

AI 技术每日分析：Liquid 发布 3.1B 端侧视觉模型，手机本地 Agent 加速落地

中国高技术产业发展促进会新质生产力工作委员会

博雅云创 & 中科创新驱动

2026 年 8 月 13 日星期四

摘要

8 月 12 日的 AI 技术更新呈现出一个很明确的变化：开放模型竞争正在从“谁的参数更大”逐渐延伸到“谁能在本地、边缘和专业任务里更容易部署”。Liquid AI 发布 LFM2.5-VL-3B，把 3.1B 参数视觉语言模型直接推向手机、笔记本和高并发 GPU 场景；Cohere Labs 发布 2.4B 参数 North Micro Vision，重点解决原生分辨率文档、图表、OCR 和多语言视觉理解；Reuters 援引 Ramp 数据称，7 月使用 OpenRouter 等提供开放权重与中国开发模型平台的 AI 付费企业占比升至 6.1%，高于一年前的 4.5%；Ai2 则为 OlmoEarth Studio 加入可导出的地球观测 Embedding，让用户无需完整微调模型，就可以进行相似性搜索、少样本分割和变化检测。AI 产业今天最值得观察的主线，不是又出现一个万能模型，而是小模型、开放权重、Embedding 和端侧推理正在形成更具体的产品入口。

目录

- 一、Liquid AI 发布 LFM2.5-VL-3B：3.1B 模型开始承担屏幕理解、工具调用和端侧视觉 2
- 二、Cohere Labs 推出 North Micro Vision：原生分辨率成为小型 VLM 的新卖点 2

三、开放权重模型的商业使用开始增长：企业从“Tokenmaxxing”转向算 ROI	3
四、Ai2 开放 OlmoEarth Embedding 导出：Foundation Model 开始把“中间能力”直接变成产品	4
五、长尾观察：Agent 工具链开始针对“生产细节”做修补	4
趋势观察	5
参考资料	5

一、Liquid AI 发布 LFM2.5-VL-3B：3.1B 模型开始承担屏幕理解、工具调用和端侧视觉

Liquid AI 在 8 月 12 日发布 LFM2.5-VL-3B。这是一款 3.1B 参数视觉语言模型，面向自有硬件和边缘部署，重点增强屏幕/UI 理解、自然语言目标定位、多图输入和 Function Calling。模型使用 SigLIP2 400M 视觉编码器，文本骨干沿用 LFM2.5 系列，预训练规模约 34 万亿 Token，视觉数据量相比上一代增加 4 倍。

这次发布更值得关注的是工程指标。Liquid AI 披露，模型在 M5 Max 上解码速度约 228 Token/s，在 Ryzen AI Max+ 395 上约 116 Token/s，占用内存约 3GB；在 Galaxy S26 Ultra 上也可达到约 20 Token/s。模型首日支持 llama.cpp、MLX、vLLM、SGLang 和 ONNX，在高并发 H100 环境中输出吞吐可达到约 1.1 万 Token/s。

这类模型代表了一个非常实际的方向：视觉 Agent 不一定依赖百亿、千亿参数模型。企业如果只需要识别设备界面、读表格、理解网页、处理文档、调用工具，3B 左右模型已经开始进入“够用且便宜”的区间。下一阶段端侧 AI 的关键，可能不再是能否运行模型，而是能否把视觉理解、工具调用和业务权限放在同一个本地工作流里。

二、Cohere Labs 推出 North Micro Vision：原生分辨率成为小型 VLM 的新卖点

Cohere Labs 同日发布 North Micro Vision Instruct，一款 2.4B 参数、Apache 2.0 许可的开放权重视觉语言模型。它由约 400M 参数视觉编码器和 2B 参数语言模型构成，支持中文、英语、德语、法语、西班牙语、日语、韩语、阿拉伯语等多种语言。

North Micro Vision 的核心不是“再做一个通用 VLM”，而是尽量保

留原始图像比例和细节。其训练过程将视觉分辨率逐步提升到能够处理约 200dpi 的 A4 文档页面，重点覆盖 OCR、表格、图表、文档理解、视觉定位和计数。

这对于企业应用很重要。大量多模态任务并不是识别一张猫狗照片，而是看财务表格、工程截图、合同扫描件、网页界面和业务单据。传统做法把图像缩成固定小方块后，细文字和布局很容易丢失。原生分辨率 VLM 如果能在 2B—3B 参数规模下稳定运行，会使“私有文档视觉理解”成为本地 AI 的一块重要市场。

三、开放权重模型的商业使用开始增长：企业从“Token-maxxing”转向算 ROI

Reuters 8 月 12 日报道，美国 AI 厂商正在重新重视开放权重模型，原因之一是企业开始更严格地审视 AI 支出。Ramp 当天发布的数据表明，7 月约 6.1% 的 AI 付费企业使用 OpenRouter 等提供开放权重和中国开发模型的平台，一年前这一比例约为 4.5%。

这一数字仍只是整体 AI 支出的一个切片，但它反映出企业采购逻辑正在变化。很多运营任务、内部工具、代码辅助和批处理任务，并不需要每次调用最昂贵的前沿模型。开放模型的优势不只在“免费”，还在于可以自托管、定制、量化和按业务需要选择推理基础设施。

这也解释了为什么最近开放模型竞争越来越强调“路由”和“部署”。企业未来更可能建立模型组合：复杂推理升级到旗舰模型，日常任务用小模型或开放模型处理。模型选择逐渐从技术偏好变成成本工程问题。

四、Ai2 开放 OlmoEarth Embedding 导出：Foundation Model 开始把“中间能力”直接变成产品

Ai2 在 8 月 12 日为 OlmoEarth Studio 增加自定义 Embedding 导出能力。用户可以选择区域、时间范围、模型编码器、空间分辨率以及 Sentinel-1/2 影像来源，直接得到 Cloud-Optimized GeoTIFF 格式的 Embedding 数据。

这项功能看似只服务遥感领域，但背后的产品思路很值得关注。过去用户使用 Foundation Model 通常有两条路：直接调用最终预测，或者拿自己的数据做微调。OlmoEarth 提供了第三条路——把预训练模型学到的中间表示直接导出。

这些 Embedding 可以用于相似性搜索、少样本分割、变化检测和无监督探索。Ai2 给出的示例中，只使用 60 个标注像素，通过线性分类器就可以得到较完整的红树林区域分割结果。

这说明 Foundation Model 未来不一定只通过 API “回答问题”创造价值。Embedding、Feature Map、工具调用能力和可复用中间表示，也可以单独成为开发者产品。对于行业 AI 来说，这种方式往往比重新训练一个完整模型更轻量。

五、长尾观察：Agent 工具链开始针对“生产细节”做修补

开源自动化项目 n8n 近期的版本更新中，集中出现多项 Agent 生产问题修复，包括阻止工具调用前的中间文本泄漏到最终回复、限制过大的工具返回结果、对 Code 节点控制台输出应用脱敏策略、按 MCP 工具声明的 JSON Schema 方言编译 Schema，以及强化 Agent Trace 和凭据使用可见性。

这些更新不是新模型，也不会登上主流科技新闻头条，但它们很能说

明 Agent 进入生产后的现实问题。真正影响企业 Agent 可靠性的，经常是工具返回太大、敏感信息被日志带出、Schema 不兼容、任务恢复失败、凭据权限混乱，而不是 Benchmark 下降了两个点。

随着模型能力趋于充足，Agent 竞争正在越来越像传统软件工程：边界、恢复、日志、权限和失败模式，比单次 Demo 更重要。

趋势观察

今天几条动态共同指向一个趋势：开放 AI 正在由“开放模型”进一步变成“开放的部署组件”。3B 级视觉模型开始进入手机和笔记本，2.4B 模型专门优化文档与 OCR，企业对开放权重平台的采用在增加，Foundation Model 则直接开放 Embedding。模型规模还会继续增长，但商业化的另一条路线已经越来越清晰——让模型更小、更容易组合、更容易本地化，也更容易纳入成本控制。

参考资料

1. Liquid AI / Hugging Face | 《LFM2.5-VL-3B for Better and Faster Vision Capabilities for the Edge》 | 2026-08-12 | 用途：核验模型参数、训练规模、工具调用能力及端侧推理速度。<https://huggingface.co/blog/LiquidAI/lfm2-5-vl-3b>
2. Liquid AI | LFM2.5-VL-3B 模型与文档 | 2026-08-12 | 用途：补充模型使用、端侧部署与 Function Calling 信息。<https://huggingface.co/LiquidAI/LFM2.5-VL-3B>
3. Cohere Labs / Hugging Face | 《Meet North Micro Vision: A 2.4B Native-Resolution Vision-Language Model》 | 2026-08-12 | 用途：核验 North Micro Vision 发布、架构、原生分辨率和训练方式。<https://huggingface.co/Cohere/north-micro-vision>

[//huggingface.co/blog/CohereLabs/meet-north-micro-vision-instruct](https://huggingface.co/blog/CohereLabs/meet-north-micro-vision-instruct)

4. Cohere Labs | North-Micro-Vision-Instruct Model Card | 2026-08-12 | 用途: 核验 2.4B 参数、Apache 2.0、多语言、128K 文本上下文等模型规格。<https://huggingface.co/CohereLabs/North-Micro-Vision-Instruct>
5. Reuters | 《American AI model makers smell an opportunity》| 2026-08-12 | 用途: 核验开放权重模型商业使用趋势及 Ramp 6.1%/4.5% 数据。<https://www.reuters.com/technology/artificial-intelligence/american-ai-model-makers-smell-an-opportunity-2026-08-12/>
6. Ai2 / Hugging Face | 《Introducing OlmoEarth embeddings: Custom embedding exports from OlmoEarth Studio for downstream analysis》| 2026-08-12 | 用途: 核验 OlmoEarth Embedding 导出、GeoTIFF 和下游任务。<https://huggingface.co/blog/allenai/olmoearth-embeddings>
7. Ai2 | OlmoEarth 技术资料 | 2026 年 | 用途: 补充地球观测 Foundation Model 和 Embedding 技术背景。<https://allenai.org/olmoearth>
8. Hugging Face Blog | Community Blog & Articles | 2026-08-12 | 用途: 筛选当日开放模型、研究与长尾项目。<https://huggingface.co/blog>
9. n8n GitHub | Releases | 2026-08-11 至 12 | 用途: 补充 Agent 工具调用、日志脱敏、MCP Schema 和 Trace 工程化动态。<https://github.com/n8n-io/n8n/releases>
10. Paperclip GitHub | 《Open-source orchestration for teams of AI agents》

| 2026-08 检索 | 用途: 补充 Agent 预算、治理、RBAC、审计与多 Agent 编排长尾生态。<https://github.com/paperclipai/paperclip>

联系我们，请扫描二维码



新质生产力工作委员会
官方公众号



工业智能算网
gyznsw.cn

新质生产力工作委员会：

中国高技术产业发展促进会新质生产力工作委员会，专注于推动工业人工智能、智能制造、数字化转型等前沿技术发展，为企业提供政策解读、技术咨询和产业对接服务。

工业智能算网：

专注于工业人工智能、新质生产力、工业软件 CAE、智能制造等前沿技术。提供每日动态分析、技术趋势解读、解决方案分享，推动工业智能化转型。

网站地址：<https://gyznsw.cn>