

MiniMax M3.1 Flash 上线：声称使用成本降低约 70%

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 10 月 20 日

摘要

10 月 19 日的几条 AI 新闻，讨论的都是智能体落地时绕不开的问题：跑得快不快、花多少钱、能否换一套芯片运行，以及谁有权让智能体执行操作。MiniMax 上线 M3.1 Flash，主打速度和调用费用；亚马逊介绍面向 Trainium 的 TorchNeuron，希望降低 PyTorch 程序迁移难度；Modal 计划采用英伟达 Vera CPU 托管长时间运行的智能体。另有 doubleAI 发布计算执行优化模型，World 公布金融业务真人验证方案。这些消息的进展程度并不一样：模型服务上线、软件拟公开发布、硬件部署计划和供应商产品方案，需要分别看待；降本数字也应保留其企业自报的口径。

Contents

一、MiniMax 正式推出 M3.1 Flash：智能体模型将速度与单任务成本摆上台面	2
二、亚马逊介绍 TorchNeuron：AI 芯片竞争开始正面争夺 PyTorch 兼容性	2
三、Modal 宣布采用英伟达 Vera CPU：智能体的持续运行开始重估 CPU 价值	3
四、doubleAI 发布计算执行优化模型：小公司瞄准既有算力的使用效率	3
五、World 发布金融真人验证方案：AI 代理需要区分使用者与被模仿者	4
趋势观察	4
参考资料	5

一、MiniMax 正式推出 M3.1 Flash：智能体模型将速度与单任务成本摆上台面

MiniMax 于 10 月 19 日上线 M3.1 Flash，面向编程、计算机操作、跨应用智能体任务和安全分析。公司称，在线推理速度接近此前版本的三倍，整体使用成本降低约 70%。公开 API 报价为每百万输入 Token 0.15 美元、输出 Token 0.60 美元、缓存输入 Token 0.015 美元。官方列出了 DeepSWE v1.1、OSWorld V2 等基准成绩，称模型已经进入 API、M Plan 和 MiniMax Code 产品。[1][2]

智能体执行一个任务，常要反复调用工具、回填上下文，有时还要从失败处重试。因此，单 Token 报价和输出速度会影响代理编程、跨系统办公的费用，但不能据此断言每项任务的总成本都下降 70%。企业更适合拿相同 workflow、工具权限和最终成功率做端到端测试。目前上线的是 M3.1 Flash 服务；模型权重计划后续开放，并非 10 月 19 日已经开源。

二、亚马逊介绍 TorchNeuron：AI 芯片竞争开始正面争夺 PyTorch 兼容性

据路透社 10 月 19 日报道，亚马逊旗下 Annapurna Labs 介绍了 TorchNeuron。这款软件旨在让原本用 PyTorch 编写 AI 程序的开发者，经过较少配置调整就在 Trainium 加速器上运行程序，无须为迁移专门重写模型代码。亚马逊称将免费开源发布，已有 30 多家早期客户参与使用；正式发布计划与当日的硅谷开发者会议相联系。[3]

换芯片的成本不只是买一批新设备。企业现有的模型算子、调试脚本、推理服务，以及工程师熟悉的开发方法，都可能绑在原平台上。PyTorch 接口兼容有助于缩短试用时间，至于各类算子、分布式训练框架和推理负载在 Trainium 上的表现，还得逐项测试。报道中的“早期客户使用”“宣

布开源计划”和“所有开发者正式可下载”是不同状态，不能混为一谈；接口兼容也不等于 CUDA 生态已被完整替代。

三、Modal 宣布采用英伟达 Vera CPU：智能体的持续运行开始重估 CPU 价值

云计算创业公司 Modal 在 10 月 19 日表示，计划首次引入英伟达 Vera 中央处理器，用于托管可长时间执行任务的 AI 智能体。此前，这类服务采用英特尔和 AMD 的 CPU。此次选择让英伟达除了在模型训练、推理所用 GPU 之外，也参与工具执行、任务协调和系统调度所依赖的中央计算部分。[4]

代码智能体运行时会启动进程、读取仓库、访问网络、执行测试。模型推理由 GPU 承担，任务等待多久、何时完成，也受 CPU、内存、存储和网络延迟影响。平台服务商评估这类服务时，可以把模型推理、工具执行、容器冷启动、CPU 占用和失败恢复的耗时分开记录。Modal 公布的是部署计划和新增硬件选择；实际吞吐量及成本改善，仍要看客户生产环境中的长期数据。

四、doubleAI 发布计算执行优化模型：小公司瞄准既有算力的使用效率

《华尔街日报》10 月 19 日报道，以色列创业公司 doubleAI 推出一种旨在降低计算开销的新型 AI 模型。Mobileye 联合创始人 Amnon Shashua 参与创办了该公司；投资者包括英伟达等机构，累计融资约 2.2 亿美元。公司称，一名早期客户的计算成本下降 67%。这是一项企业披露的单客户结果，不能当作跨硬件、跨模型的平均收益。[5]

并非每家企业都只能靠继续采购加速卡来处理算力开支。分析运行时行为、改进执行效率，也可能让既有设备多承担一些工作。doubleAI 的

方案能否稳定推广，要看它面对不同代码、推理框架和并发条件时的表现，还要确认优化不影响结果正确性。采购方应要求对方说明基准环境、节省的具体成本项、回退方式，并提供独立测试。融资规模和投资者名单不能代替这些数据。

五、World 发布金融真人验证方案：AI 代理需要区分使用者与被模仿者

World 相关机构于 10 月 19 日披露面向银行和金融服务的 World ID 验证方案。在 AI 语音、视频和自动化代理更易被用于冒用身份的情况下，这项服务试图帮助金融平台确认交互对象是真人、具备授权。供应商提出将真人证明用于账户注册、敏感业务操作或智能服务；金融机构是否采用、怎样满足监管要求，仍取决于各自落地情况。[6][7]

智能体能按指令办理业务，不等于交易对方已经确认指令来自有权操作的人。金融业务可以按风险分层使用实名核验、真人证明、行为风控和人工复核。若采用生物识别，还需要审查收集哪些个人信息、保留多久，以及撤销授权和处理误判的办法。World 这次推出的是供应商产品能力，不意味着所有银行已采用同一身份标准。

趋势观察

模型价格、PyTorch 迁移、CPU 调度、运行时优化和身份验证，看似分属不同领域，最后都会进入智能体服务的一张成本与责任清单。只看每秒 Token 数或单次调用报价，容易漏掉工具执行、重试和合规开支。选型时还应核算任务成功率、总算力费用，并分清厂商给出的性能数字与实际部署结果。

参考资料

1. MiniMax 官方研究博客 | 《MiniMax M3.1 Flash: Redefining Agent Performance and Cost》 | 2026-10-19 | 用途: 核验模型上线渠道、官方性能声称、API 价格以及权重尚待发布。

<https://www.minimax.io/blog/minimax-m31f>

2. Reuters | 《China's MiniMax launches updated AI model M3.1 Flash, claims major improvements》 | 2026-10-19 | 用途: 交叉核验创业公司发布动作, 并将速度、费用改进明确归为厂商表述。

<https://www.reuters.com/world/asia-pacific/chinas-minimax-launches-updated-ai-model-m31-flash-claims-major-improvements-2026-10-19/>

3. Reuters | 《Amazon says its AI chips will work with popular coding language, challenging Nvidia》 | 2026-10-19 | 用途: 核验 TorchNeuron、PyTorch 兼容、早期客户和开源发布计划。

<https://www.reuters.com/business/retail-consumer/amazon-says-its-ai-chips-will-work-with-popular-coding-language-challenging-2026-10-19/>

4. Reuters | 《Nvidia wins fast-growing AI startup Modal for new central processors》 | 2026-10-19 | 用途: 核验 Modal 采用 Vera CPU 的计划、既有 CPU 供应方及智能体托管场景。

<https://www.reuters.com/business/nvidia-wins-fast-growing-ai-startup-modal-new-central-processors-2026-10-19/>

5. The Wall Street Journal | 《Nvidia-Backed Startup doubleAI Launches Model to Pare Cost of Compute》 | 2026-10-19 | 用途: 核验创业公

司主体、融资背景和单一客户降本数据的性质。

<https://www.wsj.com/cio-journal/nvidia-backed-startup-doubleai-launches-model-to-pare-cost-of-compute-b780f215>

6. World / PR Newswire | 《World Launches World ID for Banking and Finance to Address AI-Driven Impersonation》 | 2026-10-19 | 用途：核验金融真人验证方案的定位与供应商公布的应用场景。

<https://www.prnewswire.com/news-releases/world-launches-world-id-for-banking-and-finance-to-address-ai-driven-imPERSONATION-302912766.html>

7. World 官方博客 | 《World ID for Banking and Finance》 | 2026-10-18, 产品背景 | 用途：区分前一日产品介绍与 19 日的金融行业对外披露。

<https://world.org/pt-br/blog/announcements/world-id-banking-finance-ai-safe>

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>