NVIDIA加速MoE微调 OpenAI推出推理芯片

中文内部早报 · 5分钟版

今日总览

今日AI产业主线聚焦硬件加速与智能体部署。NVIDIA发布NeMo AutoModel显著提升MoE模型微调效率，OpenAI与Broadcom联合推出Jalapeño推理芯片。同时，AI智能体在企业应用中快速普及，Codex在ChatGPT移动端正式可用。

Agent与Coding领域迎来重要进展。General Intuition完成3.2亿美元融资，利用游戏数据训练通用AI智能体；Runway发布Agent 2.0服务营销场景；Notion通过Cursor SDK嵌入编码智能体。OpenAI内部报告显示Codex已成为主要工作工具。

模型与多模态方面，OLMo Hybrid架构在实义词预测上展现优势，但重复短语处理能力有限。Runway推出Seedance 4K等三款新模型，OpenAI更新GPT-5.5 Instant提升对话质量。Google Finance也推出升级版本。

大厂与产业链动态频繁。IBM发布全球首款亚纳米级芯片技术，Meta加速AI内容审核部署引发员工担忧。Google Research提出线性弹性缓存优化云经济。OpenAI计划建设千兆瓦级数据中心，显示算力需求持续增长。

监管安全与研究工具板块关注模型政治倾向和开源保护。华盛顿邮报调查显示主流AI聊天机器人普遍左倾，GitHub联合开源联盟呼吁修改加州AI透明度法案以保护开源模式。Google Gemini 3.5 Flash引入computer use功能增强安全性。

头条 今日头条

1. NVIDIA NeMo AutoModel：一行代码加速Transformer MoE模型微调

Hugging Face 博客 · 06/25 00:00 · 大厂与产业链

NVIDIA发布基于Transformers v5的NeMo AutoModel开源库，集成Expert Parallelism、DeepEP融合all-to-all调度和TransformerEngine内核。在MoE模型微调中，相比原生v5训练吞吐量提升3.4-3.7倍，GPU内存减少29-32%，仅需改动一行import。在128张H100上全微调Nemotron 3 Ultra 550B时，v5因内存不足无法运行而AutoModel凭借EP=64使训练可行。

2. Codex在ChatGPT移动App正式可用

X：OpenAI Developers (@OpenAIDevs) · 06/26 08:00 · 监管、安全与版权

OpenAI宣布Codex在ChatGPT移动应用中正式开放，新增一对一设备配对实现更安全的手机与电脑连接。移动端新增通知、目标、侧边聊天、文件预览及内联审阅评论功能。用户可通过ChatGPT移动App启动新工作、审查输出、引导执行和批准下一步，而Codex实际继续在笔记本、Mac mini或开发机上后台运行。

3. General Intuition完成3.2亿美元融资用游戏数据训练通用AI智能体

TechCrunch AI · 06/26 08:00 · 智能体、编程与工作流

General Intuition以23亿美元估值完成3.2亿美元融资，累计披露融资4.54亿美元。公司从旗下游戏剪辑平台Medal获取数亿小时含精确按键动作标签的游戏操作数据，训练单一模型同时驾驭Fortnite等虚拟环境和四足机器人。演示中AI智能体在游戏中连续运行100小时，机器人仅靠8分钟真实街道数据微调即可自主探索办公室。

4. Claude Code v2.1.193发布

Claude Code：GitHub Releases · 06/26 08:00 · 开源项目与开发者工具

Claude Code v2.1.193新增autoMode.classifyAllShell设置，将全部Bash/PowerShell命令经自动模式分类器处理。自动模式拒绝原因现加入转录、拒绝提示及/permissions页面。新增claude_code.assistant_response OpenTelemetry日志事件。Bash模式支持实时文件路径自动补全，MCP服务器需认证时显示启动提示，新增空闲后台shell命令自动内存压力回收功能。

5. Runway发布Agent 2.0

Runway 新闻 · 06/26 08:00 · 智能体、编程与工作流

Runway发布Agent 2.0帮助营销人员创建、测试和优化广告、视频及营销活动。品牌营销人员可在对话中开发活动概念、生成变体并自动本地化；绩效营销人员可上传创意并导入Meta、YouTube、TikTok或Google广告数据，由Agent分析后生成下一轮待测广告。社交媒体营销人员可一次性生成一周内容，自动裁切为不同格式。

模型 前沿模型与多模态

1. OLMo Hybrid vs Transformer：混合模型在实义词上优势明显，但重复短语上几无优势

Hugging Face 博客 · 06/26 08:00

通过对比7B参数的OLMo 3（Transformer）与OLMo Hybrid（混合架构），实验发现混合模型在大多数token上预测损失更低：对名词、动词、形容词等实义词优势明显（loss gap约0.04），功能词上gap约0.02，且在需上下文推理的代词指代上更好。但在重复出现的n-gram和闭合括号（如`}`）上，混合模型的优势几乎消失，Transformer凭借注意力机制更擅长从输入中直接检索精确信息。

2. Runway推出Seedance 4K等三款新模型

X：Runway (@runwayml) · 06/24 05:36

Seedance 4K。Seedance Mini。Kling 3.0 Turbo。现已推出。 全球最佳模型，汇聚一处。 使用优惠码 30RUNWAY，前三个月可享七折优惠。 通过下方链接开始使用。

3. GPT-5.5 Instant 新版本，对话更有趣

X：OpenAI (@OpenAI) · 06/25 02:00

我们为你带来了新版 GPT-5.5 Instant，它现在聊起天来有趣多了。 我们最常用的模型现在能更好地理解问题背后的意图，并相应地调整回应。 它也能更可靠地处理复杂约束，让购物和本地推荐更加实用和连贯。 今天向付费用户推送，明天向免费用户推送。

4. Our latest Google Finance upgrades, including a new app

Google AI Blog · 06/26 00:00

Our latest Google Finance upgrades, including a new app。

智能体 智能体、编程与工作流

1. OpenAI内部报告：智能体Codex如何改变工作

OpenAI · 06/26 08:00

OpenAI 在2025年8月至2026年6月间观察到，智能体产品 Codex 取代 ChatGPT 成为主要工作工具，各部门输出 token 中 Codex 占比从不足10%升至99.8%。80.6%个体用户曾发起预计等效人类工作时间超30分钟的请求，70.2%超1小时，25.6%超8小时；99百分位用户每日生成超60小时 agent turns。非开发者用户增长迅猛：个体用户增长137倍，组织用户增长189倍。Legal、Finance、Recruiting 部门在2026年4月前后跨过 Codex 使用过半拐点，平均每位律师或招聘人员超85%输出 token 来自 Codex。

2. Notion 使用 Cursor SDK 嵌入编码智能体

Cursor Blog · 06/25 04:55

Notion 通过 Cursor SDK 在数周内将编码智能体嵌入产品。用户可在文档中@Cursor、在讨论串中提及或向数据库指派任务，Cursor 即可端到端完成规划、构建、测试、验证并自动创建 PR。集成基于一套 Provider 无关的智能体框架，Notion 的讨论串对应一个 Cursor 智能体，每条消息对应一次智能体运行；结果通过 SSE 流式传输，支持断连恢复。Cursor SDK 提供与生产环境相同的模型、运行时和远程 MCP 支持，让 Notion 无需自建智能体基础设施即可获得完整栈编码能力。用户还可自定义模板、MCP 服务器、技能和子智能体，并设置自动触发规则。

3. Figma在Config 2026押注人类判断，画布AI能力却来自第三方

The Decoder：AI News · 06/25 00:49

Figma在Config 2026将设计画布扩展至代码、动画、3D深度和着色器效果，并集成去年收购的Weave工作流系统。新功能包括Code Layers（代码与设计并存）、Motion动画、深度层、Shader及Generative Plugins。协作方面，团队可搜索复用AI提示词、保存工作流为技能、共享插件。Figma的AI功能依赖Anthropic、OpenAI和Google等外部模型，推理成本挤压利润率。同时，Anthropic等公司的竞争产品可直接生成界面，构成威胁。

4. Sail Research 构建集群感知编排，加速异步推理

Tomer Tunguz 博客（VC 分析） · 06/26 08:00

推理市场是软件中最大的市场。AI工作负载正从同步聊天转向异步、多轮智能体，运行时长可达数小时。Sail Research 为此构建了集群感知（fleet-aware）编排系统，以最大化每美元推理支出的吞吐量。

产业 大厂与产业链

1. OpenAI 与 Broadcom 联合发布 LLM 推理芯片 Jalapeño

OpenAI · 06/24 14:00

OpenAI 与 Broadcom 发布首款自研推理加速器 Jalapeño，专为当前及未来 LLM 从头设计。早期测试显示，其性能功耗比大幅优于现有 SOTA。工程样片已在实验室以目标频率和功耗运行 GPT-5.3-Codex-Spark 等负载。芯片从设计到流片仅用 9 个月，并利用 OpenAI 模型加速部分流程。OpenAI 计划从 2026 年起与 Microsoft 等合作伙伴部署千兆瓦级数据中心，推出多代计算平台。

2. IBM 首度推出亚纳米级芯片技术

Hacker News · 06/26 08:00

IBM 于 2026 年 6 月 25 日发布全球首款亚纳米级芯片技术，采用 0.7 nm（7 埃米）节点与全新三维纳米堆叠（nanostack）架构。指甲盖大小的芯片集成近 1000 亿个晶体管，密度约为 IBM 2021 年 2 nm 芯片的两倍。相比 2 nm 芯片，性能最高提升 50%，能效最高提升 70%。纳米堆叠架构还实现 SRAM 面积缩减 40%，有助于支撑先进 AI 工作负载的高带宽需求。该技术已在 VLSI 2026 会议上验证，IBM 预计 5 年内量产。

3. Meta员工警告AI内容审核部署过快

The Decoder：AI News · 06/26 08:00

Meta在2025年已用大语言模型替换约一半人工审核请求，计划年底前将部分内容类型的AI审核比例提升至90%以上，每年节省数十亿美元。Meta否认成本动机，称自3月测试显示其模型错误率比人类低13%，且多捕捉10%违规。但员工指出模型仍会移除或限流无害内容，缺乏足够监督，快速部署已导致外包裁员。此外，Meta已从使用Google Gemini转向自家新基础模型Muse Spark，该模型基于人工审核员的历史决策训练。

4. 用线性弹性缓存优化云经济

Google Research：Blog · 06/26 08:00

Google Research 与 Google Cloud 提出线性弹性缓存，将缓存管理转为线性成本优化问题，动态调整大小以最小化总拥有成本。为每条数据引入“滑雪租赁”决策框架，在租用内存（持续付费）与购买缺失（缓存未命中惩罚）间选择，并用轻量级机器学习实时优化内存占用与缺失率权衡。无服务器云场景下（每 GiB 内存每天 $3），该技术可在不牺牲性能的同时显著降本。论文发表于 CIDR。

开源 开源项目与开发者工具

1. 跨模型与任务的 GitHub Copilot agentic harness 性能与效率评估

GitHub Blog · 06/26 08:00

GitHub Copilot agentic harness 在多个基准测试中表现强劲，同时具备领先的 token 效率，并支持在 20 多个模型间灵活选择。

2. Generative AI Fizzle™：生成式AI泡沫正在缓慢消退

Gary Marcus：The Road to AI We Can Trust · 06/26 08:00

Gary Marcus 昨日提出新术语 Generative AI Fizzle™，认为生成式AI行业估值过高，投资者对 hype 与利润的落差失去热情。LLM 已商品化，价格战激烈，提供商盈利艰难。昨日一款新的中国开源模型发布，可能进一步冲击美国 LLM 公司。多数 AI 股票本月显著下跌，泡沫可能不会突然破裂，而是缓慢消退。

3. GitHub联合开源联盟呼吁修改加州AI透明度法案以保护开源

GitHub Blog · 06/23 23:48

GitHub 联合 Black Forest Labs、Hugging Face 与 Mozilla Corporation 组成开源联盟，呼吁对加州 AI 透明度法案（SB 942，拟由 SB 1000 修正）进行针对性修改。当前草案要求开发者在下游用户未履行义务时撤销开源许可证，这与开源许可证永久不可撤销的性质冲突。联盟认为该要求非必要，已有直接监管和执法机制，并建议参考欧盟 AI 法案的透明度实践规范，以向下游用户通知最佳实践文档的方式替代撤销条款。GitHub 支持这些修正，以在保持透明度目标的同时兼容开源开发模式。

4. 无限制OCR：单次长时域解析

Hacker News · 06/23 21:32

Unlimited OCR 是一个托管在 GitHub 的项目，实现单次长时域解析（One-Shot Long-Horizon Parsing），旨在一次性处理长时间跨度的 OCR 任务。

观点 KOL 与巨头言论

1. 多数主流AI聊天机器人政治立场偏左，“反觉醒”模型也不例外

The Decoder：AI News · 06/26 08:00

华盛顿邮报调查显示，多数主流AI聊天机器人在政治问题上明显偏左。OpenAI GPT-5.5在80%回答中仅呈现左派论据；DeepSeek V4 Pro为70%；Anthropic Claude Opus 4.8有43%纯左、57%给出双方观点。xAI的Grok 4.3左倾回答仍多于右倾。右翼平台Gab的Arya左倾回答是右倾的12倍。Google Gemini 3.1 Pro是例外，93%回答同时呈现双方立场。特朗普推动的“反觉醒”AI未能改变这一格局。

监管 监管、安全与版权

1. Gemini 3.5 Flash 引入 computer use 功能

Google DeepMind：Blog · 06/25 00:30

Google DeepMind 宣布，computer use 现作为内置工具集成于 Gemini 3.5 Flash，开发者可构建跨浏览器、移动端和桌面的智能体，实现视觉感知、推理与操作。此前该功能仅以独立模型形式存在于 Gemini 2.5。3.5 Flash 已支持函数调用及 Search、Maps 等内置工具，新增的 computer use 可提升持续软件测试和跨专业应用知识工作等长周期企业自动化任务的性能。安全方面采用针对性对抗训练，并可选配两项企业防护系统：要求用户确认敏感操作，以及在检测到间接 prompt 注入时自动停止任务。可通过 Gemini API 和 Gemini Enterprise Agent Platform 使用。

2. Meta 隐私感知基础设施的资产分类：混合模式将 LLM 蒸馏为确定性规则

Meta Engineering Blog · 06/26 08:00

Meta 在 Privacy-Aware Infrastructure (PAI) 的资产分类中采用混合模式：先构建含代码、血缘、语义标注的上下文证据，再调用 LLM 处理歧义、冷启动和新颖资产；人工审核标签与模型推荐严格隔离。LLM 不直接做生产决策，其稳定行为被蒸馏为版本化确定性规则用于生产执行，LLM 角色随规则积累逐步缩小。核心原则：上下文比提示词更重要、解耦评估与优化、将稳定行为规则化。

3. Mistral AI 为 Connectors 推出多项安全与可控新能力

Mistral AI：News · 06/24 23:59

2026年6月24日，Mistral AI 发布 Connectors 多项新能力：Enriched admin controls（GA）支持按工作空间设置连接器访问权限并单独开关工具；API keys with connector scopes（GA）防止自动化 AI 工作负载中身份冒充；Multi-account connectors（GA）允许单个连接器绑定多个账户；Connectors Debugger（公开预览）对 MCP 连接器进行端到端根因分析；Connectors in Vibe Code（GA）和 Connectors in Workflows（公开预览）分别允许在开发者界面复用连接器及支持长时间运行任务不中断。

研究 研究、评测与方法论

1. IBM 开源 CUGA：轻量级智能体框架，提供二十余个单文件示例应用

Hugging Face 博客 · 06/23 20:51

IBM 开源了 CUGA（Configurable Generalist Agent），一个处理规划、执行循环、工具调用和状态管理的轻量级智能体框架。开发者只需提供工具列表和提示词即可构建 CugaAgent。内置计划-执行-反思循环，在 AppWorld（2025年7月-2026年2月）和 WebArena（2025年2月-9月）基准上排名第一。支持 Fast / Balanced / Accurate 三种推理模式，代码执行可在本地、Docker 或 E2B 沙箱中运行。可互换工具支持 OpenAPI、MCP 和 LangChain 函数，通过环境变量一键切换 OpenAI、watsonx、Ollama 等提供商。随框架发布二十余个单文件示例应用，涵盖电影推荐、IBM Cloud 架构顾问等场景，每个应用仅需一个 FastAPI 文件。

2. FFASR 排行榜发布：真实远场条件下 ASR 评测

Hugging Face 博客 · 06/24 08:00

Treble Technologies 与 Hugging Face 联合推出 FFASR（Far-Field ASR）排行榜，这是首个开源社区驱动的真实远场声学条件 ASR 评测基准。传统近场评测无法反映混响、背景噪声和麦克风距离带来的性能下降。FFASR 使用混合波模拟引擎生成声学数据，涵盖 14 种房间（20-470 m3）和三个信噪比级别（远场高 SNR >14 dB、中 SNR 8-12 dB、低 SNR <6 dB），加上近场干燥条件，共四类条件决定主排名。另有实验室实测/模拟验证轨道和移动声源 beta 版。性能指标同时报告词错误率（WER）和实时因子（RTFx，在 NVIDIA L4 GPU 上评估）。未来将支持多说话人场景、麦克风阵列和回声消除。

3. OpenRouter MCP 服务器发布

OpenRouter 公告 · 06/26 08:00

OpenRouter 推出 MCP 服务器，为编程智能体提供实时模型数据、基准排名、定价和文档查询。开发者通过一键安装（支持 Claude Code、Codex CLI、Cursor 等客户端），即可在编辑器内完成模型筛选、价格对比和测试推理，无需切换标签页。服务器整合 Artificial Analysis、Design Arena 及 OpenRouter 自身排名数据，例如推荐 GLM-5.2 作为性价比最佳的编码模型。工具集包括 models-list、model-get、model-endpoints、benchmarks 等，支持通过 chat-send 发送测试提示，比较不同模型（如 Claude Opus 4.8、G