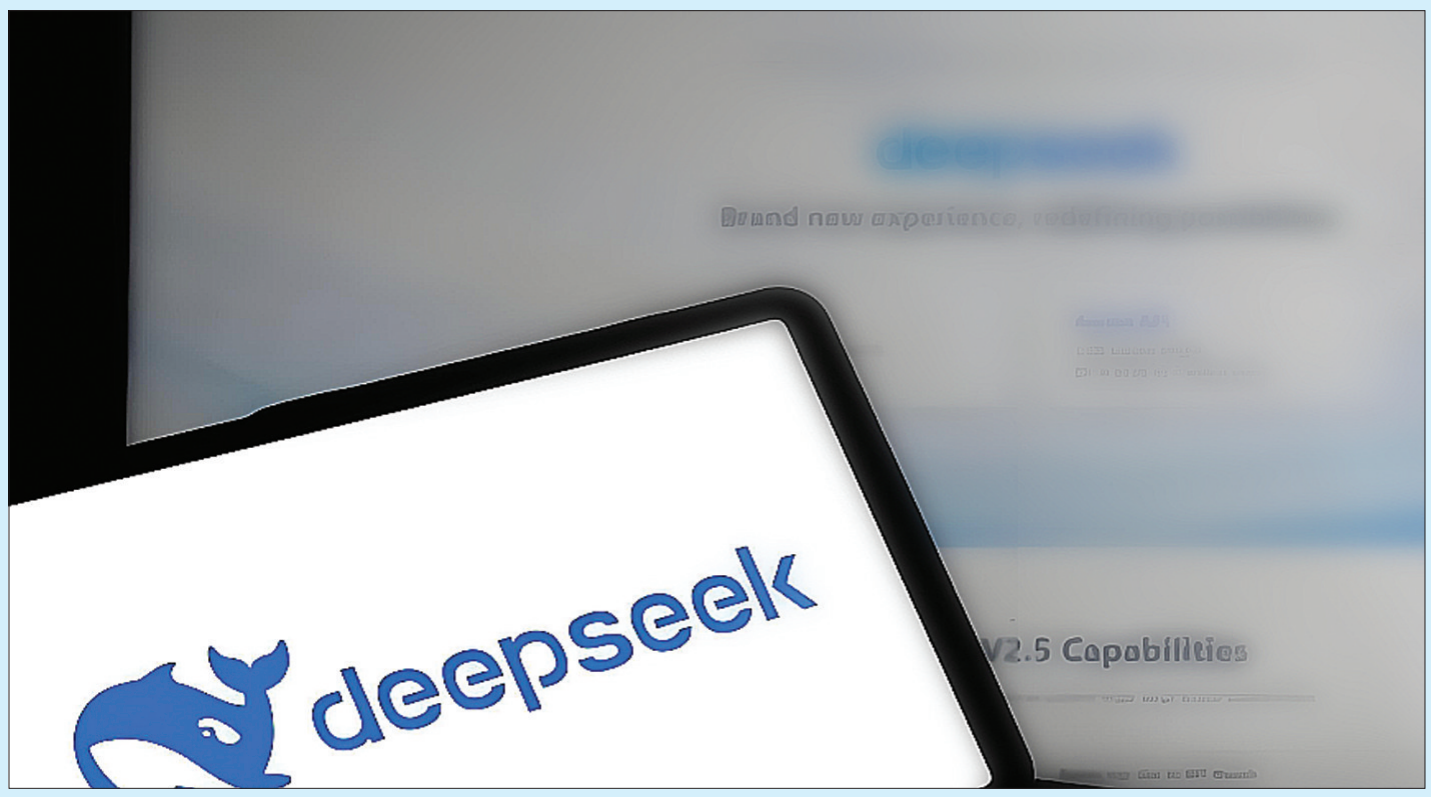
大模型再启降价潮，谁将受益？



本报记者 宋婧

继DeepSeek-R1以有限算力，凭借强大算法创新“惊艳”全球之后，大模型研发领域似乎也开始打起“价格战”。美国斯坦福大学、加利福尼亚大学伯克利分校等机构的研究团队，先后宣布仅以几十美元成本，开发出性能可媲美前沿推理模型的人工智能（AI）模型。与此同时，在DeepSeek的刺激下，阿里云、腾讯、字节跳动、智谱等厂商均宣布旗下模型API降价，多家厂商推出免费服务。AI大模型领域掀起新一轮降价潮。

AI创新成本快速下降

随着算法和硬件技术的不断进步，以及更多企业和个人参与到大模型的研发中，大模型的训练成本正逐渐降低。

在降价队伍中，尤以闭源模型领域的代表——百度和OpenAI的转变最具代表性。百度先是宣布文心一言将于4月1日0时起全面免费，仅隔一天又宣布将在未来几个月中陆续推出文心大模型4.5系列，并于6月30日起正式开源。OpenAI也在免费开放ChatGPT的搜索服务后，又宣布免费开放了基于GPT-4o mini模型的高级语音聊天模式。

这一连串“蝴蝶效应”透露出一个重要信号：AI创新成本正在快速下降。OpenAI首席执行官萨姆·奥特曼在最新的一篇博客中这样写道：“使用特定水平人工智能的成本，大约每12个月下降10倍，更低的价格还会带来更多的使用。”

“在过去，当我们谈论摩尔定律时说，每18个月性能就会翻倍、成本就会减半；但今天，当我们谈论大语言模型时，可以说每12个月，推理成本就可以降低90%以上。这比我们过去几十年经历的计算机革命要快得多。”百度CEO李彦宏坦言，“我们到处都能看到创新，必须适应这种快速变化的创新。”

从“按分计价”到“按厘计价”，再到如今“零元购”，在中国科学院科技战略咨询研究院研究员周城雄看来，上述转变的发生，与相关技术的进步和规模效应有关。他提到，随着算法和硬件技术的不断进步，以及更多企业和个人参与到大模型的研发中，大模型的训练成本正逐渐降低。DeepSeek-R1就是一个例子，它通过优化算法采用MoE（混合专家模型）架构、MLA（多头潜在注意力机制）等技术，使得GPU集群使用效率远超行业平均水平，模型训练的算力需求显著降低。

业内人士普遍认为，模型优化、算法革新以及硬件成本下降是大模型成本下降的关键因素。一方面，随着深度学习技术的不断发展，大模型的性能和效率得到了显著提升。厂商们不断通过优化模型结构、提升计算效率、降低硬件成本等手段降低大模型使用成本。另一方面，随着芯片技术的不断进步和规模化生产，用于大模型训练和推理的芯片成本不断降低，这些硬件成本的下降直接反映在了大模型的服务价格上。

厂商争夺流量入口

免费用户的行为数据可反哺模型迭代，实现数据与场景沉淀，应用场景的拓展则为未来商业化提供可能性。

实际上，能在这场风靡全球的AI竞逐赛中坚持到今天的公司，无疑都已投入了巨大的成本。自2014年起，百度在人工智能领域的投入已超过1700亿元。微软自2019年以来向OpenAI投资了超过130亿美元用于人工智能领域探索。最新消息显示，阿里未来3年将投入超过3800亿元，用于建设云和AI硬件基础设施，总额超过去10年的总和。

巨额的投入一直是引发营收焦虑的关键原因。此前，AI大模型的商业运营模式大多是围绕付费使用展开的。OpenAI最初推出付费订阅计划“ChatGPT Plus”，首次将AI大模型的C端用户使用价定在每月20美元。百度文心一言上线专业版率先在国内开启了付费模式。腾讯混元大模型、阿里通义千问等也都延续了这一模式。

DeepSeek的鲶鱼效应是“降价潮”的一个重要驱动力。DeepSeek-V2模型的横空出世，将大模型的价格降至GPT-4的1%，阿里、腾讯、字节跳动等多家企业迅速跟进，AI大模型“按厘计价”时代正式开启。DeepSeek-R1的出现更是将这一进程直接推向“零成本”的临界点。

从短期来看，大模型的降价潮无疑会给厂商带来一定的经营压力。然而从长期来看，降价和免费背后的目的还是在于争取更多的用户和流量。通过降低价格门槛，厂商有望吸引更广泛的企业用户群体，从而进一步平衡收入和成本。同时，更多C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模型服务能力。

“国内外的通用大模型在性能上趋同，比较难通过技术差异来建立产品的壁垒，那就要靠价格来竞争，免费策略能够满足大模型企业快速积累用户规模、通过流量入口抢占生态位的市场扩张需求。”周城雄分析称。同时，他指出，免费用户的行为数据可反哺模型迭代，实现数据与场景沉淀，应用场景的拓展则为未来商业化提供可能性。“这也是资本市场的常规博弈策略，头部企业可以凭借融资优势，以短期亏损换取长期市场优势地位。”周城雄表示。

AI普惠时代到来

由于DeepSeek已将其降本方法论开源，大模型的训练/推理成本有望整体降低，这将有助于AI应用迅速拓展。

近日，在AI消费硬件领域，包括DeepSeek在内的开源技术催生了一波AI眼镜、智能戒指、助听器等产品的创新浪潮。业内人士普遍认为，AI大模型开启新一轮降价甚至是免费，将进一步降低AI应用的开发门槛，使开发者可往B端（企业）和C端更多产品方向探索，加快AI应用落地和商业化进程。

光大证券研报指出，由于DeepSeek已将其降本方法论开源，大模型的训练/推理成本有望整体降低，新一轮降价潮或将来袭，这将有助于AI应用迅速拓展。长江证券研报则认为，由DeepSeek开源引发的AI平权正持续刺激xAI、OpenAI等海外大模型厂商加速模型研发，并改变其此前在开源问题上的策略，从而推动大模型产业加速迭代发展。

“受到DeepSeek的压力，全球AI行业已进入免费竞争与成本驱动的关键阶段，开源模型和性能高低决定企业生死。”快思慢想研究院院长田丰判断说道，“从目前来看，DeepSeek由于团队人数较少（150人左右）且专注科研，暂时无法提供定制化的企业级服务，短期内不会涉足ToB服务，企业级服务未来会是其他AI大模型产品的差异化落点。”

“现在对中小型AI企业来说，是一个难得的机会。”田丰表示。他认为，如今算力门槛降低，中小企业可以在国产开源大模型的基础上，低成本“蒸馏”开发适合自己硬件终端的小模型、行业应用小模型等。同时，大厂会聚焦自己核心业务的AI升级，中小创企则可以聚焦垂直领域的训练数据加工、模型调优、AI项目交付服务等，这样就可以躲开巨头的锋芒，或者成为巨头产业链、传统产业链上的一环。

中国新一代人工智能发展战略研究院首席经济学家刘刚指出，在这股大模型的免费潮流中，中国企业通过技术创新降低成本，提高效率，形成了独特的竞争优势。而国际企业在应对中国大模型崛起的过程中也展现出了多样化的策略和生态博弈能力。例如，生态绑定策略被广泛应用，Google将Gemini大模型与Workspace办公套件深度捆绑，通过提升用户迁移成本对冲价格竞争。此外，技术伦理护城河成为新的竞争维度，Anthropic等企业强化“负责任AI”认证体系，以伦理合规构建差异化壁垒。这些策略不仅反映了国际企业的应变能力，也揭示了大模型竞争的复杂性和多维度性。

“就全球模型变化而言，从研发到生产，创新活跃，大模型技术路线及能力快速迭代，开源闭源交叉领先，国产模型差距缩小，易用性大幅提升，模型平权趋势明显。”中信建投研究发展部兼国际业务部行政负责人武超则表示。她指出，之前国产模型普遍较海外晚半年，目前时间差逐渐缩小。因此类比海外，随着国内大模型能力的迭代，预计国内AI爆款应用将密集出现。

阿里巴巴发布并开源全新推理模型通义千问QwQ-32B

本报讯 记者**宋婧**报道：3月6日凌晨，阿里巴巴发布并开源全新的推理模型通义千问QwQ-32B。通过大规模强化学习，千问QwQ-32B在数学、代码及通用能力上实现质的飞跃，整体性能比肩DeepSeek-R1。在保持强劲性能的同时，千问QwQ-32B还大幅降低了部署使用成本，在消费级显卡上也能实现本地部署。目前，阿里已采用宽松的Apache2.0协议，将千问QwQ-32B模型向全球开源，所有人都可免费下载及商用。同时，用户也可通过通义APP免费体验最新的千问QwQ-32B模型。

据悉，千问QwQ-32B既能提供极强的推理能力，又能满足更低的资源消耗需求，非常适合快速响应或对数据安全要求高的应用场景，开发者和企业可以在消费级硬件上轻松将其部署到本地设备中，进一步打造高度定制化的AI解决方案。此外，千问QwQ-32B模型中还集成了与智能体Agent相关的能力，使其能够在使用工具的同时进行批判性思考，并根据环境反馈调整推理过程。

阿里通义团队表示，未来将继续探索将智能体与强化学习的集成，以实现长时推理，探索更高智能进而最终实现AGI的目标。

从性能表现来看，千问QwQ-32B模型在一系列权威基准测试中表现得异常出色，几乎完全超越了OpenAI-o1-mini，比肩最强开源推理模型DeepSeek-R1：在测试数学能力的AIME24评测集上，以及评估代码能力的LiveCodeBench中，千问QwQ-32B的表现与DeepSeek-R1相当，远胜于o1-mini及相同尺寸的R1蒸馏模型；在由Meta首席科学家杨立昆领衔的“最难LLMs评测榜”LiveBench、谷歌等提出的指令遵循能力IFEval评测集、由加州大学伯克利分校等提出的评估准确调用函数或工具方面的BFCL测试中，千问QwQ-32B的得分均超越了DeepSeek-R1。

目前，千问QwQ-32B已在魔搭社区、HuggingFace及GitHub等平台基于宽松的Apache2.0协议开源，所有人都可免费下载模型进行本地部署，或者通过阿里云百炼平台直接调用模型API服务。对于云端部署需求，用户可通过阿里云PAI平台完成快速部署，并进行模型微调、评测和应用搭建；或是选择容器服务ACK搭配阿里云GPU算力（如GPU云服务器、容器计算服务ACS等），实现模型容器化部署和高效推理。

实际上，自2023年以来，阿里通义团队已开源200多款模型，包含大语言模型千问Qwen及视觉生成模型万相Wan等两大基模系列，开源囊括文本生成模型、视觉理解/生成模型、语音理解/生成模型、文生图及视频模型等“全模态”，覆盖从0.5B到110B等参数“全尺寸”，屡次斩获Chatbot Arena、司南OpenCompass等权威榜单“全球开源冠军”“国产模型冠军”。截至目前，海内外AI开源社区中千问Qwen的衍生模型数量突破10万，超越美国Llama系列模型，成为全球最大的开源模型族群。

荣耀将投入百亿美元建设AI设备生态

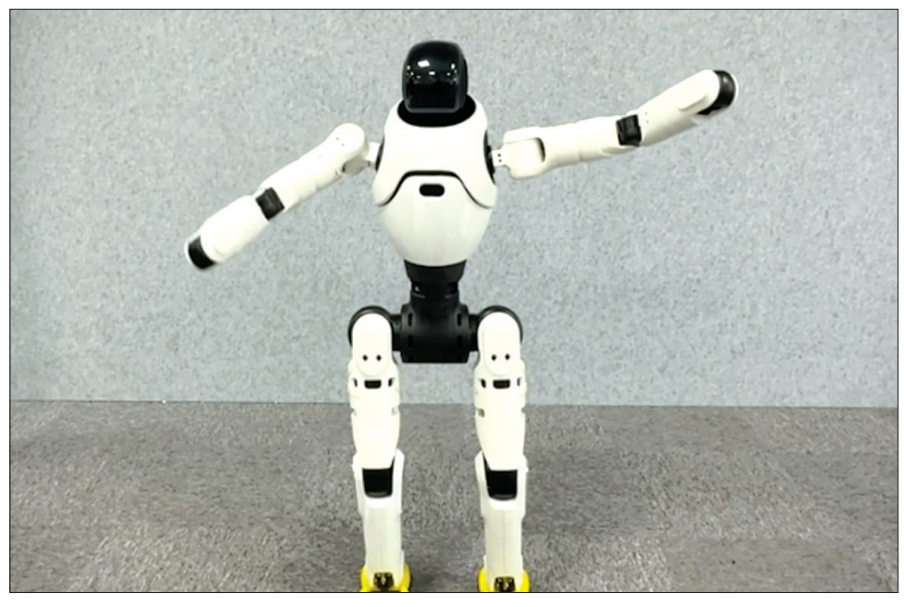
本报讯 记者**谷月**报道：近日，荣耀发布了“荣耀阿尔法战略”。公司CEO李健表示，未来，荣耀将投入百亿美元建设AI设备生态，企业将从智能手机制造商向全球领先的AI终端生态公司转型。

据荣耀方面介绍，荣耀阿尔法战略其实就是荣耀面向AI时代的战略方向。该战略落地分为三个步骤：第一步，开发AI终端；第二步，建立AI生态；第三步，共建AI世界。

李健称，这也是新CEO上任以来的首个重大决定。虽然在过去数年里荣耀并未表示自己不是智能手机制造商，但却一直都在朝着AI终端厂商的方向努力。从2016年荣耀Magic开启AI时代，到MagicOS 9.0开启手机自动驾驶时代，荣耀在AI的路上从未停止，而面向未来更为激烈的竞争，荣耀选择“all in AI”。李健呼吁，行业要开放AI能力，赋能更多设备，实现无缝协同。

目前，荣耀正联合全球合作伙伴，全力打造价值共享的生态体系，并承诺未来5年将投入超过100亿美元助力生态建设。

智元机器人发布最新双足人形机器人灵犀X2



本报讯 3月11日，智元机器人正式发布其最新研发的全能探索机器人——灵犀X2。灵犀X2具备完善的运动、交互及作业能力，展示了人工智能与人形机器人技术的完美融合。

运动方面，灵犀X2采用柔性材料外壳，全身拥有28个自由度，未使用任何并联结构。配备小脑控制器Xyber-Edge、域控制器Xyber-DCU、智能电源管理系统Xyber-BMS及核心关节模组Powerflow等核心组件，灵犀X2实现了运控算法层面的全面突破。通过结合深度强化学习和模仿算法学习的优势，灵犀X2展现了出色的运动灵活性。

交互能力方面，作为第一台真正具备复杂交互能力的“灵动机器人”，灵犀X2搭载了基于VLM的多模态交互大模型“硅光动语”，能够实现毫秒级的交互反应，通过人类的面部表情和语音语调精准判断情感状态，并做出相应的回应。研发团队还将动作模态集成到了模型当中，赋予了灵犀X2更加鲜活的“生命力”，比如模仿人类的呼吸韵律、会“暗中观察”，还有各类细小动作和肢体语言，让机器人拥有更多情绪表达的能力。

作业能力方面，基于一脑多形的智元启元大模型，X2初步具备了简单任务中对操作物体的零样本泛化能力，可在某些任务中实现多机协作，并外溢到日常生活的方方面面，同步应用于教育、医疗等多个领域。同时，灵犀X2采用轻量化设计，可模块化拓展，拥有完备的二次开发接口，以及预训练模型和“采-训-推”一站式方案，用户可根据需求自由探索，为康养、服务、家庭陪伴等各类场景打造应用。

有专家指出，这款机器人不仅在技术上实现了突破，更在人机互动的自然度和沉浸感上达到了新高度。随着技术的不断进步，灵犀X2有望成为人类生活的重要助手，为未来智能生活带来更多可能性。（志 源）