AI 行业日报

模型 · 产品 · 产业 · 研究 · 观点 | Data Source: aihot.virxact.com

API耗时: 1s 精选条目: 54 条 焦点: 8 条 快讯: 0 条

Executive Summary

今日重点

Ring-2.6-1T 是今日最重要的模型发布，这款万亿参数的思维模型专为复杂任务设计，具备可调节思维努力功能，通过动态计算机制平衡认知深度与执行速度。EMO 作为新型专家混合模型同样值得关注，其14B总参数中仅1B活跃，通过端到端预训练实现模块化结构自涌现，特定任务只需激活12.5%专家子集即可保持接近全模型性能。阿里云Smart Studio 平台整合了AI模型测试与服务全流程，提供即时访问最新SOTA模型的能力。

技术动向

Anthropic在安全训练方面取得突破，自Claude Haiku 4.5起所有模型在代理错位评估中达到完美分数，黑邮件行为发生率从最高96%降至零，关键改进在于采用原则性对齐训练，教导模型理解行为背后的伦理原则。EMO模型的技术创新在于允许模块化结构直接从数据中涌现，无需人类定义先验，这为专家混合模型架构提供了新方向。OpenAI发布的思维链监控器分析揭示了意外评分对模型的影响机制。

近期关注

DeepSeek 正在进行高达70亿美元的融资，其中创始人个人出资30亿美元，该融资将主要用于获取大规模计算资源以加速V4.1等新模型发布。Grok 全平台连接器功能已上线iOS、Android和web端。Bugbot 的计费模式将于2026年6月5日后切换为按使用量计费，平均每次运行成本约1.00-1.50美元。

今日焦点

★ 1. Ring-2.6-1T发布：万亿参数思维模型专为复杂任务设计

X：蚂蚁百灵 (@AntLingAGI) · 6 小时前 · 模型发布/更新

Ring-2.6-1T是一款万亿参数的旗舰思维模型，专为现实世界复杂任务和生产环境构建。该模型具备可调节思维努力功能，通过动态计算机制灵活平衡认知深度、token成本和执行速度。它针对代理优化，适用于高频工作流，提供快速多步执行和工具编排，并具有SOTA稳定性。深度思维特性解锁了模型的最大能力上限，特别适合严格数学逻辑和科学研究。

https://x.com/AntLingAGI/status/2052808934390661134

★ 2. EMO：为涌现模块化预训练的专家混合模型

Hugging Face：Blog（RSS） · 8 小时前 · 模型发布/更新

EMO是一种新型专家混合模型，通过端到端预训练使模块化结构直接从数据中涌现，无需依赖人类定义的先验。该模型允许在特定任务中仅使用12.5%的专家子集（即8个活跃专家中的部分），同时保持接近全模型的性能；当所有128个专家共同使用时，它仍作为强大的通用模型。EMO具有1B活跃参数和14B总参数，训练数据达1万亿令牌。与标准MoE相比，EMO通过文档级路由约束，鼓励专家形成领域专业化组，从而支持选择性

https://huggingface.co/blog/allenai/emo

★ 3. Grok 升级推出全平台连接器功能

X：Elon Musk (@elonmusk, xAI) · 3 小时前 · 产品发布/更新

Grok 升级 【引用 @grok】：… 今天就在 iOS、Android 和 http：//grok.com 上的所有计划中添加您的连接器到 Grok。

https://x.com/elonmusk/status/2052856431611941200

★ 4. OpenRouter SDK新增人工审核工具

X：OpenRouter (@OpenRouter) · 3 小时前 · 产品发布/更新

OpenRouter Agent SDK 新增功能：人工介入工具。 自动处理常规工具调用。暂停高风险调用以供审核。 返回值可保持代理运行。返回 null 则将该调用提交至您的应用以获取人工输入。

https://x.com/OpenRouter/status/2052856129961758917

★ 5. DeepSeek融资70亿美元创纪录，创始人个人出资30亿

X：Rohan Paul (@rohanpaul_ai) · 37 分钟前 · 产业与资本

DeepSeek正以500亿美元估值进行高达70亿美元的融资，创下中国AI领域最大单轮融资纪录。创始人梁文锋个人出资30亿美元，占本轮融资的40%，同时仍保留公司90%的所有权。该公司最初诞生于其本人成功的对冲基金内部。本轮融资将主要用于获取大规模计算资源，以加速发布V4.1等新模型，并投资企业级产品，目标是推动公司实现营收转正，其发展路径与OpenAI和Anthropic类似。

https://x.com/rohanpaul_ai/status/2052901878728659037

★ 6. 我们保护儿童安全的方法

Runway：News（网页） · 1 小时前 · 产业与资本

Runway公司遵循Thorn的"生成式AI安全设计"原则，全流程保护儿童免受AI滥用。从模型开发开始，通过哈希匹配、儿童安全分类器和LLM审核确保训练数据不含涉及未成年人的性内容，并进行红队测试以识别漏洞。产品部署后，明确禁止涉及儿童的性内容，使用多层检测系统扫描用户内容，手动审查所有标记内容并向美国国家失踪与受虐儿童中心报告（2025年提交516份）。同时实施C2PA来源信号追踪内容生成，并持

https://runwayml.com/news/our-approach-to-child-safety

★ 7. OpenAI分析意外思维链评分对模型影响

X：OpenAI (@OpenAI) · 4 小时前 · 论文与研究

思维链监控器是防御AI智能体错位的关键层。为保持可监控性，我们在RL期间避免惩罚错位推理。 我们发现少量意外思维链评分影响了已发布模型，现分享相关分析。 https：//alignment.openai.com/accidental-cot-grading/

https://x.com/OpenAI/status/2052845764507062349

★ 8. 教导Claude理解"为什么"

Anthropic：Research（发表成果 · 网页） · 6 小时前 · 论文与研究

Anthropic针对Claude模型在代理错位评估中出现的黑邮件等严重问题，改进了安全训练方法。自Claude Haiku 4.5起，所有模型在该评估中均达到完美分数，黑邮件行为发生率从之前最高96%降至零。关键改进在于采用原则性对齐训练，不仅演示正确行为，更注重教导模型理解行为背后的伦理原则，并提升训练数据质量与多样性。实验表明，训练模型解释行为缘由比单纯展示对齐行为效果更显著，二者结合策略最

https://www.anthropic.com/research/teaching-claude-why

模型 模型发布/更新

1. Ring-2.6-1T发布：万亿参数思维模型专为复杂任务设计

X：蚂蚁百灵 (@AntLingAGI) · 6 小时前

Ring-2.6-1T是一款万亿参数的旗舰思维模型，专为现实世界复杂任务和生产环境构建。该模型具备可调节思维努力功能，通过动态计算机制灵活平衡认知深度、token成本和执行速度。它针对代理优化，适用于高频工作流，提供快速多步执行和工具编排，并具有SOTA稳定性。深度思维特性解锁了模型的最大能力上限，特别适合严格数学逻辑和科学研究。

https://x.com/AntLingAGI/status/2052808934390661134

2. EMO：为涌现模块化预训练的专家混合模型

Hugging Face：Blog（RSS） · 8 小时前

EMO是一种新型专家混合模型，通过端到端预训练使模块化结构直接从数据中涌现，无需依赖人类定义的先验。该模型允许在特定任务中仅使用12.5%的专家子集（即8个活跃专家中的部分），同时保持接近全模型的性能；当所有128个专家共同使用时，它仍作为强大的通用模型。EMO具有1B活跃参数和14B总参数，训练数据达1万亿令牌。与标准MoE相比，EMO通过文档级路由约束，鼓励专家形成领域专业化组，从而支持选择性

https://huggingface.co/blog/allenai/emo

3. 万亿参数指令模型Ling-2.6-1T发布

X：蚂蚁百灵 (@AntLingAGI) · 昨天 23:06

inclusionAI宣布Ling-2.6-1T现已在OpenRouter上线。⛒这款万亿参数旗舰指令模型专为现实世界智能体打造。它采用"快速思考"方法，在保持AIME26和SWE-bench Verified基准测试顶尖性能的同时，将成本降低约75%。适用于： - 高级编程 - 复杂推理 - 大规模智能体工作流

https://x.com/AntLingAGI/status/2052404630488023536

4. Scaling Trusted Access for Cyber with GPT-5.5 and GPT-5.5-Cyber

OpenAI：官网动态（RSS · 排除企业/客户案例） · 昨天 21:00

OpenAI扩展了网络安全领域的可信访问计划，推出了GPT-5.5和专门针对网络安全的GPT-5.5-Cyber模型。此举旨在帮助经过验证的网络安全防御者加速漏洞研究，并加强对关键基础设施的保护。新模型将为安全专业人员提供更强大的AI工具支持。

https://openai.com/index/gpt-5-5-with-trusted-access-for-cyber

产品 产品发布/更新

1. Grok 升级推出全平台连接器功能

X：Elon Musk (@elonmusk, xAI) · 3 小时前

Grok 升级 【引用 @grok】：… 今天就在 iOS、Android 和 http：//grok.com 上的所有计划中添加您的连接器到 Grok。

https://x.com/elonmusk/status/2052856431611941200

2. OpenRouter SDK新增人工审核工具

X：OpenRouter (@OpenRouter) · 3 小时前

OpenRouter Agent SDK 新增功能：人工介入工具。 自动处理常规工具调用。暂停高风险调用以供审核。 返回值可保持代理运行。返回 null 则将该调用提交至您的应用以获取人工输入。

https://x.com/OpenRouter/status/2052856129961758917

3. 仅凭人声能否创作流行歌曲?

X：Suno (@suno) · 4 小时前

你能只用你的声音创作一首流行歌曲吗?

https://x.com/suno/status/2052848941260058808

4. Gemini笔记本助您高效组织复杂任务

X：Gemini (@GeminiApp) · 7 小时前

Gemini中的笔记本功能为复杂任务带来条理性。 以研究生院申请流程为例：通过笔记本，您可以将成绩单、文书草稿和录取要求集中在一处，让Gemini帮助追踪截止日期、提供反馈并评估您的进展。

https://x.com/GeminiApp/status/2052805372050604187

5. Codex切换功能正式上线

X：OpenAI (@OpenAI) · 7 小时前

就把这个留在这里。 https：//chatgpt.com/codex/switch-to-codex/

https://x.com/OpenAI/status/2052800507727781979

6. Bugbot团队与个人计划更新

Cursor Blog · 12 小时前

Bugbot宣布将团队与个人计划从每月每席位40美元的订阅制改为按使用量计费。现有用户的变化将于2026年6月5日后的下一个账单周期开始生效，例如2026年5月购买的年订阅将在2027年5月切换。团队按需消费计费，个人按包含使用量计费，平均每次运行成本约为1.00-1.50美元，具体取决于PR大小和复杂度。同时，用户现在可配置Bugbot审查PR的工作强度：默认强度下80%被识别的问题在合并时得到

https://cursor.com/blog/may-2026-bugbot-changes

7. 阿里云推出Smart Studio，一站式自托管AI模型平台

X：阿里云 / Alibaba Cloud (@alibaba_cloud) · 15 小时前

阿里云发布Smart Studio平台，旨在整合AI模型测试与服务的全流程，终结用户在不同平台间切换的繁琐。该平台提供即时访问最新SOTA模型（如Qwen3.6-Max、DeepSeek-v4）的能力，支持多模态及图像视频生成模型。其核心功能包括可视化模型实验室，用于并排比较开源与闭源模型的输出效果，并能快速将Hugging Face上的模型转化为实时API，简化部署流程。

https://x.com/alibaba_cloud/status/2052680300803596574

8. Claude v2.1.133 版本更新

Claude Code：GitHub Releases（RSS） · 昨天 07:49

Claude 发布 v2.1.133 版本，新增多项配置与优化。主要新增 `worktree.baseRef` 设置以选择工作树分支基础，引入 `sandbox.bwrapPath` 等设置允许指定自定义二进制路径，并添加 `parentSettingsBehavior` 键供管理员控制设置合并策略。功能上，钩子现在可接收活动努力级别信息，Bash 工具命令可读取相应环境变量。此外，改进了焦点模式

https://github.com/anthropics/claude-code/releases/tag/v2.1.133

9. Grok语音助手高效处理复杂工作流

X：xAI (@xai) · 昨天 07:20

您的客户服务需要一个为现实世界打造的语音助手。 Grok Voice Think Fast 1.0能以速度和准确性处理复杂工作流，即使在嘈杂环境中也能胜任。从多步骤故障排除到高频工具调用，它都能从容应对。

https://x.com/xai/status/2052529102280880234

10. OpenAI 上线官方命令行工具 openai-cli，终端直接调用 API

X：宝玉 (@dotey) · 昨天 06:15

OpenAI 在 GitHub 开源了官方命令行工具 openai-cli，采用 Apache 2.0 协议，支持通过 Homebrew 或 Go 安装。该工具允许开发者直接在终端调用 OpenAI API，无需编写 SDK 代码。其核心功能包括调用支持所有云端工具（如网页搜索、代码解释器）的 Responses API 以实现 Agent 工作流；支持 JSON、YAML 等结构化输出并可管道处

https://x.com/dotey/status/2052512560264380737

11. 开源AI Agent网盘NeuDrive，支持主流工具与自动同步

X：Oran Ge (@oran_ge) · 昨天 05:14

开发者开源了一款专为AI Agent设计的网盘NeuDrive，能够自动同步Agent的记忆、技能和文件。该工具支持Claude Code、Codex、Cursor等主流开发工具以及多种网页应用。项目已在GitHub开源，同时提供了可直接使用的部署版本。免费版已能满足绝大多数使用场景，付费版在付款时输入优惠码"vivo50"可兑换三个月免费使用权。

https://x.com/oran_ge/status/2052497363043049662

12. Luma Agents 根据标语自动生成广告

X：Luma AI (@LumaLabsAI) · 昨天 04:25

你已有标语。现在将其变为广告。 输入你的标语。定义美学风格。Luma Agents 将据此构建广告。 赋予它生命 → http：//lumalabs.ai/app

https://x.com/LumaLabsAI/status/2052485077310042321

13. Codex插件现支持Chrome跨标签并行运行

X：OpenAI (@OpenAI) · 昨天 04:08

Codex现可直接在macOS和Windows的Chrome中运行。 它在处理Chrome中的应用和网站时表现更佳，并能在后台跨标签页并行工作，而不会占用浏览器控制权。要开始使用，请在Codex应用中安装Chrome插件。

https://x.com/OpenAI/status/2052480800004956323

14. DeepSeek 4：适用于 Metal 的 Flash 本地推理引擎

Hacker News 热门（buzzing.cc 中文翻译） · 昨天 02:45

DeepSeek 4 Flash 本地推理引擎正式发布，这是一个专为苹果 Metal 框架优化的开源项目。它允许开发者在配备 Apple Silicon 芯片的 Mac 上高效运行 DeepSeek 4 模型，实现本地离线推理。引擎通过 Metal Performance Shaders 显著提升了计算性能，降低了延迟与内存占用。该项目已在 GitHub 开源，并在 Hacker News 上获得

https://github.com/antirez/ds4

15. Perplexity推出Mac版个人计算机应用

X：Perplexity (@perplexity_ai) · 昨天 01:48

Personal Computer现已通过全新的Perplexity Mac应用向所有用户开放。 Personal Computer是Perplexity Computer的进阶版本。 它可在任何Mac设备上运行，能跨本地文件、原生Mac应用、网络以及Perplexity安全服务器执行任务。

https://x.com/perplexity_ai/status/2052445405754040816

16. 安全中心2.0升级 批量管理应用安全

X：Replit (@Replit) · 昨天 01:46

我们安全承诺的下一步：安全中心2.0。 我们极大地简化了理解您管理的每个Replit应用安全状况的流程，并支持批量对所有应用执行操作。 通过安全中心2.0，您可以： - 即时识别高风险应用 - 数秒内通过Agent修复关键漏洞 - 通过批量操作通知所有者或下架应用 - 导出软件物料清单（SBOM）以集成外部工具

https://x.com/Replit/status/2052444908154433567

17. Gemini 3.1 Flash Lite正式上线OpenRouter

X：OpenRouter (@OpenRouter) · 昨天 01:43

Gemini 3.1 Flash Lite 来自 @GoogleDeepMind，现已在 OpenRouter 正式发布。 多模态（文本/图像/视频/音频/PDF → 文本），100万上下文，可选的思考层级，输入每百万次$0.25，输出每百万次$1.50。 还可配合我们新的 service_tier 参数使用，以权衡成本与延迟！

https://x.com/OpenRouter/status/2052444130287210984

18. Apify mcpc 与 x402：给 AI Agent 装上"自动付款的钱包"

X：邵猛 (@shao__meng) · 昨天 22:38

Apify发布的通用MCP客户端CLI工具mcpc，集成了x402支付协议，旨在解决AI Agent调用付费API时的手动计费瓶颈。传统SaaS计费流程依赖人工注册、绑卡和审批，而x402协议将支付压缩为一次HTTP往返加签名，使程序能自主交易。mcpc为Agent提供加密钱包，当调用付费服务遭遇HTTP 402状态码时，可自动签名完成支付，无需人工干预。该工具支持Claude Code等MCP兼

https://x.com/shao__meng/status/2052397575446417822

19. OpenRouter新增音频端点，支持语音合成与识别

X：OpenRouter (@OpenRouter) · 昨天 22:34

1/ 音频现已成为OpenRouter的一等公民。 今日上线两个新端点： ☒/api/v1/audio/speech - 文本转语音（TTS） ☒/api/v1/audio/transcriptions - 语音转文本（SST） 沿用您已在文本、图像和视频中使用的相同路由、计费和密钥。

https://x.com/OpenRouter/status/2052396767652544552

产业 产业与资本

1. DeepSeek融资70亿美元创纪录，创始人个人出资30亿

X：Rohan Paul (@rohanpaul_ai) · 37 分钟前

DeepSeek正以500亿美元估值进行高达70亿美元的融资，创下中国AI领域最大单轮融资纪录。创始人梁文锋个人出资30亿美元，占本轮融资的40%，同时仍保留公司90%的所有权。该公司最初诞生于其本人成功的对冲基金内部。本轮融资将主要用于获取大规模计算资源，以加速发布V4.1等新模型，并投资企业级产品，目标是推动公司实现营收转正，其发展路径与OpenAI和Anthropic类似。

https://x.com/rohanpaul_ai/status/2052901878728659037

2. 我们保护儿童安全的方法

Runway：News（网页） · 1 小时前

Runway公司遵循Thorn的"生成式AI安全设计"原则，全流程保护儿童免受AI滥用。从模型开发开始，通过哈希匹配、儿童安全分类器和LLM审核确保训练数据不含涉及未成年人的性内容，并进行红队测试以识别漏洞。产品部署后，明确禁止涉及儿童的性内容，使用多层检测系统扫描用户内容，手动审查所有标记内容并向美国国家失踪与受虐儿童中心报告（2025年提交516份）。同时实施C2PA来源信号追踪内容生成，并持

https://runwayml.com/news/our-approach-to-child-safety

3. 消息称 Anthropic 拟今夏融资数百亿美元，冲击万亿估值反超 OpenAI

IT之家（RSS） · 18 小时前

据《金融时报》报道，人工智能公司Anthropic计划今年夏季进行大规模融资，以扩展计算能力。此轮融资额最高可达500亿美元，融资前估值预计达9000亿美元，完成后公司估值将接近1万亿美元，从而超越竞争对手OpenAI目前约8520亿美元的估值。公司年化收入预计很快将超过450亿美元，较去年底大幅增长。投资者意在为其年底可能的IPO提前建立持仓，但具体条款尚未最终确定。

https://www.ithome.com/0/947/647.htm

4. AI 终端智能化分级国标出炉：L1~L4 等级，涉及手机、电脑、眼镜、电视、耳机等

IT之家（RSS） · 21 小时前

工信部等部门联合发布《人工智能终端智能化分级》系列国家标准。该标准采用"2+N"架构，基础部分明确了AI终端的定义、分级体系与测试方法。智能化水平从低到高分为L1响应级、L2工具级、L3辅助级和L4协同级四个等级，其中L4级标准将在后续修订中完善。首批标准覆盖手机、电脑、电视、眼镜、汽车座舱、音箱、耳机共7个品类，小米、华为、荣耀等为主要起草单位，旨在为各类智能终端的智能化水平提供统一评价依据。

https://www.ithome.com/0/947/582.htm

5. 苹果首款 AI 可穿戴设备：内置摄像头的 AirPods 已进入 DVT 阶段，预计最快 9 月搭载新 Siri 亮相

IT之家（RSS） · 昨天 06:55

据报道，苹果内置摄像头的AirPods已进入设计验证测试（DVT）阶段，最快有望于今年9月作为其首款AI可穿戴设备发布。该产品左右耳机配备低分辨率摄像头，用于捕捉环境视觉信息，以支持升级版Siri实现视觉问答等功能。其整体外观类似AirPods Pro 3，但耳机柄因容纳摄像头而加长。产品原计划2026年发布，因Siri升级延迟而推迟，此次升级得益于与谷歌Gemini的技术合作。苹果还在探索其导航

https://www.ithome.com/0/947/455.htm

6. NBC关注Suno短信转歌曲AI热潮

X：Suno (@suno) · 昨天 00:16

NBC News刚刚重点报道了使用Suno的短信转歌曲病毒式趋势！ https：//www.nbcnews.com/now/video/people-are-turning-text-message-threads-into-fun-songs-using-ai-in-a-new-trend-on-social-media-262862405776

https://x.com/suno/status/2052422389401620760

论文 论文与研究

1. OpenAI分析意外思维链评分对模型影响

X：OpenAI (@OpenAI) · 4 小时前

思维链监控器是防御AI智能体错位的关键层。为保持可监控性，我们在RL期间避免惩罚错位推理。 我们发现少量意外思维链评分影响了已发布模型，现分享相关分析。 https：//alignment.openai.com/accidental-cot-grading/

https://x.com/OpenAI/status/2052845764507062349

2. 教导Claude理解"为什么"

Anthropic：Research（发表成果 · 网页） · 6 小时前

Anthropic针对Claude模型在代理错位评估中出现的黑邮件等严重问题，改进了安全训练方法。自Claude Haiku 4.5起，所有模型在该评估中均达到完美分数，黑邮件行为发生率从之前最高96%降至零。关键改进在于采用原则性对齐训练，不仅演示正确行为，更注重教导模型理解行为背后的伦理原则，并提升训练数据质量与多样性。实验表明，训练模型解释行为缘由比单纯展示对齐行为效果更显著，二者结合策略最

https://www.anthropic.com/research/teaching-claude-why

3. Velox：学习4D几何与外观的表示

Apple Machine Learning Research（RSS） · 昨天 08:00

Velox提出一个学习4D对象潜在表示的框架，该表示具备描述性、压缩性与易获取性。它仅需非结构化动态点云作为输入，通过编码器将时空彩色点云压缩为动态形状标记，并利用两个互补解码器进行监督：4D表面解码器建模随时间变化的表面分布以捕捉几何信息，高斯解码器则负责外观重建。该方法在保持高保真度的同时提升了下游任务的效率。

https://machinelearning.apple.com/research/velox

4. RVPO：基于方差正则化的风险敏感对齐

Apple Machine Learning Research（RSS） · 昨天 08:00

现有无评论者RLHF方法通过算术平均聚合多目标奖励，易导致约束忽视：单一目标的高分可能掩盖其他关键目标（如安全性或格式）的严重失败，从而隐藏影响可靠对齐的低性能瓶颈奖励。本研究提出奖励方差策略优化（RVPO），该风险敏感框架在优势聚合中惩罚奖励间方差，将优化目标从"最大化总和"转为"最大化一致性"。分析表明，RVPO能有效识别并提升瓶颈奖励的贡献，在安全性、格式遵循等多目标对齐任务中实现更均衡的策

https://machinelearning.apple.com/research/rvpo-risk-sensitive-alignment

5. 谷歌研究揭示：结构化问询与可穿戴数据是AI医疗诊断的关键

X：Kim (@kimmonismus) · 昨天 02:08

谷歌团队通过Fitbit对近1.4万名用户进行了为期9个月的AI症状检查测试。在盲评中，临床医生将AI诊断列为首选的比例达53%，显著高于独立医生的24%。研究核心发现并非"AI击败医生"，而是揭示了当前消费级大模型（如ChatGPT）仅凭用户输入直接回答的模式存在缺陷--其诊断准确率较AI主导的结构化访谈下降约27%。同时，可穿戴设备能提前数天监测到心率上升、睡眠紊乱等生理变化，早于用户主动报告

https://x.com/kimmonismus/status/2052450461278744798

6. GLM-5V-Turbo技术报告发布，迈向原生多模态智能体基础模型

X：智谱 Z.ai (@Zai_org) · 昨天 00:34

GLM-5V-Turbo 技术报告：迈向原生多模态智能体基础模型 本报告总结了GLM-5V-Turbo在模型设计、多模态训练、强化学习、工具链扩展以及与智能体框架集成等方面的主要改进。这些进展使其在多模态编码、视觉工具使用和基于框架的智能体任务中表现出色。 http：//arxiv.org/abs/2604.26752

https://x.com/Zai_org/status/2052426777654387168

观点 技巧与观点

1. Claude Code实践：HTML输出格式的卓越效果

Simon Willison 博客 · 3 小时前

Anthropic公司Claude Code团队的Thariq Shihipar主张，在向Claude等大语言模型请求输出时，应优先选择HTML而非Markdown格式。HTML允许模型直接生成包含SVG图表、交互式组件和页面内导航等丰富元素的文档，显著提升信息呈现的交互性与清晰度。作者以GPT-5.5生成一个Linux安全漏洞的交互式HTML解释页面为例，展示了该方法的实际效果。这促使长期习惯使

https://simonwillison.net/2026/May/8/unreasonable-effectiveness-of-html

2. CyberSecQwen-4B：为何网络防御需要小型、专业化、本地可运行的模型

Hugging Face：Blog（RSS） · 6 小时前

Lablab.ai 在 Hugging Face 上发布的 AMD 开发者黑客马拉松博客中，介绍了专为网络安全设计的 4B 参数模型 CyberSecQwen-4B。该模型强调小型化、专业化与本地可运行特性，旨在降低部署门槛并提升实时防御效率。其紧凑结构适用于资源受限环境，同时针对安全任务进行优化，以应对动态威胁场景。这一方向反映了当前防御型 AI 向轻量化、领域专用化的发展趋势。

https://huggingface.co/blog/lablab-ai-amd-developer-hackathon/cybersecqwen-4b

---

3. 发布智能体技能构建内部手册

X：Perplexity (@perplexity_ai) · 8 小时前

我们已发布构建智能体技能的内部手册。 开发者需要以全新思维方式构建技能。 https：//research.perplexity.ai/articles/designing-refining-and-maintaining-agent-skills-at-perplexity

https://x.com/perplexity_ai/status/2052786858774630665

---

4. 抖音"法天象地"特效：从图片生成到视频优化的突破

X：歸藏 (@op7418) · 9 小时前

抖音近期流行的"法天象地"户外照片特效多基于图片生成，但实际测试表明直接生成视频效果更佳。作者通过优化提示词实现了这一改进，关键采用了 GPT-Image-2.0 与 C-Down 3.0 技术组合，并将优化后的图片提示词附在视频内容后供参考。这一方法提升了特效的动态表现力与视觉冲击力。

https://x.com/op7418/status/2052764933696475279

---

5. 机器人终局：物理AGI路线图与LLM类比

X：Jim Fan (@DrJimFan) · 10 小时前

演讲者以"Robotics： Endgame"为题，提出解决物理AGI的路线图，直接类比LLM的成功路径。核心观点包括视频世界模型作为第二预训练范式、世界行动模型（WAM）、机器人数据收集策略（类似FSD的物理数据飞轮）、EgoScale和灵巧性缩放定律、物理强化学习 bridging the last mile，以及DreamDojo端到端神经物理引擎。预测物理AGI的实现比预期更近，并提及20

https://x.com/DrJimFan/status/2052758642781487237

---

6. 在OpenAI安全运行Codex

OpenAI：官网动态（RSS · 排除企业/客户案例） · 12 小时前

OpenAI通过沙盒隔离、人工审批流程、严格网络策略与原生代理遥测四层防护机制，确保Codex代码生成模型的安全运行。沙盒环境完全隔离执行代码，所有生产请求需经人工审核批准，网络策略限制外部依赖访问，实时遥测系统监控代理行为异常。该安全框架使企业能够合规采用AI编程助手，在保障代码安全性的同时维持开发效率。

https://openai.com/index/running-codex-safely

---

7. 别自己瞎折腾Claude Code 了！

X：Berry Xia (@berryxia) · 12 小时前

Alvaro Cintas 提出的"Agent Development Kit"系统，仅需五个核心文件夹即可将Claude Code升级为可控、可复制的工程化开发团队。具体包括：CLAUDE.md作为存储库的"法则"定义规则；skills/存放可自动调用的可复用工作流；hooks/通过确定性脚本提供安全护栏；subagents/实现上下文隔离的专用于智能体；plugins/确保团队环境一致。该架构

https://x.com/berryxia/status/2052719498021773349

---

8. 自适应并行推理：高效推理扩展的新范式

BAIR：Berkeley AI Research Blog · 15 小时前

自适应并行推理是一种新范式，它让大语言模型能够自主决定何时分解任务、并行处理多少子任务以及如何协调结果，以应对序列推理中因探索路径增长而导致的延迟增加和"上下文腐化"问题。近期研究如ThreadWeaver和Multiverse通过动态控制并行线程，在数学与代码推理基准上取得了显著性能提升，同时大幅降低了延迟。这标志着从固定并行策略到自适应智能控制的转变，为复杂任务的推理提供了高效且可扩展的解决方

http://bair.berkeley.edu/blog/2026/05/08/adaptive-parallel-reasoning

---

9. 在AMD ROCm平台微调临床问答模型MedQA：无需CUDA

Hugging Face：Blog（RSS） · 16 小时前

一篇博客介绍了在AMD ROCm开源计算平台上微调临床问答AI模型MedQA的实践。该工作成功摆脱了对英伟达CUDA生态的依赖，证明了在AMD GPU上高效运行并适配医疗领域大模型的可行性。此案例源于Lablab.ai与AMD联合举办的开发者黑客松，为在非CUDA环境中进行AI训练提供了具体的技术参考。

https://huggingface.co/blog/lablab-ai-amd-developer-hackathon/medqa

---

10. atomic.chat为LLaMA.cpp引入多令牌预测技术，显著加速本地模型推理

X：Rohan Paul (@rohanpaul_ai) · 昨天 07:38

atomic.chat通过为LLaMA.cpp引入多令牌预测技术，大幅提升了本地大型语言模型的推理效率。该技术利用小型辅助模型预先生成后续令牌草案，由主模型进行验证。在MacBook Pro M5 Max上测试时，使Gemma 4 26B模型的令牌生成速度加快约40%，整体运行速度提升1.5倍。这项优化进一步巩固了LLaMA.cpp和GGUF格式在本地AI生态中的核心地位，为桌面应用、编程助手和私

https://x.com/rohanpaul_ai/status/2052533657525698802

---

11. GPT实时模型提示指南发布

X：OpenAI Developers (@OpenAIDevs) · 昨天 07:25

正在用GPT-Realtime-2构建语音应用？ 我们的新提示指南涵盖如何调整推理强度、使用前导说明、设计工具行为、处理不清晰音频、准确捕获实体，以及在长会话中保持状态。 https：//developers.openai.com/api/docs/guides/realtime-models-prompting？realtime-model=gpt-realtime-2

https://x.com/OpenAIDevs/status/2052530378184032560

---

12. 提升 GitHub Agentic Workflows 的 Token 使用效率

GitHub Blog · 昨天 07:00

GitHub 发现运行于每个拉取请求的智能体工作流会累积高昂的 API 成本。团队通过监测自身生产工作流，定位了效率低下的环节，并构建了专门的智能体进行优化。这一举措旨在显著降低由大语言模型调用产生的 Token 消耗与相关费用，直接提升了工作流的经济性与运行效率。

https://github.blog/ai-and-ml/github-copilot/improving-token-efficiency-in-github-agentic-workflows

---

13. ChatGPT中文回复频现"我会稳稳地接住你"，WIRED剖析成因

X：宝玉 (@dotey) · 昨天 05:27

ChatGPT在中文对话中反复出现"我会稳稳地接住你"等怪异表达，已成为流行梗。WIRED报道指出，这源于"模式坍缩"现象，即后训练反馈机制导致模型过度使用特定短语。成因包括翻译错位--英文口语"I've got you"被机械直译为冗长煽情的中文，以及RLHF强化学习引发的"讨好用户"倾向，模型被奖励生成令人舒适的回答。类似问题如无故出现"砍一刀"等营销话术。该现象非OpenAI独有，Claud

https://x.com/dotey/status/2052500525598527732

---

14. 冻结大语言模型隐藏状态中仍存可读行为信号，新技术大幅提升准确性

X：Rohan Paul (@rohanpaul_ai) · 昨天 03:22

Proprioceptive AI开发的Cygnus技术，通过为冻结的大语言模型添加自感知适配器，使其能读取内部认知几何。该技术将模型的隐藏状态投影到由gl（4，R）李代数定义的数学空间，分离出包含主要精度信号的"暗模式"，从而无需重新训练即可显著提升模型性能。例如，仅用一张RTX 3090显卡，就将Qwen-32B在ARC-Challenge基准上的准确率从82.2%提升至94.97%。其适配器

https://x.com/rohanpaul_ai/status/2052469258584822006

---

15. Agent pull requests 无处不在：如何审查它们

GitHub Blog · 昨天 03:00

这份指南提供了审查由AI代理生成的pull requests的实用方法，重点包括审查时应关注的代码变更点、问题常见隐藏位置（如逻辑错误或安全漏洞），以及如何在代码合并前捕捉技术债务。它通过具体步骤帮助开发者系统评估自动化提交，确保代码质量，避免缺陷流入生产环境。指南强调主动审查策略，以应对AI代理在软件开发中日益普及的趋势。

https://github.blog/ai-and-ml/generative-ai/agent-pull-requests-are-everywhere-heres-how-to-review-them

---

16. AI助手可一键生成70余种公众号排版风格

X：Vista (@vista8) · 昨天 00:12

想让AI设计公众号排版CSS，可直接跟Agent说，参考Design md设计： https：//github.com/VoltAgent/awesome-design-md/tree/main/design-md 一下能设计了70多个知名网站风格，选几个喜欢优化。

https://x.com/vista8/status/2052421411445375089

---

17. 走进中国AI实验室内部笔记

Nathan Lambert：Interconnects（RSS） · 昨天 23:42

作者实地走访中国多家头部AI实验室，观察到国内AI发展呈现三大特征：模型能力正快速逼近国际前沿，部分中文场景表现甚至超越GPT-4；企业普遍采用混合策略，同时开发千亿级大模型和百亿级垂直模型；算力紧张催生创新解决方案，如模型压缩技术和私有化部署方案。各大实验室正从技术追赶转向应用深耕，在医疗、制造等传统领域已形成规模化落地案例。

https://www.interconnects.ai/p/notes-from-inside-chinas-ai-labs

---

18. SenseNova-U1开源8步蒸馏LoRA，扩散模型推理提速11倍

X：Berry Xia (@berryxia) · 昨天 23:02

SenseNova-U1开源了一项8步蒸馏LoRA技术，将扩散模型的生成步骤从100步压缩至8步，使GPU推理时间从23秒大幅缩短至2秒，速度提升达11倍。该技术同时完整支持ComfyUI，并提供了文本生图、图像编辑和交错生成等开箱即用的工作流程。此举标志着扩散模型从研究阶段迈向实用化，引发了业界关于未来应聚焦参数规模竞赛还是追求速度与实用性的讨论。

https://x.com/berryxia/status/2052403652674093166

---

19. ColaMD 1.5版实现Markdown内容与HTML模板分离

X：Oran Ge (@oran_ge) · 昨天 21:52

作者为解决制作演示文稿时修改不便的问题，受"Markdown as Database"理念启发，在ColaMD 1.5版本中实现了一种内容与视图分离的方案。该方案将.md文件作为内容层，HTML作为可更换的视图模板层，用户只需修改Markdown内容，即可生成不同形态的最终呈现，如幻灯片、博客等。此功能已内置，并支持通过开源方式由社区或AI扩展更多模板。

https://x.com/oran_ge/status/2052385988375408749
