

GPT-5.6系列模型发布引领AI性能新标杆

中文内部早报 · 5分钟版

今日总览

OpenAI发布GPT-5.6系列模型，包括旗舰Sol、平衡型Terra和成本最优Luna三个版本，在Agents' Last Exam和Coding Agent Index等多项基准测试中超越Claude Fable 5。该系列模型已作为首选模型集成到Microsoft 365 Copilot中，标志着新一代AI能力的商业化落地。

Agent和Coding领域迎来多项重要更新。Google Research推出SensorFM，这是在超过100万亿分钟传感器数据上预训练的通用健康基础模型。微软发布Flint可视化语言，Claude Code新增应用内浏览器功能，Cursor与SpaceXAI联合推出Grok 4.5混合专家模型。

大厂产业链方面，Google推出LiteRT.js高性能Web AI推理运行时，Hugging Face宣布transformers vLLM后端达到原生速度。Meta CEO扎克伯格首度回应算力过剩质疑，确认Meta Compute云计算计划。蚂蚁集团开源SGLang高性能推理框架，显示产业链竞争加剧。

监管安全领域出现重要进展。Apple起诉OpenAI涉嫌窃取商业机密，涉及前员工跳槽过程中的机密信息泄露。研究显示博科圣地已系统性利用前沿AI技术辅助作战活动，安全风险进一步凸显。Theory Ventures分析指出AI压缩时间导致风投阶段定义失效，安全攻击面因MCP服务器等因素急剧扩大。

开源工具生态持续活跃。Claude Code发布v2.1.206版本，新增多项实用功能。GitHub分析发现Copilot代码审查中工具统一化反而导致性能下降，通过重写指令实现20%成本降低。多个开源项目获得高关注度，显示开发者社区对AI工具链的持续投入。

头条

今日头条

1. Google推出LiteRT.js高性能Web AI推理运行时

Google 开发者博客 · 07/10 05:21 · 大厂与产业链

Google发布LiteRT.js，这是LiteRT跨平台边缘AI运行时的最新成员，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。

2. Claude Code v2.1.206发布

Claude Code: GitHub Releases · 07/11 08:00 · 开源项目与开发者工具

Claude Code v2.1.206发布，主要更新包括为cd命令添加目录路径建议，新增/doctor检查以建议修剪CLAUDE.md文件中模型可从代码库推导的内容，/commit-push-pr现在自动允许git push到仓库配置的推送远程仓库，/login支持Anthropic运营的公共网关端点，后台智能体在更新后自动升级。

3. ChatGPT Sites将创意变可发布网站

X: OpenAI Developers (@OpenAIDevs) · 07/10 05:27 · 前沿模型与多模态

OpenAI团队成员展示了Sites功能，能够将一个想法变成可发布和分享的实时网站。prd_008用Sites构建了个人专注应用示例，显示了从概念到可部署网站的完整流程能力。

4. GPT-5.6系列模型发布：Sol、Terra、Luna

OpenAI · 07/09 18:00 · 前沿模型与多模态

OpenAI 推出 GPT-5.6 系列，包括旗舰 Sol、平衡型 Terra 和成本最优 Luna。在 Agents' Last Exam 上，Sol 以 53.6 分超越 Claude Fable 5（自适应推理）13.1 分；中等推理时以约四分之一成本领先 11.4 分。Terra 和 Luna 以约十六分之一成本超越 Fable 5。在 Artificial Analysis Coding Agent Index 上，Sol 以 80 分创 SOTA，高于 Fable 5 的 2.8 分，使用不到一半输出 token、一半时间，成本低约三分之一。引入 Programmatic Tool Calling 减少中间数据往返，ultra 模式协调多个智能体加速复杂任务。模型配备迄今最强安全防护，经人类红队和自动化测试验证。

5. Google Research推出SensorFM：面向可穿戴健康数据的通用基础模型

Google Research: Blog · 07/09 00:00 · 智能体、编程与 workflows

Google Research发布SensorFM，一个在超过100万亿分钟多模态传感器数据上预训练的大规模基础模型，数据来自500万同意参与者，涵盖100多个国家及20余款Fitbit和Pixel Watch设备。SensorFM可迁移至心血管、代谢、睡眠、心理健康及生活方式等35项健康预测任务。

模型

前沿模型与多模态

1. Meta 发布 Muse Spark 1.1 模型

X: AI at Meta (@AlatMeta) · 07/09 22:00

据X: AI at Meta (@AlatMeta)，Meta 发布 Muse Spark 1.1 模型。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。

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. Thinking Machines Lab：构建延伸人类意志与判断的 AI

Thinking Machines Lab：官方博客 · 07/11 08:00

Thinking Machines Lab 在官方博客中阐述其使命：构建能够延伸人类意志与判断的 AI。文章指出，当前多数 AI 在少数地方训练后便冻结，无法被使用者塑造。该实验室正致力于训练具备多模态交互和可定制化能力的强模型，开发允许用户训练模型权重的工具，并构建拓宽人机沟通渠道的界面。其核心理念是让 AI 服务于分布式的人类知识，使每个组织都能利用自身独特知识微调模型，并持续适应知识演变。

智能体

智能体、编程与工作流

1. 微软发布Flint：面向AI智能体的可视化语言

Hacker News · 07/10 05:46

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，可渲染到Vega-Lite、ECharts和Chart.js三个后端。项目通过npm安装（TypeScript/JavaScript），并提供MCP服务器用于智能体工作流集成。采用弹性布局模型自动优化图表尺寸与间距，已开源。

2. Claude Code桌面版新增应用内浏览器

X: Claude Devs (@ClaudeDevs) · 07/11 08:00

Claude Code 桌面版现在有了应用内浏览器。Claude 可以调出文档、设计稿或任何其他网站。它可以像操作本地开发服务器一样，进行阅读、点击浏览和交互。该浏览器采用沙盒机制且可配置：你可以自行选择会话是否持久保留。

3. Elon Musk 转发用户称赞 Grok Build 的反馈

X: Elon Musk (@elonmusk, xAI) · 07/11 08:00

Elon Musk 转发用户 @0x0funky 对 Grok Build 的称赞。该用户称 Grok Build 是目前唯一集大成的 coding agentic workflow，内建图像生成和图片生视频功能，生图速度快且品质不输 Codex。Agent 可直接完成图像与视频生成，无需额外串接 MCP 或外部服务。用户结合 Agent Sprite Forge 工具，利用 Grok Build 的视频生成能力，先产出 6 秒角色动作视频再反编译为 game sprite，大幅减少对齐问题，制作 2D 横版游戏耗时不到 30 分钟，而此前用 Codex 需 1-2 小时且品质更优。

4. 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。

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 · 07/09 18:00

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. 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。

3. Pulpie：用于清理网络的Pareto最优模型

Hacker News · 07/08 09:33

Pulpie是一族Pareto最优模型，用于从HTML页面提取主要内容。其最小模型pulpie-orange-small（210M参数）在WebMainBench上取得0.862的ROUGE-5 F1分数，接近600M参数的Dripper（0.864），但成本仅1/20。在NVIDIA L4 GPU上，Pulpie处理速度13.7页/秒，Dripper仅0.68页/秒。清理10亿页HTML，Pulpie成本约\$7，900，Dripper需\$159，000。模型采用编码器架构，单次前向传播即可标记每个HTML块为内容或模板，已在HuggingFace开源。

4. 扎克伯格首度回应 Meta “算力过剩”：没人会嫌算力太多，但租出去更赚钱

IT之家 · 07/11 08:00

Meta CEO 扎克伯格首次正面回应公司筹划云基础设施业务一事，否认“算力过剩”猜测，称内部算力需求依然旺盛、满负荷运转。但他同时表示，当前市场对算力出价极高，将部分 AI 基础设施对外出租在财务上更划算。Meta 正制定代号“Meta Compute”的云计算计划，包括开放模型访问权限和直接出租裸算力两条路线。扎克伯格还提及 SpaceX 以每月 12.5 亿美元将数据中心租给 Anthropic 的模式。Meta 2026 年资本支出指引达 1250 亿至 1450 亿美元，计划 2026 年 9 月量产自研 AI 芯片，目标 2027 年部署算力提升至 14 吉瓦。

1. 蚂蚁集团开源高性能大模型推理框架 SGLang

蚂蚁 inclusionAI · GitHub 新仓库 · 07/11 08:00

蚂蚁集团通过 GitHub 新仓库 inclusionAI/sglang 开源了 SGLang，这是一个面向大语言模型和多模态模型的高性能推理服务框架。

2. GitHub 发布开源工具链新进展

GitHub · 07/11 06:01

据GitHub，GitHub 发布开源工具链新进展。这类进展主要影响开发者工具链、本地部署和研发试用。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. GitHub Copilot 代码审查改用共享工具后性能下降，通过重写指令实现 20% 成本降低

GitHub Blog · 07/11 08:00

GitHub 在 Copilot 代码审查中尝试用 Copilot CLI 的共享代码探索工具（grep、glob、view）替换原有专用工具，结果导致审查成本上升、有效评论数量下降。分析 trace 发现，问题不在工具本身，而在于指令让智能体像通用编程助手一样大范围浏览仓库，而非像审查者一样从 diff 出发进行定向搜索。重写指令后，审查平均成本降低约 20%，同时保持相同审查质量。

4. GitHub 发布开源工具链新进展

GitHub · 07/10 07:56

据GitHub，GitHub 发布开源工具链新进展。这类进展主要影响开发者工具链、本地部署和研发试用。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。

1. Show 发布 AI 行业观点与判断

Hacker News · 07/10 01:36

据Hacker News，Show 发布 AI 行业观点与判断。这类信息反映行业参与者对技术路线和商业节奏的判断。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. Show 发布 AI 行业观点与判断

Hacker News · 07/10 03:21

据Hacker News，Show 发布 AI 行业观点与判断。这类信息反映行业参与者对技术路线和商业节奏的判断。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. Ask 发布 AI 行业观点与判断

Hacker News · 07/09 01:48

据Hacker News，Ask 发布 AI 行业观点与判断。这类信息反映行业参与者对技术路线和商业节奏的判断。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。

1. Theory Ventures 三周年：AI 如何重塑软件栈与风投定义

Tomer Tunguz 博客 (VC 分析) · 07/11 08:00

Theory Ventures 合伙人 Tomer Tunguz 发文总结 AI 带来的市场巨变。新模型每 41 天发布一次，公司达 1 亿美元收入速度创纪录。AI 压缩时间导致风投阶段定义失效，种子轮规模从 100 万到 5 亿美元不等。推理正取代模型成为 AI 主导市场，基础设施按视频、批处理、本地、智能体等负载类型专业化细分。推理的高成本催生了 AI 广告补贴模式，Koah 的 AI 对话原生广告点击率是展示广告基线的 4-5 倍。闭源与开源、云端与本地模型差距缩小，许多模型已“足够好”。安全攻击面因 MCP 服务器、技能、插件和编码智能体而急剧扩大。

2. 博科圣地如何利用前沿AI技术

Hacker News · 07/11 08:00

2025至2026年间对尼日利亚东北部27名前“博科圣地”成员的半结构化访谈揭示了该组织在2024年系统性地利用前沿AI技术。两大派系均使用ChatGPT、Claude、Gemini、Grok、Meta AI和DeepSeek辅助作战与日常运作，AI应用已通过专门小组和内部培训实现制度化。成员成功绕过部分安全限制，将AI用于袭击策划、武器故障排查及爆炸装置设计。相关技术通过跨国圣战网络传播，伊斯兰国特工提供了面对面培训。受访者对AI表现出强烈热情，部分人对大规模杀伤性武器持开放态度，但记录在案的使用仍限于常规手段。

3. Apple 起诉 OpenAI 窃取商业机密

TechCrunch AI · 07/11 08:00

Apple 于周五向美国加州北区联邦法院提起诉讼，指控 OpenAI 窃取商业机密并违反合同。诉状称，OpenAI 高级领导层（包括首席硬件官 Tang Tan）指使前 Apple 员工在招聘过程中窃取机密，包括使用未发布产品的项目代号、要求应聘者携带硬件组件参加面试。Apple 还指控前高级系统电气工程师 Chang Liu 在 2026 年离职加入 OpenAI 后未归还公司笔记本电脑，并用其下载机密技术文档。Apple 称其机密信息已被 OpenAI 用于开发自有硬件产品，例如一项专有金属精加工技术。Apple 要求法院禁止 OpenAI 使用或披露其商业机密，并归还相关材料。

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。