

市场研究部

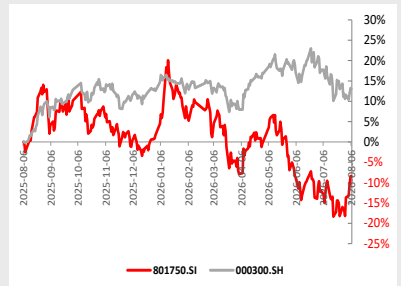
2026 年 8 月 6 日

GPT-5.6 降价与 DeepSeek 定价转向

看好

市场表现截至

2026.8.5



数据来源：Wind，国新证券整理

相关研究

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

事件

2026 年 7 月末至 8 月初，OpenAI 对 GPT-5.6 家族落地结构性调价，中低端档位大幅让渡价格空间；DeepSeek 先于 7 月 31 日上线 V4 Flash 正式版，将高频场景调用成本压至行业新低，随即在 8 月初预告 API 整体上调定价。

核心观点

大模型行业已迈过同质化降价阶段，形成中美分层、国内双轨的定价格局，核心支撑是算力成本刚性约束与场景价值分化。美国市场率先转向高端溢价，OpenAI 旗舰模型 2026 年重回提价通道，依托复杂 Agent、专业编码等高价值场景重建能力溢价。中国市场双轨并行：智谱等年内多次上调 API 定价，调用量仍实现高增长，在代码、企业智能体场景拿到局部定价权；DeepSeek 以极致成本优势抢占全球份额。2025 年末至 2026 年上半年 GPU 算力价格回升近四成，终结行业普降基础，推动竞争转向分层化。

2026 年 7 月 OpenAI 为 GPT-5.6 搭建 Sol、Terra、Luna 三层产品矩阵并快速调价，执行“高端守价、中端巩固、低端抢量”策略：Sol 维持原价并推出付费加速模式，守住能力溢价与品牌上限；Terra 承接中等复杂度需求，优化家族路由效率；Luna 大幅降价 80%，主动下沉应对国产模型竞争。此次调整标志其收入逻辑从“旗舰溢价驱动单用户收入”转向“分层路由驱动总调用量与客户留存”，中低端模型不再具备稳定提价权。

DeepSeek V4 Flash 通过架构与工程优化，能力接近 GPT-5.6Luna，高比例缓存折扣大幅降低 Agent 规模化落地门槛。随后其预告 API 整体涨价，标志策略从低价抢量转向平衡规模与盈利质量，行业价格战阶段性降温，竞争回归能力、服务与场景适配的综合比拼。

当前全球市场分层清晰：高端高失败成本场景以能力为核心，头部厂商保有强定价权；中端规模化自动化市场竞争激烈，价格为选型核心变量。智能路由与聚合平台兴起，侵蚀闭源模型广谱溢价，兼具模型、算力与生态的云厂商优势凸显。中美竞争进入定价权争夺阶段，中国模型凭借显著性价比优势，周调用量连续十四周领跑全球，市场份额持续重构。

投资线索

本轮定价转移标志产业步入商业化成熟期，利润池呈结构性迁移。基础模型赛道关注两类厂商：在高端场景具备不可替代性、可持续维持定价权的厂商，以及拥有极致工程效率、能在低价市场实现规模化盈利的厂商。云与 MaaS 平台是主要受益方，调用量增长叠加算力利用率提升，头部厂商有望实现收入利润双增。推理优化与 AI

证券研究报告

中间件确定性较强，将持续受益于企业降本诉求。垂直 AI 应用依托场景理解与 workflow 壁垒具备长期估值弹性，办公自动化、开发工具等赛道落地节奏有望加快。

 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、定价格局从普降潮到分层双轨.....	3
二、GPT-5.6 分层调价：OPENAI 的分层防守策略	3
三、DEEPSEEK 定价转向：从极致成本优势到商业化修复	4
四、全球大模型市场新生态	6
五、投资建议	7
六、风险提示	8

一、定价格局从普降潮到分层双轨

大模型行业早已迈过同质化降价的初期阶段，形成了清晰的中美市场分层、国内双轨并行的格局，背后最核心的支撑来自算力成本的刚性约束与场景价值的持续分化。

美国市场率先完成了从降价抢量到高端溢价的切换。OpenAI 旗舰模型定价在经历两年下行后，于 2026 年重回上行通道，从 GPT-5 迭代到 GPT-5.6 Sol，旗舰层输入价格稳定在每百万 Token 5 美元、输出 30 美元的水平。这一轮定价抬升的核心逻辑，是头部厂商商业目标的转向--在复杂 Agent、专业编码、高风险工作流等场景中，客户为任务成功率付费的意愿远强于对 Token 单价的敏感度，头部厂商因此得以重建高端场景的能力溢价。

中国市场则走出了双轨路线。智谱在 2026 年一季度累计上调 API 调用定价 83%，其中 2 月 GLM Coding Plan 结构性调价涨幅 30%起、3 月 GLM-5-Turbo API 价格上调 20%、4 月 GLM-5.1 再提价 10%；即便如此，一季度调用量仍增长 400%。腾讯云、阿里云等平台的算力服务与部分模型也出现不同幅度涨价，既反映出上游算力紧张向服务端的传导，也说明部分国产模型在代码、企业智能体等高价值场景中，已经拿到了局部定价权。另一部分厂商则延续规模优先的路线，此前 DeepSeek 通过多轮永久降价，以极致成本优势抢占全球市场份额。

这种分化的底层支撑是上游算力供给的刚性约束。2025 年末到 2026 年上半年，GPU 云算力租赁价格并未延续此前的下降趋势，反而出现阶段性回升，一年期 H100 合同价格从 2025 年 10 月的每卡每小时 1.7 美元，上涨至 2026 年 3 月的 2.35 美元，涨幅接近四成，新卡价格上行更为明显。算力成本的回升，让全行业无差别降价失去了基础，厂商必须根据自身定位与目标场景做出差异化选择，行业也因此从单一的 Token 单价竞争，演化成不同能力层级、不同任务场景下的分层竞争。

二、GPT-5.6 分层调价：OpenAI 的分层防守策略

2026 年 7 月，OpenAI 围绕 GPT-5.6 完成了产品与定价的系统性调整，首次在同代内搭建起 Sol、Terra、Luna 三层产品矩阵，且上线仅三周便快速实施结构性调价：旗舰 Sol 维持原价（输入\$5.00/输出\$30.00 每百万 Token），Terra 降价 20%（调整后输入\$2.00/输出\$12.00），Luna 大幅降价 80%（调整后输入\$0.20/输出\$1.20），同时为 Sol 推出双倍价格的 Fast 加速模式（处理速度提升至标准模式的 2.5 倍）。整个调整节奏明显快于过往产品周期，透露出明确的竞争应对意图。

这套三层架构标志着 OpenAI 不再用单一旗舰模型通吃全部市场，转而搭建更贴近企业预算管理的分层供给体系。其中 Sol 定位旗舰前沿能力，面向复杂推理、科研、长程 Agent 等高失败成本场景，维持原价的核心目的是守住 OpenAI 的能力

溢价、利润锚与品牌上限；同步推出的 Fast 模式则把时延从服务属性拆分为可单独售卖的商品，让旗舰层的商业逻辑从单一能力溢价，延伸为能力与时延的双重溢价。Terra 定位为主流生产层，面向日常生产 workflow，承担着将原本跑在 Sol 上的中等复杂度请求拦截在更低成本层的作用，既可以降低客户整体账单，也能提升家族内部的路由效率，是整个经营体系的核心中枢。Luna 则定位为高并发放量层，面向分类、摘要、提取、轻量 Agent 等高吞吐、价格敏感任务，大幅降价后直接切入超低价市场，与中国厂商的高性价比模型展开正面竞争。

从调价幅度的差异不难看出 OpenAI 的战略意图：高端守价、中端巩固、低端抢量，主动拉大各层之间的价格梯度，形成覆盖不同预算与场景的完整产品矩阵。官方披露的工程优化，包括针对 Hopper 架构重写的 GPU kernel、迭代后的推测解码框架，能够带动端到端成本下降两到三成，但显然不足以支撑 Luna 高达 80% 的降幅。这一轮降价的核心驱动力，本质还是来自市场竞争压力。进入 2026 年二季度后，DeepSeek 通过 V4 Pro 的永久折扣、V4 Flash 的低价与激进的缓存折扣，持续下压“足够好用”场景的市场清算价格，其他国产模型也在高性能区间持续发力，不断挤压 OpenAI 中低端产品的市场空间。如果 Luna 维持最初的定价，其所在的预算层将快速被中国厂商挤出企业的默认选型清单，OpenAI 会失去大量高频调用的流量入口。

因此 Luna 的大幅降价，本质是 OpenAI 主动下沉中低价市场的防守动作。它不需要做到绝对最低价，只需要让价格差距缩小到不足以覆盖客户的迁移、合规与生态切换成本，就能守住 API 流量与开发者入口。整体来看，这一次分层调价标志着 OpenAI 的收入模型，正式从“旗舰溢价驱动单用户收入”切换为“分层路由驱动总调用量与客户留存”。OpenAI 并没有放弃定价权，而是在重新定义定价权，把过去单一旗舰的每 Token 溢价，转化为全家族在不同任务层级上的预算占有率。这也意味着，单纯依赖代际升级持续抬价的阶段已经结束，至少中低层能力不再具备稳定的提价权，闭源龙头必须接受模型层毛利率中枢下移的现实，通过生态绑定与 workflow 绑定来巩固自身的市场地位。

三、DeepSeek 定价转向：从极致成本优势到商业化修复

7 月 31 日上线的 DeepSeek V4 Flash 0731 正式版，真正的冲击力在于以接近 GPT-5.6 Luna 的能力水平，通过架构与工程优化重构了高频场景的成本函数；而紧随其后的 API 涨价预告，则让市场看到了国产头部模型从规模扩张期进入商业化修复期的明确信号。

从能力层面看，新版本的 V4 Flash 维持了 284B 总参数、13B 激活参数与 1M 上下文窗口的基础架构，仅通过后训练就实现了能力的大幅跃升。Artificial Analysis 的数据显示，V4 Flash 0731 在智能指数（Intelligence Index）上拿到 50 分，仅比 GPT-5.6 Luna 的 51 分低 1 分，二者综合智能水平极为接近。在分项基

准上二者互有胜负：V4 Flash 在 AutomationBench (25.1%vs14.9%)、Toolathon (70.3%vs53.4%) 等 Agent 工具调用类基准上领先 Luna；而 Luna 在 Agents' Last Exam (50.3%vs25.2%)、DeepSWE (67.2%vs54.4%)、Terminal-Bench 2.1 (84.7%vs82.7%) 等深度推理与代码基准上仍保持优势，且 Luna 支持多模态输入而 V4 Flash 仅支持文本。这种能力提升没有伴随参数规模的扩大与价格的上涨，完全来自后训练对执行能力的强化，也说明 2026 年的前沿模型竞争，已经不再是单纯堆参数提价格的线性逻辑，后训练优化、推理系统降本与 API 侧定价设计的组合，正在成为更核心的竞争力。

其成本优势的底层支撑，来自架构与工程两方面的持续优化。架构层面，V4 Flash 沿用的 CSA+HCA 混合注意力机制，能够对 KV 缓存做高比例压缩，同时通过稀疏筛选只保留相关性最高的 Token，大幅削减无效计算。DeepSeek 技术报告显示，在 1M 上下文设定下，V4-Pro 仅需 DeepSeek-V3.2 的 27% 单 Token 推理 FLOPs 和 10% 的 KV Cache；V4-Flash 由于进一步通过稀疏筛选只保留相关性最高的 Token，工程测算显示其单 Token 计算与缓存占用较 V4-Pro 再度显著下降。配合 FP4/FP8 混合精度策略，进一步压缩了显存占用与算力消耗。工程层面，DeepSeek 联合北京大学发布的 DSpark 推测解码框架，在 V4 Flash 真实线上流量中，相较 MTP-1 基线实现单用户生成速度提升 60% 到 85%，严格 SLA 下系统吞吐最高可提升数倍，吞吐提升直接对应单位算力承载更多请求，进一步摊薄了单 Token 的算力成本。

反映到定价上，V4 Flash 最核心的杀伤力来自极致的缓存定价机制，其缓存命中输入价格仅为未命中价格的 2%，远高于行业普遍的 90% 折扣水平。这种差异在普通聊天场景中影响有限，但在 Agent 工作流程中会被无限放大。Agent 场景的典型特征是多轮调用、重复上下文多，每一轮交互都会反复发送相同的系统提示词、工具定义、代码上下文与历史状态，八九成的输入都属于重复内容。提示缓存会将固定内容的计算结果暂存，命中后直接跳过 Transformer 层运算，几乎不消耗算力资源，因此 DeepSeek 可以给出近乎免费的缓存定价。根据 Artificial Analysis 测算，在典型 Agent 任务下，V4 Flash 的单任务成本比 GPT-5.6 Luna 低约 60%，除了价格本身更低之外，关键原因是 DeepSeek 提供 98% 的缓存命中折扣（行业普遍 90%）。这种成本结构的重构，最直接的影响是大幅降低了 Agent 规模化落地的门槛，让原本只够做小规模试点的预算，现在可以支撑全量的生产部署。

就在 V4 Flash 正式版上线后不久，DeepSeek 发布公告，计划近期整体上调 API 服务定价，预计涨幅较大，具体方案以正式通知为准。这是其继 7 月推出峰谷定价机制后，又一次重要的定价调整。此前的峰谷分时定价，本质是用价格杠杆分流高峰算力负载，缓解服务卡顿、保障稳定性，并非全面涨价。这也标志着其策略从低价抢市场，转向平衡规模与盈利质量，告别单纯的补贴式价格战。除此之外，市场普遍认为本次涨价与即将发布的 V4 Pro 正式版相关，模型能力升级伴随算力投入提升，需要匹配相应的定价体系。

当然，即使完成涨价，DeepSeek 的定价仍显著低于海外同档位模型，其性价比优势依然存在。但价格中枢的上移，意味着国产模型的地板价阶段告一段落，行业价格战阶段性降温，未来的竞争将更多回归能力、服务与场景适配的综合比拼。

四、全球大模型市场新生态

GPT-5.6 的分层调价与 DeepSeek 的定价转向，共同推动全球大模型市场进入了清晰的分层竞争阶段，市场不再由单一的最强模型通吃，不同层级有着完全不同的竞争逻辑与定价规则。

最高一层是高失败成本的高端市场，覆盖复杂代码修复、长时知识工作、多工具联动的长链路 Agent、高风险专业决策等场景。在这些场景中，一次任务失败带来的业务损失，往往远高于不同模型之间的 Token 费用差，客户的核心诉求是任务的成功率、可靠性与稳定性，愿意为更强的能力支付显著溢价。这一层市场的参与者主要是 Claude 高阶版本、GPT-5.6 Sol、Kimi K3 等顶尖模型，它们的定价依然保持在较高水平，具备较强的定价权，竞争的核心是持续的能力领先与场景深耕，价格并非首要决策因素。

第二层是规模化自动化市场，覆盖常规文本生成、简单编码、信息提取、分类摘要、轻量 Agent、批量文档处理等高频场景。这些场景的共同特征是任务标准化程度高、调用频次高、单任务价值低，对极限推理能力的要求不高，只要模型能力达到够用的门槛，价格就会成为客户选型的核心决策变量。这一层也是当前竞争最激烈的领域，DeepSeek V4 Flash、GPT-5.6 Luna、GLM 系列、Gemini Flash 等众多产品都在这一赛道布局。DeepSeek 涨价后，该层级的价格梯度会有所收窄，但不同产品之间的工程效率差异，依然会带来数倍的单任务成本差距。

在分层市场的基础上，智能路由正在成为企业应用的默认架构。越来越多的企业不再选择单一模型作为全部场景的底座，而是搭建多模型路由体系，根据任务的复杂度、价值密度与容错要求，自动将请求调度到最合适的模型上，简单任务交给低成本模型，复杂高风险任务再升级到高端模型。这种架构的普及，直接侵蚀了闭源模型的广谱溢价，过去厂商可以对全部流量收取平均高价，现在只能对少数真正需要前沿能力的流量收取溢价，大量标准化的中低端流量会流向性价比更高的模型。

路由层的兴起也催生了聚合平台的价值，以 OpenRouter 为代表的聚合路由平台，将不同厂商的模型整合到统一接口下，提供按任务自动选模、故障容灾、多模型对比等功能，本质是将模型能力做成了可替换、可比较的标准化商品。这种模式下，模型厂商从直接触达客户的入口位置，退化为供应链中的可替换部件，平台则掌握了流量分发与定价话语权。当然跨模型路由也存在边界，在长上下文、长链路 Agent 的场景中，切换模型会损失缓存收益，反而可能推高实际成本，因此同栈路由、会话内分层会成为更优的方案，这也让同时掌握模型、缓存、算力与计费体系

的云厂商具备了更强的竞争优势。

从中美竞争的维度看，行业已经从能力追赶阶段，进入能力接近后的定价权争夺阶段。经过多年的技术迭代，中国大模型的综合能力与美国前沿模型的差距持续收窄，截至 2026 年年中，二者的能力差距已经收敛到较低水平，但价格差距依然十分显著，中国前沿模型的平均价格仅为美国前沿模型的一成左右。这种巨大的性价比优势，正在快速转化为实际的市场份额。OpenRouter 数据显示，2026 年 2 月中国模型周调用量首次超过美国模型；此后领先优势持续扩大，到 7 月 27 日-8 月 2 日，中国 AI 大模型周调用量达 28.13 万亿 Token，美国为 4.38 万亿 Token，中国达到美国的 6.4 倍，连续十四周超越美国并稳居全球首位。DeepSeek 的定价修复，并不意味着性价比优势的丧失，而是国产模型从低价渗透向质量与利润平衡升级的信号，可以预见，随着中国模型能力的进一步提升，这种优势会继续向更多中高端场景渗透，推动全球大模型的市场份额持续重构。

五、投资建议

本轮定价模式转移，标志着大模型产业从技术验证期步入商业化成熟期，行业利润池正发生结构性迁移，投资逻辑需回归商业本质，从追逐单一模型能力领先转向关注各环节盈利质量与商业可持续性。

基础模型赛道需审慎甄别盈利韧性。通用能力商品化趋势明确，中低端模型价格长期承压，依赖价格敏感型负载的厂商将持续面临毛利率下行压力。长期价值集中于两类厂商：一类在高端高失败成本场景具备不可替代性，可通过任务可靠性维持定价权；另一类拥有极致工程效率与成本控制能力，能在低价市场实现规模化盈利并构建生态壁垒。头部厂商的定价修复会阶段性改善盈利预期，但行业整体估值中枢将逐步向收入结构与利润质量靠拢。

云与 MaaS 平台是本轮价值迁移的主要承接方。模型降价带动的调用量增长将显著快于单价下滑，拥有自研模型、自有算力与完整生态的头部云厂商，可通过算力利用率提升与增值服务拓展实现收入利润双增。当前中国 MaaS 市场处于高速放量阶段，AI 收入占比高、算力调度能力领先的头部云厂商将承接更多中间层利润。

推理优化与 AI 中间件是确定性较强的受益赛道。企业 AI 支出转向精细化管理后，成本控制与投入产出比成为核心诉求，推理优化、模型路由、可观测性等中间环节需求将快速释放。这类环节不直接参与价格战，反而持续受益于企业降本需求。SemiAnalysis 数据显示，2026 年推理算力已占 AI 总算力消耗的 67%；在生产系统中，推理往往占 AI 整体基础设施预算的 60%-80%，推理优化因此具备极高的商业化确定性。垂直 AI 应用具备长期估值弹性。基础模型标准化、可替换化后，真正的壁垒在于场景理解、工作流嵌入与专有数据积累。深嵌企业核心业务流程、能量化交付价值的应用将摆脱模型价格波动影响，获得更强定价权。办公自动化、开

发工具、工业软件、信息安全等场景明确、替换成本高的赛道，商业化落地速度将持续加快。

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

钟哲元，在此声明，本人具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度，独立、客观地出具本报告。本报告清晰准确地反映了本人的研究观点。

本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿等。在本人所知情的范围内，本人所在机构、本人以及本人的利害关系人与本报告所评价或推荐的证券不存在任何利害关系。国新证券股份有限公司（已具备中国证监会批复的证券投资咨询业务资格，以下简称本公司）已在知晓范围内按照相关法律规定履行披露义务。本公司的资产管理和证券自营部门以及其他投资业务部门可能独立做出与本报告中的意见和建议不一致的投资决策。本报告仅提供给本公司客户有偿使用。

本公司不会因接收人收到本报告而视其为客户。本公司会授权相关媒体刊登研究报告，但相关媒体客户并不视为本公司客户。本报告版权归本公司所有。未获得本公司书面授权，任何人不得对本报告进行任何形式的发布、复制、传播，不得以任何形式侵害该报告版权及所有相关权利。

本报告中的信息、建议等均仅供本公司客户参考之用，不构成所述证券买卖的出价或征价。本报告并未考虑到客户的具体投资目的、财务状况以及特定需求，在任何时候均不构成对任何人的个人推荐。客户应当对本报告中的信息和意见进行独立评估，并应同时考量各自的投资目的、财务状况和特定需求，必要时可就研究报告相关问题咨询本公司的投资顾问。本公司市场研究部及其分析师认为本报告所载资料来源可靠，但本公司对这些信息的准确性和完整性均不作任何保证，也不承担任何投资者因使用本报告而产生的任何责任。本公司及其关联方可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务，敬请投资者注意可能存在的利益冲突及由此造成的对本报告客观性的影响。

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