

DeepSeek 让大模型水涨船高



deepseek

本报记者 宋婧

作为当前最为耀眼的AI明星之一,DeepSeek激起的AI热潮层层迭起,一浪接一浪。先是芯片厂商、云厂商、运营商、手机厂商、汽车厂商等竞相接入DeepSeek,掀起模型平权和应用浪潮;紧接着,ChatGPT和文心一言两大闭源模型代表同时宣布免费,并上线深度搜索功能,文心大模型4.5更是直接从闭源转为开源,DeepSeek的“鲶鱼效应”引发行业震动。然而,DeepSeek的后来居上、一骑绝尘也给豆包、Kimi、文小言等带来了巨大压力。滚滚洪流下,“豆包们”是被动跟随还是主动出招?又该如何书写新的篇章?

近日, DeepSeek一路狂飙,把曾经的AI原生APP“三强”远远甩在了身后。

DeepSeek太快了, 还是“豆包们”变慢了?

作为一款现象级大模型, DeepSeek走红的速度出乎所有人意料。在发布后短短一周内, DeepSeek就实现了用户破亿,打破了由ChatGPT保持的纪录;2025年1月27日, DeepSeek超越ChatGPT登顶美国应用商店榜首,之后在100多个国家陆续霸榜,同一天,英伟达股票狂跌17%左右,单股蒸发5890亿美元,这个数字创造了美股历史之最,没有之一;1月28日, DeepSeek日活跃用户数(约2063万)首次超越豆包(约1936万),并在3天后突破3000万大关达到了3286万,创全球最快纪录。截至目前, DeepSeek的日活跃用户数已超过4000万,达到ChatGPT日活跃用户量的74.3%。

而在此之前,字节跳动旗下的

豆包、月之暗面的Kimi和百度的文小言曾是被业界寄予厚望的AI原生APP“三强”。研究机构Quest-Mobile发布的报告显示,截至2024年12月底,中国AIGC(人工智能生成内容)应用的月活跃用户规模达到1.2亿。彼时,豆包、Kimi、文小言的月活跃用户量分别为7523万、2101万和1224万,包揽了全行业80%的流量。其中尤以豆包用户数最多,约占行业63%的市场份额; Kimi用户增长速度最快,年复合增长率达到69.4%,是豆包的3倍、文小言的15倍。

彼时, QuestMobile判断称,虽然已与其他AI原生APP拉开差距,不过和微信、抖音、微博等动辄数亿的国民级应用相比,前三大AI原生APP还远称不上是超级应用,“三

强”地位也并不稳固。如今,这一观点也得到了印证, DeepSeek一路狂飙,把曾经的AI原生APP“三强”远远甩在了身后。

“‘豆包们’增速放缓有迹可循。”成都市人工智能产业协会秘书长李娅娜在接受《中国电子报》记者采访时表示。在技术迭代上,“豆包们”普遍依赖通用性模型,复杂逻辑推理能力比DeepSeek低20%左右, Kimi在长文本处理能力上被DeepSeek超越,导致其用户增速下降了35%;在营销策略上,豆包虽然月活跃用户(MAU)达7861万,但海外用户占比仅15%,且未开放核心模型,限制了开发者生态扩展。相比之下, DeepSeek的开发者生态规模是豆包的3倍。此外,随着用户需求变

迁,职场场景对AI专业度要求更高,根据用户反馈的数据,超60%的用户认为豆包在“深度思考”能力上存在不足。

不过,现在下结论还为时尚早。Gartner高级研究总监方琦对《中国电子报》记者说道:“DeepSeek-R1接近模型第一梯队的表现,将互联网搜索和推理模型的表现,让大量的用户得到了新的体验。但是从中长期来说, DeepSeek应用存在各种不稳定性,等用户的新鲜劲儿褪去之后,比拼的还是产品的运营以及是否能够真正帮助用户解决他们的问题,并持续提供优秀的用户体验。从这点来讲,‘豆包们’没有变慢,如何做好自己的产品能力和用户留存才是最重要的问题。”

要在激烈的AI大模型竞争中分得一杯羹,“豆包们”必须主动出击,差异化破局。

只能被动跟随, 还是应主动出招?

DeepSeek头顶的光环越亮眼,就越映衬出“豆包们”的尴尬局面。“对投资人来说, DeepSeek出现以后很难再出现第二个了。豆包砸了这么多钱, DeepSeek零广告,却席卷全球。”洪泰基金创始人、资深投资银行家盛希泰感慨道。

实际上,“豆包们”也并非全无胜算。以豆包为例,背靠字节跳动,既有算力预算充足和数据高速增长带来的模型训练优势,又有以抖音为核心的自有流量池,豆包的竞争力并不弱。根据Strike Digital统计的数据,截至2025年1月, TikTok(国际版抖音)的全球用户

规模估计为20.5亿,且预计每年将继续以2%至3%左右的速度增长,这也为豆包提供了天然的用户入口。

“通过硬件撬动AI软件的流量,也是一种很好的做法。”Quest-Mobile产业研究院研究总监陈燕表示,对比硬件上市前后的流量,耳机的出货量对豆包月活起到了比较好的拉动作用。据了解,字节跳动在2024年10月推出一款名为Ola Friend的智能耳机。如果用户想去调用AI功能,就需要下载豆包APP,这种硬件绑定的方式可以大幅提升大模型应用的用户黏性。

不过,豆包也存在劣势。比如,技术方面,模型创新性不足,在垂直领域的深度比行业领先模型低30%,主要依赖热点更新维持用户黏性;生态方面,开发者生态薄弱,企业级定制化服务能力受限,目前企业级客户数量仅为行业领先者的20%;全球化方面,进程相对滞后,在海外市场特别是欧美地区市场占有率不足5%,未抓住全球AI应用爆发期。

要在激烈的AI大模型竞争中分得一杯羹,“豆包们”必须主动出击,差异化破局。李娅娜指出,一方面,要聚焦垂直场景深耕,例如在教

育领域开发针对学生个性化学习的专用模型,在医疗领域开发辅助诊断模型等,借鉴DeepSeek开源策略,开放API接口,吸引开发者,构建“模型+工具链+社区”生态。另一方面,像豆包、文小言等大模型的背后都有实力雄厚的母公司支持,应最大限度地撬动资源,打造差异化能力,比如Kimi强化情感交互、豆包探索“AI+短视频”内容生成,与DeepSeek形成场景区隔。此外,还应与华为昇腾、寒武纪等厂商合作,优化硬件适配,同时加速拓展海外市场,重点抢占东南亚、印度等新兴市场。

开源模型厂商需要找到合适的商业模式,可持续地为市场提供价值,对这类厂商来说也是一个挑战。

2025年将成为“大浪淘沙”的一年

从2023年的“百模大战”,到2024年令人眼花缭乱的行业应用,再到2025年的“大浪淘沙”,大模型行业的竞争格局正在从分散走向收敛。咨询机构弗若斯特沙利文指出,中国目前在通用基础大模型领域的竞争者已缩减至20余家,主要由互联网企业、云计算巨头及人工智能创业公司主导。

与产业链上下游相比,处于中游的模型厂商普遍面临盈利困境。一方面,从云服务商手中购买算力开支巨大;另一方面,产品商业化目前仍处于探索阶段,远未能形成稳定的和足够覆盖成本的收入。DeepSeek的出现有望改变这一局面,加速模型平权和普惠AI的

到来。

“Gartner在2023年9月就提过真正做大模型的企业不会超过10家,在这一轮洗礼后,如果一个商业化模型的性能接近或者还比不上开源模型,那它的商业化就大概率不会成功。同时,开源模型厂商也需要找到合适的商业模式,可持续地为市场提供价值,对这类厂商来说这也是一个挑战。”方琦坦言。

“AI大模型洗牌趋势是从野蛮生长到价值验证,市场竞争的关键在于技术突破、生态构建与场景落地。”李娅娜表示。能存活下来的大模型应该具备三大核心特征:一是低成本与高性能平衡,如DeepSeek通过MoE架构将推理成本降

低40%,并适配国产硬件;二是开源,开源模型将主导开发者生态,预计未来开源模型市场份额将达70%;三是专业化,教育、工业等细分领域专业化模型更具生命力,预计未来两年垂直细分领域模型市场规模将增长80%。而那些存在技术同质化问题、缺乏差异化能力、应用场景模糊、高度依赖于资本输血而非商业闭环的大模型将逐渐被市场淘汰,预计2025年将有50%以上的大模型因商业化失败而退出。

此外,全球化能力也很重要,必须做好本地化运营,避免因于单一市场。从目前国内外专业榜公布的相关数据来看,国内大模型除了

DeepSeek、Qwen-Max(阿里旗下)等少数应用能够跻身全球前十排名以外,其他绝大多数大模型仍局限于国内应用榜内。DeepSeek的冲击尽管给它们带来了巨大压力,但并未尝不能成为本土大模型的“觉醒”动力。

“当前, AI行业竞争已从技术单点突破转向‘技术+生态+场景’综合较量。企业需以用户需求为导向,解决实际问题,如提升办公效率、辅助医疗诊断、优化工业生产等,在成本、性能、差异化中找到平衡,才能在浪潮中立于不败之地。”李娅娜表示。经过新一轮“大浪淘沙”,未来将会有更多“金子”破土而出,在国际市场上绽放光彩。

微信接入DeepSeek

本报讯 记者宋婧报道:近日,微信正式内测“AI搜索”新功能,在其中接入了DeepSeek-R1的“深度思考”服务。对此,腾讯官方回应称,部分测试用户,可在微信对话框顶部搜索入口,看到“AI搜索”字样,点击进入后,可免费使用DeepSeek-R1满血版模型,获得更多元化的搜索体验。

此次接入DeepSeek的微信AI搜索功能,不仅整合了微信公众号、视频号等腾讯生态内的信息源,还支持联网搜索,能为用户提供更全面、实时、准确的回答。目前该能力还在灰度测试中,将根据用户体验和反馈持续优化。

该消息带动腾讯概念股持续走强。2月17日,腾讯控股一度涨超7.5%,市值暴增约3200亿港元(约合人民币2981亿元)。其相关概念股也随之跟涨,数据港、云赛智联、东华软件、弘信电子、科华数据等竞价涨停。

实际上,超级APP接入大模型服务,并非新鲜事。在微信之前,抖音、百度、快手等都已完成类似操作,不同之处在于,它们都是选择接入自家自研的大模型产品,微信则一反常态,成了首家在自家产品中接入外部大模型的科技大厂。

不过,从微信接入DeepSeek的现状来看,其也保留着一些审慎的态度。根据部分内测用户的体验,当前接入微信的DeepSeek-R1,与官方APP存在两处体验上的不同:一是不支持多轮对话问答;二是不支持查看历史问答记录。业内人士分析称,在DeepSeek

频频因用户涌入过多而导致服务卡顿之下,上述举措或许也可以看作是微信保持用户流畅体验的应对之举。

高盛最新研报指出,腾讯是首批将DeepSeek-R1-671B模型深度思考模式并且实现搜索功能的国内CSP(云服务提供商)企业。利用腾讯独特微信内容提供增强推理能力,并由腾讯云AI提供推理基础设施支持,高盛认为,这凸显了腾讯在AI开放平台的一种倍增战略,利用内部(混元)和外部模型(例如DeepSeek),希望在中国打造AI ToC“杀手级”应用和AI Agent智能体生态。

国盛证券研报分析称,微信生态内拥有购物消费、交通出行、生活服务、金融理财等多类九宫格服务。如果长期将搜索问答和这些服务供给之间引流打通,或有望在“搜索入口-服务供给”之间形成显著正向循环,大幅提升用户黏性和变现空间。

根据2月18日的最新消息,由于微信搜索灰度测试AI搜索过程中,用户对DeepSeek模型的使用热情远远超出预期,为了让更多用户方便稳定地体验和使用DeepSeek,腾讯内部紧急协调部署,调用元宝支持微信搜索,给用户提供更多入口和选择。即日起,尚未被灰度到AI搜索的用户可通过腾讯元宝,免费使用DeepSeek-R1满血版、DeepSeek-V3、腾讯混元深度思考模型T1以及腾讯混元通用模型Turbo。

蚂蚁集团将自研人形机器人

本报讯 记者王伟报道:2月19日,记者从猎聘网获悉,蚂蚁集团旗下全资子公司上海蚂蚁灵波科技有限公司(以下简称“蚂蚁灵波”)开放招聘具身智能人形机器人软硬件相关岗位,此举标志着蚂蚁集团正式开启自研人形机器人的新征程。

公开信息显示,蚂蚁灵波于2024年12月成立,注册资本1亿元,由蚂蚁智能(杭州)科技有限公司(蚂蚁集团全称)100%持股。

蚂蚁灵波在猎聘网发布的招聘岗位有人形机器人系统和应用软件工程师、人形机器人硬件结构工程师、人形机器人运动控制软件工程师、人形机器人整机/硬件产品专家等,工作地点在上海或者杭州,

薪酬在35K~80K之间,提供16薪。

事实上,蚂蚁集团早已涉猎机器人领域,但都是作为投资方的身份。去年11月,蚂蚁集团旗下子公司上海云场企业管理咨询有限公司就曾入股具身智能机器人开发商星海图(苏州)人工智能科技有限公司。此外蚂蚁集团还曾经参投仓储物流机器人公司极智嘉,在极智嘉1月向港交所递交的《招股书》显示,蚂蚁集团通过Accelerator VI Ltd和上海云场合计持有4.93%的股份。

截至记者发稿时,蚂蚁集团官方回应表示:“蚂蚁集团去年注册成立了蚂蚁灵波科技有限公司,聚焦具身智能技术和产品研发。蚂蚁集团在持续布局和加大AI投入。”

DeepSeek发布NSA注意力机制

本报讯 2月18日, DeepSeek团队发布了一篇新论文,介绍了一种改进的稀疏注意力机制NSA,可用于超快速的长上下文训练与推理。NSA以性价比极高的方式,罕见地在训练阶段应用稀疏性,在训练场景中均实现速度的明显提升,特别是在解码阶段实现了高达11.6倍的提升。

据DeepSeek介绍, NSA具有三大核心组件:动态分层稀疏策略、粗粒度token压缩、精粒度token选择。通过三大组件的配合,既提升了效率,也保留了模型对全局长上下文的感知能力和局部精

确性。

这一机制专门针对现代硬件进行优化设计,原生支持模型训练,在加速推理的同时降低训练成本,对性能也无明显影响。采用NSA机制的模型在通用基准、长上下文任务和基于指令的推理上,与全注意力模型相当或表现更优。

在8卡A100计算集群上, NSA的前向传播和反向传播速度分别比全注意力快9倍和6倍,由于减少了内存访问量, NSA在长序列解码时相较于全注意力模型速度显著提升。(文 编)

奇安信与东方国信达成战略合作

本报讯 近日,奇安信与东方国信达成战略合作,携手打造基于DeepSeek基座的人工智能智算平台,致力于加速推动人工智能的广泛应用,并解决AI基座自主可控、降本增效以及开源生态等关键问题,共同推动我国人工智能产业迈向新高度。

此次合作标志着两家行业龙头在技术创新上的深度融合,更预示着“AI+安全+算力”的无限想象空间。东方国信作为大数据、云计算及人工智能领域的龙头企业,其自主研发的幕僚智数、大模型训推平台等核心产品已成功与DeepSeek-R1系列大模型实现深度集成,广泛应用于

智能客服、智能经分等行业场景中。奇安信凭借其在网络安全行业的领导地位,将自主研发的QAX安全大模型全面接入DeepSeek,并率先应用于威胁研判、安全运营等多个关键领域,展示了卓越的技术实力和创新能力。

奇安信将为东方国信提供涵盖从智算基础设施到大模型及行业应用全层级的端到端网络安全防护方案,确保AI应用的安全性与可靠性。同时,双方还将结合政企客户需求,开发定制化的安全AI应用智能体融合方案,通过深度融合内生的方式提升安全运营、智能攻防对抗、威胁情报分析等复杂任务的处