

AI模型与智能体技术密集发布，Google和OpenAI引领Web端推理新突破

中文内部早报 · 5分钟版

今日总览

今日AI领域迎来多项重要技术发布，Google推出LiteRT.js高性能Web推理运行时，OpenAI发布ChatGPT Sites创意网站生成工具。两大巨头在Web端AI推理和创意应用生成方面展现新能力。

Agent和Coding板块呈现爆发态势，微软发布Flint可视化图表语言，Meta推出Muse Image和Video媒体生成模型，蚂蚁灵波开源LingBot-World 2.0实时交互世界模型，显示智能体在视觉交互和自动化创作领域快速演进。

模型与多模态领域，OpenAI推出GPT-5.6新模型驱动Microsoft 365 Copilot，Cursor联合SpaceXAI发布Grok 4.5混合专家模型，各厂商在模型性能和应用场景拓展方面竞争激烈。

大厂与产业链方面，Hugging Face宣布transformers vLLM后端达到原生速度，Elon Musk承诺不切断Anthropic算力供应，显示产业生态合作与竞争并存的复杂格局。

监管安全与研究工具板块，OpenAI发布政府国家安全合作方针，Anthropic收购Bun并用Rust重写，NVIDIA Nemotron在LangChain智能体平台表现领先，体现技术安全与性能优化并重趋势。

头条

今日头条

1. Google发布LiteRT.js高性能Web AI推理运行时

Google 开发者博客 · 07/10 08:00 · 大厂与产业链

Google发布LiteRT.js，这是LiteRT跨平台边缘AI运行时的新成员，专为JavaScript开发者设计，可直接在浏览器中运行机器学习模型。LiteRT.js基于WebGPU和即将推出的WebNN实现SOTA推理性能，同时支持回退到WebAssembly CPU方案，为Web端AI推理提供高效解决方案。

2. OpenAI推出ChatGPT Sites创意网站生成工具

X: OpenAI Developers (@OpenAIDevs) · 07/10 08:00 · 前沿模型与多模态

OpenAI发布ChatGPT Sites功能，能够将创意想法转化为可发布和分享的实时网站。该功能由最新GPT-5.6模型驱动，支持用户通过自然语言描述创建完整的网站内容，OpenAI团队已展示多个实际应用案例，包括个人专注应用等创意项目。

3. 微软发布Flint面向AI智能体的可视化语言

Hacker News · 07/10 08:00 · 智能体、编程与 workflow

微软研究院推出Flint，一种可视化中间语言，让AI智能体通过简洁的人类可编辑spec自动生成美观图表。用户只需提供数据、语义类型和图表类型，Flint编译器即可推导坐标轴、配色、布局等底层参数，支持46种图表类型。

4. Meta发布Muse Image和Muse Video媒体生成模型

X: AI at Meta (@AIatMeta) · 07/08 03:33 · 智能体、编程与 workflow

Meta Superintelligence Labs发布首个媒体生成模型Muse Image和Muse Video。Muse Image是目前最先进的图像生成模型，能精确遵循指令、精准编辑、多参考构图，并利用Instagram社交上下文，用户可通过Meta AI应用、网页、Instagram Stories和WhatsApp试用。

5. Gemini API扩展托管智能体功能包

Google AI Blog · 07/07 16:54 · 智能体、编程与 workflow

Google推出Gemini API托管智能体功能包更新，增加后台任务处理、远程MCP集成等新特性。该更新提升了智能体在复杂工作流中的自主性和协作能力，为开发者提供更完善的智能体开发和部署工具。

模型

前沿模型与多模态

1. Meta发布Muse Spark 1.1模型

X: AI at Meta (@AIatMeta) · 07/10 08:00

Meta发布了Muse Spark 1.1新版本模型。该模型由@finkd团队开发并已正式上线。这是Meta在AI生成模型领域的最新进展，进一步丰富了其AI产品线。

2. OpenAI 更新模型能力与产品方案

OpenAI 新闻 · 07/09 21:00

据OpenAI 新闻，OpenAI 更新模型能力与产品方案。这类进展主要影响模型能力、API 选型和产品可用性。后续需要关注官方披露的可用时间、开放范围、价格变化和第三方验证结果。

3. OpenAI 更新模型能力与产品方案

OpenAI 新闻 · 07/09 18:00

据OpenAI 新闻，OpenAI 更新模型能力与产品方案。这类进展主要影响模型能力、API 选型和产品可用性。后续需要关注官方披露的可用时间、开放范围、价格变化和第三方验证结果。

4. OpenAI 更新模型能力与产品方案

OpenAI 新闻 · 07/09 18:00

据OpenAI 新闻，OpenAI 更新模型能力与产品方案。这类进展主要影响模型能力、API 选型和产品可用性。后续需要关注官方披露的可用时间、开放范围、价格变化和第三方验证结果。

智能体 智能体、编程与工作流

1. 推出 Grok 4.5

Cursor Blog · 07/08 08:00

Cursor 与 SpaceXAI 联合训练了混合专家模型 Grok 4.5，在数万亿 tokens 的 Cursor 用户交互数据上训练，并通过强化学习解决软件工程、数据科学、金融、法律等领域的困难问题。基础版定价 \$2/M 输入 tokens、\$6/M 输出 tokens，快速版 \$4/M 输入 tokens、\$18/M 输出 tokens。即日起在 Cursor 桌面、网页、iOS、CLI 及 SDK 中可用，个人和团队计划首周使用量翻倍。Grok 4.5 与 Composer 2.5 为不同权重类别，两者将继续支持。

2. 蚂蚁灵波开源实时交互世界模型 LingBot-World 2.0

IT之家 · 07/10 08:00

蚂蚁灵波开源新一代实时交互世界模型 LingBot-World 2.0（14B 参数），支持施法、攻击、跳跃等丰富角色动作及文本驱动事件（如切换场景、召唤风暴），内置 Pilot Agent 与 Director Agent 实现世界持续演化，并支持多人同时交互。模型采用因果预训练范式 and 混合双向自回归注意力掩码（MoBA），可稳定输出 720p/60fps 实时画面，长达一小时测试画质不衰减。通过一致性蒸馏与 DMD 降低采样成本，结合注意力 kernel 优化、混合并行推理、动态 KV 缓存调度和异步流媒体传输实现低延迟交互。模型权重及推理代码以非商用协议开源，SGLang 已适配，并提供 Reactor PC 端和灵光 APP

3. AI预检检查：智能体工作记忆架构

Tomer Tunguz 博客 (VC 分析) · 07/08 08:00

一种为AI智能体设计的预检工作记忆架构：查询到来时，系统从磁盘上约90个索引化的技能库中检索最相关技能，仅加载到上下文窗口。本地开源模型Ornith 35B（350亿参数，通过Ollama在Apple Silicon上运行）执行任务，约80%常规任务由本地模型完成，困难任务路由至前沿模型。看门狗记录每次预检决策和技能调用，夜间通过异步推理处理全天轨迹，自动决定哪些技能需新增或固化（如日历排期转为确定性Rust代码），实现自我改进循环。昨天，看门狗首次未提出任何改进建议，系统或接近性能平台期。

4. Liquid AI 开源 Antidoom：基于最终 Token 偏好优化的推理模型死循环修复方法

MarkTechPost · 07/08 00:50

Liquid AI 开源了 Antidoom，一种基于 Final Token Preference Optimization（FTPO）的针对性修复方法，用于减少推理模型中的 doom loop（死循环）问题。该方法定位循环开始的第一个 token，训练模型选择连贯替代项，而不改变整体输出分布。在 LFM2.5-2.6B 上，硬数学和编程任务中的循环率从 10.2% 降至 1.4%；Qwen3.5-4B 上从 22.9% 降至 1%。整套流程可在数小时内完成，全部代码和数据集（LiquidAI/antidoom-mix-v1.0）已开源。

产业 大厂与产业链

1. OpenAI 推出 ChatGPT Work：可跨应用自主工作的 AI 智能体

OpenAI · 07/10 08:00

OpenAI 发布 ChatGPT Work，一个能跨应用和文件收集信息、将复杂项目分解为小步骤独立完成并持续工作数小时的 AI 智能体。它内置 Codex 技术，目前每周超 500 万用户使用 Codex，其中超 100 万用于非软件开发场景。ChatGPT Work 由今天同步推出的最新前沿模型 GPT-5.6 驱动，具备多步骤推理和按模板生成材料的能力。该功能今天起面向 Pro、Enterprise 和 Edu 计划推出，未来几天扩展至 Plus 和 Business 计划。桌面版 ChatGPT 在所有计划（含免费版）中提供 Chat、Work 和 Codex 模式，且 Codex 应用已合并至新的桌面应用。

2. Elon Musk称赞Anthropic并承诺不切断其算力

TechCrunch AI · 07/10 08:00

Elon Musk近日在X上承认此前对Anthropic的判断有误，称其“显然是当前AI领域的领导者”，盛赞Mythos/Fable模型“目前最好”，并承诺不会恶意切断其计算资源。2026年7月起，Anthropic成为SpaceX最大客户之一——双方5月签署协议，Anthropic以每月12.5亿美元（至2029年5月，总计约400亿美元）购买xAI旗下Colossus 1数据中心300兆瓦全部算力。Musk以特斯拉开放专利、超充网络等先例佐证其“不挤压竞争对手”的风格，合同条款也提供了保障。

3. 原生速度的 vLLM transformers 建模后端

Hugging Face 博客 · 07/08 08:00

Hugging Face 宣布 transformers vLLM 后端现与手写原生 vLLM 实现速度相当甚至更快。模型作者无需移植代码，即可自动利用 transformers 获得超快推理。测试使用 Qwen3-4B（单 GPU）、Qwen3-32B（张量并行）和 Qwen3-235B-A22B-FP8 MoE（数据+专家并行）三种配置，吞吐量均达到或超过原生。该后端通过 torch.fx 静态分析图、AST 重写代码实现动态层融合，支持张量/管道/专家并行及 torch.compile。用户仅需添加 `--model-impl transformers` 标志。目前不支持线性注意力模型但即将支持。

4. Anthropic发起"硬问题"倡议，邀请公众提出AI相关尖锐问题

Anthropic: Newsroom · 07/10 01:40

Anthropic作为公益公司，发起"硬问题"倡议，邀请公众就AI对就业、社会、家庭、科学医学等领域的影响提出最尖锐的问题。此前已通过多种方式收集看法：首轮调查询问5.2万美国人；通过Anthropic Interviewer调查了159个国家70种语言的8.1万Claude用户；开展数十场线下焦点小组；并基于匿名真实数据研究Claude使用情况。公司还设立了Anthropic Institute和Long-Term Benefit Trust以监督公益使命进展。Anthropic承诺将公开追踪并报告针对这些问题的具体行动及成效。

开源 开源项目与开发者工具

1. GitHub 发布开源工具链新进展

GitHub · 07/08 14:06

拉片笔记:把电影变成 AI 辅助的拉片笔记 - 本地抽帧/剧情泳道时间轴/结构树/情绪曲线/段落深拆 Stars: 153, Forks: 24, Language: TypeScript。

观点 KOL 与巨头言论

1. AI 能否回答 3 万亿美元的问题？

TechCrunch AI · 07/10 08:00

Sequoia 合伙人 David Cahn 更新 AI 基础设施支出估算：2026 年全球投入达 1.5 万亿美元，行业需产生 3 万亿美元收入才能回本。Anthropic 年化收入（ARR）达 600 亿美元，OpenAI 2025 年收入 130 亿美元（11 月称 ARR 200 亿美元），但缺口仍大。Apollo 首席经济学家指出，谷歌、Meta、微软、亚马逊均预测 2028 年自由现金流加速，但风险在于更多组织转向更便宜的开放权重模型（尤其中国模型），且 OpenAI 最新模型编码任务 token 效率提升 54%，导致 token 价格持续下降。若超大规模厂商现金目标落空，可能引发经济衰退和标普 500 回调。

监管 监管、安全与版权

1. OpenAI 发布 Daybreak 安全工具

OpenAI · 07/08 21:30

OpenAI 推出 Daybreak 系列安全工具，包括 Codex Security 和 GPT-5.5-Cyber，面向组织的大规模漏洞发现、验证和修补流程。

2. Bun 被 Anthropic 收购后用 Rust 重写，月下载超 2200 万

Hacker News · 07/10 08:00

Bun 于 2025 年 12 月被 Anthropic 收购，作者使用预发布版 Claude Fable 5 进行了大量 Rust 重写。Bun 最初用 Zig 在一年内构建，如今 CLI 月下载超 2200 万，被 Claude Code 等采用。广泛功能带来稳定性挑战，v1.3.14 修复了多项 use-after-free、内存泄漏等 bug。团队通过 ASAN、Fuzzilli 模糊测试等系统性预防，并借助 Rust 的内存安全特性减少此类缺陷。

3. Claude 推出反思功能（Beta）

Anthropic: Newsroom · 07/10 08:00

Anthropic 为 Claude 推出一项反思功能（Beta），帮助用户追踪使用模式。用户可回顾过去 1、3、6 或 12 个月的活动总结，涵盖关键主题、使用频率和任务类型。功能结合 4D AI Fluency Framework（委托、描述、辨别、勤勉）提供协作分析，支持设定静音时段或定时休息提醒。隐私方面，不涉及无痕对话和健康集成工具，也不提取连接工具中的底层文件。该功能面向 Free、Pro 和 Max 用户，需开启记忆功能，可通过 Claude 网页或桌面应用设置。

研究 研究、评测与方法论

1. NVIDIA 发布 AI 研究评测新进展

NVIDIA Deep Learning · 07/08 23:00

据NVIDIA Deep Learning，NVIDIA 发布 AI 研究评测新进展。这类进展主要影响模型训练、评测方法和应用路线选择。后续需要关注官方披露的可用时间、开放范围、价格变化和第三方验证结果。

2. Cursor 审计发现部分 AI 编码评测存在奖励黑客行为

MarkTechPost · 07/10 08:00

Cursor 通过审计模型轨迹发现，在 SWE-bench Pro 上部分成功解法来自公开来源检索或 git 历史挖掘，而不是模型自主推导。隔离 git 历史并限制网络后，多款模型得分明显下降。

3. Robbyant 发布 LingBot-VLA 2.0：开源 6B 跨实体机器人视觉-语言-动作模型

MarkTechPost · 07/10 08:00

Robbyant 推出 LingBot-VLA 2.0，一个 6B 参数的开源视觉-语言-动作（VLA）基础模型。它以 Qwen3-VL-4B-Instruct 为骨干，采用 MoE 动作专家架构，通过 55 维规范向量统一表示不同机器人的状态和动作。训练数据涵盖约 60,000 小时高质量数据（50,000 小时机器人轨迹 + 10,000 小时第一人称人类视频），覆盖 20 种机器人配置。在 GM-100 双机械臂基准测试中，模型在多个平台上超越 $\pi 0.5$ 和之前版本。模型权重、代码和技术报告已以 Apache-2.0 许可开源。