

大模型定价规则重构？



图为多模态世界大模型智源悟界·Emu3.5 文生图

本报记者 陈存

近日,当下国外炙手可热的两款大模型——Claude Fable 5和GPT-5.6模型各自官宣了最新的定价策略。Anthropic宣布Claude Fable 5订阅用户限时免费时间延长至7月12日,免费时间结束后,Claude Fable 5将退出订阅计划,用户需以输出50美元每百万Token、输入10美元每百万Token的价格按使用量购买。

同为世界顶尖级大模型,相比前代的5.5模型,GPT-5.6模型则整体呈现旗舰版价格持平、均衡版价格减半、轻量版价格大降的格局,即使是最贵的旗舰版(Sol)模型价格也仅为Claude Fable 5的一半。

同一个市场,同一个赛道,有人在涨价,有人在降价。过去的两个月里,类似的剧情也在中国大模型市场里频繁上演。

首次涨价潮 释放了什么信号

国内大模型行业似乎一直遵循着这条发展轨迹：模型越来越强，价格越来越低，低价甚至免费一度成为各家厂商抢占用户的重要依凭。不少业内人士认为，随着模型的不断迭代，大模型最终将像云计算一样进入“成熟智能通缩”阶段——即能力逐渐趋同，价格持续下降。

年初,这一规律首次失效了,国内大模型企业迎来了两年来的第一次系统性涨价潮。智谱AI年内三次上调核心模型API价格,编程套餐涨幅超30%;腾讯云混元系列模型价格大幅上调,部分版本涨幅超460%;MiniMax同步上调付费套餐和API价格;豆包专业版延续此前公布的三级订阅方案,选择收费,引发了诸多讨论。

更深层的信号来自另一则消息:7月4日,豆包和通义千问相继发布通知,宣布旗下智能体功能将于7月15日正式下线。两款国内头部AI产品几乎同时关闭智能体功能,其原因不难理解——C端智能体功能运营成本高、变现模式受限,正适合做“减法”。

近年来,大模型企业持续投入模型训练、推理服务和产品研发,算力、存储、芯片等基础设施成本居高不下。随着智能体、多模态等需求的不断增加,单次任务消耗的计算资源也明显增加。如何在扩大用户规模的同时

实现商业闭环,已成为行业需要面对的首要问题。

赛迪研究院信息化与软件产业研究所人工智能产业研究室副主任刘丽超对《中国电子报》记者表示,模型产品在推理、代码、复杂任务等方面能力的持续增强,正在驱动企业从烧钱获客转向寻求可持续商业盈利,通过价值付费模式接受市场检验。

除此之外,大模型供给井喷以来,用户对于AI付费的认知也在不断提升。相较讨论该不该收费,许多用户更关心的是定价是否合理、模型能力是否切实提升。

在遥远的未来,大模型行业或许真如科技板块知名投资者加文·贝克(Gavin Baker)所描述的那样:“前沿模型攫取90%的经济价值,而开源与廉价模型承载80%的Token消耗。”

降价的底气 源自何方

然而,这波价格战比表面看起来还要复杂得多。系统涨价潮结束不久,新一轮“跌破”地板价的大模型接踵而来。今年5月,DeepSeek宣布将V4-Pro模型API价格永久调整为原来的1/4,每百万Token输入(缓存命中)最低降至0.02元;小米也宣布Mi-Mo系列模型API永久降价,最高降幅达到99%。

在短时间内,多家企业密集调价,容易让人联想到新一轮价格战。但与此前单纯依靠烧钱争

夺市场不同,这一轮降价主要得益于模型推理效率和工程能力的提升。

摩根大通在全球中国峰会上表示,中国国产芯片和LLM持续改善,上行空间扩大。无间芯穹CTO李伯勋透露,过去一年,公司推理成本已下降约10倍,希望未来3至6个月每段时间都能持续下降50%。他认为,未来一定是单位Token成本持续下降,而Token成本规模总收益上升的。

此外,刘丽超向记者表示,降价背后还有更深的目的。大模型之争说到底还是入口之争,低价或免费API能快速吸引中小开发者,壮大开发者应用生态,在行业格局未定前卡位市场份额。

值得关注的是,行业探索的已不仅是降价,还是更加精细的定价方式。6月底,DeepSeek推出峰谷定价机制,根据不同时间段动态调整API价格,高峰时段(每日9:00-12:00及14:00-18:00)价格将翻倍。这意味着,Token计费开始进入“电价化”时代,价格成为重新配置AI资源的重要工具。

商业模式变化 从卖Token到卖能力

涨降价并存的表象之下,一个更深层的变化正在发生:大模型的定价单位,正在从Token转向任务和能力。

刘丽超指出,当前“涨降价并存”的状态并非行业混乱,而是‘统一按Token计价’模式失效,行业开启差

异化竞争的大模型定价体系重构”。她认为,一段时间后可能形成分层相对稳定的更精细的价格体系,即高价值场景基于任务执行效果,接受模型能力与可靠性溢价;通用任务强调专业化服务支撑,更为注重性价比。

这一判断正在被产业实践验证。除了买Token外,市场上已经出现了买Agent能力、买任务结果等交易模式。刘丽超分析,在买Agent或买任务结果的模式下,“用户更关心成功率、可靠性、响应延迟等实际执行效果”,这意味着竞争核心正从“参数竞赛”转向“实用能力”,在Agent框架、工具调用、自我纠错、长期记忆、推理成本控制等方面积累深厚的厂商,有望获得更高用户黏性和溢价空间。

回到定价本身,分化定价终将促成崭新的多层次竞争格局,并对中小模型厂商带来诸多挑战。对他们而言,追寻极致的智能聪明或极致的低价都是艰难的,更可行的方法是寻找自己的“帕累托前沿”,即在模型能力和性价比之间找到那个最优的平衡点;同时,努力开拓具备特色化、专业化的解决方案能力,深耕巨头难以兼顾的垂直场景、私有部署、行业定制或轻量化服务。

事实上,收费探索与持续降价并不矛盾,而是在回答同一个问题:如何让大模型商业化走得更远。免费的时代没有结束,但所有能力都免费的时代结束了。换言之,只有足够出色或者便宜的模型才能在激烈竞争中站稳脚跟。

微观察

DeepSeek 向世界头部AI模型企业看齐

陈存 崔烁

7月7日,有媒体援引知情人士消息称,中国AI大模型企业DeepSeek(深度求索)正在开发自有AI推理芯片,该芯片的研发工作大约在一年前启动,目前仍处于早期阶段。

坊间也透露,DeepSeek已与芯片设计公司、晶圆代工厂及存储厂商展开接洽。在人才队伍建设方面,近几个月,该公司加大了芯片设计工程师的招聘力度,但均通过猎头和内部推荐等非公开渠道进行,未在任何公开招聘平台发布职位信息。

虽然DeepSeek目前尚未对此事有公开信息发布,但结合Deep-Seek目前的地位和能力,此举符合逻辑。

长期以来,DeepSeek同时使用英伟达和华为芯片,其底层基础模型曾在英伟达H800上完成训练,美国对华实施出口芯片管制后,DeepSeek逐渐转向华为昇腾芯片,今年4月发布的DeepSeek-V4模型已适配华为昇腾架构。业内人士普遍认为,DeepSeek自研芯片旨在降低对外部芯片的依赖。

随着AI应用大规模铺开,行业算力消耗的重心正从模型训练转向推理运行。相比兼顾训练与推理的通用GPU,面向特定模型定制的推理专用芯片能够在功耗和单位成本上形成显著优势。单一供应商的依赖风险、高端芯片供应紧张以及高昂的推理成本,促使这家以算法优化著称的公司向上游硬件延伸,以适应自己巨大的API(应用程序编

程接口)服务量,自研芯片能直接降低推理成本、优化响应速度。况且,DeepSeek自然是对自身模型架构理解最深的那一方,将这些知识直接固化到芯片设计中,能够实现通用GPU无法达到的软硬协同效率。

当然,芯片研制不是简单“攒硬件”,要缩短从设计到流片再到量产的技术周期、突破制造工艺和生态壁垒,只有华山一条路——加大投入。幸运的是,这份必将不小的新开支,已有了第一笔资金托底。2024年,DeepSeek创始人梁文锋表示,DeepSeek“目前没有短期融资计划,我们真正的挑战从来不是资金,而是高端芯片的出口禁令”。时至今日,即使仍需面临同样的芯片难题,这家曾经拒绝外部资金进入的公司,至少改变了对于融资的态度,在6月完成了约510亿元的首轮融资。可以预见,其中很大一部分将被用于芯片研发。

从行业趋势看,AI大模型厂商自研芯片也已成为趋势,OpenAI上月刚刚发布了与博通合作开发的首款定制推理芯片Jalapeno,谷歌、Meta、Anthropic等全球AI大模型头部厂商也都在积极推进自研芯片进程,DeepSeek作为开源大模型的领军者,跟进这一趋势,符合其“江湖地位”。

DeepSeek着力夯实硬件基础,这也在一定程度上表明,中国AI企业正从模型算法竞争,逐步迈向全方位自主能力建设的新阶段。从产业的发展角度,既需要与上游企业“深度绑定”的发展模式,也欢迎有更多企业加大自研,更多“全栈自研”企业能获得成功。

河南仰韶文化AI国风XR项目 正式上线

本报讯 记者齐旭报道:7月7日,仰韶文化AI国风XR沉浸式互动项目《仰韶·华夏启元之异兽山海》在河南仰韶文化博物馆正式面向全国上线。

据了解,该项目由河南仰韶发展投资集团、三门峡市文化旅游交通发展集团、河南大象融媒体集团联合打造,以“仰韶源代码·唤醒七千年”为主题,依托AI国风XR数字技术,让沉睡七千余年的彩陶纹样在光影交互中流动重生。项目以最前沿的数字技术唤醒沉睡的文明基因,打破了传统博物馆“隔着玻璃看文物”的静态模式,是三门峡推动“科技+文旅”深度融合的标志性成果。

记者了解到,项目提取了仰韶彩陶上的鱼、鸟、蛙、虎等图腾纹样,结合《山海经》宏大叙事,运用AI+CG技术重构28只沉睡千年的神兽。每一只神兽,都是文明基因的活态载体。

据介绍,依托“AI叙事引擎+空间智能系统+文博知识库”三大技术模块,项目有效解决了遗址观赏性弱、文化内涵理解门槛高的行业痛点——让文物开口说话,让历

史照进现实。项目构建了电影级标准的完整四幕剧情线,游客不再是旁观者,而是“寻火者”,是文明的亲历者。同时布局线上游戏、动漫番剧、影视授权、文创开发等全产业链体系。

在发行战略上,该项目采取“双轨制”:针对全国成熟影院、商场等场馆,制定标准化“开机计费”模型;针对考古博物馆等核心文化地标,采取“一馆一策”联营联票模式。让仰韶从河南的仰韶,成为中国的仰韶;从中国的仰韶,成为世界的仰韶。发布会现场,项目方与5家平台企业完成战略合作签约,共同构建稳固的“1+2+2”生态矩阵。

据了解,即日起,《仰韶·华夏启元之异兽山海》体验空间于渑池仰韶时代文化园对外开放,全国各地游客均可前往沉浸式溯源华夏文明主根脉。项目同步深度联动研学教育赛道,面向青少年打造趣味化史前文明课堂,让青少年在虚实交互中触摸历史肌理。未来,市民游客前往渑池仰韶时代文化园,便可沉浸式读懂华夏文明源头,让沉寂文物走出展柜,让您远历史触手可及。



推动科技创新 和产业创新融合发展