

# DeepSeek举起涨价牌



“计划近期整体上调DeepSeek API(应用程序编程接口)服务的定价,预计涨幅较大。”8月6日,DeepSeek开放平台弹出的一行提示,在AI从业者群体中引发热议。如同科技股的行情波动,大模型API价格的起落,比行业预想来得更快。

回看DeepSeek,这家曾以“价格屠夫”姿态搅动市场的企业,近三个月内多次调整定价机制。5月,DeepSeek保留DeepSeek V4-Pro的2.5折优惠;6月预告DeepSeek V4正式版上线时间,并公布API峰谷定价方案。放眼整个国产大模型赛道,涨价已不是个例,以智谱为例,2026年一季度其API价格上涨83%。相较其他国产大模型,DeepSeek的API长期处于低位,如今抛出涨价预告背后有多重行业猜测。可以确定的是,尽管长周期来看Token成本趋势下行,但当下大模型厂商与客户,仍要承受短期价格波动带来的影响。

## 一则调价通知

8月6日,北京商报记者发现,DeepSeek官网API开放平台页面上方出现一行蓝色小字,通告了DeepSeek API近期的价格上调计划。不过该通知仅表示,预计涨幅较大,请合理安排使用,具体方案以正式通知为准。

按照DeepSeek规定,公司根据模型输入和输出的总Token数进行计量计费。扣减费用=Token消耗量×模型单价,对应的费用将直接从充值余额或赠送余额中进行扣减。

目前DeepSeek API仅支持 DeepSeek V4-Flash模型,暂不支持DeepSeek V4-Pro模型。这两个模型是DeepSeek V4的预览版本,于4月正式上线并同步开源,DeepSeek V4-Pro的参数为1.6T,DeepSeek V4-Flash的参数为284B。

DeepSeek V4拥有百万字超长上下文,在Agent(智能体)能力、世界知识和推理性能上均实现国内与开源领域的领先。

根据 Agentic Coding 当时的评测,

DeepSeek V4-Pro已达到当时开源模型最佳水平,并在其他Agent相关评测中同样表现优异。4月时,DeepSeek V4已成为DeepSeek内部员工使用的Agentic Coding模型,根据评测反馈使用体验优于Sonnet 4.5,交付质量接近Opus 4.6 非思考模式,但仍与Opus 4.6思考模式存在一定差距。

目前,DeepSeek V4-Flash模型的百万Tokens输入(缓存命中)的价格是0.02元,百万Tokens输入(缓存未命中)的价格为1元,百万Tokens输出价格为2元。DeepSeek V4-Pro模型的百万Tokens输入(缓存命中)价格是0.025元,百万Tokens输入(缓存未命中)价格为3元,百万Tokens输出价格是6元。

也就是说,参数更小的DeepSeek V4-Flash百万Tokens输入(缓存未命中)价格是大参数DeepSeek V4-Pro的80%,百万Tokens输入(缓存未命中)和百万Tokens输出价格均是DeepSeek V4-Pro的三分之一。

根据通知,DeepSeek V4-Flash正式版API已上线公测,Agent能力大幅增强。DeepSeek API将在8月初支持DeepSeek V4-Pro模型。

## 涨多少

如果将这则涨价通知放在DeepSeek的定价史上则显得格外刺眼。DeepSeek出圈后,API价格降时多、涨时少。4月,DeepSeek V4预览版上线后几天,公司就对V4-Pro模型开启2.5折限时特惠,并将原本至5月5日的优惠期延至5月31日。还未到优惠截止日,DeepSeek又宣布V4-Pro模型2.5折价格永久化。

这还没有结束,6月底,DeepSeek预告V4正式版并提出峰谷定价机制,高峰时段为每日9—12时和14—18时,在这一时段调用价格翻番。

如今看来,这正是8月7日涨价通知的苗头。“这是一个标志,DeepSeek之前的价格太低了”,文渊智库创始人王超向北京商报记者分享了自己的猜测,“这次价格应该会成倍以上地上涨,或者起码跟目前国产大模型价格持平。”

按照全球AI模型聚合平台OpenRouter数据,本周调用量前十的国产大模型包括

DeepSeek V4-Flash、Hy3、GLM-5.2、DeepSeek V4-Pro、MiniMax M3。

以同为“大模型六小虎”的模型GLM-5.2、MiniMax M3为例,智谱官网显示,上下文1M Tokens的GLM-5.2百万Tokens输入单价为8元,百万Tokens输出单价为28元,缓存存储限时免费,缓存命中百万Tokens价格为2元。MiniMax M3 API的价格分多个档次,以上下文512K—1M的版本为例,标准版定价更便宜,这款模型的百万Tokens输入价格为4.2元,输出价格是16.8元,缓存读取价是0.84元。

OpenRouter统计的本周调用量前十的模型里,还有7月底刚刚宣布降价的GPT-5.6 Luna。按照OpenAI的通知,在这次GPT-5.6系列模型降价中,GPT-5.6 Luna价格降幅最大,下调80%。调整后,GPT-5.6 Luna百万Tokens输入价格为0.2美元,输出价格是1.2美元。

在王超看来,DeepSeek这次涨价的起因是之前的“价格黑洞”：“它过于特立独行,用户被高性价比所吸引到DeepSeek,运行那些最耗资源的应用,导致用户过多,以至于DeepSeek的推理能力跟不上了。”

## 设备降价再上拍 哪吒汽车遇资产处置难题

昔日造车新势力哪吒汽车南宁工厂设备,因无人接盘,如今已降价进入二拍流程。8月6日,北京商报记者从阿里司法拍卖平台获悉,属于破产资产的“存放于南宁市青秀区伶俐镇昌隆路88号机械设备一批”,起拍价6030.79万元,但因无人竞买,一拍以流拍收场。

目前,上述设备已进入二拍程序,开拍时间为8月12日,竞价周期为1天,起拍价则下调至约4824.63万元,降幅达20%。

相关资产评估报告显示,该批设备包括冲压车间11项11台套;总装车间177项177台套;模具22项51件;工具228项228件,合计438项467件,评估价为6030.79万元。辅拍机构相关工作人员向北京商报记者表示,拍卖的是流水线的部分设备,半成品、成品还有原料不在拍卖范围内。

值得注意的是,竞买公告显示的拍卖标的地址,正是哪吒汽车南宁工厂所在地。本次资产拍卖对应的执行案件信息也显示,被执行人为南宁宁达新能源汽车有限公司、广西宁达汽车科技有限公司、合众新能源汽车股份有限公司(以下简称“合众新能源”)、南宁雍合投资有限公司四家企业,执行标的金额为4568.66万元,案由为买卖合同欠款纠纷。其中,合众新能源为哪吒汽车母公司,其余三家公司也均为哪吒汽车关联企业。

据广西日报此前报道,2019年,南宁市政府与合众新能源签署战略合作协议,合众新能源将在南宁投资建设年产10万辆纯电动乘用车项目。根据协议,合众新能源将投资35亿元,预计实现年产10万辆整车的生产规模,年产值约120亿元,旨在带动区域汽车产业链升级。

然而,在上述合作落地后,企业经营状况却持续恶化。公开信息显示,2021—2023年,合众新能源累计亏损高达183亿元,卖一辆车平均亏损超8万元,持续大额亏损为南宁工

## 多面Token

其实,业界对于DeepSeek的低价早有体会,但这个低调的公司从未公开自己的定价逻辑。王超站在算力和生态的角度思考这次涨价:“如果DeepSeek算力不够用,用户体验变差,会影响接入DeepSeek API的云厂商的收入,这样不利于生态扩散。要让部署该模型的云厂商获益,模型的使用体验是前提。更何况DeepSeek本身定价不高,为大模型付费的人都是拿它当作生产力工具,对价格并不敏感。”

仅这半年,因调用量影响模型体验的案例不止一起。7月上旬,Kimi就深夜致歉,称自Kimi K3发布以来,Kimi收获了远超预期的支持,但也面临始料未及的算力挑战。为了保障已有订阅用户的体验,Kimi决定即日起,暂停C端新用户订阅,将已有算力全部投入服务已订阅用户。

长期来看,业内人士从不同角度分享了自己对Token价格走向的判断。

与北京商报记者交流GPT-5.6系列模型降价时,智源研究院院长王仲远认为,“王坚院士希望Token变得像一张纸、一支笔一样廉价,这是有可能的。Token降价基于算力提升、工程和算法优化,以及任务不断分级,简单的任务不再需要复杂的模型。再加上算力芯片的能力提升,价格自然会变得更便宜”。

涂鸦智能联合创始人杨懿从另一个维度给出了判断,他告诉北京商报记者,“Token贵不贵,关键看用得对不对”。

更宏观点看,不论涨跌,围绕Token的讨论不止一面。在Kimi创始人兼CEO杨植麟看来,“生产力会变成Agent,Agent会产生Token,Token在一定程度上等价于GDP”。站在评价体系的角度,百度创始人李彦宏则提出以“日活智能体数”(DAA)作为AI时代的核心度量衡。他认为,Token不一定代表结局,只代表成本,并不代表收益,衡量的是投入而非产出,DAA则大致对应移动互联网时代通用的度量衡日活用户数(DAU)。北京商报记者 魏蔚 图片来源:视觉中国

## 小瓶白酒的渠道之争

周六傍晚,北京王府井喜悦KKV门店的酒水区货架前站着几位年轻顾客。他们在一排排小瓶酒前驻足,拿起一瓶100毫升迷你五粮液端详片刻,又转头看向旁边的梅见青梅酒。

小瓶酒并非新鲜事物。过去很长一段时间,小酒大多作为大瓶产品的“迷你版”或是试饮装,销售渠道也以传统烟酒店为主。如今,这一局面正在被改写。

### 货架角力: 从便利店到商超

KKV酒水专区内,50毫升、100毫升规格的国产小瓶白酒已成为货架主流,头部酒企的核心产品也均推出迷你版本,单瓶售价49.9—199元,受到年轻消费者关注。

便利店货架同样发生明显变化。在便利店、潮流零售门店中,50—250毫升规格的小瓶白酒悄然走红,与小瓶装洋酒争抢货架黄金位置。这股趋势正从便利店、潮流零售店,逐步渗透至社区商超。据中国酒业协会调研,京客隆、物美等北京主流社区商超,已将50毫升小酒纳入常态化采购体系,北京商报记者走访的11家商超、便利店终端均印证了这一变化。

酒类营销专家肖竹青对北京商报记者表示,过去小酒更多承担促销品功能,未来有望成为酒企重要的文创符号载体。以往酒企推出100毫升小酒,主要用作品鉴赠饮,维护与烟酒店、小超市、餐饮终端的客情;在行业“围绕终端做C端”的大趋势下,小酒的价值将持续释放。

各大酒企密集布局小酒赛道,行业协会也牵头推动名优白酒小瓶化。今年5月,中国酒业协会联合美团闪购,携手茅台、五

粮液、汾酒、泸州老窖、洋河、古井贡酒、郎酒、贵州习酒、剑南春九家头部酒企,共同推出“T9 50毫升小酌瓶”。酒类资深营销专家蔡学飞认为,T9小酌瓶瞄准的不是传统白酒消费、宴请礼赠场景,而是年轻人居家小酌、好友小聚、露营野餐、一人微醺等私人化、悦己型消费场景。

### 消费转向: 年轻人捧热迷你酒

从KKV、嗨特购、连锁便利店再到社区超市,小酒的市场版图持续扩张,背后是消费需求迭代、渠道效率升级两大驱动因素。

中国酒业协会《2025中国白酒中期研究报告》显示,随着年轻消费群体崛起,白酒消费结构正在转变,年轻化成为行业确定趋势,1985—1994年出生的白领、职场新人,成为酒企挖掘细分市场的重要突破口。另据《年轻人饮酒习惯调查问卷》统计,33.9%的消费者会选择迷你小瓶装尝试白酒,女性消费者中这一比例高达47.2%。

香颂资本董事沈萌分析称,当前消费市场正在发生结构性调整,消费者价格敏感度提升,小规格酒水对年轻群体的吸引力更强。

消费习惯变化,直接体现在年轻人的酒水消费预算上,价格优势也成为小瓶白酒快速铺市的重要原因。《年轻人饮酒习惯调查问卷》数据显示,73.9%的年轻消费者的酒类单次购买预算不超过300元,其中46.8%的女性消费者、26.4%的男性消费者倾向选择百元以下产品。

白酒行业从业者表示,千元级名优白酒,和年轻人日常饮用场景、消费能力的匹配度有限。小酒降低了年轻人体验高端白酒的门槛,更容易被接受;对酒企而言,

依靠小酒完成消费者教育、建立品牌认知,推广成本也相对更低。

### 红利消退: 从铺货走向深度运营

货架之上的小酒竞争,已经告别网红单品短期混战阶段,形成国产名酒小酒与进口烈酒小酒长期并存的格局。头部名优白酒与进口酒水争夺有限的货架空间,依托各自客群、消费逻辑分割市场。但抢占货架只是起步,渠道铺货完成之后,小酒赛道亟待破解的核心命题,是深度融入年轻人日常生活,把一次性铺货转化为稳定复购。

肖竹青判断,大量品牌涌入之后,小酒赛道即将迎来结构性洗牌,未来呈现五大趋势:产品低度化;消费场景碎片化,渠道进一步向便利店、即时零售倾斜;市场格局两极分化;赛道分化为大众平价自饮、高端名酒品鉴收藏两大板块;行业竞争由外包装营销,转向场景深耕、渠道精细化运营。

即时零售卖出的小酒,并非全部用于即饮。以整套T9小酌瓶为例,不少消费者将其用于收藏、小型礼赠,潮玩化为小酒打开全新增长空间。业内人士认为,如果酒企持续打磨外观设计、包装美学,推进IP跨界联名,小酒可以成为年轻人社交载体与潮玩品类。借鉴盲盒、手办的收藏展示属性,带动复购增长。

当涌入赛道的品牌越来越多,行业竞争已经走出粗放铺货阶段,进入比拼产品力、场景洞察的精细化运营时代。肖竹青坦言,小酒是白酒培育新生代消费者的重要载体,但赛道红利不再普惠所有人局者;只有精准匹配消费场景、贴合新生代口感、打通新型零售渠道的产品,才能在竞争中突围。北京商报记者 刘一博 冯若男