

人工智能大模型“价格战”的原因及影响分析

近期，以互联网“老兵”为主的人工智能大模型企业连番掀起“价格战”，希望复刻互联网时代的成功商业路径，引发产业界普遍关注。本文将对人工智能大模型“价格战”的原因及影响进行分析。

文 | **王菲** 赛迪智库信息化与软件产业研究所软件产业研究室副主任
王宇霞 赛迪智库信息化与软件产业研究所人工智能产业研究室主任

2024 年，寻找商业化落地出口成为国内外人工智能大模型企业的共同目标。近期，以互联网“老兵”为主的人工智能大模型企业连番掀起“价格战”，希望复刻互联网时代的成功商业路径，引发产业界普遍关注。

赛迪智库认为，在降价背后，成本回落、获取数据、争夺用户是主要原因，同时也将带来创业企业生存空间受挤压、应用生态加快繁荣、智算云业务快速增长等影响，建议加快提升大模型核心能

力，持续丰富大模型应用生态，推动市场健康有序地发展。

人工智能大模型“价格战”正酣

当前，大模型商业化定价主要通过文本最小语义单元（tokens）使用量计费、订阅制会员费、模型微调服务和开发者应用程序接口（API）调用服务、模型定制化服务和赋能用户业务等多种方式。为加快商业化落地，国内外企业纷纷加快降价步伐，以期通过“价格战”换取



赛迪网官方微信



数字经济官方微信

大模型应用市场。

从降价力度看，从轻量模型到主力模型均有不同程度降价。当前免费模型主要为轻量级模型。如，百度文心大模型系列中轻量化 ERNIE-Speed、ERNIE-Lite、ERNIE-Tiny 系列模型可直接免费、长期开放使用，无须训练与维护。科大讯飞宣布讯飞星火 Lite 版轻量化大模型 API 永久免费开放。主力模型中，通用类大模型价格显著下降至近于免费，高性能、垂直类多模态大模型价格并未调整。如，字节跳动豆包通用模型 pro-32K 版较行业价格低 99.3%；阿里云通义千问 GPT4 级通用模型 Qwen-Long 的 API 输入价格直降 97%，约为 GPT-4 价格的 1/400。而智谱 AI 的入门级产品 GLM-3 Turbo 模型降价，但企业级产品 GLM-4/GLM-4V 价格保持 0.1 元 / 千 tokens 未变。此轮降价仅涉及 tokens 使用量计费、订阅制会员费及 API 调用，而模型的精调、训练、部署价格并未调整。

从降价主体看，本轮降价主要集中在国外顶尖 AI 企业及国内互联网大厂和初创公司。在国外，人工智能巨头 OpenAI 基于庞大的 ChatGPT 用户数量及技术成本优势已进行 4 次降价，并于 5 月 30 日宣布所有用户全面免费、有限次开放 GPT-4o 模型。在国内，阿里、百度、腾讯、字节跳动等互联网大厂基于云业务“造血”能力支撑大模型“低价探路”，降价力度更大。以百度云千帆大模型平台为例，模型部署、精调、评估、数据管理、插件调用等云服务均按量付费，但大模型可以免费调用。智谱、深度求索等人

工智能初创企业也纷纷卷入“价格战”。微软、谷歌、Meta、亚马逊等国际科技公司，以及国内华为公司暂未降价。

原因分析

成本驱动：模型和芯片成本双向降低是本轮降价的重要前提。一方面，模型优化和算法革新带动模型消耗成本降低。过去一年 GPT-4 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。百度文心大模型的推理性能提升了 105 倍，推理成本则降到了原来的 1%。火山引擎可以通过混合专家模型（MoE）、分布式推理等技术手段，优化豆包大模型的推理成本。另一方面，用于大模型训练和推理的芯片成本不断降低。如英伟达开发的基于 Blackwell 的 GPU 架构，其成本和能耗比上一代的 Hopper GPU 架构降低了 25 倍。

性能驱动：烧钱换取高价值数据、提升大模型服务能力是本轮降价的内在驱动。一方面，通过低价换取用户更丰富、更多维数据集支持大模型进化升级。如 GPT 从 1.0 到 4.0，数据参数从亿量级猛涨至万亿量级，对数据需求量大幅提升。与此同时，通过低价吸引各行各业开发者注入垂直领域的数据集，可以加快大模型在垂直领域的商业化进程。另一方面，通过低价获取的用户动态交互数据集提高了大模型的“类人化”能力。相较于静态数据、合成数据，动态交互数据更有助于提升大模型的输出精准度和感知用户情绪的能力，尤其对具备反馈机制的大模型更甚。如，LLaMA-2

和 InstructGPT 采用反馈强化学习技术，(RLHF)将用户偏好数据转化为数字信号，进一步提升了大模型用户服务能力。

运营驱动：争夺各类用户、拓展落地场景是本轮降价的最终目的。一方面，意在复刻互联网时代的商业模式，降价是争夺用户的经验选择。互联网时代，各大领域的兴起和胜者的决出都伴随着降价、补贴等大战的身影，如滴滴、快的、Uber 的打车大战，美团、饿了么、百度、口碑的外卖大战，摩拜、OFO、青桔的单车大战，以及近三年阿里、京东、拼多多的百亿补贴大战。另一方面，各大模型进入同质化竞争阶段，降价是抢占场景的无奈打法。数据显示，GPT-4 发布以来，多个模型性能和实力相当，当前尚没有明显领先的模型。各大模型在能力趋同的情况下，难以通过技术进步带动业务正向发展。在此背景下，不得不放弃收费模式，通过降价，甚至免费，通过用户“免费试用”探索商业落地场景。据悉，几家大模型厂商都给了不同程度的注册免费送额度，如智谱 AI 宣布新注册开放平台用户赠送 2500 万 tokens。

影响分析

从短期看，对创业企业生存空间形成挤压。一方面，对于一些无法承受持续低价竞争的企业，可能会被迫退出市场。企业通过“价格战”争夺市场份额，往往会导致企业现金流承压加大、利润下降，进而影响人才招聘、基础设施升级、数据中心部署、网络升级等投资，长期将影响模型进化与迭代。从当前参与价

格战的企业来看，主要以阿里、百度、腾讯、字节等资金雄厚的互联网大厂为主。如果创新创业企业卷入，一旦资金跟进乏力，将失去持续发展的能力。另一方面，对于不参与价格战的创业企业，可能因缺少用户被排挤出市场。各大企业通过价格战可以快速抢占市场，并通过持续补贴和个性化服务形成用户黏性，打造生态壁垒。而对于不参与价格战的创业企业，如没有强大的技术优势，获客将成为其落地的一大难点。长期无法形成商业闭环将使得创业企业营收及获得进一步投资的空间变小，进而被排挤出市场。创新创业企业的被迫退出，恐降低我国大模型领域技术的创新活力。

从中期看，加速大模型应用生态繁荣。一方面，技术普惠将推动应用领域的创新成本降低。随着大模型推理成本的降低，依托大模型开发应用的创新成本将随之降低，或将迎来新应用开发和新场景试错的高潮期。数据显示，百万 tokens 大约相当于 70 万个英文单词或 50 万个汉字，按照现在大模型 API 调用价格计算，1 元可以购买超过 100 万个汉字。对于中小企业来说，部署并应用针对本企业的大模型应用成本降低，或将刺激企业开发应用的意愿。

另一方面，应用创新成本降低将进一步降低消费端使用成本，促进应用生态进一步繁荣。类比移动互联网时期，从 3G 到 4G 再到 5G 时代，套餐外流量费用从高达 10000 元/GB，降低到 100 元/GB，最后到 1 元/GB，甚至无限流量，随着流量成本降低和技术能力提升，移动互联

应用生态快速繁荣，如微信崛起、抖音诞生、直播涌现、电商爆发。可以预见，随着大模型应用开发成本不断降低，将加快涌现爆款超级应用，推动大模型应用生态加速繁荣。

从长期看，进一步催生智算云新增长点。一方面，大模型的发展将带来更高的智能计算需求。大模型训练与推理依赖大规模 GPU 集群及高效的网络存储系统支持。当前，云计算算力成本和存储成本的大幅降低带来大模型计算成本下降，越来越多的大模型应用转向云端，“价格战”带来的大模型应用潮将进一步激发智能计算需求。另一方面，“价格战”将进一步加速智算云服务变现速度。相比于通用大模型，用户开发专用大模型对大模型定制与应用开发工具链需求较高，迫切需要云服务商搭建从选模型、调模型、搭应用到对外服务的一站式服务平台，以及在多模型调用时提供更高级别数据安全保障。相比于常规的 B 端、C 端用户云服务付费，用户更愿意对可提供定制化开发、数据安全保障等服务进行付费。

对策建议

提升人工智能大模型核心能力。一是引导人工智能企业从“卷价格”向“卷技术”转变，加大生成式人工智能相关核心技术研发投入，持续提升产品性能和特色化水平，打造企业核心竞争力。二是引导人工智能企业与高校等科研机构加强协作，加大大模型架构创新等原创性研究力度，从重视模型调优转向基

础框架和底层算法创新。三是支持第三方机构加快构建通用型、专业型高质量数据集，为人工智能大模型创新提供充足“弹药”。

加强人工智能大模型落地引导。一是搭建行业企业与人工智能企业供需合作对接平台，促进供需信息共享与资源融合，降低行业企业与人工智能企业合作门槛。二是遴选一批成效显著的人工智能应用典型案例，形成可复制、可推广的商业化模式。三是鼓励国内大模型厂商与云厂商在技术创新、行业解决方案等方面展开全面合作，共同构建大模型应用软件工具生态。

引导市场良性规范发展。一是加快开展人工智能大模型性能评测标准研制，依托第三方机构开展人工智能大模型能力测评，为用户“货比三家”选择大模型提供依据。二是加大人工智能领域数据治理研究，加快制定数据权属确认、收益分配等相关指引，推动企业合理合法接入和使用用户数据。三是加大对人工智能创新创业企业支持力度，保持市场创新活力。

责任编辑：金烨 jinye@staff.ccidnet.com