

仅靠模型突破成就不了AI大产业



中国工业互联网研究院高级工程师 朱国伟 姜宝正

上个月,月之暗面公司发布新一代旗舰模型Kimi K3。据月之暗面官方信息,Kimi K3总参数规模为2.8万亿,采用稀疏混合专家架构,原生支持视觉理解,最高支持100万Tokens上下文窗口,重点面向长程编程、知识工作和智能体任务。模型发布后市场反响强烈,短时间内用户请求量快速增长。7月19日,月之暗面发布公告称,过去48小时用户请求量远超预估并逼近现有集群承载极限,为保障存量用户体验,暂停C端新增付费订阅,并加快推进算力扩容。此次暂停订阅虽已过去一个月,但事件所反映出的问题并没有发生本质改变;我国大模型技术能力、产品吸引力和国际影响力持续提升,也暴露出模型迭代速度与算力供给、商业承载、开源生态和安全治理能力尚不匹配。

Kimi K3暂停新增订阅凸显四方面挑战

一是模型能力快速跃升与高质量算力供给不匹配。此次算力紧缺既可能与企业容量预估和发布节奏有关,也折射出前沿模型进入规模化应用后对推理基础设施提出的更高要求。当前,国内算力资源仍存在高端芯片供给受限、区域布局不均、异构算力调度困难、软硬件适配不足等问题,部分智算中心重建建设轻运营,跨主体、跨区域、跨平台流通效率不高,难以快速响应突发性、高并发需求。算力建设不能只看规模,更要提高利用效率、弹性供给和推理服务保障水平。

二是模型开放程度与自主开源生态支撑能力不匹配。模型权重开放并不等同于整个技术生态实现自主可控。超大参数模型对芯片、内存、网络和推理框架要求较高,大多数中小企业和开发者难以独立部署,可能形成“权重开放、算力集中、服务平台化”的新格局。如果模型运行仍高度依赖国外芯片、算力库、开发工具和社区平台,可能出现模型影响力与平台生态收益相分离的风险,即中国模型吸引开发者,而开发工具、分发渠道和社区资产更多沉淀于境外平台。开源模型要形成

紧抓开放模型快速发展窗口期

为了在下一阶段竞争中把短板补长、把弱项做强,产业应该通过各种举措,将模型能力突破转化为产业体系优势。

一是加快建设自主可控、弹性高效的智能算力供给体系。统筹推进人工智能芯片、先进存储、高速互连、超节点、推理引擎和绿色低碳数据中心建设,强化模型企业、芯片企业、云服务商和基础电信企业协同。依托全国一体化算力网,完善异构算力接入、调度和交易机制,提高跨区域资源统筹能力。支持建设面向中小企业和开发者的公共推理服务平台,通过算力券、模型券等方式降低先进模型使用门槛。

二是推动模型、芯片、框架和平台协同创新。组织国内基础模型与国产软硬件开展联合适配,面向超大规模稀疏模型、长上下文、多模态和智能体任务,加强高性能通信、专家并行、内存管理、量化压缩和推理加速等关键技术攻关。建立统一的适配测试和评价体系,将算效、能效、稳定性和服务成本作为重点评价指标。支持模型企业加强用户需求和算力负载预测,提高应对突发流量和实施弹性扩容的能力。

三是加快构建具有国际影响力的人工智能开源生态。支持企业开放先进模型、推理工具、智能体框架和应用组件,建设高水平开源社区、模型评测基准、数据集和工具库。鼓励龙头企业向中小企业开放模型能力和开发平台,形成大中小企业融通创新生态。完善开源许可证、数据来源、知识产权和供应链安全规则,推动开源由

10类终端将纳入智能化分级标准体系

本报讯 继今年5月《人工智能终端智能化分级》(GB/Z 177—2026)系列国家标准发布后,近期,人工智能终端工作组联合全国信息技术标准化技术委员会(SAC/TC28)、全国音视频标委会(SAC/TC242),正式启动《人工智能终端智能化分级》第二批共10项国家标准编制工作,腕式可穿戴设备、学习机、智能玩具、交互式电子白板、投影显示设备、工业控制计算机、照明系统、扫地机器人、智能门锁、翻译笔等10类产品包含在内。

今年5月,工业和信息化部、

商务部、市场监管总局等部门联合启动实施《人工智能终端智能化分级》(GB/Z 177—2026)系列国家标准。

国标采用“2+N”架构:“2”指《第1部分:参考框架》和《第2部分:总体要求》,明确了智能化的概念、等级划分和测试方法,阐述“什么是人工智能终端、怎么分级、怎么判定”,是所有品类标准的基础。

终端智能化的分级体系从L1响应级、L2工具级、L3辅助级到L4协同级,智能化水平依次提高,终端更“聪明”。“N”是面向不同产品的

算力建设不能只看规模,更要提高利用效率、弹性供给和推理服务保障水平。

算力成本、产品价格和企业收入之间建立平衡。

四是智能体能力提升与安全治理水平不匹配。人工智能从内容生成走向工具调用和自主执行后,风险将由错误回答扩展到数据泄露、权限滥用、错误操作、系统失控和责任不清。从更广泛的开放模型治理看,权重开放可能增加恶意改造和滥用风险,需要同步完善许可约束、安全评测和发布管理机制。当前治理仍较多聚焦生成内容,对权限控制、工具调用、异常熔断、人工接管和供应链安全等制度储备不足。

抓住开放模型加快发展的窗口期,补齐智能算力、自主技术栈、商业化和安全治理等短板。

准规范。对具有自主执行能力的系统,建立最小权限、操作确认、异常熔断、人工接管和全过程留痕机制。推动模型企业公开必要的安全评测、风险说明和适用边界。积极参与人工智能、智能体和开源治理国际规则制定,将我国模型技术优势转化为标准和治理话语权。

总体看,Kimi K3发布后用户请求量快速增长,并因集群承载压力暂停C端新用户订阅,表明前沿模型进入规模化应用后,产业竞争重点正由单一模型能力比拼,逐步扩展至算力供给、平台运营、商业模式、开放生态和安全治理等多个维度。我国应抓住开放模型加快发展的窗口期,补齐智能算力、自主技术栈、商业化和安全治理等短板,推动模型能力优势更好转化为产业链、平台生态和国际竞争优势。

中国AI模型公司近期大事

阿里巴巴

8月14日,阿里云千问团队正式宣布开源Qwen3.8系列模型,所有开发者、科研机构和企业均可自由下载、部署及商用。其中,备受全球AI社区期待的Qwen3.8-27B(270亿参数)模型率先上线,成为本次开源的核心亮点。Qwen3.8-27B是一款原生多模态稠密(Dense)模型,原生支持262K上下文,通过YaRN技术可外推至1M Tokens。在性能表现上,该模型较前代Qwen3.6-27B在编程和办公场景中实现大幅提升,整体表现甚至超越了Qwen3.7-Plus。模型新增的reasoning_effort功能可根据任务难度动态控制思考深度,有效节省计算资源。

智谱AI

8月14日,智谱发布GLM-5.3。智谱称GLM-5.3为“编程能力最强的开源模型”,在前一代GLM-5.2相同的基座模型基础上,通过后训练大幅提高了模型的编程能力和网络安全能力,在内部评测中较GLM-5.2提升50%。

DeepSeek

7月31日,DeepSeek在API文档更新日志中低调宣布V4-Flash正式版,该模型总参数2840亿、激活参数130亿,定位高效低成本推理场景。与常规认知中“Pro强、Flash弱”的分层逻辑不同,此次升级最引人关注的是,V4-Flash正式版在智能体(Agent)能力上实现了对自家V4-Pro预览版的全面反超。8月13日,DeepSeek推出DeepSeek-V4-Pro-0813正式版,同时宣布自8月17日起实行峰谷分时定价机制。

字节跳动

7月31日,字节跳动正式发布新一代视频创作模型Seedance 2.5。该模型延续了Seedance 2.0统一的多模态音视频联合生成架构,重点突破长叙事能力、多模态参考能力和编辑能力。

华为

7月31日,华为正式宣布开源盘古5050亿参数的openPangu-2.0-Pro模型。openPangu是华为开源AI模型品牌,致力于通过昇腾原生训练与推理技术打造Agent时代的智能底座。

MiniMax

7月17日,MiniMax首次公开发布旗舰全模态大模型M3,并预热新一代多模态生成模型H3。7月31日,MiniMax发布新一代多模态生成模型MiniMax H3,支持文本、图像、视频、声音组成的多模态上下文的统一理解能力,能够输出具备原生双声道的音视频,最高可支持15秒2K分辨率。

月之暗面

7月16日,月之暗面公司发布新一代旗舰模型Kimi K3。模型发布后市场反响强烈,短时间内用户请求量快速增长。过去48小时后,由于用户请求量远超预估并逼近现有集群承载极限,月之暗面公告暂停C端新增付费订阅,并加快推进算力扩容。7月27日,Kimi K3正式全面开源,同步放出完整模型权重、技术报告及三项核心训练基础设施技术。

腾讯混元

7月6日,腾讯混元Hy3正式发布。7月21日,腾讯混元宣布推出Hyra(Hunyuans Research Agent)首个版本Hyra-1.0,次日上线Miora,定位为有记忆、能理解需求、能调用多智能体(Agent)协作的AI创意智能体。8月13日,腾讯微信团队在X平台公布大语言模型家族WeLM系列,包含80B(800亿)和617B(6170亿)两个参数数量的版本,后者在参数数量上已经超过混元Hy3正式版。腾讯总裁刘炽平在财报电话会上表示,正在训练“更大参数的模型”Hy4,计划在年内推出。

(本报记者谷月整理)

李乐成调研盐湖产业和工业绿色低碳发展

(上接第1版)要强化绿色价值创造,加快发展氢能、新能源装备、新型储能、节能环保装备等代表绿色生产力方向的先进技术和产业,攻坚绿色低碳关键核心技术,培育壮大绿色低碳产业,推动产业体系加速向全球中高端迈进。要强化绿色发展能力提升,用好绿色能源资源,拓展工业领域清洁低碳氢应用场景,推进资源循环利用、变废为宝,激活企业绿色发展内生动力,将绿色低碳优势切实转化为产品附加值、产业竞争力。要强化智绿融合,推动绿色能源与人工智能技术的共同进步、融合发展,让人工智能技术赋能车间产

线,与工业生产、能耗和碳排放管控、资源循环利用等绿色低碳应用场景紧密耦合,以“数转智改”牵引生产方式绿色变革,催生新的经济增长动能。

其间,李乐成还调研了青海通信管理局、甘肃通信管理局、宁夏通信管理局并看望慰问干部职工。他强调,要深入学习贯彻习近平新时代中国特色社会主义思想,巩固拓展树立和践行正确政绩观学习教育成果,着力推动“十五五”时期信息通信业高质量发展,全力做好当前防汛关键期应急通信保障工作。

工业和信息化部有关司局负责同志参加调研。(耀文)