

AI模型竞赛白热化：Qwen3.8挑战Fable 5，Kimi K3登顶前端编码

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，阿里巴巴Qwen3.8以2.4T参数量声称仅次于Fable 5，月之暗面Kimi K3在前端编码领域登顶榜首。同时，NVIDIA在日本部署大型AI基础设施，AI内存需求推高硬件成本。

Agent与Coding领域呈现分化态势，Google Deepmind通过GenCception验证视频生成器具备计算机视觉能力，Kimi K3在前端编码表现优异但复杂数学能力薄弱。开源工具方面出现新的编码代理和隐写技术项目。

模型与多模态技术持续突破，NVIDIA发布Nemotron 3 Embed系列在RTEB基准排名第一，面壁智能开源具身AI模型MiniCPM-Robot。昆仑万维宣布2026年为世界模型元年，发布Matrix-Game 3.5。

大厂与产业链动态密集，黄仁勋访日宣布建设Vera Rubin AI工厂，内存需求激增推高电脑价格。面壁智能发布端侧模型MiniCPM5-2B，transcribe.cpp发布支持16个ASR模型族的语音转录库。

监管安全与研究工具板块关注AI可靠性问题，Apple起诉OpenAI引发关注，研究显示AI读片系统在错误时仍过度自信。美团发布LoHoSearch搜索智能体基准，Schema框架在ARC-AGI-3上取得突破性成绩。

头条

今日头条

1. 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 · 07/20 08:00 · 大厂与产业链

transcribe.cpp v0.1.0发布，这是一个基于ggml的语音转录库，支持16个ASR模型族超过60个模型，并通过Vulkan、Metal、CUDA和TinyBLAS实现GPU加速。

2. OpenAI推出AI评分卡衡量ROI指标

OpenAI 新闻 · 07/17 18:00 · 前沿模型与多模态

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。

3. NVIDIA发布Nemotron 3 Embed系列，8B版本在RTEB基准排名第一

MarkTechPost · 07/17 15:53 · 研究、评测与方法论

NVIDIA发布Nemotron 3 Embed系列，包含三个开源checkpoint，其中8B-BF16版本在RTEB基准上以78.46的平均NDCG@10排名第一。1B-NVFP4版本在Blackwell上吞吐量比BF16高2倍，精度保留99.5%，所有模型最大序列长度32768 tokens。

4. Google Deepmind论证视频生成器已包含计算机视觉缺失的世界模型

The Decoder · 07/19 18:17 · 智能体、编程与工作流

Google Deepmind的GenCception将视频生成器重新用于深度估计和分割等经典视觉任务，匹配最先进的系统但训练数据更少。该模型几乎完全在合成视频上训练，结果加剧了关于视频生成器是否已包含世界模型的辩论。

5. 月之暗面Kimi K3在前端代码方面超越Fable 5，但在复杂数学方面落后

The Decoder · 07/19 17:32 · 智能体、编程与工作流

月之暗面的Kimi K3成为首个登顶Code Arena前端排名的中国模型，击败Claude Fable 5和GPT-5.6 Sol。但在高级数学方面差距显著：Kimi K3在FrontierMath Tier 4上仅获得约39%分数，而OpenAI和Anthropic的模型接近90%。

模型

前沿模型与多模态

1. 面壁智能开源MiniCPM-Robot具身智能模型系列

X: 面壁智能 OpenBMB (@OpenBMB) · 07/19 21:49

面壁智能开源首个具身AI模型系列MiniCPM-Robot，包含1.5B参数的通用视觉-语言-动作模型MiniCPM-RobotManip和用于目标跟踪的MiniCPM-RobotTrack。同时发布高性能推理框架PhyAI，旨在让机器人实现理解、记忆与行动。

2. 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。

3. 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。团队称其是目前最强大的模型之一，可与领先前沿AI模型媲美，仅次于Fable 5。Qwen3.8-Max-Preview已在阿里巴巴的Token Plan、Qoder和QoderWork上首次亮相。

4. 阿里巴巴Qwen挑战Kimi K3，发布开放权重Qwen 3.8称仅次于Fable 5

The Decoder · 07/19 20:07

阿里巴巴发布Qwen 3.8，这是一款拥有2.4万亿参数的多模态AI模型，Qwen团队称其可匹敌领先模型且仅落后于Fable 5。预览版现已可用，该模型直接对标月之暗面Kimi K3。

智能体 智能体、编程与工作流

1. Kimi K3登顶前端编码榜，开放权重挑战闭源双巨头

X: 阿易 AI Notes (@AYi_AInotes) · 07/17 12:57

Kimi K3在Frontend Code Arena以1679分登顶，力压Claude Fable 5与GPT-5.6 Sol，7个前端细分赛道拿下6个第一。该模型为2.8万亿参数MoE架构，百万上下文窗口，7月27日开放权重。API定价为输入每百万tokens 15美元。

2. 《奥德赛》导演Christopher Nolan称AI为明显的特洛伊木马

TechCrunch AI · 07/19 22:52

据TechCrunch AI，TechCrunch 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。

3. 非营利组织Current AI竞力建设AI万维网，供所有人免费使用

TechCrunch AI · 07/19 22:00

非营利组织Current AI致力于构建不让任何文化落后的AI，已在设备、AI聊天等方面取得显著进展，目标是建设供所有人免费使用的AI万维网。

4. 月之暗面在GTC 2026披露Kimi K2.5技术路线：用MuonClip、线性注意力与Agent Swarm重构三大基础组件

X: 宝玉 (@dotey) · 07/18 01:38

月之暗面CEO杨植麟在GTC 2026演讲中提出用MuonClip优化器替代Adam，可将数据利用效率提升近一倍。同时推出Kimi Linear线性注意力，在百万Token上下文下全面超越全注意力；Agent Swarm已支持300个Agent并行工作。

产业 大厂与产业链

1. 黄仁勋访日：Nvidia联手日本打造物理AI时代，Noetra主权AI工厂与Cosmos机器人联盟落地

TechCrunch AI · 07/20 08:00

Nvidia CEO黄仁勋在7月15-16日访日期间，宣布为日本建设Vera Rubin AI工厂，配备13750颗Vera CPU和27500颗Rubin GPU，预计2028年投运。

2. MiniCPM5-2B发布：4B以下全球性能第一，适配9款芯片

公众号: 面壁智能 (MiniCPM) · 07/20 08:00

面壁智能联合OpenBMB发布端侧模型MiniCPM5-2B，以2B参数量在AA-Index榜单取得4B以下模型最高分17分，平均分54.26，超越Qwen3.5-2B等竞品。模型原生支持混合思考与512K上下文，已完成华为昇腾、英伟达等9款芯片的Day0适配，即将开源。

3. NVIDIA 推出 AI 产业部署新进展

Hugging Face Blog · 07/17 23:57

NVIDIA与Hugging Face合作推出NeMo Automodel，支持大规模微调视频和图像模型，提供Diffusers框架的优化支持。

4. 生成式人工智能是一场工程灾难：AI公司抢购70%高端内存，推高电脑价格

Hacker News · 07/17 08:40

AI公司为维持大语言模型运行，可能已购买全球70%的高端计算机内存，导致内存与存储价格飙升。两年前350美元的硬盘现已涨至800美元且缺货，部分笔记本电脑涨价50%。预测称平价入门级电脑可能在2028年前消失，内存短缺预计持续数年。

开源 开源项目与开发者工具

1. 新开源项目 KlaatAI/klaatcode 上线

GitHub · 07/17 19:38

开源AI编码代理终端工具，具有Claude Code级别的准确性，通过智能模型路由使用适合任务的AI模型，降低成本10倍。支持Claude、GPT、Gemini、DeepSeek等。Stars: 147, Forks: 23, 语言: TypeScript。

2. 新开源项目 nethical6/conversation-steganography 上线

GitHub · 07/17 12:58

GitHub 新仓库 nethical6/conversation-steganography 在本次覆盖窗口内创建。仓库简介显示，该项目面向 AI 开发者工具或自动化场景。当前公开数据约为 822 Stars、48 Forks，主要语言为 Go。该条目代表新项目出现，不把普通代码提交描述为版本发布。

3. 新开源项目 MlgHTy-alleN/Trading-Bot 上线

GitHub · 07/17 17:45

GitHub 新仓库 MlgHTy-alleN/Trading-Bot 在本次覆盖窗口内创建。仓库简介显示，该项目面向 AI 开发者工具或自动化场景。当前公开数据约为 351 Stars、193 Forks，主要语言为 Solidity。该条目代表新项目出现，不把普通代码提交描述为版本发布。

1. 展示HN：OpenCareLoop - 解决家庭健康问题，一次一个循环

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。

1. Apple起诉OpenAI：诉讼背后是竞争焦虑还是时机博弈？

The Verge: AI · 07/18 01:41

Apple对OpenAI提起诉讼，指控其存在多项不当行为，尽管许多专家认为部分指控属于行业惯例。此举正值Apple发布新版软件公测版以新Siri AI为核心之际，外界猜测Apple究竟是担忧OpenAI成为潜在竞争对手，还是想利用OpenAI的弱势期获利。

2. 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。

3. Apple诉讼能否阻碍OpenAI硬件计划？

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。

1. AI聊天机器人读片即使错误时也可能过于自信

The Decoder · 07/19 15:35

RadLE 2.0基准测试AI模型在放射学中何时应将诊断留给医生的能力。许多模型在错误时给出完全自信的结果，人类放射科医生仍然领先。在AI能够独立诊断之前，需要学会何时保持沉默。

2. 美团LongCat发布LoHoSearch：更难搜索智能体基准

X: 美团 LongCat (@Meituan_LongCat) · 07/17 22:08

美团LongCat推出LoHoSearch，一个基于762万实体维基百科知识图谱自动生成问题的搜索智能体基准，旨在解决BrowseComp等现有基准趋于饱和的问题。在11个前沿模型测试中，最佳得分仅34.74%，远低于当前模型在BrowseComp上约90%的成绩；上下文策略仅带来+6.8个百分点的提升。该基准包含544道问题、11个领域，采用树与图结构，已开源。

3. Schema Harness在ARC-AGI-3公开集上取得约99%成绩

Hacker News · 07/17 16:41

Schema框架在ARC-AGI-3公开集上，使用Claude Opus 4.8和Fable 5达到99% RHA E分数，使用GPT-5.6 Sol达到95.35%。该框架不修改模型权重，而是将原始观测转化为可编辑程序，联合解决状态归因和机制发现问题。此前最强模型GPT-5.6 Sol在半私有集上仅得7.78%。