

高性价比 加速圈粉全球

近年来,全球AI市场的竞争格局正在悄然生变。全球AI模型聚合平台OpenRouter公开运行数据显示,2026年以来,中国大模型周调用量稳居全球前列,多个头部模型长期占据调用量榜单前排。在2026年7月下旬的一个统计周,该平台调用量榜单前五均为中国企业模型。

开源社区的表现同样亮眼。在全球最大AI开源社区之一“抱抱脸”(Hugging Face)上,中国已成为平台月下载量最大的模型来源国。该社区2026年春季报告显示,中国研发的开源模型下载量已占全球总量的41%,累计下载量突破百亿次,深度融入全球开源创新体系。过去一年里,多款中国开源模型长期位居平台下载量前列,从技术追赶者成长为全球开源生态的重要供给方与驱动力量。

海外科技界的态度最能说明问题。英伟达首席执行官黄

仁勋多次公开称赞中国开源AI模型“非常优秀”;特斯拉首席执行官马斯克也为中国人工智能点赞。中国AI技术的全球认可度持续提升。

用户的选择标准向来直白:一是模型能力是否过硬、响应是否高效,能不能真正解决实际问题;二是使用成本是否合理,能不能长期稳定使用。词元(Token)调用量的本质是真实使用量,背后折射的是AI大模型在生产生活场景中的渗透深度与应用广度。中国AI大模型被全球用户高频选择,就是对这个选择标准最好的回应。

仁勋

多次公

开称赞中国

开源AI模型“非常

优秀”;特斯拉首席执行官

马斯克也为中国人工智能点赞。中国AI技术的全球认可度

持续提升。

从产品落地来看,中国大模型早已走出国门,融入全球千行百业。新加坡工程师基于中国开源大模型,训练出适配本地需求的大模型,服务当地数字经济发展;美国创作者使用中国的视频生成模型,制作出达到专业级水准的视觉内容,大幅降低了创意生产门槛。从东南亚的政务服务到中东的能源行业,从欧美的初创企业到拉美的电商平台,中国大模型正在用实实在在的能力赢得全球用户的信赖。

架构革新跑出“中国效率”

中国大模型能实现性能与成本的双重优势,秘诀在于技术路线的创新突破,其中混合专家(MoE)稀疏架构的规模化应用是关键一环。

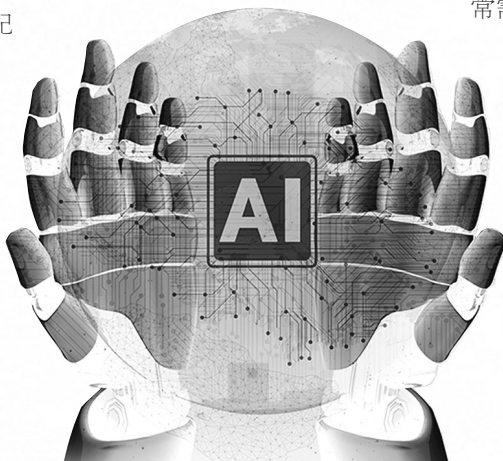
不同于传统稠密模型每次推理都要激活全部参数的模式,混合专家架构通过“门控网络”按需调用少量“专家”子网络,大部分参数处于休眠状态,在保障输出质量的前提下,大幅降低了算力消耗。国内不少头部大模型团队已布局这一技术路线,针对路由负载不均、相关通信瓶颈等行业共性工程难题开展优化,不断提升算力利用效率,力求以更少算力实现同等输出质量。

在此基础上,国内团队在全链路工程优化上持续深耕。原生低精度量化技术大幅降低模型显存占用;KV缓存(键—值缓存)与多级缓存技术显著减少重复计算;系统级通信与流水线

调度优化,进一步释放算力集群潜能。国内大模型逐步形成“技术迭代—成本下降—应用爆发”的正向循环,依靠算法创新与工程优化不断压低推理成本,构建起突出的性价比优势。

稳定的能源供给与完善的算力基础设施,为中国大模型形成成本优势提供保障。中国智能算力规模位居世界前列,绿电供给能力持续增强,规模化数据中心建设依托规模效应助力降低单位算力成本。叠加国产算力芯片持续适配与量产落地,国内AI产业逐步构建更具韧性与竞争力的产业体系,为向全球用户提供普惠AI服务打下坚实基础。

为什么中国大模型便宜又好用?



开源生态 汇聚全球迭代合力

中国大模型的快速进步,离不开开源开放的发展思路。与海外部分头部企业坚持闭源路线不同,中国主流大模型团队普遍选择了开放权重或部分开源的策略,主动融入全球开源创新生态。

开源带来的最直接效应是加速迭代。模型权重开放后,全球数百万开发者可以排查缺陷、贡献优化方案,海量真实场景的反馈源源不断地回流到模型团队,推动能力快速进化。这种全

球众包式迭代的速度,远快于闭源团队闭门打磨的节奏。业内普遍认为,开源不是单向的技术输出,而是最高效的创新方式,全世界开发者共同参与建设,最终会推动整个行业生态持续繁荣。

开源生态的繁荣也反过来放大了中国模型的全球影响力。国际主流AI开发平台第一时间接入中国大模型,不少海外推理平台将中国新模型的上线视为行业重要进展。越来越多的海外开发者基于中国大模型开展二次开发,适配本地语言、行业场景与文化需求,催生出一批新技术、新应用,进一步丰富中国模型的生态版图,形成良性双向循环。

更重要的是,开源降低了AI技术的使用门槛。对于广大发展中国家和中小企业而言,无需投入巨额资金从零训练模型,依托成熟的中国开源模型便可快速搭建自身AI应用。这也印证一项行业趋势:中国大模型正从单一产品出海,迈向技术底座出海,为全球数字经济发展赋能。

场景实训练就“实用派”硬核本领

海外用户觉得中国大模型“好用”,不只体现在实验室跑分上,更体现在解决实际问题的能力上。而这种实际应用能力的提升,离不开中国丰富应用场景的持续打磨。

中国拥有全球最庞大的互联网用户群体与门类完备的产业应用生态。电商客服、文档处理、代码开发、内容创作、工业质检、政务服务……海量真实应用场景持续对大模型提出现实需求,推动模型在实用性层面快速进化。以长文本处理为例,国内企业常

常需要处理上百页合同、整套代码库,这类需求推动国产模型普遍配备超大上下文窗口,而这也是海外企业用户面临的突出痛点。不少海外用户反馈,在处理长文档、复杂

任务时,中国模型的实际表现往往超出预期。

智能体、代码生成等前沿能力的突破,同样源于现实场景的淬炼。国内互联网、制造业数字化转型需求,推动国产大模型在工具调用、任务规划、代码生成等维度持续迭代。不少海外开发者实测表明,中国模型在终端任务执行、代码调试、多工具协同等复杂工作场景中表现稳定,能够有效提升开发效率,同时综合使用成本具备较强吸引力。事实上,实验室跑分与真实场景的适配度本就存在差异,不少模型在标准评测中分数亮眼,但面对真实场景下的模糊需求、异常情形,往往容易出现适配偏差。中国模型经过海量用户实战场景打磨,处理生产生活等各类现实任务表现更加顺手,这种贴近实际需求、也是获得全球用户认可的重要原因。

(记者 杨俊峰)

来源:新华网