

AI模型与智能体技术密集发布：Google DeepMind推出Gemini 3.7...

中文内部早报 · 5分钟版

今日总览

AI模型竞争进入白热化阶段，Google DeepMind和OpenAI在编程与智能体领域密集发布新产品。Google推出面向编程的Gemini 3.7 Flash，定价仅为前代一半，同时在数据库迁移服务中集成AI辅助功能。OpenAI则预览Ultrafast模式，声称GPT-5.6 Sol速度提升最高14倍。

智能体与编程工具生态持续演进。Cursor收购Firetiger团队并推出builds功能，云智能体启动速度提升3倍。Claude Code更新至v2.1.232版本，默认启用subagent forking，新增GitLab支持。GPT-5.6系列模型在ARC-AGI-3测试中表现显著提升。

大厂产业链合作深化，NVIDIA发布Nemotron 3.5 Lightning开源模型，SGLang宣布Day-0支持。Google Cloud在Database Migration Service中集成Gemini驱动的代码转换功能。微软发布首个自研推理模型MAI-Thinking-1，Hugging Face与AWS合作推进机器人智能体循环。

安全与合规成为焦点，Cursor获得AIUC-1认证，通过独立审查验证智能体安全防护机制。Google发布开源C++库Credentio用于C2PA内容凭证验证。GitHub Secure Open Source Fund第四期支持50个开源项目提升AI时代安全性。

前沿模型创新不断，Google Sheets推出Sheets canvas功能，MiniMax发布Music 3.0音乐生成模型。Runway上线Seedance 2.5，支持50个角色参考。OpenAI任命新首席营收官，推动企业AI应用落地。

头条

今日头条

1. Claude Code v2.1.232发布：默认启用Subagent forking，新增GitLab支持与多项安全修复

Claude Code：GitHub Releases

08/14 08:00

开源项目与开发者工具

Claude Code v2.1.232默认启用subagent forking，子代理可继承完整对话与提示缓存，交互会话中的非队友代理默认后台运行。新增GitLab token密钥脱敏、插件市场GitLab仓库克隆支持，并修复PowerShell与Windows权限绕过、嵌套git仓库信任继承等多项安全漏洞。

2. Google DeepMind推出Gemini 3.7 Flash：面向编程与智能体的最快模型

Google DeepMind：Blog

08/14 08:00

智能体、编程与工作流

Google DeepMind发布Gemini 3.7 Flash，距3.6 Flash仅三周，主打编程与智能体任务，输入/输出价格分别为每百万token \$0.75和\$3.75，为原3.6 Flash的一半。该模型针对编程和智能体应用场景进行了专门优化。

3. OpenAI预览Ultrafast模式：GPT-5.6 Sol速度提升最高14倍

OpenAI

08/14 08:00

大厂与产业链

OpenAI推出新的API服务层级Ultrafast，由Cerebras提供算力，运行GPT-5.6 Sol的速度最高提升14倍，输出速率可达每秒750 tokens。该模式目前处于预览阶段，主要面向需要高速响应的应用场景。

4. Google 推出 AI 产业部署新进展

Google Cloud：Databases

08/12 00:00

大厂与产业链

据Google Cloud：Databases，Google 推出 AI 产业部署新进展。这类进展主要影响云算力、硬件投入、企业采购和服务成本。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。

5. NVIDIA推出Nemotron 3.5 Lightning，加速本地智能体任务

NVIDIA Blog

08/11 21:00

大厂与产业链

NVIDIA发布Nemotron 3.5 Lightning，一款可定制的开源30B混合专家模型，专为常驻智能体设计。相比同类开源模型，其token生成速度最高提升4倍，任务完成时间缩短30%。该模型采用开放权重，支持用户微调以匹配特定任务，并可在RTX PC、DGX Spark及Jetson等设备上运行。

模型

前沿模型与多模态

1. Google Sheets 推出 Sheets canvas：用 Gemini 将表格数据变为交互式迷你应用

Google Blog：AI

08/14 08:00

Google Sheets 发布新功能 Sheets canvas，基于 Gemini 构建，用户只需用自然语言提示词即可将表格数据转化为交互式仪表盘、学习追踪器、座位表等“迷你应用”。

2. MiniMax Music 3.0 发布：新一代开源权重、生产级全能音乐模型

MiniMax：Blog

08/14 00:52

MiniMax 推出 Music 3.0，新一代音乐生成模型，可根据创意概念和可选歌词一次性完成整首歌的作曲、编曲、演奏与制作，最长支持五分钟。

3. OpenAI 更新模型能力与产品方案

OpenAI 新闻

08/13 17:00

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. Runway Seedance 2.5 上线，支持50角色参考

X: [Runway \(@runwayml\)](#) · 08/12 02:53

完整阵容，完整曲目，一次生成一个。Seedance 2.5 已在 Runway 上线，支持 50 个独特角色参考，以及最长 30 秒、与音乐同步的片段。点击下方链接即可开始使用。

智能体

智能体、编程与工作流

1. Cursor 更新 AI 工作流与智能体方案

[Cursor Blog](#) · 08/14 08:00

Firetiger 团队正式加入 Cursor。该公司构建面向生产环境的智能体，可监控发布、捕获回归、调查事件并将发现反馈给编码智能体。Firetiger 由 Rustam Lalkaka 和 Achille Roussel 于 2024 年创立，此前曾在 Cloudflare、Twitch、Segment 和 Twilio 构建大型生产系统。

2. Cursor 推出 builds：云智能体启动速度提升至 3 倍

[Cursor Blog](#) · 08/14 08:00

Cursor 推出 builds 功能，在后台持续准备就绪的开发环境副本，让云智能体启动时无需从零搭建，响应速度最高提升 3 倍。内部环境启动快 10 倍，首个 token 生成快 3 倍；智能体始终从最近一次成功的 build 启动，依赖更新或安装脚本出错时不会影响运行。8 月 17 日起所有环境默认启用 builds，无需额外费用。

3. GPT-5.6 构建者指南：如何以更低成本实现前沿智能体性能

[OpenAI](#) · 08/14 08:00

GPT-5.6 模型家族以更低成本实现前沿级智能体性能，并新增推理持久化、原生多智能体编排和程序化工具调用等 API 能力。在 ARC-AGI-3 上，启用保留推理和压缩后，Sol 得分从 13.3% 跃升至 38.3%，且输出 token 减少约 6 倍。Luna 在 BrowseComp 上以 84.04% 的得分追平 GPT-5.5（84.36%），成本从 \$33.27 降至 \$1.33。

4. DeepSeek-V4-Pro 正式版上线，Agent 能力大幅增强

[DeepSeek: API 更新日志](#) · 08/13 19:16

据DeepSeek: API 更新日志，DeepSeek-V4-Pro 正式版上线，Agent 能力大幅增强。这类进展主要影响智能体编排、代码生成和企业工作流落地。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。

产业

大厂与产业链

1. Strands Robots 如何用 Hugging Face Storage Buckets 实现记录、训练与部署一体化

[Hugging Face 博客](#) · 08/14 08:00

AWS 开源的 Strands Robots（Apache 2.0）通过单一智能体循环，将机器人演示记录、基于 Hub 数据流的策略训练及硬件部署整合在一起，全程保持 LeRobot 磁盘格式不变。

2. 微软首发自研推理模型MAI-Thinking-1

X: [Mustafa Suleyman \(Microsoft AI CEO\)](#) (@mustafasuleyman) · 08/13 00:00

我们的首个推理模型 MAI-Thinking-1 从零开始构建，现已在 Microsoft Foundry 上线。为团队点赞！更多详情如下。

3. SGLang 宣布 Day-0 支持 NVIDIA Nemotron 3.5 Lightning

[LMSYS: Blog \(Chatbot Arena 团队\)](#) · 08/11 21:51

SGLang 宣布对 NVIDIA Nemotron 3.5 Lightning 提供 Day-0 支持，该开源模型为 30B 总参数、3B 激活参数的混合专家架构，支持最长 1M token 上下文，可从 Hugging Face 下载 BF16 和 NVFP4 权重。模型支持 MTP、DFlash、DSpark 三种投机解码技术，并可通过 OpenAI 兼容 API 接入智能体工作流。

4. OpenAI 推出 AI 产业部署新进展

[OpenAI 新闻](#) · 08/12 14:00

据OpenAI 新闻，OpenAI 推出 AI 产业部署新进展。这类进展主要影响云算力、硬件投入、企业采购和服务成本。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。

开源

开源项目与开发者工具

1. GitHub Secure Open Source Fund 第四期 50 个开源项目如何提升 AI 时代安全性

[GitHub Blog](#) · 08/14 08:00

GitHub Secure Open Source Fund 第四期 50 个开源项目结合 AI 辅助工作流、维护者经验、GitHub 安全工具、专家指导与资金支持，系统性提升项目安全性。该计划展示了开源生态在 AI 时代应对安全挑战的实践路径，为维护者提供了可复用的安全加固模式。

2. AutoGPT 如何用 AGENTS.md 和技能门控管理 AI 生成的拉取请求

[GitHub Blog](#) · 08/13 02:00

AutoGPT 维护者发现，AI 智能体不会主动阅读文档，因此将指令放在 AGENTS.md 和技能文件中，并置于代码目录旁。他们通过强制 PR 模板、测试计划、CI 覆盖率门槛和 CLA 签名等门控机制，将智能体提交的 PR 从 "不可用" 转变为 "可用但不符合路线图"。其中 CLA 签名因需浏览器和 OAuth 流程，被用作区分人类与智能体的 "人类探测器"。

3. Hugging Face 发布新动态：《What We Learned by Reproducing 2,200 papers from ICML》

[Hugging Face Blog](#) · 08/13 08:00

据Hugging Face Blog，本条原始主题为《What We Learned by Reproducing 2,200 papers from ICML》。该材料涉及开发者工具、模型路由或开源生态，需继续核验代码、许可、维护频率和部署条件。

4. OlmoEarth推出定制嵌入导出功能用于下游分析

[Hugging Face Blog](#) · 08/13 00:14

OlmoEarth Studio推出定制嵌入导出功能，用户可以将自定义嵌入导出用于下游分析任务。该功能来自Allen Institute for AI的研究成果，旨在为地球科学数据分析提供更灵活的工具支持。

观点

KOL 与巨头言论

1. Hacker 发布新动态：《Show HN: LymeScribe – one computer on your network transcribes for the rest》

[Hacker News](#) · 08/14 02:18

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. Hacker 发布新动态：《Show HN: Tried some experiments with architecture for Long term memory for LLM》

[Hacker News](#) · 08/13 04:03

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. Hacker 发布新动态：《Show HN: Pane – An open-source browser that turns your work into living websites》

[Hacker News](#) · 08/13 00:25

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。

监管

监管、安全与版权

1. BigQuery Graph 新增 measures 支持，为智能体工作负载提供可信关系推理

[Google Cloud: Databases](#) · 08/14 08:00

Google Cloud 在 BigQuery Graph（预览版）中引入 measures 支持，将治理指标与关系映射统一，使 AI 智能体能在图结构上基于精确指标进行推理，解决传统表格无法追踪多跳业务关系导致错误决策的问题。

2. Google 发布开源 C++ 库 Credentio，用于 C2PA 内容凭证验证

[Google 开发者博客](#) · 08/14 08:00

Google 发布开源 C++ 库 Credentio，支持在客户端和服务端应用中集成高性能、本地优先的 C2PA 内容凭证验证。该库以优化的内存占用完全本地处理资产，可为数 GB 级媒体文件提供即时验证结果，避免云延迟、带宽成本与数据隐私风险。目前支持深度清单解析与可配置信任列表集成，已在 Google Source 上线，未来计划支持完整的凭证生成与嵌入。

3. Cursor 获得 AIUC-1 认证，通过智能体安全与可靠性独立审查

[Cursor Blog](#) · 08/14 08:00

Cursor 通过独立审查与对抗性测试，正式获得 AIUC-1 认证，该标准由 100 多家财富 500 强 CISO 参与制定，并获 MITRE、云安全联盟及斯坦福研究者的技术支持。测试覆盖 IDE 和云端智能体，涉及规则、hooks 与 Auto-review 等防护机制，Cursor 在数千个场景中通过全部要求。维持认证需至少每季度复测一次，并每年接受全面审计。

研究

研究、评测与方法论

1. JetBrains CTO 谈如何评估并部署 Claude Fable 5：私有仓库评测、效率提升与安全策略

[Claude 博客](#) · 08/14 08:00

JetBrains CTO Vladislav Tankov 详解其团队如何用私有仓库评测前沿模型，并决定何时采用 Claude Fable 5。该模型在其评测中 Python 通过率达 44.3%，较 Opus 4.8 的 28.2% 提升 16 个百分点，且解题步骤减少约 22%。JetBrains 将安全与数据保留视为部署核心，偏好零数据保留，但接受为调查最严重问题而进行的有限审查。