

NVIDIA推出MoE微调加速工具，IBM开源轻量级智能体框架

中文内部早报 · 5分钟版

今日总览

今日AI领域聚焦模型优化与智能体开发两大主线。NVIDIA发布NeMo AutoModel显著提升MoE模型微调效率，IBM开源CUGA框架专注轻量级智能体构建。

Agent与Coding板块活跃，Notion集成Cursor SDK实现编码智能体嵌入，Figma在Config 2026扩展画布AI能力。豆包推出专业版产品，Grok Build新增/goal模式支持长时间任务执行。

模型与多模态方面，OpenAI发布新版GPT-5.5 Instant增强对话体验，ChatGPT语音模型Bidi 1上线测试支持双向交互。Runway推出Seedance 4K等三款新模型，Google将计算机使用功能整合至Gemini 3.5 Flash。

大厂与产业链动态频繁，OpenAI与Broadcom联合推出Jalapeño定制推理芯片。工程岗位展现韧性，数据显示2025年工程仍是最具韧性的岗位类别。Sky Computing Lab发布FastWan-QAD视频生成模型。

监管安全与研究工具板块，Mistral AI为Connectors增加安全控制能力，Google强化Gemini安全防护。研究方面出现AgentCIBench评估智能体隐私保护问题，FFASR排行榜关注真实远场语音识别性能。

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 08:00 · 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，AutoModel凭借EP=64使训练可行而v5因内存不足无法运行。

2. IBM开源CUGA轻量级智能体框架

Hugging Face 博客 · 06/23 20:51 · 研究、评测与方法论

IBM开源CUGA(Configurable Generalist Agent)轻量级智能体框架，处理规划、执行循环、工具调用和状态管理。内置计划-执行-反思循环，在AppWorld和WebArena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。随框架发布二十多个单文件示例应用。

3. Figma在Config 2026扩展画布AI能力

The Decoder: AI News · 06/25 08:00 · 智能体、编程与 workflows

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、Motion动画、深度层、Shader及Generative Plugins。协作方面支持搜索复用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。

4. Notion通过Cursor SDK嵌入编码智能体

Cursor Blog · 06/25 08:00 · 智能体、编程与 workflows

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Provider无关的智能体框架，支持断连恢复，用户可自定义模板、MCP服务器、技能和子智能体。

5. FFASR排行榜发布真实远场语音识别评测

Hugging Face 博客 · 06/25 08:00 · 研究、评测与方法论

Treble Technologies与Hugging Face联合推出FFASR(Far-Field ASR)排行榜，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种房间和三个信噪比级别，同时报告词错误率和实时因子。

模型

前沿模型与多模态

1. 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。

2. GPT-5.5 Instant 新版本，对话更有趣

X: OpenAI (@OpenAI) · 06/25 08:00

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。

3. How GPT-5 helped immunologist Derya Unutmaz solve a 3-year-old mystery

OpenAI 新闻 · 06/24 01:00

How GPT-5 helped immunologist Derya Unutmaz solve a 3-year-old mystery。

4. OpenAI ChatGPT 语音最大规模升级：双向AI语音模型 Bidi 1 已上线测试

IT之家 · 06/25 08:00

6月23日，部分用户反馈 ChatGPT 网页版和 App 版上线了双向 AI 语音模型 Bidi 1，位于设置模型选择器中，与标准语音和高级语音并列。该模型支持边说话边监听，用户可在对话中途打断并发出新指令，例如要求从1数到10时中途喊停倒数，模型会立即切换执行。OpenAI 尚未官宣，预计本周启动更大范围测试。

智能体 智能体、编程与工作流

1. Show HN：Oak - 面向 AI 智能体的 Git 替代方案

Hacker News · 06/23 05:00

Oak 是面向 AI 智能体的开源版本控制系统，服务 Claude Code、Codex、Cursor 等工具。它以分支-会话为基本工作单元，使用内容哈希、分块、diff/merge 和 Blob/Manifest/Commit/Tree 数据模型，支持 SQLite 和 git 后端。

2. huggingface_hub 实现每周发布：AI、开源工具、人工审核闭环

Hugging Face 博客 · 06/23 08:00

Hugging Face 将 huggingface_hub 的发布周期从每 4-6 周缩短至每周，全部由单个 GitHub Actions 工作流自动完成。流程依赖开源工具和开权重模型（当前为 Z.ai 的 GLM-5.2）来起草发布说明和 Slack 公告，但保留人类在最终审核环节的决策权。自动步骤包括版本号更新、提交标签推送、PyPI 发布、下游测试分支创建、发布说明草稿、Slack 公告草稿、归档、后置版本提升以及对合入 PR 的评论。所有组件均基于开源生态构建，任何维护者都可直接复制使用。

3. Grok Build 推出 /goal 模式

xAI 新闻 · 06/22 08:00

xAI 在 Grok Build 中引入 /goal 模式。用户用一行命令设定目标后，agent 会自动规划方案、拆解任务并持续执行，直到目标完成并通过验证；期间用户仍可追加指令和监督任务进展。

4. 今天，豆包正式推出专业版

公众号：豆包（字节） · 06/25 08:00

豆包专业版基于豆包2.1系列大模型上线，面向复杂办公与生产力场景。办公任务模式接入可执行Agent任务的豆包2.1模型，支持操作本地电脑、浏览器、调用Skills技能、定时任务，内置Office办公套件，并可生成带后端数据库的在线应用。免费用户可体验豆包2.1 Turbo版办公任务模式，专业版接入豆包2.1 Pro模型。定价：标准套餐68元/月（连续包月），加强套餐200元/月，高级套餐500元/月。大学生认证后标准套餐38元/月，持续6个月。

产业 大厂与产业链

1. OpenAI 与 Broadcom 发布面向 LLM 推理的定制芯片 Jalapeño

OpenAI · 06/25 08:00

OpenAI 与 Broadcom 联合推出 Jalapeño，一款专为大语言模型（LLM）推理优化的定制 AI 芯片，旨在提升 AI 系统的性能、效率与规模。

2. 我们用免费本地模型对 OpenClaw 仓库进行实时分类

Hugging Face 博客 · 06/22 08:00

Hugging Face 在 OpenClaw 仓库上测试用 Gemma 和 Qwen 等本地模型实时分类 issue 和 PR。他们使用 Pi agent harness 驱动模型，配合 reposhell 只允许读操作防止提示词注入。测试的模型包括 gemma-4-26b-a4b 和 qwen3.6-35b-a3b，经性能优化后均可在本地生成数百 token/s。该方案运行在 NVIDIA GB10（128 GB 统一内存）上，相比每月 200 美元的 ChatGPT Pro 订阅，可实现近乎实时的通知且仅消耗电费。

3. AI被认为会取代工程岗位，但新数据显示工程是2025年最具韧性的岗位

TechCrunch AI · 06/25 08:00

风投机构SignalFire追踪8000万家公司数百万员工数据发现，工程是2025年最具韧性的岗位。大型科技公司总招聘较2019年下降25%，工程岗仅降11%；工程岗占 Alphabet、Meta等12家“Tech Majors”新招员工的55%（2019年为46%）。早期初创公司2025年工程师招聘比2019年增长7%。Anthropic CEO警告AI可能消灭一半入门级白领，但该公司经济主管称尚未看到显著影响。NVIDIA CEO黄仁勋表示AI让工程师更忙碌，是杰文斯悖论的典型例证。

4. FastWan-QAD：单卡5090上1.8秒生成5秒视频

X: Sky Computing Lab (@haoailab) · 06/24 02:52

Sky Computing Lab 发布 FastWan-QAD 视频生成模型系列，基于 FastVideo 的量化感知蒸馏（QAD）方案训练。在单张 NVIDIA GeForce RTX 5090 上，端到端生成一段 5 秒 480P 视频仅需 1.8 秒。模型、代码及博客已开源。

开源 开源项目与开发者工具

1. 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。

2. 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。

3. benchflow-ai/awesome-evals

GitHub · 06/24 16:10

benchflow-ai/awesome-evals。

监管 监管、安全与版权

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

Mistral AI: News · 06/25 08:00

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。

2. Gemini 3.5 Flash 中的计算机使用

Hacker News · 06/25 08:00

Google 将计算机使用（Computer use）作为内置工具集成至 Gemini 3.5 Flash，使开发者能构建跨浏览器、移动端和桌面环境的智能体。此前该功能仅作为独立模型在 Gemini 2.5 中提供，现已原生整合至主 Flash 模型。开发者可通过 Gemini API 及 Gemini Enterprise Agent Platform 调用。安全方面，模型采用针对性对抗训练降低提示注入风险，并新增两项可选企业级保护：要求用户确认敏感操作、检测到间接提示注入时自动停止。该能力在持续软件测试、跨应用知识工作等长周期企业自动化场景中表现更优。（198字）

3. OpenAI 发布 Daybreak 安全工具

OpenAI · 06/22 18:00

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。

研究 研究、评测与方法论

1. 能力强但粗心：计算机使用智能体是否遵循情境完整性？

HuggingFace Daily Papers（社区热门论文） · 06/22 08:00

AgentCIBench评估计算机使用智能体（CUA）是否遵循情境完整性。它针对三种常见失败模式：视觉共置（智能体拉取任务目标旁边被禁止的项目）、任务模糊性过度分享（在提示不明确时泄露个人状态）以及收件人错配（向不适当的收件人发送内容）。对15个前沿CUA的评测显示平均泄漏率67.9%，其中11个在超过50%的场景中泄漏，这些失败在端到端任务中同样存在。AgentCIBench已发布，旨在推动开发更安全的计算机使用智能体。

2. 思考即回忆：推理如何解锁LLM中的参数化知识

Google Research: Blog · 06/25 08:00

Google Research研究发现，推理（chain-of-thought）能帮助大语言模型（LLM）回忆简单事实，即使这些事实无需复杂推导。在 Gemini-2.5 Flash和Pro以及Qwen3-32B 上，启用推理后模型能够回答原本无法直接回答的简单问题，pass@k显示正确事实存在于输出分布中。该现象由两个机制驱动：一是生成的推理token充当计算缓冲，允许模型进行隐藏计算以提取参数化知识；二是推理过程中产生的相关事实起到启动效应（factual priming），帮助模型激活正确答案。

3. Qwen-AgentWorld 开源：让 Agent 学会“先预测，再行动”

公众号：通义实验室（千问） · 06/25 08:00

通义千问推出首个原生语言世界模型 Qwen-AgentWorld，覆盖 MCP、Search、Terminal、SWE、Web、OS、Android 七大领域。模型基于超 1000 万条真实交互轨迹，经 CPT→SFT→RL 三阶段训练，在 AgentWorldBench 上超越 GPT-5.4（58.25）和 Claude Opus 4.8，Qwen-AgentWorld-397B-A17B 取得 58.71 分。两种范式已验证其能力：作为解耦环境模拟器实现可控 Sim RL，在 WideSearch 上超越真实环境 RL（F1 50.3% vs 45.6%）；作为智能体基础模型，LWM 预热可迁移至七个基准（三个完全未出现在训练