

AI早报：Agent workflows前移，治理与安全同步升温

中文内部早报 · 5分钟版

今日总览

今天的AI新闻主线集中在智能体工作流、企业部署和治理安全。头条里最重要的几条分别是NVIDIA 发布 Cosmos 3 Edge：4B 参数开源世界模型，为机器人及边缘 AI 提供实时推理与动作生成、Claude Code v2.1.216 发布：修复长会话卡顿与多项 Agent 行为问题、LoRA Speedrun 公开排行榜：6分05秒微调Qwen2.5-1.5B达GSM8K 61.1%准确率，整体呈现出从模型能力展示转向真实工作流和企业环境落地的趋势。

智能体、编程与工作流：最重要的是上海科学智能研究院开放科学多模态基础模型“神珍”。这类新闻的共同点是，AI 工具正在进入任务拆解、跨系统协作、版本控制和长期执行等环节，不再只是停留在代码补全或单轮问答。同板块还包括不会代码也能做产品：一份从0开始的Vibe Coding保姆级教程、Hugging Face 更新 AI 工作流与智能体方案。

前沿模型与多模态：最重要的是LoRA Speedrun 公开排行榜：6分05秒微调Qwen2.5-1.5B达GSM8K 61.1%准确率。模型能力的亮点更多体现在产品嵌入和场景化调用上，视频编辑、旅行对话、模型聚合入口等消息都指向同一个方向：能力需要被放进具体工作流，读者更应关注实际可用性和成本。

大厂与产业链：最重要的是NVIDIA 发布 Cosmos 3 Edge：4B 参数开源世界模型，为机器人及边缘 AI 提供实时推理与动作生成。企业部署、云上交付、内容合作和基础设施投入仍是产业侧重点；算力、水电、版权和企业采购会继续影响 AI 服务成本与可落地范围。

监管、安全与研究：监管安全侧最重要的是Cursor 测试新型 AI 智能体集群：规划者+执行者分工，4小时通过80% SQL测试。今天的信号集中在漏洞修复、治理控制面和数据驻留等主题，说明企业在采用模型时需要同步处理审计、权限和供应商边界。研究评测方面，Ollama 获 8800 万美元融资，加速开放模型生态发展值得放入方法论更新。开发者工具方面，Claude Code v2.1.216 发布：修复长会话卡顿与多项 Agent 行为问题可作为近期试用入口。

头条

今日头条

1. NVIDIA 发布 Cosmos 3 Edge：4B 参数开源世界模型，为机器人及边缘 AI 提供实时推理与动作生成

Hugging Face 博客 · 07/21 08:00 · 大厂与产业链

NVIDIA 在 Hugging Face 上开源了 Cosmos 3 Edge，一个 40 亿参数的世界模型，旨在帮助机器人和视觉 AI 智能体在边缘设备上理解环境、实时推理并生成动作。

2. Claude Code v2.1.216 发布：修复长会话卡顿与多项 Agent 行为问题

Claude Code: GitHub Releases · 07/21 08:00 · 开源项目与开发者工具

Claude Code v2.1.216 修复了长会话中消息归一化成本随轮次二次增长导致的数秒停顿问题，并修正了 OAuth token 过期后自动模式误判 HTTP 401、后台子智能体重启后恢复默认配置、以及工作树隔离子智能体重定向 git 目录等多项缺陷。新增 `sandbox.filesystem.disabled` 设置，可在保留网络出口控制的同时跳过文件系统隔离。

3. LoRA Speedrun 公开排行榜：6分05秒微调Qwen2.5-1.5B达GSM8K 61.1%准确率

Hacker News · 07/21 08:00 · 前沿模型与多模态

LoRA Speedrun项目推出公开排行榜，在固定硬件（单张L40S）上比拼Qwen2.5-1.5B的微调运行时间。当前纪录由@Saivineeth147以6分05秒保持，采用序列打包与仅完成损失掩码技术，相比基线11分57秒提速约2倍且准确率更高（61.1%）。项目提供免费Modal沙箱验证，任何提交需经3次独立复现确认。

4. Cursor 测试新型 AI 智能体集群：规划者+执行者分工，4小时通过80% SQL测试

Cursor Blog · 07/21 08:00 · 监管、安全与版权

Cursor 测试了新型 AI 智能体集群，将任务分解为规划者（使用最强模型）和执行者（使用快速廉价模型）。使用 Grok 4.5 时，新集群在 4 小时内通过了 80% 的 SQL 测试套件，而旧集群在第二小时前失败。该系统已用于构建浏览器、修复漏洞和生成数十亿 token 合成训练数据。

5. 《第九区》导演Neill Blomkamp发布首部完全由AI生成的短片《Nightborne》

The Decoder: AI News · 07/21 08:00 · 监管、安全与版权

Neill Blomkamp发布了13分钟科幻恐怖短片《Nightborne》，完全使用Seedance 2.0视频生成模型通过文本提示逐帧创作。影片采用纪录片风格，使用了32位真实人物的面部和声音（已获授权），人类艺术家负责概念艺术。Blomkamp表示计划以相同格式拍摄一部长片，并已创立AI电影工作室Barley Studios。

模型

前沿模型与多模态

1. 字节跳动发布 Seed Audio 1.0 音频创作模型，统一建模人声与音效实现端到端影视级音频生成

字节 Seed: Research Feed (网页内嵌数据) · 07/20 00:00

字节跳动 Seed 团队发布音频创作模型 Seed Audio 1.0，在统一框架下联合建模人声、音效和环境声，支持 100ms 间隔精度的时间控制、单次约 2 分钟音频的延展生成及 20+ 语种自然生成。该模型已在火山方舟体验中心上线，多数场景音频可用率达 90% 以上，多语言音频自然度 MOS 超 4 分。

2. 通义实验室发布 Qwen-Audio-3.0-TTS 实时语音合成模型

公众号: 通义实验室 (千问) · 07/21 08:00

通义实验室发布 Qwen-Audio-3.0-TTS，含 Flash（首包延迟约300ms）和 Plus 两个版本。Plus 版本在 Artificial Analysis 榜单夺冠，支持16种语言和20种中文方言，平均 WER/CER 低至3.87（Flash），说话人相似度最高达82.75（Plus）。

3. 昆仑万维宣布2026为"世界模型元年"，发布Matrix-Game 3.5等模型

公众号：昆仑万维（天工） · 07/19 20:35

昆仑万维董事长方汉在WAIC上宣布2026年为"世界模型元年"，并发布Matrix-Game 3.5世界模型、Mureka v9.5与O3音乐模型。Matrix-Game 3.5实现Patch级记忆注入，5B模型在720p分辨率下单卡可达20FPS实时生成，核心架构已开源。

4. Qwen3.8 开源发布，2.4T 参数模型上线

X: 通义千问 / Qwen (@Alibaba_Qwen) · 07/19 16:10

Qwen3.8 即将发布并很快开源权重！该模型拥有 2.4T 海量参数，正在持续进化。我们相信它是目前最强大的模型之一，可与领先的前沿 AI 模型媲美，仅次于 Fable 5。你无需等待即可测试。就在刚才，Qwen3.8-Max-Preview 已在阿里巴巴的 Token Plan、Qoder 和 QoderWork 上首次亮相。快来抢先体验吧。期待你的作品。敬请关注！

智能体

智能体、编程与工作流

1. 上海科学智能研究院开放科学多模态基础模型“神珍”

IT之家 · 07/21 08:00

上海科学智能研究院开放科学多模态基础模型“神珍”，总参数约110亿，可处理DNA、RNA、蛋白质、小分子、地球系统和医学影像六类数据。在生物序列20项任务中，该模型9项取得最优结果；医学影像分割平均Dice得分91.20，在7种参评方法中最优。模型权重及代码已在星河启智、Hugging Face和GitHub开放。

2. 不会代码也能做产品：一份从0开始的Vibe Coding保姆级教程

公众号：数字生命卡兹克 · 07/21 08:00

本文面向零代码用户，提供一套使用国产大模型（Kimi、GLM、Qwen等）从零开发并上线产品的完整流程。核心步骤包括购买Coding Plan、下载官方Agent编程产品、注册域名与服务器并同步做ICP备案，然后通过Agent的Plan模式描述需求并让AI自动执行开发。上线后建议建立分支保护与测试流程，并强调即使不懂代码，也必须对系统架构了如指掌。

3. Hugging Face 更新 AI 工作流与智能体方案

The Decoder · 07/20 20:12

据The Decoder，Hugging Face 更新 AI 工作流与智能体方案。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。

4. China 更新 AI 工作流与智能体方案

MIT Technology Review AI · 07/21 02:00

据MIT Technology Review AI，China 更新 AI 工作流与智能体方案。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. Ray 2.55 正式支持 Google Cloud TPU，通过 KubeRay 自动编排多主机切片

Google 开发者博客 · 07/21 08:00

Ray 2.55 首次为 Google Cloud TPU 提供一等支持，开发者可通过 Ray 任务与 Actor API 在 TPU 上运行分布式 Python 负载。

2. 乐天用 Claude Fable 5 构建可自主运行数小时的智能体

Claude 博客 · 07/21 08:00

乐天AI业务总经理Yusuke Kaji测试Claude Fable 5后表示，该模型能自主运行更长时间，首次实现夜间无人值守完成复杂任务。与之前模型不同，Fable 5在运行中持续自我验证和纠错，避免早期错误假设导致整次运行失败。乐天已在一周内将Claude Managed Agents部署到产品、销售、市场和财务等部门，各领域问题解决速度提升约10倍。

3. OpenAI 推出 AI 产业部署新进展

OpenAI 新闻 · 07/20 18:00

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。

4. transcribe.cpp 发布：基于 ggml 的跨平台语音转录库，支持 16 个 ASR 模型族

Hacker News · 07/19 10:49

据Hacker News，transcribe.cpp 发布：基于 ggml 的跨平台语音转录库，支持 16 个 ASR 模型族。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. Grok for Excel 发布：在 Microsoft Excel 中用自然语言提问、写公式和运行场景

xAI 新闻 · 07/21 08:00

xAI 将 Grok 引入 Microsoft Excel，推出免费 Microsoft 365 加载项。用户可在工作表中用自然语言提问、根据描述编写公式或运行场景，答案会引用具体单元格，图表可直接插入工作表。该加载项还支持连接 SharePoint 或 Google Drive 获取上下文，并已同步支持 Word 和 PowerPoint。

2. 中国AI几乎追平美国，Kimi K3开源模型引发市场震荡

Gary Marcus: The Road to AI We Can Trust · 07/21 08:00

中国公司月之暗面（Moonshot.AI）发布Kimi K3模型，性能与最佳美国模型相当，且为开源权重模型，用户可免费下载本地运行。受此消息影响，美国股市上周五下跌，OpenAI和Anthropic的商业模式及IPO前景受到严重质疑。美国在AI软件领域的护城河已不如预期，AI竞赛正演变为工业系统竞争。

3. 开源权重模型逼近前沿，闭源仍领先

Tomer Tunguz 博客 (VC 分析) · 07/21 08:00

开源权重模型已多次达到与闭源前沿模型相当的水平，但闭源模型如 GPT-5.2 仍持续创造阶跃式突破。Kimi K3 为 2.8T 参数开源模型，Qwen 3.8 为 2.4T 参数模型。按 90/10 输入输出比例计算，开源前沿模型价格中位数比 GPT-5.2 便宜约 15%，最便宜的 DeepSeek V4 Flash 便宜约 90%。

4. 面壁智能发布首个具身智能成果 MiniCPM-Robot 系列模型，含 1.5B VLA 与 0.9B 跟踪模型

公众号：面壁智能（MiniCPM） · 07/21 08:00

据公众号：面壁智能（MiniCPM），本条原始主题为《面壁智能发布首个具身智能成果 MiniCPM-Robot 系列模型，含 1.5B VLA 与 0.9B 跟踪模型》。原文摘要为：面壁智能联合 OpenBMB 发布并开源 MiniCPM-Robot 系列，包括通用 VLA 模型 MiniCPM-RobotManip（1.5B 参数）与移动跟踪模型 MiniCPM-RobotTrack（0.9B 参数）。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。

观点

KOL 与巨头言论

1. Show 发布 AI 行业观点与判断

Hacker News · 07/19 00:17

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. Index Ventures 联合创始人 Neil Rimer 认为 AI 财富将面临"再分配"

TechCrunch AI · 07/18 12:47

Index Ventures 联合创始人 Neil Rimer 表示，围绕 AI 积累的巨额财富将面临"某种形式的再分配"，无论是自愿还是强制。他呼吁科技领袖在推动自愿再分配中发挥主导作用。与此同时，美国慈善捐赠总额虽创新高，但捐赠人数持续下降，而加州正考虑对亿万富翁征收 5% 的一次性财富税。

2. Trump 更新 AI 治理与安全动态

The Decoder · 07/20 22:03

据The Decoder，Trump 更新 AI 治理与安全动态。这类进展主要影响数据、版权、安全和合规部署边界。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 AI 治理与安全动态

TechCrunch AI · 07/20 03:24

据TechCrunch AI，OpenAI 更新 AI 治理与安全动态。这类进展主要影响数据、版权、安全和合规部署边界。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。

研究

研究、评测与方法论

1. Ollama 获 8800 万美元融资，加速开放模型生态发展

Hacker News · 07/21 08:00

Ollama 宣布完成 8800 万美元融资，由 Benchmark、Theory Ventures 和 8VC 等领投。该平台已服务 890 万开发者，并被 85% 的财富 500 强企业使用，其云端 token 用量月均翻倍。资金将用于支持无缝混合推理、新模型发布当日集成，以及让开发者在不牺牲所有权和隐私的前提下使用最强开放模型。

2. LVSum 基准：评估多模态大模型的长视频时间感知摘要能力

Apple Machine Learning Research · 07/21 08:00

Apple 推出 LVSum，一个含细粒度时间对齐的人工标注长视频摘要基准，包含 13 个领域的 72 个视频（平均时长 16 分钟），每个视频配有最多 10 条含时间引用的人类摘要。评估显示，转录文本对摘要质量的贡献远大于视觉帧，当前多模态大模型在时间定位、指令遵循和跨模态一致性上存在系统性缺陷，与人类摘要仍有显著差距。

3. Apple 发布 AI 研究评测新进展

Apple Machine Learning Research · 07/21 08:00

Apple 研究团队提出 LenVM，一种在每步解码时预测剩余生成长度的 token 级框架，将长度建模转化为无需标注的价值估计问题。在 LIFE Bench 精确长度匹配任务上，LenVM 将 7B 模型的长度分数从 30.9 提升至 64.8，超越前沿闭源模型；在 GSM8K 上以 200 token 预算维持 63% 准确率（基线仅 6%）。