从以分计价到以厘计价

大模型为何纷纷降价

AI世界

◎本报记者 吴叶凡 崔 爽

近日，各大电商平台开启了年中购物促销活动。但令人们始料未及的是，许多国内的人工智能大模型厂商也掀起了一波“降价潮”。记者梳理发现，仅上个月，就有字节跳动、阿里、百度、腾讯等多家知名企业旗下的大模型降价。

国家数据局局长刘烈宏在今年3月透露，中国10亿参数规模以上的大模型数量已超过100个。在业内人士看来，“百模大战”已然打响，国产大模型的降价，似乎是这场“战事”愈演愈烈的佐证。

国产大模型为何纷纷下调价格？降价会对行业产生哪些影响？大模型厂商应如何应对？带着这些问题，记者采访了有关专家。

观众正在了解大模型相关产品。 郭海鹏/视觉中国

技术优化让成本降低

记者梳理发现，较早发布降价信息大模型厂商的是深度求索（DeepSeek）。5月6日，其旗下的DeepSeek-V2 API（应用程序编程接口）定价下调至每百万tokens输入1元、输出2元。

厂商所说的“tokens”，是大模型用来表示自然语言文本的单位，可以直观地理解为“字”或“词”。通义千问官网相关说明显示，对于中文文本，1token通常对应一个汉字或词语；对于英文文本，1token通常对应3至4个字母或1个单词。阿里云智能集团资深副总裁刘伟光形象地将通义千问旗下一款模型的降价幅度描述为：1块钱可以买200万tokens，相当于5本《新华字典》的文字量。

目前，国内大模型厂商普遍根据输入和输出的token数量，分别进行计费。根据通义千问官网发布的解释，这是因为模型在推理过程中，输入和输出的资源消耗不同。输入计费是针对用户向模型提交的请求数据进行计费，包括用户提交给模型的文本、图像、音频等原始数据。输出计费则是针对模型返回给用户的输出结果进行计费，包括模型生成的文本、图像、音频等处理结果。

从降价幅度看，此次大模型总体降价幅度较大，其中输入价格较输出价格降幅更大。例如，通义千问的主力模型Qwen-Long的API输入价格从0.02元/千tokens降至0.0005元/千tokens，降幅达97%；API输出价格从0.02元/千tokens降至0.002元/千tokens，降幅达90%。腾讯旗下的混元-standard模型的API输入价格从0.01元/千tokens降至0.0045元/千tokens，下降55%；API输出价格从0.01元/千tokens降至0.005元/千tokens，下降50%。

大模型价格是如何被“打下来”的？火山引擎总裁谭待曾公开表示，降价是通过技术优化实现的，并非只是补贴、用亏损换收入。

从技术上看，通过优化算法和模型训练过程来提升算力的使用效率，是大多数大模型厂商得以降低成本的原因之一。一些厂商也使用了分布式推理和混合调度等手段来提升整体算力资源利用效率。此外，在AI基础设施层面，一些大模型厂商同时也是大型云服务商，其在公共云方面的技术红利和规模效应，带来了成本优势。

拓展市场成降价诱因

在艾媒咨询CEO兼首席分析师张毅看来，大模型降价除了技术提升带来的红利，更多是基于商业逻辑、市场竞争等方面的考量。他坦言，虽然大模型未来发展依旧面临较多挑战，但对于大模型厂商而言，技术并不能构建起绝对的“门槛”。大模型“大厂”间的技术竞争会越来越白热化，差异也会越来越小。

“因此厂商们选择在商业逻辑上展开竞争。”张毅说，大模型用户可分为两类：一类是普通消费者，即个人用户端，也就是所谓的C端；一类是企业用户，即以商家为代表的B端。

目前，对于大多数普通消费者来说，通过官网或手机应用来体验、使用大模型产品大多无需付费。张毅认为对C端的免费将是必然趋势。对于B端用户，厂商则为他们提供封闭式大模型解决方案，供其开发垂直或行业型应用。此次大模型降价，正是为拓展潜在的B端用户，这也是大模型主流的商业路径。

“无论是面向C端还是B端，因为无法在技术上具有绝对的领先优势，大家拼的就是时间。这意味着厂商的融资能力会决定一切，而融资能力又是由用户规模决定的。”在张毅眼中，这类似20年前搜索引擎的商业模式。大模型厂商需要依靠免费、低价来吸引用户，扩大用户规模，进而获得融资，在大浪淘沙后生存下去。

追求价廉还需兼顾物美

人工智能作为一项战略性新兴技术，日益成为科技创新、产业升级和生产力提升的重要驱动力量。目前，大部分企业对于应用AI来提升生产效率都有需求。但想要在应用场景中真正使大模型落地，降低落地应用成本和试错成本是关键。

此次降价，对于各个领域的企业来说，都是推动人工智能在业务场景中应用的机遇。正如谭待所说：“大模型从以分计价到以厘计价，将助力企业以更低成本加速业务创新。”创新工场董事长兼零一万物CEO李开复也预测，在明年下半年，人们将迎来大模型普惠应用的井喷期。

乐观声音出现的同时，也有一些人认为，降价可能带来一些潜在的问题，比如服务质量下降、技术研发投入减少、市场恶性竞争等。事实上，此次降价的大模型企业，大多为实力雄厚的“大厂”。大部分创业公司都没有“跟风”降价，比如李开复就曾公开表示：“我们绝对不会跟这样的一个定价来做对标。”对于大模型创业公司，如果以往的价格优势不复存在，就意味着其B端业务可能受到影响。“这次大降价基本宣告了大模型创业公司必须寻找新的商业模式。”猎豹移动董事长兼CEO傅盛在个人社交媒体上表示。

落地后是否能给企业带来实惠与便利，是大模型能否成功获客的关键。因此，大模型厂商更应不断“修炼内功”，提升服务质量和水平，在追求“价廉”的同时做到兼顾“物美”。

首个支持30种方言混说 语音大模型亮相

科技日报讯（记者**崔爽**）记者6月16日获悉，中国电信人工智能研究院发布业内首个支持30种方言自由混说的语音识别大模型——星辰超多方言语音识别大模型。它打破了单一模型只能识别特定单一方言的困境，可同时识别理解粤语、上海话、四川话、温州话等30多种方言，是国内支持方言种类最多的语音识别大模型。

基于几亿用户和丰富应用场景优势，中国电信人工智能研究院构建了超30种、超30万小时的高质量方言数据库，推出星辰超多方言语音识别大模型。研发团队通过超大规模语音预训练和多方言联合建模，率先实现单一模型支持30种方言自由混说语音识别，是国内支持方言种类最多、覆盖人口最多的语音大模型。

团队首创“蒸馏+膨胀”联合训练算法，解决超大规模多场景数据集和大规模参数条件下，预训练坍缩的问题，实现1B参数80层模型稳定训练。星辰超多方言语音识别大模型也是业内首个开源的基于离散语音表征的语音识别大模型，将推理时语音传输比特率降低数十倍。

据悉，星辰超多方言语音识别大模型已在福建、江西、广西等地的智能客服试点应用。接入大模型后，智能客服能秒懂30种方言，日均处理约200万通电话。星辰超多方言语音识别大模型还落地多地市12345平台，为客服人员赋能，提升沟通效率，助力政务工作智能化升级。

愚公YUKON矿山大模型 正式发布

科技日报讯（记者**崔爽**）记者6月17日获悉，在近日召开的新质生产力和智能产业发展平行会议上，中国科学院自动化研究所、中国矿业大学（北京）、中科慧拓（北京）科技有限公司（以下简称中科慧拓）联合发布愚公YUKON矿山大模型。

中科慧拓CEO陈龙介绍，愚公YUKON矿山大模型依托丰富的行业落地经验和高质量矿山场景数据形成。以该模型为底座，中科慧拓推出首批产品，包括矿山AI助手“矿宝”、生成式平行智能数据体系GenDS、矿山端到端自动驾驶大模型GenAD等。它们将成为矿山行业工作者的得力助手。

据悉，“矿宝”是首个面向矿山从业人员的AI助手。它可以轻松理解自然语言指令，完成工作人员下达的任务。

生成式平行智能数据体系GenDS是首个针对矿区的数据体系，能够指数级生成高质量数据。

端到端自动驾驶大模型GenAD可赋予自动驾驶矿车人类驾驶员的深度理解和决策能力，可提升在矿山自动驾驶的安全性和可靠性。通过深度学习和强化学习等技术手段，自动驾驶大模型能不断进行学习和优化，提高自身的泛化能力和鲁棒性，更好地应对未知和复杂的矿山环境。

陈龙透露，愚公YUKON矿山大模型及其基础设施智能体系2025年将全面落地，赋能全国超过40座智慧矿山的建设；2030年，将全面保障全球超过500座矿山的绿色安全高效运营。

自动驾驶汽车驶上广州高快速路

◎本报记者 叶 青

近日，自动驾驶科技公司文远知行WeRide获得广州市颁布的远程测试（无人）牌照和载货测试牌照，旗下自动驾驶货运车Robovan获准在广州市开展自动驾驶城市货运车“纯无人测试”及“载货测试”，测试范围覆盖白云、花都、番禺、黄埔、南沙、海珠6个行政区共797条测试道路，双向里程3247公里，其中包括南沙区全域。

这是中国首个城市开放道路场景下L4级自动驾驶货运车纯无人远程测试许可，也是中国首个支持7×24全天时自动驾驶货运车载货测试活动。

与此同时，小马智行旗下自动驾驶出租车和自动驾驶卡车各有一款车型获得广州首家智能网联汽车高快速路测试许可。这意味着，广州的自动驾驶车辆终于能驶上高快速路。

“高快速路测试道路的开放，丰富了广州智能网联汽车路测的道路类型。”小马智行副总裁、广深研发中心负责人莫璐怡表示，随着测试经验的不断累积，全域开放的时代势必加速到来，届时用户将享受到更全面的自动驾驶出行服务。

高快速路对于智能网联汽车有着更高的技术要求。以小马智行本次获得许可的其中一款丰田赛那车型为例，要取得广州高快速路测试许可，需获得机动车安全技术检验报告、封闭场地测试报告，以及此前获取的其他道路类型测试许可、以往累积的道路测试里程和平均脱离间隔里程等资料，满足相关条件方可申请。

记者了解到，小马智行入选的车型属于广州L4级别自动驾驶高快速路测试许可的车型。按我国实施的《汽车驾驶自动化分级》，驾驶自动化分6级：L0至L2为驾驶辅助，驾驶员需全程监控驾驶；L3是有条件自动驾驶，驾驶员在紧急情况执行接管；L4为高度自动驾驶，自动驾驶系统在特定条件下能够完全控制车辆的驾驶任务，并且能够实现与人类驾驶相似的安全、舒适、高效的驾驶体验；L5为完全自动驾驶。

随着技术的快速发展，自动驾驶汽车大规模上路是否指日可待？中国电动汽车百人会副理事长兼秘书长张永伟认为，尽管自动驾驶的技术水平在一些场景接近甚至超越人类驾驶水平，但其大规模社会化、商业化应用之路还很长。

高新兴科技集团股份有限公司高级副总裁吴冬升表示，自动驾驶技术的广泛应用需要不断攻克技术难题，如提升传感器精度、优化自动驾驶算法以及确保网络通信稳定等；同时，也需要进一步完善相关法律法规，确保该技术的安全性和可靠性。

中国算力网粤港澳大湾区 算力服务平台上线

科技日报讯（记者**龙跃梅**）记者6月15日获悉，在近日召开的第三届粤港澳大湾区（广东）算力产业大会暨第二届中国算力网大会上，中国算力网粤港澳大湾区算力服务平台在广东韶关上线。

平台将由韶关市数据产业研究院、中国联通、韶关数投公司联合运营。目前，平台已成功接入包括鹏城云脑、广州超算等在内的多家算力资源节点。

中国工程院院士、鹏城实验室主任高文表示，鹏城实验室已携手韶关打造中国算力网“高速公路”节点，希望把东西部资源优势充分、合理地结合起来，以科技创新为中国数字经济发展注入“算力动能”。

据了解，2023年10月20日，中国算力网粤港澳大湾区调度中心正式上线。这是中国算力网项目全国首个区域级资源调度中心。上线服务半年来，中心已初步实现粤港澳大湾区算力、数据、网络等资源云际互联、融会贯通，有力地促进了战略性产业创新发展。为进一步强化中心的资源调度、交易等服务能力，韶关市政府与鹏城实验室共建了中国算力网粤港澳大湾区算力服务平台。

韶关市委书记陈少荣表示，韶关将举全市之力大力发展大数据全产业链，以数据和算力赋能新质生产力，引领高质量发展。

大模型发展仍处初级阶段

规模化落地还需打造“超级应用”

◎本报记者 操秀英

“一年前我跟ChatGPT对话可能还有一点郑重其事的仪式感，现在这种对话已经无缝嵌入到了思考之中。”近日，物理学家、科普作家万维钢在一篇文章中这样写道。

确如他所说，第一批擅长使用AI工具的人已经离不开它们了。也正如业界此前判断的，2023年是大模型元年，2024年是AI超级应用的爆发之年。即便如此，业界普遍认为，AI大模型行业仍处于初级阶段。

大模型产品渗透率仍偏低

截至今年4月，中国的大模型数量已近200个，其中通用大模型数量在40个左右。同时，各大模型厂商都在努力打造大模型个人终端，许多大模型厂商已经推出了独立的C端应用。比如，谷歌发布文生图大模型Imagen 3和视频生成模型Veo；字节跳动发布“豆包大模型家族”，统一使用“豆包”品牌，“豆包”是目前字节跳动最大的C端AI应用。

数据显示，当前人们使用大模型相关产品时，超65%的需求集中在工作、学习等场景，但相关的AI产品解决方案尚不成熟。近期一项由路透社新闻研究所发布的在线调查结果显示，生成式人工智能工具的频繁使用率仍偏低。

“国内移动互联网用户有12亿多，但每家大模型产品可能只有几十万或几百万的日活用户。合在一起看，大模型日活用户总数可能就是百万量级。”腾讯云副总裁、腾讯混元大模型负责人刘煜宏分析，“但对于庞大的互联网用户数量来讲，这一比例非常低，可能还不到1%。”

看似火热的技术和产品，渗透率为何如此之低？

在刘煜宏看来，这有两个原因。首先，大模型处于发展早期，产品能力不足，而它距企业和用户需求又很远，导致其落地能力较弱。其次，公众对大模型的认知度不够，虽然很多企业、开发者和用户都知道大模型很厉害，但具体怎么用它，大部分人都不知道。

以应用驱动技术创新

即便当前大模型渗透率较低，但相关人士认为，我国大模型产业发展前景依旧广阔。

百度创始人、董事长兼首席执行官李彦宏日前在法国巴黎举办的“欧洲科技创新展览会”主论坛上表示，应用驱动了中国AI产业快速发展。现在，人们越来越多地在讨论什么是AI时代的“超级应用”。

业内专家分析，AI时代的“超级应用”要人人可用，能够帮助用户解决复杂问题，并在可实际应用中展现相应价值。“超级应用”是多方面综合作用的产物，其出现需要相关方在技术、数据、算力等多方面加强支持。

对此，腾讯迈出了探索的步伐。5月底，公司宣布，基于腾讯混元大模型的App“腾讯元宝”正式上线。它不仅面向工作提效场景提供了AI搜索、AI总结、AI写作等核心功能，还面向日常生活提供了多个特色AI应用，以及创建个人智能体等新功能。

“从去年到现在，腾讯混元大模型整体技术架构有了很大升级，采用了比较先进的MoE混合专家架构。腾讯混元大模型参数量从千亿级上升到万亿级，‘喂’给它的知识含量也提高了很多。”刘煜宏介绍。

记者体验发现，腾讯混元大模型的升级体现在多个方面。例如，为解决过去大模型对实时信息的掌握能力较差、可用性不高的问题，腾讯对该大模型的搜索功能进行增强，让它更了解时事。在长文本能力方面，该大模型可以理解数十万字的小说。此外，该大模型的文生图、文生视频能力也取得了进步。上述技术突破让“腾讯元宝”的“可玩性”大大增加。

在刘煜宏看来，大模型要成为用户生活中的好伙伴、好帮手，最终服务于每个人的生活，才是其价值所在。“希望应用能反过来驱动底层技术的创新和演进。”他说。

图说智能

机器人“挥毫泼墨”

近日，人工智能造福人类全球峰会开幕。会议聚焦人工智能全球合作、监管与治理、标准建设，以及人工智能在教育、健康、通信、金融、气候变化方面的应用等议题。图为峰会上，绘画艺术家机器人Ai-Da在画画。 新华社记者 连漪摄