

## 淘汰赛开打:大模型免费的表与里

本报记者 秦淼 北京报道

大模型问世仅仅一年,就迎来了免费时代。

在字节跳动的大模型喊出比行业便宜99.3%之后,其“行业最低价”正在被不断打破。近日,阿里云、百度等巨头公司纷纷宣布将大模型价格降到更低,百度甚至直接宣布免费。科大讯飞更是宣布,其讯飞星火API能力正式免费开放。其中,讯飞星火Lite API永久免费

## 没有最低,只有更低

短短数日,各家大模型厂商均已投入到这场“价格战”中。

5月22日,腾讯云正式对外宣布了全新的大模型升级方案,其核心模型之一的混元-lite模型价格策略发生重要调整,由原先的0.008元/千tokens调整为全面免费使用。同时,最高配置的万亿参数模型混元-pro也进行了价格优化,从原有的0.1元/千tokens下降至0.03元/千tokens。

短短数日,百度、阿里云、腾讯、字节跳动、科大讯飞等各家大模型厂商均已投入到这场“价格战”中。而国内这场“价格战”最早可追溯到5月6日,幻方量化旗下DeepSeek发布第二代MoE(专家模型)DeepSeek-V2,API定价为每百万tokens输入1元、输出2元(32K上下文),价格为GPT-4-Turbo的近百分之一。

token既是大模型文本处理的最小单位,也是最基础的计量计费单位,用于衡量产品服务使用量和商业化收费标准。

大模型在生成内容的过程中,需要先把人们对话输入的文本转化成一个个的token,然后根据对上下文tokens信息的理解和分析,预测接下来应该生成的tokens内容,并将这些生成的tokens转换成人们熟悉的文本内容进行输出。

国内外的计算标准并不统一,按已有惯例,1个token≈1—1.8个汉字,而在英文文本中,1个token≈

开放。

多位业内人士在接受《中国经营报》记者采访时表示,尽管宣布降价的模型产品众多,但真正大规模、高性能并且支持高并发的大模型推理仍然需要收费,降价幅度有限。不过,可以明确的是,这场关于大模型免费使用的价格战才刚刚开始。也许很快人们就会看到,在大模型这场比赛中,大多数的大模型公司都将被淘汰。

3—4个字母。

从各家公布的价格来看,一元就能买到字节跳动豆包主力模型的125万tokens,大约是200万个汉字,相当于3本《三国演义》。同样,花一元在通义千问上最多可以买到200万tokens,相当于5本新华字典。

“由于各家不同模型的规格和支持tokens的长度各不相同,直接进行价格的横向对比过于片面。”一位业内人士对记者表示:“尽管此次降价为我们带来了一定的经济优惠,但可以看到,对于一些在特定使用需求下的模型或功能,一些厂家并没有采取免费,只是通过不同程度的降价手段来尽量保持市场地位。”

九三学社中央科技委委员陈根坦言:“这背后更多的是无奈与挑战。由于大模型构建在强大的算力基础之上,而这些算力的硬件投入与日常运维,都需要庞大的资本支出。因此,大部分公司都难以满足算力的支出需求。为了尽可能减少亏损,只能是最大程度地减少免费使用的范围。”

北京社科院副研究员王鹏认为,高性能模型推理服务主要面向企业级用户,他们对服务的稳定性、可靠性和性能有更高要求。企业级用户通常愿意为这些增值服务支付额外费用。

## “幕后推手”

互联网企业的案例在此刻隐隐若现。

实际上,真正让国内大模型神经紧张的是引领了此次AI大模型时代的OpenAI。

OpenAI从2023年年初发布GPT-3至今,已经进行了4次降价。5月13日发布的GPT-4o不仅实现了性能跃升,价格也下降了50%。5月22日凌晨,在微软Build开发者大会上,微软CEO纳德拉便表示,过去一年GPT-4的性能提升了6倍,但成本降低到了之前的1/12,对应性能/成本提升了70倍。

“国产大模型大幅降价的核心在于OpenAI最近不断推出的各种免费模式,并且以超乎大部分大模型公司预计的速度迭代了技术。这就意味着,更强大的模型采取了更开放的策略,并且便捷到不用注册就能使用。”陈根表示:“当然OpenAI推动大模型免费的策略,主要的目的有两方面:一方面是为了更大程度地垄断用户数据与资源,以迭代其技术,并获取更多用户行为数据;另一方面是为了提高并保持领先的用

## “终点”在哪儿

通过降低价格门槛,大模型厂商有望吸引更多广泛的企业用户群体,从而进一步平衡收入和成本。

“站在整个AI业界的角度,我想说,降价是个拍脑袋就可以做的简单策略,但是真正的to B商业成功更难。”Lepton.ai创始人贾扬清公开表示,“今天不是说API贵才没有人用,而是因为要搞清楚,到底怎么用起来产生业务价值。”

2024年是大模型商业化落地的元年,应用落地成为行业竞争焦点。经过过去一年的发展,目前国内大模型的头部梯队已接近GPT-4的能力,打磨“内功”同时将目光瞄向用户,通过降价来换取更大的市场空间。而这场“价格战”的目的被归结为加速AI应用。

户数量。国产大模型厂家面对OpenAI的免费策略,如果要留住用户,并且获得用户的人机交换数据来推动技术升级,就必然要跟进免费的策略。”

然而,在上述业内人士看来:“大模型最终的结果会像之前的互联网之战一样,只剩下两三家而已。”

中国民协元宇宙工委创办人吴高斌表示,这种简单粗暴的“价格战”打法,可能会在一定程度上扰乱市场秩序,对大模型行业的健康发展造成影响。尤其对于中小企业及初创企业来说,面临巨大的竞争压力,可能会在价格战中逐渐消失。

天使投资人、资深人工智能专家郭涛认为,从目前的趋势来看,大模型行业已经进入了洗牌期。在这轮价格战中,中小企业和初创企业都面临着巨大的压力,部分企业可能会被淘汰。然而,这并不意味着这些企业一定会消失。一些具有核心竞争力和创新能力的企业仍然有机会在竞



随着国内外大模型厂商技术角逐进一步激烈,大模型行业开启价格战信号明显。视觉中国/图

争中突围,成为行业的新领军者。

零一万物CEO李开复也强调:“大模型从训练到服务都很昂贵,算力紧缺是这一赛道的集体挑战,行业应当共同避免陷入不理性的of0式流血烧钱打法,让大模型能够健康地蓄能长跑。”

的确如此,“千团大战”“O2O混战”“网约车补贴大战”等互联网企业的案例在此刻隐隐若现。

但在百川智能创始人王小川看来:“千团大战”“网约车补贴大战”等背后都带有网络效应,那种商业模式改变了生产关系,改变了司机和乘客之间、外卖员和用户之间的关系。这次价格战并没有改变生产关系,而是直接做生产力供

的关键期。麦肯锡的报告也指出,在其研究的63种应用中使用生成式AI,将为全球经济带来每年2.6万亿—4.4万亿美元的增长。此轮国内大模型轮番降价,既是“百模大战”竞争的进一步加剧,也是大模型商业化拐点将至。

华龙证券在研报中表示,通过降低价格门槛,大模型厂商有望吸引更多广泛的企业用户群体。同时,更多消费端用户有望免费使用基础AI应用,庞大的访问量有助于企业进一步提升大模型的服务能力,完成良性循环。

郭涛表示,大模型的“终点”可能是实现真正的通用人工智能(AGI),即一种能够像人类一样在

给,是B端市场的价格战。”

陈根也认为,这与互联网的概念炒作有本质的区别。大模型本身就是一项重资产、重资本的产业变革技术,并且在人和财两方面都是重资产模式。

他认为,从这个角度来看,免费对于大模型而言是一件好事,至少可以加速淘汰市场上各种杂乱无序的投机与炒作公司,让资源集中到头部公司中,从而加速中文大模型的完善。从目前的技术发展趋势来看,OpenAI在加速实现通用AI的这一目标。这对于国产大模型而言,危与机并存,核心就取决于我们是否能率先在通用AI的技术上实现突破。

## 暑期档“前哨战”打响 13部影片挤进端午档

本报记者 张靖超 北京报道

截至5月30日,已有13部影片定档今年端午档。从数量上看,这是有史以来“最挤”的一届端午档。

## 端午档接着“挤”

档期拥挤成为2024年电影市场的一大特点,首先是在3天时间的清明档,十余部影片展开角逐;而后在“五一”档,9部新片定档,数量为历史第二高。

为何今年频频出现“未热先挤”的现象?

在题为《观众购票决策的变化趋势》的报告中,阿里影业指出,热门档期屡破纪录,对头部影片的票房表现有明显档期加成作用。

端午档由于时间在6月,是暑期档的一个子档期,而暑期档恰恰又是一年当中对年度总票房贡献最大的档期。在2023年,暑期档票房占

今年的清明档、“五一”档,就曾出现十部左右新片蜂拥上映的情形,但在残酷的竞争下,最终突围而出的影片仅有两部。不过,与清明档、“五一”档不同的是,端午档虽然时间

短、体量小,却是暑期档的一个子档期,这也意味着端午档上映的新片在档期结束后,可以凭借口碑继续享受暑期档带来的红利加成。

灯塔专业版分析师陈晋向《中她》与《我爱你!》也合计拿下约6.5亿元的票房。

“今年以来电影市场档期效应依然显著,春节档、清明档均创造历史最高票房成绩,“五一”档票房也进入影史前三,这其中喜剧片、动画片表现亮眼,深受家庭观众的喜爱,这说明合家欢观影的刚需在大档期更显著,今年端午档供应的家庭片、喜剧片、动画片数量充足,相信会出现票房黑马。”陈晋说,“6月7日上映的两部好莱坞电影,动作科幻片《疯狂的麦克斯:狂暴女神》和剧情战争片《美国内战》,它们的一线城市男性观众想看占比

点映。

记者了解到,在之前一轮点映结束后,两部影片片方也迅速追加了点映安排。记者通过淘票票、猫眼两大平台发现,从5月31日开始,包括《扫黑·决不放弃》《谈判专家》在北京、上海等全国多个城市都安排了点映场次,其中不乏18时至21时之间的黄金场次。

这种大规模的点映抢跑,也让更多影片加入其中。记者注意到,计划在6月7日上映的引进片《美国内战》也将在6月1日开始,在全国多个城市开始大范围的点映。

点映。

记者了解到,在之前一轮点映结束后,两部影片片方也迅速追加了点映安排。记者通过淘票票、猫眼两大平台发现,从5月31日开始,包括《扫黑·决不放弃》《谈判专家》在北京、上海等全国多个城市都安排了点映场次,其中不乏18时至21时之间的黄金场次。

这种大规模的点映抢跑,也让更多影片加入其中。记者注意到,计划在6月7日上映的引进片《美国内战》也将在6月1日开始,在全国多个城市开始大范围的点映。

国经营报》记者表示:“整体上半年端午档延续了今年‘五一’档市场影片供给充足、丰富的特点,较大程度上满足不同观众的观影需求,有望给今年暑期档打下良好开端。”

突出,6月8日上映剧情家庭片《走走停停》《来福大酒店》,是往年端午档热门类型,同一天上映的犯罪喜剧片《扫黑·决不放弃》和青春喜剧片《我才不要和你做朋友呢》深受24岁以下年轻女性观众喜爱,另外当日还有港片《谈判专家》以及获得第96届奥斯卡金像奖最佳动画长片提名的《机器人之梦》上映。今年端午假期在时间上与儿童节比较接近,预计儿童节上映的动画新片《哆啦A梦:大雄的地球交响乐》《加菲猫家族》等也将在端午档有一定的市场空间,满足带小孩的家庭观众观影需求。”

时间下午,许多影院的黄金时段都已安排了上述影片的点映。

值得注意的是,在刚刚过去的“五一”档,就曾发生点映及预售成绩不佳而宣布撤档的案例。那么在如此拥挤的端午档,是否还会发生这一现象?

从灯塔专业版显示的端午档新片票房榜来看,在端午档的国产影片中,目前预售成绩排名靠后的有《来福大酒店》,截至5月30日12时,点映及预售总票房约50万元。除此之外,还有《离伤记》《藏地生死恋》《西瓜队》《梦想之地》《日夜江河》等影片未开启预售,也没有安排点映。

## 能否提前点燃暑期档

无论是与过往的端午档进行纵向对比,抑或与今年的春节档、清明档、“五一”档的横向比较,端午档影片的数量都是最高的。

之所以会如此拥挤,从事电影制片的金妍给出了自己的理解:“端午档的档期容量其实不大,但他与清明、中秋等其他同样时间为3天的小档期不同的是,端午档被包含在暑期档内,即使影片在端午档没能‘杀出重围’,只要片方对自己的内容有信心,那么就有可能通过口碑发酵,获得不错的长尾效应。而且从时间上看,端午档的位置处在暑期档的前期,长尾效应还能获得暑期档的加成。”

灯塔专业版显示,自2016年至今,除2020—2022年外,有4部在端午档上映的影片票房超过10亿元,分别是2016年的《魔兽》(14.68亿元)、2017年的《加勒比海盜5:死无对证》(11.8亿元)、2018年的《侏罗纪世界2》(16.95亿元)、2023年的《消失的她》(35.23亿元)。其中,除了《加勒比海盜5:死无对证》之外,其余3部影片在端午档所获得票房占影片总票房的比例均在四成以下。

“暑期档的特点就是6月启动、7月爆发、8月平稳,在个别年份,余热甚至会持续到9月。”金妍说。

记者根据灯塔专业版提供的数据整理发现,2018年6月—8月,月度票房分别为

35.77亿元、69.64亿元、68.36亿元;2019年6月—8月分别为,41.86亿元、57.56亿元、78.35亿元;2021年6月—8月则是21.03亿元、32.28亿元、20.51亿元;2022年6月—8月为19.2亿元、35.08亿元、37.08亿元;2023年同期为41.34亿元、87.25亿元、77.71亿元。

从中可明显看出,端午档所在的6月为暑期档横跨的3个月中的最低值,通常7月份为最高。即使是在受疫情影响较大的2021年和2022年,7月票房也高于6月。

此外,记者还发现,自2018年至今,除2019年与2022年外,端午假期均处于6月中旬或下旬。

“一直以来,端午档都被视为暑期档的晴雨表。而且距离暑期档最核心的时间段7月份越近,越能反映出暑期档的观众热情。比如去年的端午档,票房、观影人次、上座率等数据都呈现出非常好的恢复态势,这也进一步增强了其他正持观望态度的电影公司的信心,许多大片纷纷趁热宣布定档暑期档,最终创造了有史以来最高的暑期档票房纪录。”金妍说。

公开资料显示,在端午档结束后,7月份之前,目前已官宣定档的影片包括引进片《排球少年!垃圾场决战》、彭于晏与佟丽娅主演的《狗阵》、迪士尼的动画电影《头脑特工队2》等。