发布信息图生成模型升级，增强多项核心能力

X：商汤 SenseTime (@SenseTime_AI) · 9 小时前 · 模型与工具能力

商汤科技介绍了其升级后的信息图生成模型 SenseNova-U1-8B-MoT-Infographic。该模型参数为8B，在四个关键维度进行了优化：文本准确性与可读性增强，减少了重复和不当放大；布局的一致性与合理性提升，背景更稳定；图表与示意图的质量提高；并新增了学术内容的渲染支持。推文提供了在 Hugging Face 上的模型页面链接及能力展示页面。

能力进展 新发布

https://x.com/SenseTime_AI/status/2060015749826240724

★ 7. 发布 Search Toolkit

Mistral AI：News（网页） · 12 小时前 · 应用与商业化

Mistral AI 发布了 Search Toolkit 的公共预览版。这是一个用于构建 AI 应用生产级搜索管道的可组合框架。该框架旨在解决团队在搭建搜索基础设施时，因数据摄取、检索和评估工具分散而耗费过多工程时间的问题。Search Toolkit 将这三者整合到单一框架与共享接口中，使团队能更专注于提升搜索质量。该工具开源，可部署在云端、本地或边缘环境，并支持企业搜索、RAG 等多种检索场

能力进展 基础设施 新发布

https://mistral.ai/news/search-toolkit

★ 8. Qwen3.7-Max 登顶 OpenRouter 热门大模型榜

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 16 小时前 · 产业与基础设施

Qwen3.7-Max 以 77.3B tokens 的使用量登顶 @OpenRouter 热门大语言模型榜单。 而我们才刚刚开始。 ☒https：//int.alibabacloud.com/m/1000413314/

能力进展 基础设施 新发布

https://x.com/alibaba_cloud/status/2059918150997623004

能力 模型与工具能力

1. Grok Build 0.1 on API

xAI：News（网页） · 昨天 08:00

xAI 的最新编码模型 Grok Build 0.1 已通过 xAI API 进入公开测试阶段。该模型专为智能体编码任务训练，支持网页开发、调试和 MCP，同时也是驱动 Grok Build CLI 的同一模型。其推理速度超过 100 tokens/秒，定价为输入 $1/m tokens，输出 $2/m tokens。除编码外，它也适用于通用智能体及工具调用场景，并可通过 OpenRouter 和

能力进展 基础设施 新发布

https://x.ai/news/grok-build-0-1

2. 阶跃星辰 Step 3.7 Flash 发布，聚焦智能体效率

X：阶跃星辰 StepFun (@StepFun_ai) · 48 分钟前

阶跃星辰（Step）发布了开源大模型 Step 3.7 Flash，主打智能体（Agent）工作流的效率。该模型在 ClawEval-1.1（67.1分）和 SimpleVQA Search（79.2分）评测中排名第一。其架构为 198B 参数的 MoE，约 11B 为活跃参数，支持 256K 上下文。模型具备多模态理解能力，能处理图像、文档并生成代码或调用工具执行任务。在工具使用方面，它致力于高

能力进展 新发布

https://x.com/StepFun_ai/status/2060149124117475791

3. Claude Opus 4.8 发布：在编码、智能体技能与推理方面实现全面升级

Anthropic：Newsroom（网页） · 7 小时前

Anthropic 发布了新一代模型 Claude Opus 4.8，作为 Opus 4.7 的升级版本，其在编码、智能体技能、推理和实用知识工作等各项基准测试中均取得进步。Claude Opus 4.8 现已可用，价格与前代相同。同步推出的新功能包括：用户可控制任务投入程度、Claude Code 新增"动态工作流"特性，以及 Opus 4.8 的 2.5 倍速模式价格降低为以往的三分之一。早期

能力进展 新发布

https://www.anthropic.com/news/claude-opus-4-8

4. Nano Banana Pro与Nano Banana 2正式发布

X：Google AI for Developers (@googleaidevs) · 7 小时前

☒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/2060049962356916377

5. 商汤发布信息图生成模型升级，增强多项核心能力

X：商汤 SenseTime (@SenseTime_AI) · 9 小时前

商汤科技介绍了其升级后的信息图生成模型 SenseNova-U1-8B-MoT-Infographic。该模型参数为8B，在四个关键维度进行了优化：文本准确性与可读性增强，减少了重复和不当放大；布局的一致性与合理性提升，背景更稳定；图表与示意图的质量提高；并新增了学术内容的渲染支持。推文提供了在 Hugging Face 上的模型页面链接及能力展示页面。

能力进展 新发布

https://x.com/SenseTime_AI/status/2060015749826240724

产业 产业与基础设施

1. Anthropic 完成 650 亿美元 H 轮融资，估值达 9650 亿美元

Anthropic：Newsroom（网页） · 6 小时前

Anthropic 宣布完成由 Altimeter Capital 等领投的 650 亿美元 H 轮融资，投后估值达 9650 亿美元。公司表示其旗舰模型 Claude 的企业部署持续增长，年化收入已突破 470 亿美元。此轮融资将用于推进 AI 安全与可解释性研究、扩展算力以满足 Claude 的需求，并规模化产品与合作伙伴关系。Anthropic 近期已显著扩大计算容量，并宣布 Claude

能力进展 基础设施 监管/资本

https://www.anthropic.com/news/series-h

2. Qwen3.7-Max 登顶 OpenRouter 热门大模型榜

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 16 小时前

Qwen3.7-Max 以 77.3B tokens 的使用量登顶 @OpenRouter 热门大语言模型榜单。 而我们才刚刚开始。 https：//int.alibabacloud.com/m/1000413314/

能力进展 基础设施 新发布

https://x.com/alibaba_cloud/status/2059918150997623004

3. Cognition成为全球最大独立智能体实验室

X：swyx (@swyx) · 昨天 03:23

Cognition宣布已成为全球最大的独立智能体实验室。公司完成超10亿美元融资，估值达260亿美元，由Lux Capital、General Catalyst等领投。其企业使用量自年初增长超10倍，年化收入增至4.92亿美元。Cognition于两年前推出Devin，定位为首个AI软件工程师。公司强调其拥有多项领先优势，包括首个编码智能体、顶级代码审查能力等，并得到了Peter Thiel的重大

能力进展 监管/资本 新发布

https://x.com/swyx/status/2059717021944926238

4. OpenRouter 获得1.13亿美元B轮融资

OpenRouter：Announcements（RSS） · 10 小时前

AI模型聚合平台OpenRouter宣布完成1.13亿美元B轮融资。本轮融资由CapitalG领投，NVentures、ServiceNow Ventures等多家机构参投，现有投资者Andreessen Horowitz与Menlo Ventures也参与了本轮融资。

能力进展 监管/资本 新发布

https://openrouter.ai/announcements/series-b

5. OpenAI 的前沿治理框架

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

OpenAI 发布了"前沿治理框架"，阐述其 AI 安全、安全与风险管理实践如何与欧盟和加州新出台的法规要求对齐。该框架旨在规范其前沿模型的开发与部署流程。

能力进展 监管/资本 新发布

https://openai.com/index/openai-frontier-governance-framework

6. Apple 正努力将庞大的 Gemini 模型塞进 iPhone 以驱动新 Siri

Ars Technica：AI（RSS） · 6 小时前

Apple 正尝试将大型 Gemini 模型集成到 iPhone 中，以支持全新的 Siri 功能。由于模型规模庞大，本地处理可能无法完全实现，因此一个云端组件很可能是必然的选择。

能力进展 基础设施

https://arstechnica.com/ai/2026/05/apple-reportedly-trying-to-distill-googles-multi-trillion-parameter-gemini-ai-to-run-on-iphone

7. DeepSeek计划在完成融资后立即申请科创板IPO

X：X.PIN (@thexpin) · 14 小时前

独家：DeepSeek计划在完成当前约500亿美元（3500亿人民币）融资轮后，立即申请科创板（A股）IPO。 来源：参与本轮融资的一位大型基金经理。

能力进展 监管/资本

https://x.com/thexpin/status/2059947998302679059

8. 萨姆·阿尔特曼和达里奥·阿莫代伊都纷纷收回了关于AI将引发就业危机的预测

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

Hacker News 热门（buzzing.cc 中文翻译）披露：萨姆·阿尔特曼和达里奥·阿莫代伊都纷纷收回了关于AI将引发就业危机的预测。该条属于产业与基础设施方向，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

https://fortune.com/2026/05/26/sam-altman-dario-amodei-walking-back-ai-jobs-apocalypse-prophecies-ipo

9. Replit入选Redpoint 2026 InfraRed 100榜单

X：Replit (@Replit) · 昨天 01:30

Replit 被列入 @Redpoint 的 2026 InfraRed 100 榜单。 这是一份构建 AI 运行基础设施的公司名单。很荣幸能入选，与 Stripe、Snowflake 和 HashiCorp 等往届入选者并列。 致每一位在 Replit 上发布产品的构建者：这份荣誉属于你们。 https：//www.redpoint.com/infrared/report/

新发布

https://x.com/Replit/status/2059688584899154134

10. 黄仁勋展示英伟达台湾新园区

X：Rohan Paul (@rohanpaul_ai) · 昨天 01:33

黄仁勋展示了新的台湾园区。 英伟达计划每年在台湾投资约1500亿美元。 就在竞争对手AMD宣布将向台湾AI领域投资超过100亿美元一周后。

基础设施

https://x.com/rohanpaul_ai/status/2059689400267939925

11. 可灵AI将在AI电影节展示20部4K原创短片

X：可灵 Kling AI (@Kling_ai) · 16 小时前

可灵AI将在全球最大AI影视会议AI on the Lot的社区日上，展示由Prompt Club的电影制作人创作的20部原创AI短片。所有短片均为原生4K分辨率，旨在探索AI电影的边界。该展示将于5月29日在加州卡尔弗市的卡尔弗剧院举行。

https://x.com/Kling_ai/status/2059908126611292645

应用 应用与商业化

1. 发布 Search Toolkit

Mistral AI：News（网页） · 12 小时前

Mistral AI 发布了 Search Toolkit 的公共预览版。这是一个用于构建 AI 应用生产级搜索管道的可组合框架。该框架旨在解决团队在搭建搜索基础设施时，因数据摄取、检索和评估工具分散而耗费过多工程时间的问题。Search Toolkit 将这三者整合到单一框架与共享接口中，使团队能更专注于提升搜索质量。该工具开源，可部署在云端、本地或边缘环境，并支持企业搜索、RAG 等多种检索场

能力进展 基础设施 新发布

https://mistral.ai/news/search-toolkit

2. 阿里云DataWorks推出AI数据智能体

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 21 小时前

认识 DataWorks Data Agent--阿里云的AI数据智能体！ 借助AI简化数据工作流，加速洞察，让数据管理更智能。 了解更多：https：//int.alibabacloud.com/m/1000413560/ #AlibabaCloud #DataWorks #AI #DataAgent #BigData #DataAnalytics

能力进展 基础设施 新发布

https://x.com/alibaba_cloud/status/2059840407618396391

3. 开源FastVideo Dreamverse实时视频生成工具

X：Sky Computing Lab (@haoailab) · 昨天 01:58

仅需7秒即可生成30秒1080p视频！ 我们开源了FastVideo Dreamverse：基于单张NVIDIA B200 GPU和LTX-2模型，实现实时视频生成的氛围引导工具。 Repo：https：//github.com/hao-ai-lab/FastVideo/tree/main/apps/dreamverse Blog： https：//haoailab.com/blogs/f

能力进展 基础设施 新发布

https://x.com/haoailab/status/2059695648103112946

4. Perplexity开源Unigram分词器降低CPU占用

X：Perplexity (@perplexity_ai) · 昨天 23:55

我们开源了重新构建的Unigram分词器，可将CPU占用降低5-6倍。 小型重排序器和嵌入模型在GPU上运行时间仅为个位数毫秒，使得CPU分词成为总延迟的重要组成部分。 http：//github.com/perplexityai/pplx-garden

能力进展 基础设施 新发布

https://x.com/perplexity_ai/status/2059664738087469511

5. 使用 Google Pay & Wallet Developer MCP server 加速你的集成工作流

Google Developers Blog（RSS） · 6 小时前

Google 推出 Google Pay & Wallet Developer MCP server，这是一款开放标准工具，旨在将 AI 开发助手和 IDE 安全连接到实时的 API 与账户上下文。开发者无需离开开发环境，即可搜索官方文档、验证 Wallet pass 定义、检查集成状态以及管理商户账户。该集成旨在通过减少上下文切换并提供实时、可靠的 AI 支持来减少开发摩擦，从而加速开发工作流。

能力进展 监管/资本 新发布

https://developers.googleblog.com/supercharge-your-integration-workflow-with-the-google-pay-wallet-developer-mcp-server

6. OpenClaw 2026.5.27 版本发布

X：OpenClaw (@openclaw) · 11 小时前

OpenClaw 2026.5.27 已上线 更严格的运行时/安全边界 ⚡ 更快的网关 + 回复路径 更稳定的 Codex/应用服务器内存 更好的频道、提供商、Pixverse 视频 更少阻碍，更多掌控。 https：//github.com/openclaw/openclaw/releases/tag/v2026.5.27

能力进展 监管/资本 新发布

https://x.com/openclaw/status/2059985767456231751

7. OpenAI产品支持私有MCP服务器安全连接

X：OpenAI Developers (@OpenAIDevs) · 昨天 02:29

您的团队可以在内部网络中保留MCP服务器，同时ChatGPT、Codex和Responses API通过仅出站HTTPS进行连接。

能力进展 监管/资本 新发布

https://x.com/OpenAIDevs/status/2059703536825565499

8. Runway 推出 Model Context Protocol 服务器

Runway：News（网页） · 昨天 22:09

Runway 正式推出 Runway MCP 服务器，允许任何兼容 MCP 的 AI 智能体（如 Claude、ChatGPT、Cursor）在对话界面中直接生成图像与视频，无需切换工作流。该服务器接入了 Runway 最新的多款 SOTA 模型，包括 Gen-4.5、Seedance 2.0、GPT Image 2、Kling 3.0 及 Nano Banana Pro。其应用场景涵盖为产品制作

能力进展 新发布

https://runwayml.com/news/mcp

---

9. 在Claude Code中引入动态工作流

Claude：Blog（网页） · 7 小时前

Claude Code 推出"动态工作流"功能，使 Claude 能端到端处理复杂任务。该功能通过动态编写脚本，在单个会话中并行运行数十到数百个子智能体来完成工作，并会在结果呈现前进行验证。它适用于跨代码库的 bug 查找、大规模迁移（如将 Bun 从 Zig 移植到 Rust）等需要多角度分析的任务。该功能现已在研究预览阶段可用，支持 Claude Code CLI、桌面端、VS Code 扩展

能力进展 新发布

https://claude.com/blog/introducing-dynamic-workflows-in-claude-code

---

10. MuleRun登陆阿里云市场，提供全天候AI劳动力

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 22 小时前

在阿里云市场遇见 MuleRun--一个全天候的AI劳动力，用于研究、报告、代码、设计等。功能强大，适合个人使用；企业就绪，适合团队协作--支持SSO、RBAC、私有网络、团队知识管理和无缝集成。 想得更大。让 MuleRun 处理其余事务。 方案起价 $20/月 → https：//int.alibabacloud.com/m/1000413520/ #AlibabaCloud #AIAg

能力进展 基础设施

https://x.com/alibaba_cloud/status/2059821825140367565

---

11. llm-anthropic 0.25.1

Simon Willison 博客 · 53 分钟前

llm-anthropic 发布 0.25.1 版本。主要更新包括：新增 Claude Opus 4.8 （`claude-opus-4.8`） 模型；为账户启用了该功能的组织新增了 `-o fast 1` 选项以使用快速模式；调整了各模型的默认 `max_tokens` 值，使其直接使用模型的最大输出长度，而非固定的 8，192。

能力进展 新发布

https://simonwillison.net/2026/May/28/llm-anthropic

---

12. 别只看基准测试，要看全面表现

X：OpenRouter (@OpenRouter) · 1 小时前

不要只依赖基准测试；要看全面情况！ 试试我们的新比较页面，它还能让你可视化模型性能：https：//openrouter.ai/compare/openai/gpt-5.5/anthropic/claude-opus-4.7/anthropic/claude-opus-4.8

能力进展 新发布

https://x.com/OpenRouter/status/2060142412408717518

---

13. Data Formulator 推出企业数据 AI 分析工具

X：Microsoft Research (@MSFTResearch) · 8 小时前

Data Formulator 为企业数据工作流引入了 AI 驱动的分析功能。数据团队可以轻松将企业数据带入一个 AI 就绪的工作空间，用户可以使用 AI 智能体来探索、分析和可视化数据，将原始数据转化为可操作的洞察：https：//msft.it/6013vZzUl

能力进展 新发布

https://x.com/MSFTResearch/status/2060028999229735199

---

14. Krea 2 API发布，支持多平台与智能体

X：Krea AI (@krea_ai) · 昨天 22:59

今天，我们发布了 Krea 2 的 API。 现已在 @fal 或 @ComfyUI 等平台可用，通过 @NousResearch 的 Hermes 等智能体使用，并全面支持 Claude、Codex 或 OpenClaw。 了解如何设置 ☒

能力进展 新发布

https://x.com/krea_ai/status/2059650622203515143

---

15. Grok Build 0.2.7 发布，新增多项功能

X：xAI (@xai) · 3 小时前

Grok Build 0.2.7 现已发布，包含 /usage、/login、跨子智能体共享终端，以及改进的图像理解功能。 所有更新请查看 https：//x.ai/build/changelog

能力进展 新发布

https://x.com/xai/status/2060102590122385460

---

16. Replit Canvas：智能体设计工具发布

X：Replit (@Replit) · 4 小时前

最好的设计工作不会在聊天框里发生。你需要空间来探索想法、创建变体并进行迭代。 认识新的 Replit Canvas。 你的智能体设计工具，用于构建精美的网站、应用、营销资产等。

能力进展 新发布

https://x.com/Replit/status/2060097656207413613

---

17. Gemini Omni向印度用户开放视频编辑功能

X：Gemini (@GeminiApp) · 5 小时前

好消息！印度用户现在可以上传视频（来自相册或已保存文件），并使用Gemini Omni进行编辑和转换。快来试试，并告诉我们你的想法。

能力进展 新发布

https://x.com/GeminiApp/status/2060074415304610168

---

18. Sesame，这家由Oculus创始人创办的对话式AI初创公司，发布其iOS应用

TechCrunch：AI（RSS） · 9 小时前

由Oculus创始人创办的AI初创公司Sesame发布了其iOS应用，该应用将对话式AI智能体带给公众。应用提供更自然的来回交互体验，设计上区别于传统聊天机器人，旨在让用户感觉更像在和真人对话。

能力进展 新发布

https://techcrunch.com/2026/05/28/sesame-the-conversational-ai-startup-from-oculus-founders-launches-its-ios-app

---

19. Google I/O 2026 一文速览：12 大重要时刻

Google Blog：AI（RSS） · 9 小时前

Google I/O 2026 发布会上披露了 12 个重要时刻，其中包括 Gemini Omni 和 Gemini 3.5 Flash 等产品的相关消息。

能力进展 新发布

https://blog.google/innovation-and-ai/technology/ai/io-2026-keynote-moment-videos

---

20. Web 更新

Midjourney：Updates（RSS） · 昨天 02:44

对话模式在文本和语音输入方面进行了改进。语音会话开始时，可访问用户的图像提示、风格参考、侧边栏设置和最近任务。图像提示功能现可从托盘和侧边栏直接使用。在语音提交过程中，托盘中的图像将保持不变，直至用户手动移除。

能力进展 新发布

https://updates.midjourney.com/web-updates-5

---

21. MiniMax M2.7 免费智能体编程限时开放

X：MiniMax (@MiniMax_AI) · 5 小时前

在 @OpenHandsDev 上使用 MiniMax M2.7 进行免费智能体编程？ 是的，请给我！限时提供 ⚡☒

能力进展 新发布

https://x.com/MiniMax_AI/status/2060071852970844377

---

22. OpenCode与MiMo V2.5限时免费开放

X：opencode (@opencode) · 昨天 01:59

OpenCode x MiMo V2.5 - 限时免费 1M 上下文 · 推理 · 文本 · 图像

能力进展 新发布

https://x.com/opencode/status/2059696100626297225

---

23. AI生成短片《昨夜》探索记忆碎片中的东京之夜

X：Runway (@runwayml) · 10 小时前

昨夜。一部完全由AI生成的短片，通过破碎记忆的视角，探索了在东京改变人生的一个夜晚。由一人使用Runway在一天内创作完成。 这是Project Luxo的一部分：一个探索AI生成视频如何跨越恐怖谷的新项目。 通过下方链接了解更多关于《昨夜》和Project Luxo的信息。

能力进展

https://x.com/runwayml/status/2059998751742128635

---

24. Claude Marketplace 新增五家合作伙伴

X：Claude (@claudeai) · 昨天 23:48

Claude Marketplace 新增成员：@augmentcode、@boltdotnew、@coderabbitai、@hebbia 和 @WeAreLegora。 您现有的 Anthropic 消费承诺可用于购买其 Claude 驱动的产品。 了解更多：http：//claude.com/platform/marketplace

能力进展

https://x.com/claudeai/status/2059662933924123044

---

25. Perplexity Computer现已集成微软Office套件

X：Perplexity (@perplexity_ai) · 9 小时前

Perplexity Computer现已登陆Microsoft Excel、Word、PowerPoint和Outlook。 您可以在应用程序的侧边栏中直接使用Computer来协调工作，起草文档、建模、制作演示文稿并处理电子邮件。 现已推出：https：//www.perplexity.ai/hub/products/integrations/microsoft

新发布

https://x.com/perplexity_ai/status/2060013442720010598

研究 研究与开源进展

1. hexoai开源SIA框架：AI智能体实现递归自我改进

X：Rohan Paul (@rohanpaul_ai) · 6 小时前

hexoai开源了SIA（自我改进AI）框架。该框架展示了AI智能体不仅能优化其外部工作流（harness），还能通过任务反馈直接更新自身的模型权重，从而在领域知识和能力上实现自主提升，而非仅依赖人类提供的提示或工具改进。论文报告显示，SIA在LawBench基准上性能提升56.6%，在GPU kernels运行上耗时减少91.9%，在单细胞RNA去噪任务中相比基线提升显著提升。

能力进展 基础设施 新发布

https://x.com/rohanpaul_ai/status/2060063592448446778

2. Fast， faster， Qwen. ☒

X：通义千问 / Qwen (@Alibaba_Qwen) · 昨天 00:34

Qwen3.5在TokenSpeed推理引擎上，针对智能体工作负载达到了创纪录的580 tokens per second （tps）速度。这一成果由通义千问推理团队、lightseekorg Foundation TokenSpeed团队、NVIDIA及Mooncake团队共同实现，并采用了tri_dao的FlashAttention-4 （FA4） 优化。此里程碑标志着开源大语言模型推理性能的

能力进展 基础设施 新发布

https://x.com/Alibaba_Qwen/status/2059674574397313277

3. SGLang 团队与 AMD 合作，使 AMD InstinctTM MI355X GPU 的大规模 DeepSeek-R1 分离式推理在总拥有成本上具备竞争力

LMSYS：Blog（Chatbot Arena 团队） · 8 小时前

SGLang 与 AMD 团队合作，通过一系列全栈优化，使 AMD InstinctTM MI355X GPU 在运行 DeepSeek-R1 大模型推理时实现了极具竞争力的总拥有成本。在 129 tok/s/user 的交互延迟下，其成本为每百万 token $0.169，比 NVIDIA B200（Dynamo TRT-LLM）方案低 5%，比 B200（SGLang）方案低 40%。吞吐量方

能力进展 基础设施

https://www.lmsys.org/blog/2026-05-28-mori

4. ITBench-AA：前沿大模型在首个智能体企业IT任务基准测试中得分均低于50%

Hugging Face：Blog（RSS） · 昨天 01:20

由Artificial Analysis和IBM推出的ITBench-AA SRE基准测试显示，所有前沿大模型得分均未超过50%。Claude Opus 4.7（自适应推理，最大努力）以47%领先，GPT-5.5（xhigh）和Qwen3.7 Max分别得46%和42%。该测试包含59个需要通过Shell命令调查Kubernetes事件快照并提交根因诊断的智能体任务。关键发现是模型推理轮次差异近3

能力进展 新发布

https://huggingface.co/blog/ibm-research/itbench-aa

格局 观点、资本与监管

1. 英伟达推出 AI 框架 Polar，让 Codex 跑分暴涨 显著提升

IT之家（RSS） · 22 小时前

英伟达研究团队开源了智能体强化学习框架 Polar。该框架无需重写现有智能体执行框架（如 Codex CLI、Claude Code、Qwen Code、Pi），通过在模型 API 边界放置智能体来接入 GRPO 训练。实验显示，基于 Qwen3.5-4B 模型，Polar 将 Codex 在 SWE-Bench Verified 上的 pass@1 分数从 3.8% 提升至 26.4%（增加 2

能力进展 基础设施 新发布

https://www.ithome.com/0/956/293.htm

2. 社区如何利用Tunix和TPU训练Gemma学会"思考"

Google Developers Blog（RSS） · 9 小时前

Google在Kaggle举办的Tunix黑客马拉松，挑战开发者利用TPU和有限算力，将小型基础模型转变为通用推理引擎。获胜团队通过多阶段后训练流程实现了这一目标，该流程结合了监督微调（SFT）与GRPO、SimPO等先进对齐技术。比赛结果表明，社区能够借助开源资源成功训练出高能力的结构化推理模型。

能力进展 基础设施 新发布

https://developers.googleblog.com/how-the-community-trained-gemma-to-think-with-tunix-and-tpus

3. 我认为 Anthropic 和 OpenAI 找到了产品市场契合点

Simon Willison 博客 · 昨天 00:38

Anthropic 与 OpenAI 通过编程智能体找到了产品市场契合点，这导致企业客户成本显著上升。两家公司已于 2026 年 4 月前后调整了企业套餐定价，从原先的高额折扣改为与 API 用量挂钩。Anthropic Enterprise 套餐变为每席位 20 美元/月外加 API 费用，OpenAI Codex 则按 API token 用量计费。同期发布的新模型 GPT-5.5（4月23日

能力进展 新发布

https://simonwillison.net/2026/May/27/product-market-fit

4. OpenRouter 支持模型现可选 Flex 与 Priority 服务层级

X：OpenRouter (@OpenRouter) · 10 小时前

提示：您可以为支持的模型（OpenAI、Google Vertex 等）使用 Flex 和 Priority 层级。 定价信息请查看各模型页面。文档：https：//openrouter.ai/docs/guides/features/service-tiers

能力进展 新发布

https://x.com/OpenRouter/status/2060007875590697088

5. AI智能体时代下的安全变革

Tomer Tunguz 博客（VC 分析） · 昨天 08:00

Lemonade的CISO Jonathan Jaffe探讨了AI智能体时代的安全新挑战。他指出，AI对攻击者和防御者同样强大，但可被利用的漏洞窗口正在缩小，因为AI能更快地生成、审查和修补代码。为此，安全团队正向工程团队转型，例如Lemonade的安全部门均由工程师组成，并构建了包含智能体的内部AI平台。同时，每个智能体（单个终端上可能运行200到10000个）都需要被赋予身份，并在操作点由策略

能力进展 监管/资本

https://www.tomtunguz.com/jonathan-jaffe-office-hours-post-event

6. 人民日报专访华为何庭波：今年秋季的新麒麟手机芯片，性能等相比去年是"跳跃性"提升

IT之家（RSS） · 23 小时前

华为何庭波提出半导体新演进路径"韬（τ）定律"，以"时间缩微"（如逻辑折叠）替代"几何缩微"作为新指导原则。她表示，过去6年华为已基于此自主研发381款芯片。今年秋季将发布新的麒麟手机芯片，这是首个完整的"韬芯片"，其性能、集成度相比去年是"跳跃性"提升。

基础设施 新发布

https://www.ithome.com/0/956/274.htm

7. 用好 Coding Agent，重点是两头，尤其是开头的部分，如果一开始就走偏了后面怎么改都改不好。

X：宝玉 (@dotey) · 昨天 07:09

用好 Coding Agent 的关键在于初始规划。方法是先将需求整理后，用最强模型（如 GPT-5.5、Claude Opus 4.7）分别在 Codex、Claude Code、Cursor 的 Plan 模式下生成设计方案，选择最优方案并借鉴其他版本。对于复杂计划，可将其拆分为多个 Phases 并明确要求与验证标准，形成 Markdown 文档。执行时按 Phases 进行，并辅以人工审核

能力进展

https://x.com/dotey/status/2059773942500298934

8. 四步保障AI生成应用安全

X：Replit (@Replit) · 7 小时前

如何用四步保障你的vibecoded应用安全 速度若无安全加持，便是隐患。以下是使用Replit发布应用时，如何避免留下后门的方法。 展开阅读 ↓

监管/资本 新发布

https://x.com/Replit/status/2060052289155375532

9. 与Google搜索产品副总裁Robby Stein的访谈：AI原生搜索时代

X：Kim (@kimmonismus) · 昨天 00:12

本文记录了与Google搜索产品副总裁Robby Stein在Google I/O的访谈，核心探讨Google Search向"AI原生"模式的重大转变。讨论话题包括AI Mode是进化还是重塑、如何将复杂问题拆解为多轮搜索、AI搜索的高运行成本、Google TPU及基础设施的优势、AI时代搜索量不减反增的原因，以及优质AI回答与出版商流量之间的张力。访谈还涉及Google决定展示哪些信息源与链

基础设施

https://x.com/kimmonismus/status/2059668961181004275

10. pgvector驱动的语义、混合、稀疏与量化向量搜索系统构建编码指南

MarkTechPost（RSS） · 16 小时前

本教程在Google Colab中构建一个完整的pgvector实验环境，展示PostgreSQL如何作为向量数据库服务于现代AI应用。内容涵盖安装PostgreSQL、编译pgvector扩展、通过Psycopg建立连接，并注册向量类型以实现与Python的平滑集成。最后使用SentenceTransformers创建并存储嵌入向量。

https://www.marktechpost.com/2026/05/28/a-coding-guide-to-implement-a-pgvector-powered-semantic-hybrid-sparse-and-quantized-vector-search-system