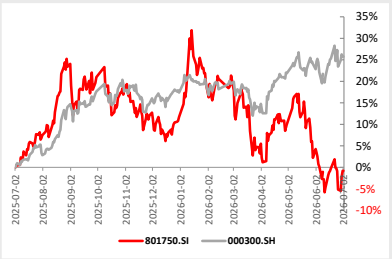
市场研究部 2026 年 7 月 2 日

中美大模型商业化进程

看好

市场表现截至 2026.7.1



数据来源：Wind，国新证券整理

事件

DeepSeek 宣布，DeepSeek V4 正式版计划于 7 月中旬正式上线，正式版发布后将同步调整 API 定价策略，引入峰谷定价机制。API 高峰时段价格将是平时价格的 2 倍，平时价格与 DeepSeek V4 API 目前定价相同。

核心观点

2026 年上半年，中美大模型价差沿能力与场景形成清晰分层，海外厂商守住高端生产场景溢价，国内厂商在普惠赛道构建全球成本优势，头部厂商同步向高价值场景突破。

海外闭源阵营维持价格刚性并抬升高端价位。Anthropic 以加量不加价巩固高端壁垒，再推旗舰新品拉升商用模型定价天花板；OpenAI 旗舰档位定价不变，通过批处理折扣、缓存机制优化实际调用成本；Google 完成产品矩阵价格梳理，形成多档位价格梯队覆盖全层级需求。国内市场分化鲜明，DeepSeek 推出永久低价与行业首个峰谷定价机制，转向精细化算力运营；智谱、月之暗面、阿里通义等多轮提价验证通过，在编码等核心生产力场景站稳高端，实现量价齐升。

全行业商业锚点正从售卖对话能力转向交付任务价值，计费模式向资源调度体系演进。美国厂商依托前沿能力深度绑定企业工作流，凭借高切换成本赚取确定性溢价；中国厂商双线并行，普惠路线靠规模化调用摊薄算力成本，国产模型全球周调用量已持续领先，高端路线深耕核心场景逐步提价，商业化重心向 B 端企业服务转移。

财务表现上，Anthropic 预计 2026 年二季度成为首家单季盈利的主流大模型厂商，ARR 实现爆发式增长；OpenAI 收入规模领先，受高强度研发投入影响暂未整体盈利。国内厂商处于 ARR 高速爬坡阶段，智谱、MiniMax、Kimi 收入与 API 占比同步提升，编码是当前行业变现的核心赛道。

投资线索

行业估值逻辑已从技术叙事转向经营叙事，建议围绕三条主线配置：优先选择商业化兑现能力已验证的头部模型厂商，规避依赖低价补贴、缺乏企业服务能力的中小厂商；布局先进封装、高带宽存储、高速光模块、液冷散热等算力产业链系统级瓶颈环节，以及国产 AI 加速器、算力运营服务商；聚焦企业 Agent、编码工具、私有化部署、端侧 AI 等具备真实落地场景与明确 ROI 的应用标的，规避泛化 AI 概念炒作。

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。

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

目录

一、中美大模型定价调整

2026 年上半年，中美大模型的价差不再是全面错位，而是沿着能力带与场景带形成清晰分层。海外厂商牢牢守住高端生产场景的溢价壁垒，国内厂商在普惠赛道形成全球成本优势，同时头部厂商开始向高价值场景发起突破。

海外闭源阵营本轮调整整体保持价格刚性，头部厂商反而通过能力分级进一步抬升了高端产品的价格天花板。Anthropic 是这一策略的典型代表。5 月底推出的 Claude Opus 4.8 维持输入 5 美元、输出 25 美元/百万 token 的定价，但上下文窗口从 200K 大幅扩展至 1M，相较 Opus 4 前代档位的 15/75 美元实际实现“降价上位”，同时新增 Dynamic Work flows 与推理力度控制功能，以加量不加价的方式巩固高端价值壁垒。仅时隔一月，Anthropic 再推出 Claude Fable 5 与 Mythos 5，将公开商用模型的定价上限拉升至输入 10 美元、输出 50 美元/百万 token。其中 Fable 5 面向公众开放，Mythos 5 仅对通过安全审核的特定合作伙伴开放。这套定价的底层支撑是前沿推理能力的稀缺性与安全审查带来的供给约束，顶级能力本身天然具备溢价基础。

OpenAI 二季度同样未下调旗舰产品定价。GPT-5.5 与 6 月进入有限预览的 GPT-5.6 旗舰档 Sol，均维持输入 5 美元、输出 30 美元/百万 token 的价位。5.6 系列新增 Terra 与 Luna 两个更低档位，但目前仅对美国本土少数合作伙伴开放。OpenAI 并未选择直接降价，而是通过批处理折扣、缓存命中机制优化部分场景的实际调用成本，其在复杂推理与长程 Agent 任务上的能力壁垒，足以支撑旗舰价位的稳定。

Google 则完成了全产品矩阵的价格梳理。Gemini 3.1 Pro 保持中高端定位，采用上下文阶梯计价：200K 以内输入 2 美元、输出 12 美元/百万 token，超过 200K 则上浮至 4/18 美元。5 月上线的 Gemini 3.5 Flash 定价为 1.5/9 美元/百万 token，个人端免费使用，由此形成从 Flash-Lite 到 3.5 Flash 再到 3.1P ro 的完整价格梯队，实现对不同层级市场需求的全覆盖。

国内市场的定价分化更为鲜明，极致性价比与高端定价两条路线同步成型，动态计费机制也首次实现本土化落地。

走极致性价比路线的厂商持续刷新高性能模型的价格下限，DeepSeek 是这条路线的核心代表。5 月下旬，DeepSeek 将 V4-Pro 预览期的 2.5 折优惠转为永久降价，缓存命中场景的输入价格低至 0.025 元/百万 token，处于全球最低水平。6 月底，公司宣布 V4 正式版将于 7 月中旬上线，并推出行业首个峰谷定价机制。V4-Pro 在工作日高峰时段价格翻倍，平时段延续优惠价，未命中输入约 3 元、输出约 6 元/百万 token，仅为海外同档旗舰模型的十分之一左右。这套机制的本质是将算力供需的时间差商品化，通过动态调度提升 GPU 利用率，标志着国内厂商已从单纯拼低价转向精细化算力运营。同期，MiniMax 将 M3 模型的限时五折改为永久优

惠，512K 上下文以内的输入、输出分别降至 2.1 元和 8.4 元/百万 token，在多模态 Agent 赛道持续强化性价比优势。

走高端定价路线的厂商则通过多轮提价验证了市场接受度，证明高品质模型在核心场景同样拥有定价权。智谱上半年连续三次调价，从 Coding Plan 到 GLM-5-Turbo 再到 GLM-5.2 升级，累计涨幅显著，当前输入 8 元、输出 28 元/百万 token 的定价约合 1.10/3.86 美元。提价后需求并未萎缩，反而出现供不应求的局面，说明国产模型在编码等关键生产力场景已站稳高端市场。月之暗面的 Kimi K2.6 较前代也有明显涨价，输入端涨幅超六成，输出端接近三成，约合 0.90/3.72 美元/百万 token，能力迭代支撑下的价格上涨并未对调用量造成明显冲击。阿里通义 Qwen3.7-Max 则在 5 月以输入 12 元、输出 36 元/百万 token 的定价锚定国产闭源旗舰档位，补齐了国内长程 Agent 场景的高端价格带。

二、大模型商业模式差异

定价体系的分化，本质是中美大模型商业模式底层逻辑的差异。当前全行业的商业锚点正从“售卖模型对话能力”全面转向“交付任务完成价值”，计费单位也从单 Token 单价向完整任务的总拥有成本延伸，这一转变进一步放大了两国厂商各自的比较优势。

从全行业视角看，定价范式的转变有三个共通趋势，背后是推理成本结构与商业价值分布的自然体现：

一是输出定价高于输入、缓存价远低于原价。输入处理对应预填充阶段，计算量大但吞吐效率高；输出生成对应解码阶段，需逐 Token 串行生成，GPU 有效计算占比低，单位成本天然更高。而输出 Token 承载着最终的推理、创作与代码生成价值，是用户付费的核心标的；缓存命中则是复用已有计算成果，边际成本极低。

二是隐藏推理 Token 成为高端模型的核心溢价载体。头部模型如 OpenAI o1/o3、DeepSeek R1 已率先将思维链、深度推理产生的隐藏 Token 计入输出计费。用户无需看到完整思考过程，但能获得更强的推理深度与更低的重试率。对金融合规、复杂代码生成等高价值任务而言，更强的模型能显著减少调试成本，单位任务的总拥有成本反而更低。

三是计费维度持续精细化。缓存命中与写入、批处理折扣、峰谷定价、上下文长度阶梯价、推理深度档位等机制逐步落地，大模型计费越来越接近云计算与电力的资源调度模式，厂商的核心竞争力也从模型能力延伸至算力资源的运营效率。

美国闭源厂商的商业模式核心，是依托前沿能力壁垒深度绑定企业级工作流，赚取高确定性溢价。其壁垒本质是工作流嵌入带来的高切换成本，而非单纯的参数领先。

Anthropic、OpenAI 均构建了从轻量到旗舰的完整产品矩阵，各档位对应不同价值任务，通过分层定价实现收益最大化。商业闭环的关键是把模型能力嵌入企业真实生产流程，而非提供独立工具。比如 Claude Code 深度融入软件开发全链路，成为工程师日常生产工具，切换成本和留存率都极高。这种模式下，客户付费的核心是“任务完成的确定性”，而非 Token 调用量本身，因此企业愿意为更高的准确率、更低的重试率、更完善的安全合规支付溢价。这也解释了海外高端模型为何能维持价格刚性甚至继续上探，其价值已脱离基础推理服务范畴，成为企业生产系统的组成部分。

中国厂商则呈现双线并行的格局，是算力约束与本土市场特征下走出的差异化路径。开源普惠路线靠规模效应摊薄成本，高端定价路线靠场景深耕提升单价。

普惠规模线以 DeepSeek、Qwen、MiniMax 为代表，依托 MoE、稀疏注意力等效率优先的技术路线将推理成本压到极致，再通过开源开放快速渗透全球开发者生态。Agent 与批量代码生成场景中，单次任务 Token 消耗是普通聊天的十数倍，成本敏感度被成倍放大，极致性价比使国产模型在成本敏感场景成为全球开发者的首选底座。OpenRouter 数据显示，中国模型周调用量自 2026 年 2 月中旬首超美国以来多次位居前列，5 月下旬 DeepSeekV4-Flash 登顶后，中国模型周调用量一度达到美国侧的 2-3 倍，正是这条路线的直接成果。其商业逻辑不是靠单 Token 高价盈利，而是靠全球规模化调用摊薄算力成本，再通过缓存、批处理、峰谷调度等运营手段提升单位算力产出。

高端定价线则以智谱、Kimi 为代表，聚焦编码、长上下文、企业 Agent 等核心生产力场景，通过定向能力迭代追上全球第一梯队，再逐步提升定价，验证了国产模型在核心场景的商业价值。当前国内厂商商业化重心正从 C 端流量快速转向 B 端企业服务，API 收入占比持续提升，本质是从流量生意转向生产力服务生意的转型。

三、大模型厂商财务表现

商业模式的差异直接投射到财务表现上，当前中美大模型行业处于截然不同的商业化验证阶段。判断大模型公司价值的核心框架，已从单一的调用量热度，转向“Token-ARR-毛利”三层验证体系。Token 是先行需求指标，ARR 是商业化成熟度指标，毛利率与现金流则是衡量商业模式可持续性的最终标尺。

海外阵营中，Anthropic 的盈利突破具有行业标志性意义。5 月融资材料披露，公司预计 2026 年第二季度营收达 109 亿美元，一季度为 48 亿美元，环比增长 127%，并实现约 5.59 亿美元营业利润，将成为全球首家单季盈利的主流大模型厂商，这一节点较内部此前预测的 2028 年扭亏提前约两年。

其盈利驱动力并非单纯的收入扩张，而是收入结构与成本效率的双重优化。收

入端，企业级 API 与服务占比 83%-85%，年支出超百万美元的企业客户突破 1000 家；成本端，推理工程优化将整体毛利率从 2025 年的 40%抬升至 70%以上，高端旗舰毛利可达 75%-80%，逐步接近成熟企业级软件水平。其 ARR 增长曲线更为陡峭：2024 年 1 月仅 0.87 亿美元，2025 年底升至 90 亿美元，2026 年 5 月达到 470 亿美元，27 个月涨幅约 5400 倍。其中 2025 年底到 2026 年 5 月的 5 个月间从 90 亿跃升至 470 亿，增长超 4 倍，且核心收入均来自长周期企业合作，切换成本远高于通用 API。

OpenAI 则保持收入规模领先，但受研发投入与业务结构影响尚未实现整体盈利。其增长逻辑是“C 端流量打底+B 端价值抬升”，规模优先的策略使其盈利节奏慢于 Anthropic。

2026 年 2 月底，OpenAI ARR 突破 250 亿美元，14 个月增长超 4 倍；2025 年营收约 131 亿美元，2026 年全年预期 290-300 亿美元，较 2025 年实现翻倍。收入结构更均衡，C 端订阅仍是基本盘，企业端占比已超 40%，计划 2026 年底 B/C 端占比持平。盈利层面，2025 年整体毛利率约 33%，较 2024 年有所下滑，主因推理算力需求激增与前沿模型高强度投入；预计 2026 年亏损约 140 亿美元，最早 2030 年现金流转正。随着企业端占比提升与推理效率优化，盈利水平将逐步改善，但大规模盈利仍需时间。

国内厂商整体处于 ARR 高速爬坡阶段，Token 爆发正逐步转化为真实收入，但绝对体量仍与海外龙头存在明显差距。对国内行业而言，ARR 高增的同时，API 占比提升、提价不伤量，才是真正的商业化拐点。

智谱是国内商业化提速的典型。2025 年营收 7.24 亿元，同比增长 131.9%；开放平台及 API 收入从 0.48 亿元增至 1.90 亿元，增幅达 292.6%。截至 2026 年 3 月，公司 MaaSAPI 业务 ARR 达 17 亿元，毛利率从低基数提升近 5 倍至 18.9%。更值得关注的是，2026 年 Q1 智谱 API 涨价 83%后，付费 Token 调用量反而增长 4 倍，印证国产模型在核心生产力场景的价值已被客户接受，行业正从“低价换量”进入“量价齐升”阶段。

MiniMax 与 Kimi 同样保持高增，且收入结构持续优化。MiniMax 2025 年营收 7903.8 万美元，同比增长 158.9%，截至 2026 年 4 月 ARR 达 3 亿美元。C 端原生产品占收入 67.2%、海外市场占 73%，但 C 端毛利率仅个位数，B 端企业服务毛利率约 70%，是利润核心，2025 年整体毛利率 25.4%。Kimi 的 ARR 在 2026 年上半年实现三级跳，3 月突破 1 亿美元、5 月突破 2 亿美元、6 月中旬突破 3 亿美元，API 收入占比超七成，完成从 C 端网红产品到 B 端技术服务商的转型。海外付费用户增长 400%、API 收入增长 400%，产品进入 200 多个国家和地区。

需要客观看待的是，ARR 并非完美衡量指标。大模型 API 的客户切换成本显著低于传统 SaaS，合同锁定性更弱，单纯的 ARR 增速不能完全代表商业化质量。从当前行业变现规律看，编码是核心胜负手。该场景付费意愿强、反馈闭环清晰，

客户本身具备高付费能力，愿意为提效工具支付溢价，这也是中美头部厂商均把 Coding 作为核心发力方向的根本原因。

四、投资建议

综合行业格局、商业模式演进与财务表现来看，当前大模型行业的投资逻辑已从“技术叙事”全面转向“经营叙事”，估值锚从模型排行榜名次切换至收入兑现能力、盈利改善空间与产业链卡位价值，建议围绕三条核心主线进行配置。

第一条主线是聚焦决赛圈的头部模型厂商，优先选择商业化兑现能力已验证的标的。经过多轮技术迭代与市场竞争，大模型行业已进入决赛圈阶段，单纯的参数规模与榜单排名不再支撑估值，真实的商业化能力成为核心定价依据。

配置上优先关注两类厂商：一类是在编码、企业 Agent 等核心生产力场景建立优势，且已验证提价能力的头部厂商，这类厂商能够在基础能力价格通缩的大环境中守住产品溢价，实现量价齐升，收入与盈利的弹性更强；另一类是具备全球化布局能力、依托开源生态掌握全球开发者心智的厂商，这类厂商能够享受开源生态扩散的长期红利，通过规模效应构建成本壁垒。需规避仅靠补贴式低价获客、缺乏企业级服务能力与算力保障的中小厂商，这类厂商在价格战与成本上行的双重压力下，盈利质量存在持续劣化风险。

第二条主线是算力产业链的系统级瓶颈环节，把握需求真实释放与供给持续偏紧的景气红利。大模型与 Agent 带来的 Token 需求增长已得到充分验证，推理算力需求持续超预期，而产业链的供给约束已从单一的 GPU 缺货，延伸至先进封装、HBM、高速互联、液冷散热、电力配套等系统级环节，供需紧平衡的状态仍将持续。

配置上优先关注三个方向：一是国产 AI 加速器及其配套测试验证产业链，受益于国产替代从可用验证转向规模上量，推理场景的渗透率正在快速提升；二是先进封装、高带宽存储、高速光模块、液冷散热等核心瓶颈环节，这类环节具备量价齐升的业绩确定性；三是算力调度、AI 基础软件与算力运营服务商，将受益于行业精细化运营的趋势，在算力供需紧平衡中体现平台价值。

第三条主线是具备真实落地场景与 ROI 支撑的应用层标的，规避泛化的 AI 概念炒作。应用层的投资逻辑已从概念验证转向业绩兑现，只有真正嵌入企业工作流、具备明确投入产出比的应用，才能获得持续的预算支持。

重点关注三大方向：一是企业 Agent 与编码工具，这是当前增量最明确、付费意愿最强的场景，垂直领域的智能体应用具备更强的客户粘性与定价能力；二是 RAG 与知识增强平台，以及私有化部署服务，国内金融、政务、能源等强监管行业对数据安全与本地化部署的需求刚性，私有化与混合部署是具备本土特色的核心赛道；三是端侧 AI 产业链，AI 手机、AIPC 带动的硬件升级与端侧模型生态正在加

速渗透，具备长期成长空间。

五、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**