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: 2s 精选条目: 57 条 焦点: 8 条 快讯: 0 条

Executive Summary

今日核心进展：商汤科技发布SenseNova U1技术报告并开源基于MoE架构的模型权重，推动AI透明度与可复现性发展。蚂蚁包容性AI团队发布ARGenSeg-8B和Ring-2.6-1T两款新模型，其中Ring-2.6-1T在Claw-Eval基准中表现优异。IBM Granite团队推出Granite Embedding Multilingual R2多语言嵌入模型，具备32K上下文长度和领先检索性能。Kimi K2.6登顶金融智能体基准榜首，显示特定领域智能体能力持续提升。

行业结构变化：微软对OpenAI累计投入超1000亿美元，其中包含130亿美元原始投资及大量Azure基础设施成本，已为微软带来约300亿美元营收，体现了大型科技公司深度参与AI生态的战略布局。百度发布面向大规模智能体应用的全栈AI云，基于自研昆仑芯构建专用集群，支持ERNIE 5.1系列模型。普华永道全球部署Claude，将在数十万员工中推广AI工具，标志着企业级AI应用进入规模化落地阶段。

后续观察：需要跟踪SenseNova U1开源模型的实际应用效果和社区反馈，以及商汤在MoE架构上的技术迭代节奏。微软与OpenAI续签非独家协议后双方合作模式的变化值得关注。企业级AI应用的商业化进程，特别是普华永道等专业服务机构的大规模部署案例将为行业提供重要参考。算力基础设施需求增长对云厂商定价策略和供给能力的影响，以及开源模型在企业场景中的采纳率将是关键变量。

重点 今日核心进展

★ 1. SenseNova U1技术报告发布，基于MoE架构开放模型权重

X：商汤 SenseTime (@SenseTime_AI) · 21 小时前 · 模型与工具能力

由联合创始人兼首席科学家李沐博士领导的团队发布了SenseNova U1技术报告，详细阐述了其架构、训练方案与创新突破。此次开源同步发布了基于混合专家模型（MoE）的新权重，旨在推动AI领域的透明度、可复现性与进一步创新。团队希望通过开放共享促进整个社区的技术发展。

能力进展 基础设施 新发布

https://x.com/SenseTime_AI/status/2054876860711149780

★ 2. inclusionAI/ARGenSeg-8B

蚂蚁 inclusionAI：HuggingFace 新模型 · 5 小时前 · 模型与工具能力

包容性AI团队发布ARGenSeg-8B模型，致力于通过开源和开放科学推动人工智能的进步与普及。该举措强调技术民主化，使更广泛的社区能够参与AI研发与应用。开源策略将促进协作创新，加速AI工具在多元场景中的落地，降低技术门槛，推动产业生态的开放发展。

能力进展 新发布

https://huggingface.co/inclusionAI/ARGenSeg-8B

★ 3. Granite Embedding Multilingual R2：开源多语言嵌入模型，具备32K上下文与领先检索性能

Hugging Face：Blog（RSS） · 13 小时前 · 模型与工具能力

IBM Granite团队在Hugging Face发布了Granite Embedding Multilingual R2多语言文本嵌入模型。该模型采用Apache 2.0开源协议，上下文长度扩展至32K令牌，参数量低于1亿。在MTEB基准的检索评估中，它取得了同规模模型的最佳性能，尤其擅长多语言混合检索，为开发者提供了高效、轻量且可商用的嵌入解决方案。

能力进展 新发布

https://huggingface.co/blog/ibm-granite/granite-embedding-multilingual-r2

★ 4. inclusionAI/Ring-2.6-1T

蚂蚁 inclusionAI：HuggingFace 新模型 · 23 小时前 · 模型与工具能力

inclusionAI发布了Ring-2.6-1T模型。该模型在Claw-Eval基准测试中取得了两项关键评估结果：在通用能力评估（General）上得分为58.4，在多轮对话评估（Multi Turn）上得分为86.8。这些分数已记录在相应的评估结果排行榜中。

能力进展 新发布

https://huggingface.co/inclusionAI/Ring-2.6-1T

★ 5. 为 OpenAI 累计投入超 1000 亿美元，纳德拉称微软当年投资时"没人愿意下注"

IT之家（RSS） · 5 小时前 · 产业与基础设施

在"马斯克诉奥尔特曼"庭审中，微软企业发展负责人确认，微软对OpenAI的累计投入已超过1000亿美元，其中包括130亿美元原始投资及大量Azure基础设施成本。此次合作已为微软带来约300亿美元营收。CEO纳德拉表示，微软是在"没人愿意下注"时承担了风险。双方已续签非独家协议，微软不再支付收入分成，并将OpenAI的分成上限设为到2030年累计380亿美元，此举较原协议节省约970亿美元。此外，

能力进展 监管/资本 新发布

https://www.ithome.com/0/950/728.htm

★ 6. Kimi K2.6登顶金融智能体基准榜首

X：Kimi.ai (@Kimi_Moonshot) · 昨天 13:57 · 模型与工具能力

Kimi K2.6 现已成为 Finance Agent Benchmark V2 开源权重排名第一。 【引用 @ValsAI】： AI 能胜任金融分析师的工作吗?

能力进展 新发布

https://x.com/Kimi_Moonshot/status/2054803169994272819

★ 7. 百度发布面向大规模智能体应用的全栈AI云

X：百度 Baidu (@Baidu_Inc) · 17 小时前 · 应用与商业化

随着智能体应用迈向更大规模部署，其背后的云技术栈也需同步扩展。 在百度Create大会上，百度集团执行副总裁、百度智能云事业群总裁沈抖宣布推出专为大规模智能体应用打造的全新全栈AI云，其升级涵盖智能体基础设施与AI基础设施。 基于我们自研的昆仑芯AI芯片构建的专用集群，已支持ERNIE 5.1系列中一个关键模型的训练。

能力进展 基础设施 新发布

https://x.com/Baidu_Inc/status/2054927298403688927

★ 8. OpenCode与Qwen 3.6 Plus再度免费开放

X：opencode (@opencode) · 8 小时前 · 应用与商业化

OpenCode x Qwen 3.6 Plus - 再次免费 上次各位把我们的容量当成了自助餐。 我们找到了更多GPU。第二轮。

能力进展 基础设施 新发布

https://x.com/opencode/status/2055068702538612784

能力 模型与工具能力

1. SenseNova U1技术报告发布，基于MoE架构开放模型权重

X：商汤 SenseTime (@SenseTime_AI) · 21 小时前

由联合创始人兼首席科学家李沐博士领导的团队发布了SenseNova U1技术报告，详细阐述了其架构、训练方案与创新突破。此次开源同步发布了基于混合专家模型（MoE）的新权重，旨在推动AI领域的透明度、可复现性与进一步创新。团队希望通过开放共享促进整个社区的技术发展。

能力进展 基础设施 新发布

https://x.com/SenseTime_AI/status/2054876860711149780

2. inclusionAI/ARGenSeg-8B

蚂蚁 inclusionAI：HuggingFace 新模型 · 5 小时前

包容性AI团队发布ARGenSeg-8B模型，致力于通过开源和开放科学推动人工智能的进步与普及。该举措强调技术民主化，使更广泛的社区能够参与AI研发与应用。开源策略将促进协作创新，加速AI工具在多元场景中的落地，降低技术门槛，推动产业生态的开放发展。

能力进展 新发布

https://huggingface.co/inclusionAI/ARGenSeg-8B

3. Granite Embedding Multilingual R2：开源多语言嵌入模型，具备32K上下文与领先检索性能

Hugging Face：Blog（RSS） · 13 小时前

IBM Granite团队在Hugging Face发布了Granite Embedding Multilingual R2多语言文本嵌入模型。该模型采用Apache 2.0开源协议，上下文长度扩展至32K令牌，参数量低于1亿。在MTEB基准的检索评估中，它取得了同规模模型的最佳性能，尤其擅长多语言混合检索，为开发者提供了高效、轻量且可商用的嵌入解决方案。

能力进展 新发布

https://huggingface.co/blog/ibm-granite/granite-embedding-multilingual-r2

4. inclusionAI/Ring-2.6-1T

蚂蚁 inclusionAI：HuggingFace 新模型 · 23 小时前

inclusionAI发布了Ring-2.6-1T模型。该模型在Claw-Eval基准测试中取得了两项关键评估结果：在通用能力评估（General）上得分为58.4，在多轮对话评估（Multi Turn）上得分为86.8。这些分数已记录在相应的评估结果排行榜中。

能力进展 新发布

https://huggingface.co/inclusionAI/Ring-2.6-1T

5. Kimi K2.6登顶金融智能体基准榜首

X：Kimi.ai (@Kimi_Moonshot) · 昨天 13:57

Kimi K2.6 现已成为 Finance Agent Benchmark V2 开源权重排名第一。 【引用 @ValsAI】： AI 能胜任金融分析师的工作吗?

能力进展 新发布

https://x.com/Kimi_Moonshot/status/2054803169994272819

产业 产业与基础设施

1. 为 OpenAI 累计投入超 1000 亿美元，纳德拉称微软当年投资时"没人愿意下注"

IT之家（RSS） · 5 小时前

在"马斯克诉奥尔特曼"庭审中，微软企业发展负责人确认，微软对OpenAI的累计投入已超过1000亿美元，其中包括130亿美元原始投资及大量Azure基础设施成本。此次合作已为微软带来约300亿美元营收。CEO纳德拉表示，微软是在"没人愿意下注"时承担了风险。双方已续签非独家协议，微软不再支付收入分成，并将OpenAI的分成上限设为到2030年累计380亿美元，此举较原协议节省约970亿美元。此外，

能力进展 监管/资本 新发布

https://www.ithome.com/0/950/728.htm

2. Anthropic与盖茨基金会达成2亿美元合作，聚焦全球健康与教育

Anthropic：Newsroom（网页） · 16 小时前

Anthropic与盖茨基金会建立为期四年、总额2亿美元的合作，通过资金、Claude使用额度及技术支持，共同推进全球健康、生命科学、教育及经济流动项目。合作重点包括改善中低收入国家约46亿人口的基本医疗，利用AI加速疫苗与疗法研发，并开发公共卫生数据集等公共产品。在教育领域，双方将为美国、撒哈拉以南非洲和印度的K-12学生开发AI教学工具。经济流动方面则关注提升小农户生产力及美国职业技能认证。首

能力进展 新发布

https://www.anthropic.com/news/gates-foundation-partnership

3. OpenAI遭集体诉讼，被指通过追踪代码向Meta等泄露用户查询隐私

X：阿易 AI Notes (@AYi_AInotes) · 22 小时前

南加州联邦法院已受理针对OpenAI的集体诉讼，指控其在ChatGPT网站中嵌入Facebook Pixel等代码，侵犯用户隐私。当用户提交查询时，查询主题会作为浏览器标题与含Facebook唯一ID的cookies一并实时发送给Meta。OpenAI虽称仅分享"有限标识符"用于广告，但原告认为查询主题本身即高度敏感的个人信息。此案揭示免费AI服务的潜在代价：用户每一次查询及数字身份可能成为被交易

能力进展 新发布

https://x.com/AYi_AInotes/status/2054856518185439662

4. 百度推进智能体布局，以日活为关键指标

X：百度 Baidu (@Baidu_Inc) · 昨天 14:48

百度推进智能体组合以拥抱智能体时代，主张将日活跃智能体作为关键指标 https：//www.prnewswire.com/news-releases/baidu-advances-agent-portfolio-to-embrace-the-agent-era-champions-daily-active-agents-as-key-metric-302771383.html

能力进展 新发布

https://x.com/Baidu_Inc/status/2054815977477738549

5. 普华永道全球部署Claude，助力客户构建技术、执行交易并重塑企业职能

Anthropic：Newsroom（网页） · 6 小时前

普华永道与Anthropic宣布扩大战略联盟，将在全球数十万员工中部署Claude AI工具。双方将联合建立卓越中心，并培训认证3万名专业人员。合作聚焦三大高杠杆领域：智能体技术构建、AI原生交易执行以及企业职能重塑。普华永道已率先成立基于Claude的财务业务组。实际应用显示，Claude在保险承保、网络安全等多个领域能将交付时间缩短最高达70%，例如将保险承保周期从十周压缩至十天。

能力进展 监管/资本

https://www.anthropic.com/news/pwc-expanded-partnership

6. Anthropic的Mythos AI在五天内协助发现并利用两个未知macOS内核漏洞

X：Rohan Paul (@rohanpaul_ai) · 8 小时前

据《华尔街日报》报道，Anthropic的Mythos AI工具在短短五天内，成功帮助研究人员发现了两个此前未知的macOS内核漏洞，并将其串联成一个完整的权限提升攻击链。该攻击针对操作系统最底层的核心，通过组合多个漏洞和技术，绕过了苹果的内存完整性保护机制，访问了本应受保护的系统区域。这凸显出现代macOS的防御重点已从单纯防止漏洞发现，转向增加漏洞利用难度。Mythos在此类研究中展现出强大能

能力进展

https://x.com/rohanpaul_ai/status/2055071152511594832

7. AI 热潮引发民怨：七成美国民众反对家门口建数据中心

IT之家（RSS） · 19 小时前

盖洛普调查显示，高达七成美国民众反对在住宅附近建设数据中心，反对率较去年大幅上升，抵触情绪甚至超过对核电站的接受度。全美已有69个辖区出台暂停令，多地项目因抗议和监管纠纷延期。数据中心建设推高批发电价，导致用电成本激增，并引发空气污染、水资源消耗等担忧。尽管白宫要求AI企业承担配套成本，但仅为无约束力承诺，未来审批将更严苛。

基础设施 监管/资本

https://www.ithome.com/0/950/598.htm

8. MiMo V2.5 Pro 获设计竞技场季军

X：小米 MiMo (@XiaomiMiMo) · 昨天 13:33

MiMo V2.5 Pro 在 @DesignArena 上刚刚获得第三名！MiMo V2.5 Pro （Thinking） 在总排行榜上比 MiMo-V2.5 提升了 8 个名次，在前端编码任务中达到与 Claude Sonnet 4.6 相同的性能水平。 衷心祝贺 @XiaomiMiMo 团队取得这些进步！

能力进展

https://x.com/XiaomiMiMo/status/2054797221250683199

9. OpenEvidence覆盖65%美国医生，shadow AI模式引关注

X：小北 (@frxiaobei) · 14 小时前

OpenEvidence已覆盖65%的美国医生，4月单月临床场景使用达2700万次，平均每位医生每月使用41次。平台由医生个人通过执业编号在手机上注册，医院最初不知情，Mount Sinai的AI负责人称此为shadow AI，表示其早在基层普及。医院后来才追签企业合作，OpenEvidence强调这是美国医疗史上首次让大多数医生自愿采用单一技术平台的突破。合作伙伴包括NEJM、JAMA、NCCN

新发布

https://x.com/frxiaobei/status/2054981573150449754

10. Runway正式进军日本市场，在东京设立总部并投入4000万美元

Runway：News（网页） · 7 小时前

生成式AI公司Runway宣布在日本东京设立总部，正式进军日本市场，并计划投入4000万美元初始资金拓展业务。日本已成为Runway增长最快的市场之一，是其全球企业及自助客户的第三大市场。过去一年，日本企业客户数量增长300%，贡献了Runway亚洲总销售额的三分之一。软银、雅马哈等企业已在营销与创意流程中使用其服务。公司此次扩张旨在贴近日本领先的创意、机器人及制造行业客户，并已开始招募日本市场负

https://runwayml.com/news/runway-is-coming-to-japan

应用 应用与商业化

1. 百度发布面向大规模智能体应用的全栈AI云

X：百度 Baidu (@Baidu_Inc) · 17 小时前

随着智能体应用迈向更大规模部署，其背后的云技术栈也需同步扩展。 在百度Create大会上，百度集团执行副总裁、百度智能云事业群总裁沈抖宣布推出专为大规模智能体应用打造的全新全栈AI云，其升级涵盖智能体基础设施与AI基础设施。 基于我们自研的昆仑芯AI芯片构建的专用集群，已支持ERNIE 5.1系列中一个关键模型的训练。

能力进展 基础设施 新发布

https://x.com/Baidu_Inc/status/2054927298403688927

2. OpenCode与Qwen 3.6 Plus再度免费开放

X：opencode (@opencode) · 8 小时前

OpenCode x Qwen 3.6 Plus - 再次免费 上次各位把我们的容量当成了自助餐。 我们找到了更多GPU。第二轮。

能力进展 基础设施 新发布

https://x.com/opencode/status/2055068702538612784

3. Claude 工具 v2.1.141 版本更新

Claude Code：GitHub Releases（RSS） · 昨天 07:19

Claude 工具发布 v2.1.141 版本，带来多项功能新增与优化。主要更新包括：为钩子输出添加 `terminalSequence` 字段以支持无控制终端的桌面通知；新增 `CLAUDE_CODE_PLUGIN_PREFER_HTTPS` 环境变量，便于通过 HTTPS 克隆插件源码；引入 `ANTHROPIC_WORKSPACE_ID` 变量以在多工作区联盟中限定令牌范围。会话管理方面，`

能力进展 新发布

https://github.com/anthropics/claude-code/releases/tag/v2.1.141

4. 开源3D生成工具包：单张图片快速构建可交互3D世界

X：Berry Xia (@berryxia) · 7 小时前

开发者@neilsonks开源了一套专为Claude Code设计的完整3D生成工具包。该工具能将输入的单张图片自动拆解，生成包含环境、网格、物理、灯光和音频的全套可交互3D场景。其流程首先利用图像与3D生成技术提取物体并生成高质量网格，随后移除物体以得到静态背景，最后为整个场景添加物理模拟、实时灯光和环境音效。配套查看器支持对生成物体的点击编辑与一键导出。此工具将以往需数天的2D转3D工作流程缩

能力进展 新发布

https://x.com/berryxia/status/2055086814009180316

5. Claude 代理工具 v2.1.142 版本更新

Claude Code：GitHub Releases（RSS） · 9 小时前

Claude 代理工具发布 v2.1.142 版本。本次更新新增了 `--add-dir`、`--settings`、`--model` 等 8 个用于配置后台会话的命令行标志，并将 Fast 模式的默认模型升级为 Opus 4.7。插件功能得到增强，拥有根目录 `SKILL.md` 的插件现可被识别为技能，插件详情面板会显示其提供的 LSP 服务器。此外，版本修复了超过 15 项问题，包括 MC

能力进展 新发布

https://github.com/anthropics/claude-code/releases/tag/v2.1.142

6. Genkit 推出中间件系统：增强智能体AI应用的可控性与可靠性

Google Developers Blog（RSS） · 14 小时前

Google开源框架Genkit近日推出其核心中间件系统，旨在提升智能体AI应用的可靠性与可控性。该系统允许开发者在生成调用、模型及工具层进行拦截，以注入自定义行为，如重试机制、模型回退以及人工介入的工具审批流程。通过创建并堆叠自定义中间件，开发者能够实现对模型输出的确定性控制。所有中间件的执行流程均可通过专用的开发者界面进行实时查看与调试，有效支持使用TypeScript、Go、Dart和Pyt

能力进展 新发布

https://developers.googleblog.com/announcing-genkit-middleware-intercept-extend-and-harden-your-agentic-apps

7. Recraft AI V4.1上线OpenRouter平台

X：OpenRouter (@OpenRouter) · 15 小时前

现已在 OpenRouter 上线：@recraftai V4.1！ 包含六款新图像生成模型：追求高美学的 V4.1 和 V4.1 Pro，用于 SVG 插画的 V4.1 Vector 和 V4.1 Pro Vector，以及优先考虑克制风格的产品图像的 V4.1 Utility 和 V4.1 Utility Pro。 照片级真实感更自然，渐变更平滑，简短提示词能更准确地命中目标，无需过多手动

能力进展 新发布

https://x.com/OpenRouter/status/2054957185982177504

---

8. xAI 推出 Grok Build 早期测试版

xAI：News（网页） · 昨天 08:00

xAI 面向 SuperGrok Heavy 订阅用户推出 Grok Build 早期测试版。这是一个直接在终端运行的新型编程智能体与命令行工具，专为专业软件工程和复杂任务设计。其核心功能包括：支持"计划模式"，允许用户在代码执行前审阅和修改详细步骤；能无缝集成现有开发工具链；可将大型任务分解，交由并行运行的专用子智能体处理。此外，该工具提供无头模式，便于脚本和自动化流程集成。用户可通过单行命令安

能力进展 新发布

https://x.ai/news/grok-build-cli

---

9. Codex推出自动化钩子与程序化令牌

X：OpenAI Developers (@OpenAIDevs) · 10 小时前

Codex正变得更易于围绕用户代码实现自动化与定制。其核心更新包括"钩子"功能，允许在任务关键节点运行脚本，以进行工作验证、扫描密钥、记录对话或按仓库定制行为。同时，面向商业和企业团队推出"程序化访问令牌"，提供范围化凭证，可从ChatGPT工作区设置创建，用于CI/CD、发布流程和内部自动化，支持设置过期或撤销，并将使用情况关联回工作区。

能力进展 新发布

https://x.com/OpenAIDevs/status/2055032115964870838

---

10. 开源工具html-anything助力Agent生成高质量HTML

X：小互 (@xiaohu) · 18 小时前

用户分享了对开源项目html-anything的积极体验。该项目旨在让AI Agent能将任何数据转换为具有世界级设计水准的HTML代码。该项目历时三天开发，包含约一万五千行代码，支持75套Skills和9种导出格式，并能兼容包括Claude Code、Codex、OpenClaw、Hermes在内的多种代码生成Agent。

能力进展 新发布

https://x.com/xiaohu/status/2054925632061231431

---

11. Kimi推出网页桥接扩展 支持多平台交互

X：Kimi.ai (@Kimi_Moonshot) · 18 小时前

认识Kimi网页桥接--Kimi的浏览器扩展。 现在智能体可以像人类一样与网站互动：搜索、滚动、点击、输入并完成任务。 支持Kimi Code CLI、Claude Code、Cursor、Codex、Hermes等平台。 现已在http：//kimi.com/features/webbridge和Chrome应用商店上线。

能力进展 新发布

https://x.com/Kimi_Moonshot/status/2054918374837322140

---

12. 随时随地使用 Codex

OpenAI：官网动态（RSS · 排除企业/客户案例） · 19 小时前

用户现可通过 ChatGPT 移动应用随时随地使用 Codex。该功能支持跨设备和远程环境实时监控、引导及批准编码任务，实现了对编程工作的无缝移动端管理。

能力进展 新发布

https://openai.com/index/work-with-codex-from-anywhere

---

13. Luma Agents高效生成电商素材全流程

X：Luma AI (@LumaLabsAI) · 9 小时前

更多产品。更多市场。更多格式。再无瓶颈。 定义需求。设定风格。Luma Agents 从此处理所有电商活动素材。 立即扩展 → http：//lumalabs.ai/app

能力进展

https://x.com/LumaLabsAI/status/2055046873740984429

---

14. Mixpanel集成Replit MCP，开发流程内嵌数据分析

X：Replit (@Replit) · 11 小时前

发布产品。衡量效果。全在一个流程中完成。 @Mixpanel 现已登陆 Replit MCP。下周伦敦黑客松现场演示

新发布

https://x.com/Replit/status/2055018223431454850

---

15. SuperGrok Heavy限时六折，Grok Build开放测试

X：cb_doge (@cb_doge) · 11 小时前

SuperGrok Heavy 现提供约67%的半年折扣，即每月仅需99美元（原价300美元）。 强烈建议升级至Heavy版本，并试用Grok Build的测试版。

新发布

https://x.com/cb_doge/status/2055017857352913319

---

16. Suno应用更新亮点

X：Suno (@suno) · 16 小时前

Suno应用刚刚焕然一新。 过去几周我们进行了一些更新。 以下是我们喜爱的部分亮点 ☒

新发布

https://x.com/suno/status/2054948668625559902

17. 计算机直连Snowflake实现实时数据洞察

X：Perplexity (@perplexity_ai) · 16 小时前

计算机现已连接至Snowflake。 可基于实时仓库数据开展端到端工作，通过SQL、源表、筛选器和指标获取答案。 这就像一支随时待命的个人数据科学团队，从公司实时数据中提供精准答案。

https://x.com/perplexity_ai/status/2054941905633612059

18. 一键生成F1进站时刻肖像特效

X：PixVerse (@PixVerse_) · 17 小时前

不容错过PitCrewMoment潮流。 一键将任何肖像转化为F1直播电视进站时刻。 立即在PixVerse网页端尝试！

https://x.com/PixVerse_/status/2054939801435177238

研究 研究与开源进展

1. AI自主研究实现突破：智能体在nanoGPT优化赛道上超越人类基准

X：Berry Xia (@berryxia) · 8 小时前

Prime Intellect 近期在AI研究自动化领域取得重要进展。他们让Claude Code与Codex智能体完全自主运行于nanoGPT速度挑战的优化器赛道，利用闲置算力完成了近万次实验，消耗约1.4万H200小时。最终，Claude Code将记录提升至2930步，超越了2990步的人类基准。实验显示，智能体在系统整合社区主流优化方法、进行超参数扫描和策略组合方面效率极高，但在要求真正创

能力进展 基础设施 新发布

https://x.com/berryxia/status/2055074608261578949

2. NousResearch推出Token Superposition Training技术，显著加速大语言模型预训练

X：硅基流动 SiliconFlow (@SiliconFlowAI) · 昨天 10:48

NousResearch发布了Token Superposition Training（TST），这是一种改进标准大语言模型预训练流程的方法。该技术无需改变模型架构、优化器、分词器或训练数据，即可在相同计算量（FLOPs）下实现2-3倍的训练时间加速。其核心是在训练的前三分之一阶段，让模型读取并预测连续的token包，对输入嵌入进行平均，并使用改进的交叉熵损失预测下一个token包；剩余训练时间则

能力进展 基础设施 新发布

https://x.com/SiliconFlowAI/status/2054755824309076359

3. MemLens：大型视觉语言模型多模态长时记忆基准测试

HuggingFace Daily Papers（社区热门论文） · 昨天 08:00

研究团队推出MEMLENS基准，系统评估大型视觉语言模型在多模态多轮对话中的长时记忆能力。该基准包含789个问题，涵盖五大记忆能力，并在四种标准上下文长度下测试。评估27个长上下文模型和7个记忆增强代理后发现：长上下文模型在短对话中表现良好但随对话延长性能下降；记忆代理长度稳定性好但损失视觉保真度。多轮推理任务将多数系统性能限制在30%以下，表明需结合长上下文注意力与结构化多模态检索的混合架构。

能力进展 新发布

https://arxiv.org/abs/2605.14906

4. 迈向自我进化的智能文献检索系统

HuggingFace Daily Papers（社区热门论文） · 昨天 08:00

针对传统检索无法理解复杂意图、而前沿大语言模型成本高且存在幻觉的问题，研究团队提出了自我进化的智能文献检索系统PaSaMaster。该系统通过迭代式意图分析、检索与排序，将文献检索转变为动态演进的过程，并采用三项关键设计：利用排序证据揭示信息缺口以优化搜索；将检索定义为意图-论文相关性排序任务，从根本上杜绝虚假文献；通过分离规划与检索来提升效率，仅用大模型理解意图，而将大规模检索与评分交由轻量模型

能力进展

https://arxiv.org/abs/2605.14306

5. 教视觉-语言模型说"电影语言"

CMU：Machine Learning Blog · 昨天 11:06

研究团队与百余名专业创作者历时一年，构建了一个视频描述生成流程，其核心在于扩展精细化的人类-AI协同监督，而非单纯扩大模型规模。该研究（入选CVPR 2026亮点论文）指出，当前主流视频生成模型在理解和生成具有电影感的专业运镜（如希区柯克式滑动变焦、精确的焦点转移或荷兰角镜头）时存在明显不足，常产出通用或焦点错误的画面。这项工作揭示了一条通过提升监督质量来增强模型"电影语言"表达能力的新路径。

能力进展

https://blog.ml.cmu.edu/2026/05/13/teaching-vision-language-models-to-speak-cinema

格局 观点、资本与监管

1. "让 Token 消耗降低 61%"：腾讯开源 Agent Memory

IT之家（RSS） · 昨天 15:12

腾讯云开源了TencentDB Agent Memory，旨在解决Agent长任务中上下文窗口易满、Token成本高的问题。该方案采用"上下文卸载"与"Mermaid任务画布"两项核心技术，将完整信息卸载至外部存储，同时用结构化任务图保留关键状态与执行路径。实验显示，该方案在多任务连续会话中最高可降低61%的Token消耗，并提升任务成功率。项目已适配OpenClaw等主流框架，支持一键集成与本地

能力进展 基础设施 新发布

https://www.ithome.com/0/950/415.htm

2. UnslothAI发布Qwen3.6 MTP GGUF模型，实现推理速度大幅提升

X：Berry Xia (@berryxia) · 昨天 10:24

UnslothAI创始人Daniel Han发布了实验性的Qwen3.6 MTP GGUF模型，显著提升了推理速度。其中，27B模型在单GPU上达到每秒140个token，35B-A3B版本更是高达每秒220个token，相比原版GGUF速度提升超过1.4倍且精度无损。关键优化在于将draft tokens设置为2，这是性能与接受率的最佳平衡点。这项MTP投机解码技术极大提升了消费级显卡运行大模型

能力进展 基础设施 新发布

https://x.com/berryxia/status/2054749585520890314

3. 克劳德代码与《代码书》技能：有针对性的技能培养

Hacker News 热门（buzzing.cc 中文翻译） · 18 小时前

开发者发布了一款名为"克劳德代码与《代码书》技能"的GitHub工具，旨在通过刻意练习提升编程技能。该工具利用AI模型生成特定主题的代码示例与解释，帮助用户进行针对性学习。项目在Hacker News上获得104点热度，关注度较高。其核心变化在于将传统的广泛学习转化为聚焦、可重复的技能训练模式，通过结构化练习提升学习效率。

能力进展 基础设施 新发布

https://github.com/DrCatHicks/learning-opportunities

4. 创始人手册：构建AI原生初创公司

Claude：Blog（网页） · 14 小时前

Anthropic公司发布了一份面向AI原生初创企业的实用指南，旨在重塑2026年创业生命周期的构思、最小可行产品、发布和规模化四个核心阶段。该手册为每个阶段提供了具体目标、退出标准、常见失败模式及AI驱动练习，涵盖如何利用Claude进行问题验证与客户发现、避免AI生成代码的技术债务、区分真实产品市场契合度与早期炒作，并引入智能工作流替代创始人手动操作。指南还整合了多家初创企业的实践案例，为从零

能力进展 监管/资本 新发布

https://claude.com/blog/the-founders-playbook

5. 开源项目OpenSquilla：智能路由与本地检索，大幅降低LLM使用成本

X：Vista (@vista8) · 昨天 10:55

开源项目OpenSquilla针对大语言模型应用Token消耗过高的问题，提出了智能模型路由与本地向量检索相结合的解决方案。系统能自动判断任务复杂度，将简单问题路由至廉价模型，复杂任务则分配给更强模型，且路由决策在本地完成，不消耗Token。通过增量发送与缓存命中机制，实际传输Token减少了90%以上。其记忆系统能在上下文将满时自动筛选并压缩关键信息，支持混合检索。项目还具备成本统计、安全沙箱、

能力进展 监管/资本 新发布

https://x.com/vista8/status/2054757474100760626

6. BestBlogs早报：AI智能体工程化实战与安全架构

X：洪明 (@hongming731) · 昨天 07:15

BestBlogs早报聚焦AI智能体的工程化落地。Anthropic官方指南详解Claude Computer Use最佳实践，包括解决点击偏移的根本原因、推荐分辨率策略及必须采用虚拟机隔离与人工确认门控的安全原则。OpenAI工程师分享了为Codex构建Windows安全沙箱的历程，其最终方案通过专属安全标识符和写受限令牌，实现了操作系统层面的强制文件系统隔离。早报同时指出，基准测试优异的RAG

能力进展 监管/资本 新发布

https://x.com/hongming731/status/2054701978924859865

7. 在 Windows 上构建安全有效的沙箱以启用 Codex

OpenAI：官网动态（RSS · 排除企业/客户案例） · 8 小时前

OpenAI 为 Windows 平台上的 Codex 构建了一个安全沙箱环境。该沙箱通过严格控制文件访问权限和实施网络限制，确保了代码生成与执行过程的安全性。这一举措使得基于 Codex 的编码助手能够以高效且受控的方式运行，在提供强大编程辅助功能的同时，有效隔离了潜在风险，保障了用户系统的安全。

能力进展 监管/资本 新发布

https://openai.com/index/building-codex-windows-sandbox

8. 加速设备端AI：Arm与Google AI Edge的优化实践

Google Developers Blog（RSS） · 15 小时前

Arm第二代可扩展矩阵扩展（SME2）与Google AI Edge软件栈集成，将CPU转变为强大的矩阵计算加速器，从而实现高性能的设备端生成式AI。本文以Stability AI的"stable-audio-open-small"模型为例，阐述了利用LiteRT、XNNPACK和KleidiAI构建的"转换、优化、部署"自动化硬件加速流程。该方案在基于Arm架构的移动设备和笔记本电脑上，成功实现

能力进展 新发布

https://developers.googleblog.com/accelerating-on-device-ai-a-look-at-arm-and-google-ai-edge-optimization

9. 牛津大学博士后开源视频翻译工具Violin，支持多语言翻译与视频对话

X：Berry Xia (@berryxia) · 6 小时前

牛津大学博士后Kevin Lin开源视频翻译工具Violin，旨在打破高质量视频内容的语言壁垒。该工具将语音识别、大语言模型翻译与语音合成整合为自动化流水线，支持多语言互译与个性化翻译风格调整，例如将学术报告转化为儿童易懂版本。用户还能直接与视频内容进行对话并获取相关答案。Violin提供Web应用、命令行界面和Agent Skill三种使用方式，所有功能基于MIT协议开源，由Together C

能力进展 新发布

https://x.com/berryxia/status/2055093068085547507

---

10. AI电子邮件的成本分析

Tomer Tunguz 博客（VC 分析） · 昨天 08:00

使用顶尖AI模型处理邮件的月度成本约为22至130美元，中位数26美元。若软件公司以75%毛利率定价，年费可能高达350美元，加上托管服务后标价或达500美元，约为Google企业邮箱费用的两倍。采用小型模型可降低成本10至20倍，而通过本地运行利用用户GPU，更能将成本削减至接近零。结合基础启发式方法和技术优化，总成本有望降低100倍。这种针对不同工作负载匹配模型并进行成本分层的推理市场细分，将

能力进展 基础设施

https://www.tomtunguz.com/cost-of-ai-email

---

11. 解锁连续批处理中的异步性

Hugging Face：Blog（RSS） · 昨天 08:00

在连续批处理中，同步方式导致CPU与GPU交替工作，造成闲置浪费。测试显示，使用8B模型生成8K令牌时，GPU有24%的时间处于空闲状态。异步批处理通过分离工作负载，让CPU准备下一批次（N+1）的同时，GPU计算当前批次（N），从而消除闲置间隙。这可通过CUDA流实现操作并发，无需更改内核或模型，仅需协调硬件执行顺序。理论上，该方法可将总生成时间从300.6秒减少至228秒，实现24%的免费加速

能力进展 基础设施

https://huggingface.co/blog/continuous_async

---

12. Yetone发布Native Feel桌面应用开发Agent Skill

X：Berry Xia (@berryxia) · 8 小时前

开发者Yetone将一篇关于桌面应用开发"最佳实践"的文章转化为一个名为"native-feel-skill"的Agent Skill。该Skill旨在帮助开发者利用Coding Agent，轻松地重构或开发跨平台桌面应用，并使其获得极其接近Native原生应用的性能体验。项目代码已开源在GitHub上。

能力进展 新发布

https://x.com/berryxia/status/2055074211715301830

---

13. Moonshot AI创始人杨植麟最近放出了一个40分钟视频

X：Berry Xia (@berryxia) · 昨天 09:19

杨植麟在视频中拆解Kimi K2模型的训练，仅花费460万美元便在编程大战中击败GPT-5.5等对手。其通过极致优化、线性注意力等架构创新，抹平资源差距，标志AI竞赛规则改变，小团队以聪明设计颠覆大厂传统玩法。

能力进展 基础设施

https://x.com/berryxia/status/2054733412846690443

---

14. 为什么资深开发者讲不清自己的专业能力

X：宝玉 (@dotey) · 6 小时前

资深开发者与业务团队存在根本认知差异。业务团队生活在"消除不确定性"的循环中，追求快速试错验证，核心是速度。而资深开发者身处"管理复杂性"的循环，核心职责是保障付费服务的长期稳定，因此对增加系统复杂性的行为极为警惕。沟通失败在于，开发者用"控制复杂性"的理由拒绝需求，却未回应业务端"消除不确定性"的迫切诉求。解决方案是，开发者应将其精简需求、复用代码等专业能力，包装成能帮助业务"更快获得答案"的方

能力进展

https://x.com/dotey/status/2055097242755706984

---

15. 微信群聊总结Skill新增，依赖wx-cli配置

X：宝玉 (@dotey) · 昨天 11:38

baoyu-skills 新加了一个 Skill：微信群聊总结 Skill：https：//github.com/JimLiu/baoyu-skills/tree/main/skills/baoyu-wechat-summary 依赖于 wx-cli：https：//github.com/jackwener/wx-cli 如何配置使用 wx-cli 请看项目文档，无法提供帮助。另外目前只是借助

能力进展

https://x.com/dotey/status/2054768295534846239

---

16. 2028年全球AI领导地位的两种情景

Anthropic：Research（发表成果 · 网页） · 13 小时前

报告展望2028年中美AI竞争的两种前景。若美国及盟友维持并扩大在关键计算芯片上的优势，通过加强出口管制、遏制技术窃取并加速AI应用，民主国家可确立12-24个月的技术领先，主导AI规则制定。反之，若政策松动，中国可能借助人才优势、利用管制漏洞迅速逼近甚至反超，使威权政权获得大规模自动化压制能力。当前民主国家在计算领域优势显著，但窗口期有限，需立即行动锁定胜局。

基础设施 监管/资本

https://www.anthropic.com/research/2028-ai-leadership

---

17. 在大型代码库中高效运用Claude Code：最佳实践与入门指南

Claude：Blog（网页） · 14 小时前

Claude Code已成功部署于数百万行的单体仓库、遗留系统及分布式架构中。其核心在于围绕模型构建的"工具套件"，而非仅依赖模型本身。该套件包含五个关键扩展点：提供代码库概览的CLAUDE.md文件、实现持续改进的钩子、按需加载专业知识的技能、插件以及MCP服务器。它采用智能体搜索模式，直接在开发者本地实时代码库上操作，无需构建和维护集中式索引，从而避免了传统RAG系统在活跃大型代码库中索引过时

能力进展

https://claude.com/blog/how-claude-code-works-in-large-codebases-best-practices-and-where-to-start

---

18. 编写循环运行Codex审查的代码技能

X：Peter Steinberger (@steipete) · 22 小时前

编写了一个技能，可以循环运行codex /review直到没有错误为止。 注意事项：它不会为你修复系统架构，所以你仍然需要将BRAIN作为主模型。https：//github.com/steipete/agent-scripts/blob/main/skills/codex-review/SKILL.md

能力进展

https://x.com/steipete/status/2054850632067019173

---

19. API提示预缓存加速首令牌生成

X：Claude Devs (@ClaudeDevs) · 8 小时前

减少API长提示首令牌生成时间的实用技巧：预热提示缓存。 在用户提示前发送系统提示。Claude会将其写入缓存，但跳过生成任何输出。 当真实用户请求到达时，将直接命中预热缓存。

能力进展

https://x.com/ClaudeDevs/status/2055069548672631218
