DeepSeek何以拥有定价权

编者按/ AI行业迎来一场极具戏剧性的“同台竞技”。

4月24日凌晨，OpenAI正式发布了GPT-5.5，意在稳固其在AI大模型领域的领先地位。几个小时后，DeepSeek-V4预览版上线并同步开源——其以1.6万亿参数的MoE架构、开源免费的“价格屠夫”姿态，再次向全球AI产业格局发起挑战。

这是DeepSeek在2025年1月R1模型惊艳全球、引发英伟达市值单日蒸发近6000亿美元后，时隔15个月迎来的第一次旗舰模型大迭代。这15个月，对一家AI创业公司而言，不短。V4的发布，让DeepSeek一直以来以技术突破实现极致性价比的策略再一次引爆，实现了对模型、资本市场和Token的三重定价权。

在此期间，DeepSeek身上聚集了诸多争议：模型一再跳票，核心研究员流失，创始人梁文锋从“永不融资”转身拥抱资本……种种传闻，让这家公司同时承载着“中国AI领跑者”和“理想主义困局”两个不同的标签。

就在V4发布的前一周，DeepSeek被曝寻求首次外部融资，目标估值从100亿美元迅速上调至逾200亿美元。时间巧合透露出关键信号：DeepSeek向市场递交技术答卷的同时，开始向投资人讲述商业故事。

这一次，我们需要解开的谜团不止一个：DeepSeek-V4究竟带来了什么、还差什么？DeepSeek真正的创新和突破点是什么？其崛起背后的逻辑是怎样的？一个长期以技术理想示人的团队，如何扛住融资、留人和商业化的三重压力？对这些问题，《中国经营报》记者通过深入采访调研，尝试从多个维度给出答案。

DeepSeek推动AI价格战：百万Token输入0.25元。　视觉中国/图

1. 技术解构

V4的双面镜像

距离DeepSeek-V3的发布已经过去484天，这一次，V4的发布比V3迭代周期几乎长了一倍。V4之所以迟到这么久，一份近60页的技术报告给出了近乎“残酷”的坦诚：训练数据量翻倍（从14.8T Token跃升至32T—33T），参数翻倍，在超大规模集群上，硬件细微误差被无限放大，训练稳定性遭遇“工程无人区”式的挑战。但恰恰是这种“啃硬骨头”的能力，构建了V4的技术门槛。

DeepSeek-V4最直观的技术突破，是将100万Token超长上下文作为所有官方服务的默认标配——无需额外付费，不分Pro与Flash版本。

这意味着一家律所可以将三年的合同打包扔给V4进行条款检索，一家科技公司可以将整个代码仓库交给模型独立维护，一个研究者则可以一次性输入数百篇论文完成文献综述。这不是“秀肌肉”，而是将长文本能力从旗舰专属变成人人都可低成本使用的基础设施。

而这一能力的实现，主要得益于架构层面的升级。V4引入的混合注意力机制，在Token维度进行压缩，结合DSA稀疏注意力，大幅降低了对计算和显存的需求。据官方数据，两款模型均大幅降低了每个标记的推理FLOP需求（降低73%），并将KV缓存内存占用降低90%。

一直深度关注AI大模型进展的北京问答智能科技创始人舒卫兵告诉记者，开源、1M上下文、国产算力适配、低成本Flash版本叠加，使V4可能成为第一个能让中国中小企业在自有数据上构建深度AI应用的开源旗舰模型——数据不出域、成本可预测。

V4发布中最具行业影响力的信号，是对国产算力底座的深度拥抱。

它说明DeepSeek的低价已经不再只是模型工程优化的结果，而是开始和国产算力的供给节奏绑定。过去，模型公司降价，外界通常理解为算法效率提升、厂商补贴或新一轮价格战。但这一次，DeepSeek把未来降价的前提，直接指向了昇腾950超节点的规模化部署。

这也意味着，中国大模型的竞争正在从“谁的模型能力更强”，进入“谁能把模型、芯片、工程系统和商业组织连成闭环”的阶段。DeepSeek-V4的发布，不仅是技术层面的突破，更是中国AI产业从跟跑向领跑转变的关键节点。

“V4亮点突出，技术上有显著进步（百万字上下文），而且算力消耗只有上一代的1/4。分了两个版本，Pro版打高端，Flash版走性价比，产品思路很成熟。同时还有战略意义给国产芯片‘站台’，也是给自己多了选择。”瞻胜传播创始人庞瑞对记者表示。

英伟达CEO黄仁勋在近期一次播客受访中说过一句分量很重的话：“如果DeepSeek先在华为平台上发布，那对我们国家来说将是灾难性的。”如今V4上线，海外媒体立刻将此话重提。

然而，国产适配的代价同样存在。DeepSeek为适配华为昇腾芯片，需要大规模重写底层代码、算子库乃至整体调度逻辑。据36氪报道，V4延期的重要原因就是V4将训练框架从英伟达迁移到华为昇腾上，以及内部决策意见不一。“当时，DeepSeek面临重新适配芯片的问题。梁文锋提出了一些自己的要求，但在执行层面很难折中。”据一名知情者透露。

一位国产芯片业内人士向记者坦言：这不是简单的“换卡”，本质上是一次从CUDA生态到CANN框架的全栈切换。CUDA发展了近二十年，积累了海量的开发者经验、算子库和优化技巧。DeepSeek要在这个领域从零做起完成迁移，面临的工程量堪称宏大——而且是在算力敏感的时间窗口内完成的。

同时，V4本身的短板也不容回避。第一，多模态的缺席，V4依然是纯文本模型。第二，幻觉率显著上升。Artificial Analysis测出，V4-Pro与Flash的幻觉率分别达94%和96%，远高于V3.2的82%。第三，高Token消耗带来额外成本。Artificial Analysis的测评发现，V4-Pro输出Token消耗量高达1.9亿，Flash版本进一步攀升到2.4亿Token。有开发者评价：“V4是一辆马力强悍的超跑，但油耗也确实不小。”

庞瑞直言，V4发布速度较慢，市场耐心被消耗。而且C端一直免费，B端API价格战打得凶，商业化如何上路还是个问号。

尽管存在不足，DeepSeek-V4的价值和意义远超技术本身。“V4证明了开源模型利用精妙的算法和工程优化，完全有能力比肩甚至超越闭源巨头，大大地激发了全球的AI创新活力。”一位开源社区贡献者对记者说，在技术封锁的背景下，DeepSeek的成功带来了宝贵的突围希望。

北京大学人工智能研究院多智能体与社会智能中心主任彭一杰则告诉记者，DeepSeek的成功标志着中国AI技术自主创新的重大突破。其启示我们，即便在不占有绝对资源优势的条件下，通过底层技术和研发思路的不断创新，同样能够跻身国际前沿。

2. 模式解码

DeepSeek出圈的深层逻辑

“全世界开发者的态度非常直接：好用就要，开源更要。”2025年2月，DeepSeek成为史上最快突破3000万日活App；在140个国家的苹果App Store下载榜中占据首位；DeepSeek的火爆出圈不言而喻。

这些数字背后，是DeepSeek自去年年初上线以来持续爆火的深层逻辑。首先是开源极大降低了开发者和中小企业的使用门槛。在DeepSeek之前，很多中小企业因为成本问题无法使用大模型技术。而现在，他们可以免费下载、自由使用，甚至基于DeepSeek进行二次开发。

其次是极致性价比，形成了“价格屠夫”的市场冲击。2025年，DeepSeek以557万美元的低成本实现对标GPT-4的性能，这一“性价比神话”成为其爆火的关键引爆点。而V4-Flash的输入成本定为缓存命中时0.2元/百万Token，再次将行业价格打到地板。

“这不是在打价格战，而是在重新定义价值。”上述开源社区贡献者解释道，当模型成本降低到一定程度，很多以前不可能的应用场景就变成了可能。

为此，DeepSeek的火爆出圈理所当然。“首先产品力是根本：DeepSeek在性能、响应速度等方面表现出色；其次开源策略形成病毒式扩散：开源降低了使用门槛，开发者生态迅速膨胀；最后‘十分之一成本做到同级’的故事打破了顶级模型必须靠资源堆叠的认知，给行业带来强烈震撼。”舒卫兵说，此外，脱胎于幻方量化，让DeepSeek自带“用算法优化对抗算力约束”的基因，走出了中国版AI创新的新路径。

庞瑞也认为，DeepSeek出圈背后，一是起步主打性价比。当时DeepSeek以低成本实现高性能出圈，训练成本才500多万美元，做到了别人几亿美元的事；二是开源，跟OpenAI路线不同，全球开发者都帮它免费宣传；三是赶上了好时候，中美科技博弈背景下，市场和媒体需要一个“国产突围”的故事，关注度极高。

而崛起背后是DeepSeek原创架构与商业模式上探索出了“第三条路”，即真正的创新和突破点在于底层架构的原创性。

分布式MoE（混合专家）架构是DeepSeek最具代表性的原创性设计，保持总参数规模的巨大容量的同时，大幅降低了单位推理成本。mHC架构解决了HyperConnection架构数据传输效率高但稳定性差的问题；Engram模块为超大模型的规模化发展铺平道路；而其对华为昇腾等国产芯片的深度适配，则打破了对英伟达CUDA生态的依赖。

“这不仅仅是技术优化，而是从0到1的原创。”前述芯片行业专家对记者说。这种创新的价值和意义在于，它标志着中国AI产业从跟跑向领跑的转变可能性，为全球开源生态注入了新的活力，也为国产AI产业链的自主可控提供了关键支撑。

而从商业逻辑看，DeepSeek也走出了一条独特的第三条路。一位长期跟踪AI赛道的投资人对记者这样分析：它没有走OpenAI那样的‘闭源+订阅’模式。DeepSeek选择的是‘开源模型权重+开源技术报告+收费API’的混合模式。开源降低准入门槛、快速建立社区信任，收费API为持续迭代提供现金流——这套闭环的飞轮效应一旦启动，竞争对手很难复制，因为它同时考验算法能力、工程能力和社区运营能力。

观察

DeepSeek崛起的真正意义

在数日的采访过程中，几乎所有受访者都重复了一个共同判断：DeepSeek真正的价值，不在于某个跑分或融资数字；而在于它以一家不到200人的企业，撬动了全球AI产业在多个关键赛道上的基本定价权和路线选择空间。这或许正是中国AI突围最需要的“鲇鱼效应”。它以模型效率突破算力壁垒，以开源换生态，以极致性价比切入全球开发者的日常工作和企业级成长链路。

如果用一句话总结DeepSeek崛起的深层价值和意义，或许可以说：DeepSeek带来的不是一家公司的成功，而是一场围绕AI基础设施的“范式转移”。

首先，DeepSeek改变了全球对中国AI创新能力的认知。去年1月R1发布时，英伟达市值单日蒸发近6000亿美元，创美股史上最大单日市值损失。德意志银行发布报告称其为中国的“斯普特尼克时刻”。V4进一步巩固了这种认知：即便在高性能GPU供应受限的环境下，依然可以通过原创架构与系统级工程优化，在全球开源生态中走在前列。

其次，DeepSeek以开源与低价策略将顶尖AI能力从“少数巨头的奢侈品”变为“人人可用的基础设施”。《财富》称V4的极致性价比可能彻底打破美国领先AI实验室的竞争护城河。

最重要的是，DeepSeek的崛起为中国探索“技术—资本—产业”良性互动的自主创新模式提供了鲜活范例。它成功将廉价电力、国产芯片产能、算法创新、开源生态和市场化定价等多个优势要素完整地串联起来，形成了一条可以模拟演进的自主链条。

V4发布当日，DeepSeek在官方推文的最末，引用了《荀子·非十二子》中的一句话：“不诱于誉，不恐于诽”。如今，DeepSeek依然需要保持清醒：其依然面临严峻的人才、商业路径、技术短板和资本竞争，当顶尖闭源模型的功能维度不断扩展时，纯文本旗舰模型与多模态现实之间也存在技术“裂谷”，而在商业化的彼岸，腾讯、阿里等巨头的注资虽然能提供资金与战略加持，但也将带来新的利益与战略决策的复杂权衡。

实际上，DeepSeek的名字已经成为中国科技史的一个坐标。这或许就是DeepSeek崛起真正的价值和意义——它不仅是一家公司的成功，更是中国AI产业从跟跑向领跑转变的缩影，并向世界展示着中国AI另一种可能性——一种强调基础创新、开源共享、系统闭环和商业可持续的中国AI突破之路。

DeepSeek是那个引领者和破晓者，未来中国的DeepSeek们会更多。对于中国AI和科技产业来说，更精彩的故事才刚刚开始。

本版文章由中经记者吴清采写

3. 机遇挑战

十字路口的未来抉择

摆在DeepSeek面前的，是一个远比跑分更沉重的交叉格局，机遇挑战并存。

在机遇端，目前国家自主可控算力战略与国产芯片的适配需求日益明确，DeepSeek与华为昇腾的深度合作恰好对接了这一“大算力时代的国家需求”，有望长期在国产算力生态中占据关键位置。

“作为首个在旗舰模型上完成国产算力闭环的玩家，DeepSeek在政企市场、国资背景项目中拥有独特的自主可控叙事优势。”舒卫兵说，V4针对多种Agent框架进行专门优化，在代码生成、文档处理等场景表现突出。2026年是智能体爆发元年，DeepSeek有机会成为能够执行复杂任务的AI Agent基础设施。

记者注意到，目前，华为昇腾、天数智芯、寒武纪等国产芯片厂商均已实现对DeepSeek-V4新模型的支持。同时，金山办公、360等众多企业已通过华为云接入DeepSeek新模型。

在挑战端，一是国内智谱、科大讯飞、MiniMax等大模型厂商加速商业化与IPO进程，市场资源高度聚集；国外Anthropic、OpenAI、谷歌也不断拉高资本壁垒。二是多模态能力缺失使得DeepSeek在业务覆盖、应用场景和技术护城河上被竞争对手逐步拉开缺口。三是互联网大厂和创业公司的“人才磁吸能力”持续放大，DeepSeek如果不能尽快落实长效激励、稳定核心队伍，可能面临人员流失带来的二次冲击。

“V4的万亿参数架构、国产芯片适配、下一代Blackwell并行训练，资金消耗成倍放大。幻方量化年收入约50亿元人民币，在万卡级/十万卡级算力投入面前已显吃力。”舒卫兵表示，没有清晰的估值锚点和股权激励兑现路径，核心人才的稳定性面临持续挑战。

据不完全统计，DeepSeek已有多名核心研发成员被挖走：DeepSeek-V2架构的关键贡献者罗福莉加盟小米，另一位核心人物郭达雅加盟字节跳动Seed团队，多模态核心研究员阮翀加盟智能驾驶解决方案商元戎启行。

一位长期从事科技人才招聘的猎头告诉记者：国内AI人才市场的竞争已经白热化。DeepSeek以理想主义吸引了一大批优秀青年，但市场最终会给人才定价。大厂不仅出价更高，还有更大的算力集群、更丰富的落地场景、更大的团队规模。对一线AI人才来说，这是一道平衡现实与理想的考题，DeepSeek想要长期留住他们，光靠技术使命感是不够的。

还有一个现实挑战是商业化理想与现实的两难。V4的发布虽加强了代码与Agent能力，但开源与商业化之间的矛盾也随之暴露——一旦API定价过低而调用量暴增，推理成本可能急剧攀升。如果DeepSeek大幅度涨价或转为闭源，那么它此前在开源社区积累的良好声誉可能受到冲击。

这一道“商业化的窄门”，梁文锋团队还需要用更多精细化的产品和生态设计来跨越。近日有消息称，DeepSeek正以200亿美元以上的估值与腾讯、阿里等洽谈融资，但截至发稿，三方均未回应。

对此，庞瑞直言，从不融资到要融资，不是创始人变卦了，而是市场环境在发生变化，一是算力军备竞赛太烧钱；二是如果没钱，优秀的人留不住；三是得通过融资绑定所需商业资源。

而在舒卫兵看来，引入外部资本意味着公司治理结构走向规范化，成为一个要面对市场检验、要向投资人交代的商业主体。这是DeepSeek正式加入AI生态下半场竞争的入场券。为公司补充海量资金补齐IDE、Coding工具、Agent产品等终端能力的同时，融资确权将股权从“纸上富贵”变为可量化、可交易的资产，是现阶段稳定核心团队的最关键举措。

同时，一个研究型团队最擅长的是突破。可是当它变成全球关注的基础模型公司之后，要面对的就不只是模型能力了。用户会涌进来，API调用会增加，企业客户会提出稳定性要求，开发者会期待生态支持，同行会快速追赶，巨头会加码，人才会被争抢，监管和国际环境也会变得更复杂。

在采访中，诸多AI从业者和专家对DeepSeek深层突破提出了建设性思路。

舒卫兵认为，面对变局，DeepSeek亟须做好以下三方面，首先是补齐应用层短板：建议尽快推出自有IDE、编程助手、Agent框架等终端产品，直接触达用户并获取反馈数据。其次是构建“场景Token”的生态闭环：DeepSeek应寻找拥有高壁垒场景数据（如金融、医疗、法律、制造等领域），将模型能力嵌入具体工作流，而非仅提供API调用。最后是保持技术敏捷性与商业规范性的平衡：完成外部融资后，DeepSeek需要在建立财务内控、合规审计等体系的同时，保护原有的技术极客文化和研发敏捷性。

还有更现实的问题是，如何保持技术普惠和商业盈利间的平衡。DeepSeek带动的开源开放生态，让中国模型在2025年内，快速在全球建立知名度和技术口碑，但在基模研发仍然“吞金”的当下，如何将口碑转化成真金白银，也是很重要的生存命题。

融资传闻的出现，恰恰说明DeepSeek正在从一个能打仗的研究团队，转变成一家要长期守成的基础设施公司。

“如果说V2证明了DeepSeek的效率，V3证明了DeepSeek的全球竞争力，R1证明了DeepSeek的破圈能力，那么V4和融资传闻共同说明的是：DeepSeek正在告别‘孤胆英雄’阶段。”上述投资人说，它正在变成一个必须处理商业化、生态化、资本化和平台关系的玩家。

这也是DeepSeek故事最有意思的地方。它一开始像是大模型时代的反叛者：不靠巨头，不靠VC，不靠铺天盖地的营销，而是靠模型本身把牌桌掀了一下。但现在，它也不得不承认，AI战争越往后，越不是单点突破，而是系统竞争。

同时我们需要从更宏观的角度深入思考并实践加速人工智能自主创新的路径。彭一杰建议，首先，进一步鼓励面向核心技术的联合攻关。其次，着力培养和吸引高水平人才。人工智能的竞争归根结底是人才的竞争。再次，大力推进人工智能与实体经济的深度融合。我国在工业制造、物流运输、医疗教育等领域都有巨大的智能化改造空间。最后，在坚持自主可控的前提下，不断拓展国际合作与规则对话。

正如某知名科技评论人所分析的：DeepSeek真正的稀缺性，来自它还是一家未被资本定义的、能做出世界级原创成果的公司。如果它能够顺利走完从理想实验室到商业基础设施的转型路，它的技术硬实力、组织敏捷度和商业潜力，都将在未来五到十年释放出更惊人的能量。