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: 61 条 焦点: 8 条 快讯: 0 条

Executive Summary

阶跃星辰发布Step 3.7 Flash开源模型，采用198B参数MoE架构支持256K上下文，专为智能体工作流效率优化，在ClawEval-1.1和SimpleVQA Search评测中均排名第一。OpenAI推出实时翻译模型，支持70多种语言输入翻译至13种输出语言，已在智能眼镜上部署运行。小米开源ControlFoley可控视频音效生成模型，统一支持文本引导、文本控制和参考音频控制三类任务，在VGGSound-Test等多个benchmark取得开源SOTA表现。Mistral AI发布Search Toolkit公共预览版，整合数据摄取、检索和评估工具于单一框架。

OpenRouter完成1.13亿美元B轮融资，由CapitalG领投，同时推出Guardrails安全治理工具套件保护智能体数据与成本控制。阿里云开源百炼CLI使Agent可调用全套模型和应用能力，ComfyUI现已支持OpenRouter模型直接调用访问20多个模型。Apple正尝试将大型Gemini模型集成到iPhone以驱动新Siri功能，可能采用云端组件配合本地处理方案。基础设施层面，模型聚合平台获得资本认可，安全治理工具成为智能体部署必需品。

后续需跟踪Step 3.7 Flash的开源生态发展和实际应用落地情况，OpenRouter在模型聚合市场的竞争策略及其安全工具的行业采纳度，以及Apple-Gemini集成的具体技术实现路径。同时关注Mistral AI搜索工具链的商业化进展和企业客户接受度，各大厂商在智能体安全治理方面的技术演进。

重点 今日核心进展

★ 1. 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) · 昨天 08:00 · 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

★ 2. OpenAI推出实时翻译模型，支持70+语言输入

X：Greg Brockman (@gdb) · 4 小时前 · 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. 小米开源可控视频音效生成模型 ControlFoley，让声音"按你想要的来"

IT之家（RSS） · 15 小时前 · 模型与工具能力

小米大模型应用团队发布开源可控视频音效生成模型 ControlFoley，旨在解决创作中的可控性难题。该模型统一支持文本引导视频配音、文本控制视频配音和参考音频控制视频配音三类任务。ControlFoley 在 VGGSound-Test 等多个 benchmark 上取得开源 SOTA 表现，其代码、模型权重和在线 Demo 均已开放。

能力进展 新发布

https://www.ithome.com/0/957/282.htm

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

X：Google AI for Developers (@googleaidevs) · 昨天 01:25 · 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

★ 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) · 昨天 23:10 · 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

★ 6. 发布 Search Toolkit

Mistral AI：News（网页） · 昨天 20:47 · 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

★ 7. OpenRouter 获得1.13亿美元B轮融资

OpenRouter：Announcements（RSS） · 昨天 22:00 · 产业与基础设施

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

★ 8. 阿里云开源百炼 CLI，Agent 可调用全套模型和应用能力

IT之家（RSS） · 17 小时前 · 应用与商业化

IT之家（RSS）披露：阿里云开源百炼 CLI，Agent 可调用全套模型和应用能力。该条属于应用与商业化方向，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 基础设施 新发布

https://www.ithome.com/0/957/149.htm

能力 模型与工具能力

1. 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) · 昨天 08:00

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

2. OpenAI推出实时翻译模型，支持70+语言输入

X：Greg Brockman (@gdb) · 4 小时前

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. 小米开源可控视频音效生成模型 ControlFoley，让声音"按你想要的来"

IT之家（RSS） · 15 小时前

小米大模型应用团队发布开源可控视频音效生成模型 ControlFoley，旨在解决创作中的可控性难题。该模型统一支持文本引导视频配音、文本控制视频配音和参考音频控制视频配音三类任务。ControlFoley 在 VGGSound-Test 等多个 benchmark 上取得开源 SOTA 表现，其代码、模型权重和在线 Demo 均已开放。

能力进展 新发布

https://www.ithome.com/0/957/282.htm

4. Nano Banana Pro与Nano Banana 2正式发布

X：Google AI for Developers (@googleaidevs) · 昨天 01:25

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) · 昨天 23:10

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

6. Qwen-VLA：从理解世界到在其中行动

Qwen：Blog Retrieval（API） · 15 小时前

Qwen Studio提供全面功能，涵盖聊天机器人、图像与视频理解、图像生成、文档处理、网络搜索集成、工具利用及Artifacts。

能力进展

https://qwen.ai/blog?id=qwenvla

产业 产业与基础设施

1. OpenRouter 获得1.13亿美元B轮融资

OpenRouter：Announcements（RSS） · 昨天 22:00

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

---

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

Ars Technica：AI（RSS） · 昨天 02:30

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

---

3. 三星电子业内率先出样 HBM4E 内存

IT之家（RSS） · 23 小时前

IT之家（RSS）披露：三星电子业内率先出样 HBM4E 内存。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://www.ithome.com/0/956/851.htm

---

4. 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） · 昨天 05:49

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

---

5. 波士顿儿童医院利用AI解锁新诊断

OpenAI：官网动态（RSS · 排除企业/客户案例） · 12 小时前

波士顿儿童医院通过部署OpenAI技术，用于改善患者护理并减轻运营负担，成功帮助诊断了超过40种罕见病病例。

新发布

https://openai.com/index/boston-childrens-hospital

---

6. 中央网信办等四部门：提升全民人工智能素养，加快人才培育、深化普及应用

IT之家（RSS） · 14 小时前

中央网信办等四部门联合印发《2026年提升全民数字素养与技能工作要点》，部署了六项重点任务。其中明确要求"提升全民人工智能素养"，具体包括强化人工智能赋能教育、加快人工智能人才培育、深化人工智能普及应用。其他任务涵盖数字资源供给、应用场景建设、普惠包容发展、安全有序网络空间以及协同联动机制。

监管/资本

https://www.ithome.com/0/957/319.htm

应用 应用与商业化

1. 发布 Search Toolkit

Mistral AI：News（网页） · 昨天 20:47

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. 阿里云开源百炼 CLI，Agent 可调用全套模型和应用能力

IT之家（RSS） · 17 小时前

IT之家（RSS）披露：阿里云开源百炼 CLI，Agent 可调用全套模型和应用能力。该条属于应用与商业化方向，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 基础设施 新发布

https://www.ithome.com/0/957/149.htm

---

3. OpenClaw 2026.5.27 版本发布

X：OpenClaw (@openclaw) · 昨天 21:10

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

---

4. Guardrails：保护你的智能体、数据与成本

OpenRouter：Announcements（RSS） · 12 小时前

Guardrails 是一套可配置的安全与治理工具，提供预算执行、零数据保留、模型与提供商限制、提示词注入防御及数据丢失预防等功能，旨在保护智能体（Agents）、数据与控制成本。

能力进展 监管/资本 新发布

https://openrouter.ai/announcements/guardrails

---

5. Codex现已支持Windows端计算机使用功能

X：OpenAI (@OpenAI) · 5 小时前

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

---

6. ComfyUI现已支持OpenRouter模型直接调用

X：OpenRouter (@OpenRouter) · 23 分钟前

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

---

7. OpenRouter支持模型生成文件补丁

X：OpenRouter (@OpenRouter) · 8 小时前

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

---

8. llm-anthropic 0.25.1

Simon Willison 博客 · 昨天 07:54

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

---

9. 别只看基准测试，要看全面表现

X：OpenRouter (@OpenRouter) · 昨天 07:33

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

---

10. Data Formulator 推出企业数据 AI 分析工具

X：Microsoft Research (@MSFTResearch) · 昨天 00:02

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

---

11. Gemini 本月更新：全新界面与智能体助手

X：Gemini (@GeminiApp) · 8 小时前

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

---

12. 用 Rosalind Biodefense 增强社会韧性

OpenAI：官网动态（RSS · 排除企业/客户案例） · 21 小时前

OpenAI 推出 Rosalind Biodefense，为通过审核的开发者和美国政府伙伴提供 GPT-Rosalind 的可信访问，以推动前沿 AI 在生物防御、公共卫生和大流行病准备方面的应用。

能力进展 新发布

https://openai.com/index/strengthening-societal-resilience-with-rosalind-biodefense

---

13. Grok Build 0.2.7 发布，新增多项功能

X：xAI (@xai) · 昨天 04:55

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

---

14. Replit Canvas：智能体设计工具发布

X：Replit (@Replit) · 昨天 04:35

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

---

15. Gemini Omni向印度用户开放视频编辑功能

X：Gemini (@GeminiApp) · 昨天 03:03

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

---

16. 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） · 昨天 23:35

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

---

17. Google I/O 2026 一文速览：12 大重要时刻

Google Blog：AI（RSS） · 昨天 23:00

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

---

18. ChatGPT对话目录功能现已上线

X：ChatGPT (@ChatGPTapp) · 3 小时前

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

---

19. MiniMax M2.7 免费智能体编程限时开放

X：MiniMax (@MiniMax_AI) · 昨天 02:52

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

---

20. Runway API持续扩展模型与端点支持

X：Runway (@runwayml) · 4 小时前

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

---

21. AI生成短片《昨夜》探索记忆碎片中的东京之夜

X：Runway (@runwayml) · 昨天 22:02

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

22. Luma Agents 自动生成宣传图，输入内容即可

X：Luma AI (@LumaLabsAI) · 3 小时前

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

23. Gemini Omni可将草图变为现实

X：Gemini (@GeminiApp) · 5 小时前

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

24. Perplexity Computer现已集成微软Office套件

X：Perplexity (@perplexity_ai) · 昨天 23:00

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

25. Codex可自主管理对话线程与并行任务

X：Greg Brockman (@gdb) · 2 小时前

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

26. Canvas新功能与Clerk自定义登录介绍

X：Replit (@Replit) · 8 小时前

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. hexoai开源SIA框架：AI智能体实现递归自我改进

X：Rohan Paul (@rohanpaul_ai) · 昨天 02:20

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. GPIC：大规模视觉生成基准数据集发布

X：Fei-Fei Li (@drfeifei, World Labs) · 7 小时前

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. 社区如何利用Tunix和TPU训练Gemma学会"思考"

Google Developers Blog（RSS） · 昨天 23:41

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

2. 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) · 1 小时前

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

---

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

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

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

---

4. 亲测为实：难以置信的推理速度

X：Rohan Paul (@rohanpaul_ai) · 7 小时前

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

---

5. Cursor 团队发布《开发者习惯报告》

X：邵猛 (@shao__meng) · 23 小时前

报告显示，AI正深刻改变开发工作形态。开发者周均代码产出从约3.6K行增至8.6K行，更大规模的PR（千行以上）占比上升。AI智能体在单次会话中的工具调用数增加约30%，正在处理更复杂的任务。同时，被接受的AI代码在60分钟后的留存率从约76%提升至约81%，表明更多AI生成内容进入了实际代码库。这些趋势共同指向AI已从个人辅助工具，演进为推动开发向更大规模任务与自动化基础设施发展的核心力量。

能力进展 新发布

https://x.com/shao__meng/status/2060167182777249886

---

6. 技能提炼

Tomer Tunguz 博客（VC 分析） · 昨天 08:00

"技能提炼"是一种知识转移方法，由前沿大模型（如 Opus 4.7、GPT-5.1、Gemini 3 Pro）负责撰写并优化标准化的 SKILL.md 流程文件。然后，本地运行的小模型（如 Qwen 35B、Gemma 26B）直接执行这些文件。此过程不同于压缩模型权重的知识蒸馏、训练权重的指令微调或检索事实的 RAG，其核心是提取并转移操作流程，让小模型按步骤执行，从而形成前沿模型作教师、小模型

能力进展 基础设施

https://www.tomtunguz.com/the-pi-agent-skill-distillation

---

7. Gemini架构师分享AI前沿探索幕后故事

X：Google AI (@GoogleAI) · 8 小时前

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

---

8. Adam's Law：用高频词写Prompt效果更好

X：Berry Xia (@berryxia) · 20 小时前

FaceMind团队用100种语言和四大核心任务实验发现，在语义不变的前提下，使用预训练语料中出现频率更高的词汇（高频表达）来撰写提示词或进行微调，可以显著提升大语言模型的表现。这被总结为Adam's Law（文本频率定律），它为数据工程补上了"频率"这一新维度。原理在于高频表达能让模型在它最熟悉的概率空间内工作，从而优化输出质量。

能力进展 基础设施

https://x.com/berryxia/status/2060212428584202428

---

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

X：OpenRouter (@OpenRouter) · 昨天 22:38

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

---

10. PyTorch 中的性能分析（第一部分）：torch.profiler 初学者指南

Hugging Face：Blog（RSS） · 昨天 08:00

该指南旨在介绍如何在 PyTorch 中使用 torch.profiler 进行性能分析。文章是系列教程的第一部分，面向初学者，讲解如何通过该工具分析模型训练与推理过程的性能瓶颈。

能力进展 基础设施

https://huggingface.co/blog/torch-profiler

---

11. Braintrust如何用Codex将客户请求转化为代码

OpenAI：官网动态（RSS · 排除企业/客户案例） · 12 小时前

Braintrust的工程师正在使用Codex结合GPT-5.5模型，以加速其实验运行与代码编写的流程。

能力进展 新发布

https://openai.com/index/braintrust

---

12. 谷歌 DeepMind CEO 哈萨比斯：AGI 最快三年内到来，研发速度远超预期

IT之家（RSS） · 17 小时前

谷歌 DeepMind 首席执行官德米斯·哈萨比斯预测，AGI 研发速度远超预期，最快可能在 2029 年至 2030 年前后出现。作为 AlphaGo、AlphaFold 的主导者，他认为当前 AI 智能体是未来更强智能的预演，随着多模态和自主决策能力成熟，三年内迎来 AGI 关键突破已非科幻。但他同时警示，全球社会对 AGI 到来的准备严重不足，必须提前建立规则与防护机制。

能力进展

https://www.ithome.com/0/957/154.htm

---

13. 特斯拉 FSD 安全性宣称遭质疑

IT之家（RSS） · 23 小时前

特斯拉声称其全自动驾驶软件（FSD）安全性最高可达人类的10倍，但路透社调查发现此数据经不起推敲。参与训练FSD的员工表示该技术远未成熟，其安全演示高度依赖人工。统计方法被11位交通安全研究人员指出存在缺陷，例如与更广泛的联邦事故数据进行不恰当比较。相比之下，竞争对手Waymo采用了更严谨的统计方法。目前，特斯拉FSD仍需驾驶员主动监督，安全部署可能还需数年。

基础设施 监管/资本

https://www.ithome.com/0/956/864.htm

---

14. 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） · 5 小时前

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

---

15. 当公司过于"AI上瘾"时会发生什么?

TechCrunch：AI（RSS） · 6 小时前

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

---

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

TechCrunch：AI（RSS） · 8 小时前

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

---

17. Claude Code--文档中未提及的所有可配置选项

Hacker News 热门（buzzing.cc 中文翻译） · 13 小时前

该篇文章标题涉及"Claude Code"的可配置选项，但提供的正文内容仅包含一张图片和一个外部链接，未给出任何关于模型版本、参数、性能、价格或功能的具体信息。根据规则，无法在摘要中提及原文不存在的细节。

能力进展

https://buildingbetter.tech/p/i-read-the-claude-code-source-code

---

18. 15秒动画IP预告片制作全流程分享

X：PixVerse (@PixVerse_) · 22 小时前

我们制作了一个15秒的动画IP预告片--从角色设定到最终视频。认识一下MILO和BUMBLE：地下邮政骑手 转发+关注+回复=我们会私信发送工作流程和提示词

能力进展

https://x.com/PixVerse_/status/2060184048920826340

---

19. 可信第三方评估的共享操作手册

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

OpenAI 发布了一份关于第三方 AI 评估的指导框架，内容涵盖如何评估前沿系统的能力、安全防护措施及评估本身的有效性。

监管/资本 新发布

https://openai.com/index/trustworthy-third-party-evaluations-foundations

---

20. 四步保障AI生成应用安全

X：Replit (@Replit) · 昨天 01:35

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

---

21. Kling AI助力电影RAPHAEL创作全流程揭秘

X：可灵 Kling AI (@Kling_ai) · 9 小时前

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
