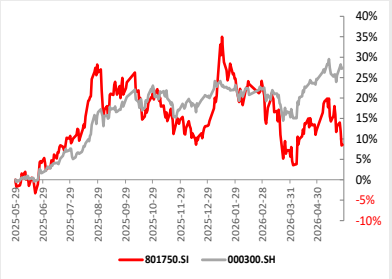
市场研究部 2026 年 5 月 29 日

DeepSeek 永久降价：模型成本曲线重构

看好

市场表现截至 2026.5.28



数据来源：Wind，国新证券整理

事件

2026 年 4-5 月，DeepSeek 将 V4-Pro API 价格降至原定价 1/4 并永久执行，成为国产大模型价格战关键转折点。小米随后跟进 MiMo-V2.5 系列 API 永久降价。

核心观点

DeepSeek 本轮降价将临时优惠转为永久价格体系，改写了全球大模型 API 定价基准。其 V4-Pro 各计费项统一降价 75%，标准输入价格仅为 GPT-5.5 Pro 的约 1/72，高缓存命中场景下企业使用成本可降 90%以上。降价未牺牲模型能力，Openrouter 数据显示 5 月至今 DeepSeek V4 Flash 调用量排名第一。降价核心源于底层技术架构优化，V4 系列采用混合注意力架构与多 token 预测技术，单 token 推理浮点运算量仅为前代的 27%，KV 缓存大小降至前代的 10%。同时，DeepSeek V4 已与华为昇腾完成深度适配，国产算力生态成熟提供了供应链支撑。Gartner 预测，到 2030 年大模型推理成本将较 2025 年降低超 90%，本次降价是这一长期趋势的阶段性体现。

DeepSeek 的永久降价打破了行业原有竞争平衡，基座模型进入 K 型分化时代。中间层通用模型厂商面临最大压力，既无闭源前沿模型的能力护城河，又失去了性价比优势；私有化部署能力强的厂商相对受益，金融、政务等行业有刚性本地化需求；C 端/多模态产品厂商冲击有限，收入不完全依赖 API；闭源前沿模型在复杂任务上仍有护城河。高端模型市场具备独立定价能力，智谱 GLM-5.1 等厂商 2026 年 Q1 涨价 83%后调用量反而增长 400%，验证高价值场景下能力优先于价格。

降价将大幅刺激大模型总调用量爆发式增长，带来产业链系统性价值重分配。云厂商迎来结构性利好，AI 推理需求持续增长推高云服务价格中枢；国产 AI 软硬件生态闭环加速形成，为国产半导体产业链打开替代窗口；应用层“Token 自由”时代到来，长文档分析、代码生成等场景将实现规模化运行，AI 应用从试点阶段进入全面商业化兑现期。

投资线索

重点把握三大投资方向：一是优先布局 AI 应用与 Agent 赛道，具备清晰场景闭环和变现能力的厂商有望率先实现业绩兑现；二是布局国产推理算力产业链，关注国产 AI 芯片、算力租赁及基础软件企业；三是关注具备强私有化部署能力的行业模型厂商。同时需警惕缺乏核心技术壁垒的中间层通用模型厂商风险。资本市场估值逻辑已转向认可成本曲线重构能力，长期看好具备成本重构与生态协同优势的产业链核心标的。

相关研究

分析师：钟哲元
登记编码：S1490523030001
邮箱：zhongzheyuan@crsec.com.cn

证券研究报告

风险提示

1、技术发展不及预期；2、市场竞争加剧；3、地缘政治影响。

目录

一、从阶段性补贴到长期价格锚的确立

2026 年 4-5 月，DeepSeek 先后推出大模型 API 阶段性降价与永久降价政策，将 V4-Pro API 价格降至原定价的 1/4，成为国产大模型价格战的关键转折点。本次降价并非短期市场促销行为，而是基于底层工程效率突破的结构性价格调整，将大模型从“高价稀缺服务”推向“可规模化交付的基础设施”。随后小米跟进 MiMo-V2.5 系列大模型 API 永久降价，进一步巩固了普惠大模型的价格锚，推动国产基座模型进入清晰的 K 型分化阶段。行业的估值逻辑正从“模型稀缺性叙事”向“成本效率+真实 ROI”切换，产业链上下游迎来价值重分配，AI 应用商业化落地进程将显著加速。

DeepSeek 本轮降价的标志性意义在于将临时优惠转化为永久价格体系，改写了全球大模型 API 的定价基准，其降价动作呈现清晰的递进式节奏，逐步确立普惠价格锚。

图表 1：DeepSeek 降价核心时间线

时间节点	核心事件
2026 年 4 月 24 日	DeepSeek-V4 预览版正式发布并同步开源
2026 年 4 月 25 日	推出 V4-Pro 模型 API 限时 2.5 折优惠（原定 5 月 5 日截止）
2026 年 4 月 26 日	全系 API 输入缓存命中价降至首发价 1/10，V4-Pro 叠加 2.5 折后总降幅达 97.5%
2026 年 4 月 28 日	将 V4-Pro 的 2.5 折限时优惠从 5 月 5 日延长至 5 月 31 日
2026 年 5 月 22 日	官宣 5 月 31 日优惠到期后永久维持原价 1/4，不再恢复原价

资料来源：官网及媒体报道，国新证券整理

永久降价后，DeepSeek-V4-Pro 形成了极具竞争力的定价体系，各计费项降幅统一为 75%，打破了此前行业的定价惯性。

图表 2：DeepSeek 降价核心时间线

计费项	原价（元/百万 Tokens）	永久价（元/百万 Tokens）	降幅
输入（缓存命中）	0.1	0.025	75%
输入（缓存未命中）	12	3	75%
输出	24	6	75%

资料来源：官网及媒体报道，国新证券整理

从价格对比来看，DeepSeek V4 系列与全球主流大模型形成了数量级的差距，其中 DeepSeek V4-Pro 在相同计费项目下仍具有显著价格优势：其标准输入价格（3 元/百万 Tokens）仅为 GPT-5.5 Pro 标准输入价格（216 元/百万 Tokens）的约 1/72，形成数十倍的价格差距，在 RAG 知识库、智能客服、文档分析等高缓存命中场景中，企业使用成本可下降 90%以上。5 月 27 日，小米宣布 MiMo-V2.5 系列 API 永久降价，最高降幅达 99%，其定价完全对标 DeepSeek V4 系列，而 MiMo-

V2-Pro 旗舰基座与全模态模型则维持原价不变，这一动作进一步印证了行业分层定价的趋势，也让普惠大模型的价格锚更加稳固。

特别的，本次降价并未以牺牲模型能力为代价，DeepSeek 在维持旗舰级性能的前提下实现了价格的断崖式下降，实现了“接近前沿可用性前提下的价格重构”。Openrouter 数据显示，5 月至今 DeepSeek V4 Flash 调用量排名第一。

二、工程效率突破而非短期补贴

DeepSeek 能够实现可持续的大幅降价，核心源于底层技术架构的大幅优化，而非单纯的资本补贴。V4 系列引入压缩稀疏注意力(CSA)与重度压缩注意力(HCA)结合的混合架构，并采用多 token 预测技术，使得 V4-Pro 单 token 推理浮点运算量仅为前代 V3 的 27%，KV 缓存大小更是降至前代的 10%。从长上下文场景的计算表现来看，随着 token 位置拉长至 1024k，V3.2 的单 token 计算量快速攀升至 1.2 TFLOPs，而 V4-Pro 仅为 0.3 TFLOPs，V4-Flash 更是低至 0.12 TFLOPs，这些技术突破直接让长上下文处理、Agent 工作流、自动化编程等原本成本高昂的场景具备了大规模商业化的基础。

与此同时，国产算力生态的成熟为降价提供了供应链支撑，DeepSeek V4 已与华为昇腾完成深度适配，Ascend Super Node 成功实现 V4 推理部署，CANN 承担了类似 CUDA 的底层适配角色，尽管当前吞吐瓶颈仍需等待 2026 年下半年昇腾 950 批量出货后解决，但已充分验证了国产芯片在推理侧的可用性。从行业长期趋势来看，推理成本持续下行是必然规律，Gartner 预测，到 2030 年，生成式 AI 提供商对 1 万亿参数的大语言模型的推理成本将较 2025 年降低超过 90%，本次 DeepSeek 的降价只是这一长期趋势的阶段性体现，其本质是通过重写行业成本曲线，提前打开大模型的市场空间。存储带宽成为推理性能瓶颈，“以存代算”成为降本关键路径，英伟达 H200 GPU 相比前代 H100，其 HBM 容量提升了 76%，同时带宽提升了 43%，便实现推理吞吐量的大幅提升，这是 DeepSeek 实现 KV 缓存大幅优化的核心产业背景。

三、基座模型进入 K 型分化时代

DeepSeek 的永久降价彻底打破了国产大模型原有的竞争平衡，行业呈现出 K 型分化格局，不同阵营的定价策略与商业模式走向截然不同的方向。本轮降价最直接的冲击对象是处于中间层的通用模型厂商，这类厂商既没有闭源前沿模型在高复杂度、长周期任务上的可靠性护城河，又失去了此前赖以生存的性价比优势，若下一代模型无法在成本或质量任一维度实现突破，将面临严重的双向挤压。

不过不同收入结构与业务模式的厂商受影响程度存在显著差异，降价带来的冲

击呈现明显的结构性特征。

图表 3：DeepSeek 降价对主要模型厂商的结构性影响

厂商类型	受影响方向	主要原因	可能的应对路径
中间层通用模型厂商	压力最大	性能不显著领先，价格优势被 DeepSeek 抹平	下一代模型拉开能力差，或退守私有化部署
私有化部署能力强厂商	相对受益	金融、政务、能源等行业有刚性本地化需求	提升行业渗透率、强化客户粘性
C 端/多模态产品厂商	冲击相对有限	收入不完全来自 API，用户迁移不只看 token 价格	强化产品体验与场景闭环
闭源前沿模型	基础层承压但仍有护城河	在复杂长周期任务、可靠性与服务上仍占优	提升价值密度，维持高端客户溢价

资料来源：国新证券整理

市场对中间层厂商冲击存在结构性分歧，以 MiniMax 为例，2025 年全年面向 C 端的 AI 原生产品收入为 5307.5 万美元，占总营收的 67.2%，企业收入多面向国际客户，公有云 API 依赖度低，叠加多模态与强化学习引擎壁垒，受降价冲击有限，表明价格战仅显著影响公有云调用收入占比高的企业。在此背景下，大模型行业的估值逻辑发生了转变，此前基于“模型稀缺性”的估值叙事被打破，市场开始重新计算真实投资回报率，模型公司的估值逐渐拆分为两部分：公有云 API 业务看成本效率与规模效应，私有化部署业务看行业渗透率与客户粘性。

小米的跟进降价进一步强化了这一分化，行业最终形成了两条并行不悖的发展赛道，不同阵营的商业诉求与核心壁垒存在本质差异。高端模型市场具备独立定价能力，比如智谱 GLM-5.1 等厂商 2026 年 Q1 累计涨价 83%，调用量反而增长 400%，验证高价值场景下能力优先于价格；大厂与创业模型厂商商业模式存在本质差异，大厂以 API 为生态入口，可接受长期微亏，创业厂商以 API 为利润中心，必须靠自身收入覆盖成本，这是 K 型分化的底层逻辑。

四、需求扩容与价值重分配

DeepSeek 的降价将通过降低单次调用成本，大幅刺激总调用量的爆发式增长。尤其是 Agent、多步推理、长文本处理等任务，其使用频率与 token 消耗量本身具备极强的价格弹性，成本的断崖式下降将显著提升企业与开发者的采用意愿，2026 年 3 月中国大模型日均词元调用量已突破 140 万亿，三个月内增长超过 40%，降价后这一增速还将进一步加快。

对云厂商而言，这种变化带来的是结构性利好，一方面 DeepSeek 等模型厂商压低了 AI 使用门槛，推动更多企业接入云服务，另一方面 AI 推理需求的持续增长正在推高云服务的价格中枢，即使模型 API 价格下行，云侧的总需求规模仍在不断扩大。

同时，DeepSeek 与国产芯片的深度适配，正在加速国产 AI 软硬件生态的闭

环形成，当下游推理需求大规模放量时，应用厂商为控制成本与保障供应链稳定，将更倾向于选择性价比更高的国产算力方案，这为国产半导体产业链打开了从能用到好用的替代窗口。算力结构加速重构，Agent 时代 CPU/GPU 配比从 1:4 向 1:2 甚至 1:1 演进，CPU 迎来价值重估。阿里云、腾讯云、百度智能云等国内头部云厂商相继上调 AI 算力产品价格，涨幅普遍在 5%至 50%之间。

在应用层，“Token 自由”时代的到来解决了此前制约 AI 应用落地的核心成本问题，长文档分析、代码生成、智能客服、企业自动化工作流等场景将实现规模化运行，企业也将形成“分层使用”的模型选型模式，高频简单任务采用低价普惠模型，低频复杂任务采用高端高性能模型，这将推动 AI 应用从试点阶段进入全面商业化兑现期。

五、投资建议

投资层面，应重点把握 DeepSeek 降价带来的产业链系统性机会，按照受益顺序优先布局三大方向。首先优先关注 AI 应用与 Agent 赛道，推理成本的大幅下行直接降低了应用开发与运营成本，此前受成本制约的长文本处理、多步骤交互、企业自动化等细分场景将迎来商业化爆发，具备清晰场景闭环、稳定用户群体与较强变现能力的应用厂商有望率先实现业绩兑现。其次布局国产推理算力产业链，重点关注国产 AI 芯片厂商、算力租赁服务商以及适配国产芯片的基础软件企业，随着 2026 年下半年昇腾 950 等新一代国产芯片批量出货，推理侧算力供需紧张的格局将逐步缓解，相关厂商将充分受益于推理需求的持续放量。再者关注具备强私有化部署能力的行业模型厂商，金融、政务、能源等领域的本地化与数据安全需求具有刚性，这类厂商受公有云价格战冲击较小，能够通过深耕行业场景、提升客户粘性实现稳健增长。

同时需要警惕缺乏核心技术壁垒，既无法在模型能力上形成差异化，又不具备极致成本控制能力的中间层通用模型厂商，这类企业的市场份额与估值水平可能面临持续压缩的风险。资本市场估值逻辑已切换，DeepSeek 拟融资 700 亿元、投前估值 450 亿美元，资金优先投向 AGI 前沿研究，表明资本从追捧参数叙事转向认可成本曲线重构能力，这一估值导向可作为行业投资判断的重要参照，长期看好具备成本重构能力与生态协同优势的产业链核心标的。

六、风险提示

1、技术发展不及预期；

2、市场竞争加剧；

3、地缘政治影响。

投资评级定义

公司评级		行业评级	
强烈推荐	预期未来 6 个月内股价相对市场基准指数升幅在 15%以上	看好	预期未来 6 个月内行业指数优于市场指数 5%以上
推荐	预期未来 6 个月内股价相对市场基准指数升幅在 5%到 15%	中性	预期未来 6 个月内行业指数相对市场指数持平
中性	预期未来 6 个月内股价相对市场基准指数变动在-5%到 5%内	看淡	预期未来 6 个月内行业指数弱于市场指数 5%以上
卖出	预期未来 6 个月内股价相对市场基准指数跌幅在 15%以上		

免责声明

国新证券股份有限公司市场研究部
地址：北京市朝阳区朝阳门北大街 18 号中国人保寿险大厦 11 层（100020）
传真：010－85556155 网址：**www.crsec.com.cn**