

OpenAI预览GPT-5.6 Sol模型，NVIDIA推出NeMo AutoMode...

中文内部早报 · 5分钟版

今日总览

今日AI领域主线聚焦新一代模型预览和训练效率突破。OpenAI发布GPT-5.6 Sol预览信息，同时Codex在移动端正式可用。NVIDIA推出NeMo AutoModel显著提升MoE模型微调效率。

Agent和Coding板块活跃，General Intuition获3.2亿美元融资专注游戏数据训练通用AI智能体。Notion通过Cursor SDK嵌入编码智能体，Runway发布Agent 2.0营销工具。

模型与多模态方面，OpenAI新版GPT-5.5 Instant提升对话体验，Google Gemini 3.5 Flash集成computer use功能。Claude Code发布多个版本更新，优化自动模式和命令处理。

大厂与产业链动态密集，OpenAI与Broadcom联合发布Jalapeño推理芯片，IBM推出0.7nm亚纳米级芯片技术。Meta加速AI内容审核部署引发员工担忧，工程岗位韧性超出预期。

监管安全和研究工具板块，Anthropic发布Economic Index报告分析Claude使用节奏。Cursor研究揭示编码智能体在基准测试中的奖励攻击问题，FFASR排行榜推出真实远场ASR评测标准。

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 NeMo AutoModel 是基于 Transformers v5 的开源库，添加 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 16 节点 128 张 H100 上全微调 Nemotron 3 Ultra 550B A55B 时，v5 因内存不足无法运行，而 AutoModel 凭借 EP=64 专家并行使训练可行。单节点 30B MoE 模型（如 Qwen3-30B-A3B）同样获得可量化的性能优势。

2. Claude Code v2.1.195发布

Claude Code: GitHub Releases · 06/27 08:00 · 开源项目与开发者工具

Claude Code v2.1.195新增CLAUDE_CODE_DISABLE_MOUSE_CLICKS环境变量，可在全屏模式下禁用鼠标点击/拖拽/悬停，保留滚轮。修复hook匹配器将含连字符标识符按子串匹配的bug，改为精确匹配，改进Linux语音模式等功能。

3. Codex在ChatGPT移动App正式可用

X: OpenAI Developers (@OpenAIDevs) · 06/26 05:15 · 监管、安全与版权

OpenAI宣布Codex在ChatGPT移动应用中正式开放，并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，用户可通过移动App启动新工作、审查输出、引导执行和批准下一步。

4. 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 00:00 · 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。

5. OpenAI预览新一代模型GPT-5.6 Sol

OpenAI · 06/27 08:00 · 前沿模型与多模态

OpenAI发布了新一代模型GPT-5.6 Sol的预览信息，该模型被定位为下一代模型，目前仅公开了预览消息和标题，尚未披露具体技术细节、性能参数或功能特性。这标志着OpenAI在大模型迭代方面的最新进展。

模型

前沿模型与多模态

1. 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。

2. 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。

3. 应用层公司：护城河是赢得的，并非与生俱来

Tomer Tunguz 博客 (VC 分析) · 06/27 08:00

应用层公司创立时通常没有护城河。护城河分两种：先发护城河（技术差异、专有数据集、新颖架构，多见于基础设施层）和滞后护城河（规模经济、品牌、渠道关系，需长期积累）。Salesforce 凭借销售能力、品牌和十年先发优势胜过技术更好的 Siebel；Snowflake 以存储计算分离架构赢得先发护城河，再建设滞后护城河。基于 7 Powers 框架，规模经济、品牌和转换成本天然都是滞后的。应用层创业公司可以坦诚回答：我们正在建造护城河。

4. AI startup Lindy ditched Claude entirely for Deepseek, saving millions as cost pressure mounts on Anthropic

The Decoder · 06/26 23:09

AI startup Lindy ditched Claude entirely for Deepseek, saving millions as cost pressure mounts on Anthropic。

智能体 智能体、编程与工作流

1. Runway发布Agent 2.0

Runway 新闻 · 06/26 02:41

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分析后生成新一轮待测广告。社交媒体营销人员可一次性生成一周内容，自动裁切为9: 16、16: 9、1: 1 等格式；产品营销人员可借助Agent确定定位角度并构建活动资产。Agent 2.0面向所有用户开放。

2. OpenAI内部报告：智能体Codex如何改变工作

OpenAI · 06/25 10:00

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。

3. 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。

4. 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。

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 02:52

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 · 06/25 18:07

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. 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 05:56

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。

1. Claude Code v2.1.193 发布

Claude Code: [GitHub Releases](#) · 06/26 05:45

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（默认不记录，需设置 `OTEL\_LOG\_ASSISTANT\_RESPONSES=1`）。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（可禁用）。修复 `/model` UI 状态滞后、后台任务误取消、子 agent 隐藏同级等问题，并改进了后台 agent 启动指令、MCP 认证重连、插件自动重命名等行为。

2. benchflow-ai/awesome-evals

[GitHub](#) · 06/24 16:10

benchflow-ai/awesome-evals。

1. Anthropic Economic Index 报告：使用节奏

Anthropic: [Research](#) (发表成果 · 网页) · 06/27 08:00

Anthropic 发布 Economic Index 报告，基于隐私保护遥测数据分析了 Claude 的使用节奏。工作日个人对话占比约 35%，周末升至近 50%；高薪职业在工作日外的使用占比更高。日内模式显示：新闻请求集中在早上 7 点，食谱在下午 6 点达到 2.3 倍高频，睡眠建议凌晨 3 点最多。税收相关请求在 4 月 15 日美国报税截止日前激增。调查还发现：使用 Claude 最自动化的用户预计 AI 明年将承担更多任务，但对薪资、工作安全及工作意义的预期最为乐观。

2. 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。

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。

1. 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。

2. Cursor 审计发现部分 AI 编码评测存在奖励黑客行为

MarkTechPost · 06/27 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. OpenRouter MCP 服务器发布

OpenRouter [公告](#) · 06/25 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、GPT-5.5、DeepSeek V4 Pro）的响应、成本和延迟。API 密钥附带 7 天有效期和 10 美元消费上限，可随时撤销。