

## 海外科技

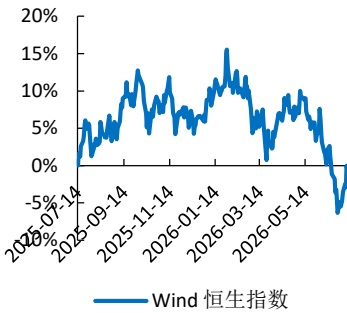
2026 年 07 月 12 日

投资评级：看好（维持）

# 云计算需求持续验证，Opus 4.7 级别模型价格带或面临松动

——行业周报

### 行业走势图



数据来源：聚源

### 相关研究报告

- 《AI 存储扩产与算力变现共振硬件链——行业周报》-2026.7.5
- 《功率半导体单柜价值量有望提升 6-8 倍，Bloom Energy 与 Brookfield 合作框架翻 5 倍——行业周报》-2026.7.5
- 《AI 需求与大厂减产驱动晶圆厂二轮涨价，高端陶瓷基板景气度上行——行业周报》-2026.6.28

初敏（分析师）

chumin@kysec.cn

证书编号：S0790522080008

荀月（分析师）

yangzhe@kysec.cn

证书编号：S0790524100001

杨哲（分析师）

xunyu@kysec.cn

证书编号：S0790524110001

### ● 互联网：市场或重新转向定价具备全栈 AI 能力的巨头

腾讯大模型战略从通用能力追赶转向高频生产力场景落地，阿里巴巴业绩预期偏向正面，具备超级 APP 入口与全链路生态协同能力的厂商有望持续扩大领先优势。首选 AI 落地领先、具备全场景生态壁垒的龙头公司，推荐阿里巴巴-W、百度集团-SW，受益标的腾讯控股。

### ● 云：Meta 计划进军云基础设施市场，阿里巴巴云业务收入增长超预期

**（1）北美：**据彭博，Meta 计划进军云基础设施市场，有望摊薄自建超大型数据中心固定成本并缩短 Capex 回收周期。对云计算竞争格局的冲击短期看 Meta 自 2024 年初以来大部分的电力容量增长是通过第三方实现的，或驱动 Coreweave、Nebius 等 RPO 增长。长期看云计算核心环节仍是 AI 芯片，随着 AI 需求从训练逐步转向推理，新增需求及估值溢价向具备 GPU/TPU/ASIC 异构算力、模型能力、积压订单可见性及资本开支投入能力的公司扩散，受益标的 Meta、谷歌、亚马逊、微软、Nebius、CoreWeave。**（2）国内：**据 36Kr 及时代周报，阿里巴巴 FY2027Q1 业绩前瞻云收入同比增速加速至 45% 左右，高于市场预期，EBITA margin 提升至低双位数，考虑到 Agent 带动推理算力需求持续增长，国内云计算增长有望加速，推荐阿里巴巴，受益标的金山云。

### ● 大模型：多款大模型同步发布，模型市场竞争或将加剧

本周 OpenAI、Grok、Meta 最新模型相继发布。根据 TrakToken 支出指数来看，全市场指数呈下行趋势，这其中闭源模型相对坚挺，开源模型在国产模型的驱动下出现阶段性下行。模型厂商仍然需要持续向 SOTA 级模型推进，当前第一梯队的模型门槛或已达到 Opus 4.7/4.8 的水平，后续模型市场竞争或将进一步加剧，模型是否能够维持与 Anthropic 的身位差将是关键观察要素。

### ● 投资建议：

**AI：**云服务需求与推理算力需求持续增长，云计算高景气延续，国产大模型流量提升带动商业化变现能力提升；AI 智能体产业进入快速增长期，受益标的谷歌、亚马逊、智谱、minimax、Nebius、Coreweave、讯策、德适-B、金山云。

**互联网：**关注 AI 商业化及应用场景拓展，随多模型能力持续提升、Agent 技术与应用生态快速成熟，token 消耗量有望延续高速增长，推理需求有望带动云计算行业摆脱低价内卷，供给端 DRAM 价格与产能紧约束，CPU、服务器等成本同步上行，共同推动云计算进入以 AI 算力溢价为核心的新阶段；国产 AI 芯片增长迅速；全球化出海有望由粗放扩张转向精细化盈利，推荐阿里巴巴-W、拼多多，百度集团-SW，受益标的腾讯控股。

**风险提示：**软件及产品推出不及预期，产能及供应链风险，监管政策变动，宏观经济增长放缓，地缘政治风险。

## 目 录

|                                         |    |
|-----------------------------------------|----|
| 1、 互联网：阿里巴巴云增长超预期，腾讯 AI 落地高频生产力场景.....  | 3  |
| 2、 Meta 或进军云，从纯 C 端流量变现向 B 端基础设施演进..... | 5  |
| 3、 大模型：多款大模型同步发布，模型市场竞争或将加剧.....        | 7  |
| 4、 周度数据更新.....                          | 11 |
| 5、 投资建议.....                            | 15 |
| 6、 风险提示.....                            | 15 |

## 图表目录

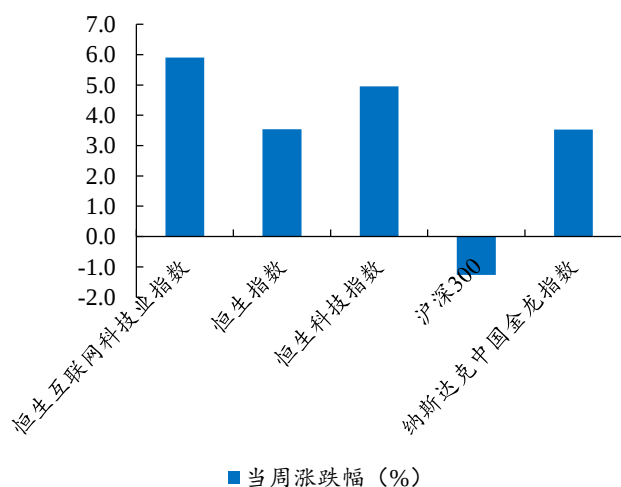
|                                                                            |    |
|----------------------------------------------------------------------------|----|
| 图 1： 20260706-20260710 恒生互联网科技指数涨 5.9%.....                                | 3  |
| 图 2： 20260706-20260710 阿里巴巴-W 涨 17.1%.....                                 | 3  |
| 图 3： 阿里云收入同比增速持续加快.....                                                    | 4  |
| 图 4： 2025 年阿里云 EBITA 利润率基本稳定.....                                          | 4  |
| 图 5： Hy3 盲测表现超越同级.....                                                     | 5  |
| 图 6： Meta 资本开支自 2024 年起维持快速增长.....                                         | 5  |
| 图 7： 2026Q1 及 Q2 Meta 新签署算力合同超过 5GW.....                                   | 6  |
| 图 8： Meta 目前的算力主要分配给 Meta Superintelligence.....                           | 6  |
| 图 9： Terminal-Bench 2.1 榜单上 GPT-5.6 Sol 优于竞品.....                          | 7  |
| 图 10： ExploitBench 榜单上，GPT-5.6 Sol 与 Claude Mythos Preview 持平，但成本明显更低..... | 8  |
| 图 11： ExploitGym 榜单上，GPT-5.6 Sol 优于前代 GPT-5.5.....                         | 8  |
| 图 12： Muse Spark 1.1 在多个榜单上与 Opus 4.8 与 GPT 5.5 相近.....                    | 9  |
| 图 13： DeepSWE 1.1 排名上，Grok 4.5 优于 GLM 5.2，或已处于 Opus 4.7/4.8 级别.....        | 10 |
| 图 14： 在 TrakToken 指数中，闭源模型价格相对坚挺.....                                      | 10 |
| 图 15： 本周恒生指数（2026.07.06-2026.07.10）在全球主要市场中跌幅靠后（单位：%）.....                 | 11 |
| 图 16： 本周（2026.07.06-2026.07.10）恒生科技跌幅较大（%）.....                            | 11 |
| 图 17： 本周（2026.07.06-2026.07.10）金融、消费类涨幅居前（%）.....                          | 11 |
| 图 18： 2026 年 6 月以来港股通当日买入成交净额有所波动（亿元）.....                                 | 12 |
| 图 19： 恒生沪港通 AH 溢价指数或已触底.....                                               | 13 |
| 图 20： OpenRouter 的 token 调用量保持高位.....                                      | 14 |
| 图 21： 国产 AI 在 ZenMux 的 Token 市场份额占比接近 50%.....                             | 15 |
| 表 1： 本周（2026.07.06-2026.07.10）港股通头部活跃个股资金流向电子等标的（仅计算前十大活跃个股明细）.....        | 12 |

## 1、互联网：阿里巴巴云增长超预期，腾讯 AI 落地高频生产力

### 场景

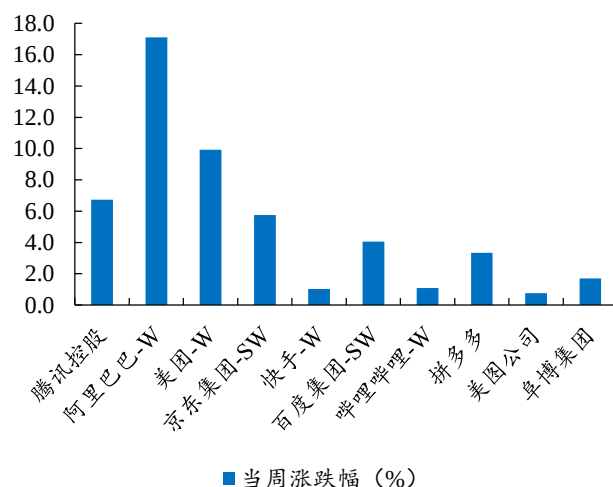
本周（2026 年 7 月 6 日-2026 年 7 月 10 日）恒生互联网科技业指数涨 5.9%，跑赢恒生指数（涨 3.5%）、恒生科技指数（涨 4.9%）、沪深 300（跌 1.3%）、纳斯达克中国金龙指数（涨 3.5%）。本周阿里巴巴-W 涨 17.1%，美团-W 涨 9.9%，腾讯控股涨 6.7%。

图1：20260706-0710 恒生互联网科技指数涨 5.9%



数据来源：Wind、开源证券研究所

图2：20260706-0710 阿里巴巴-W 涨 17.1%

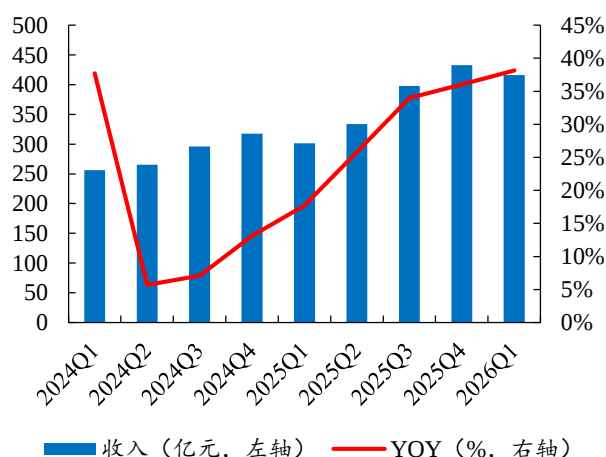


数据来源：Wind、开源证券研究所

据时代周报及 36Kr，阿里巴巴 FY2027Q1 业绩前瞻阿里云收入同比增长加速至 45%左右，EBITA margin 提升至低双位数，高于市场预期；整体电商业务（中国电商+AIDC）利润同比持平，好于市场预期；淘宝闪购减亏快于市场预期，与竞对 UE 差距收窄并有望继续改善，在补贴收窄同时市场份额保持稳定。

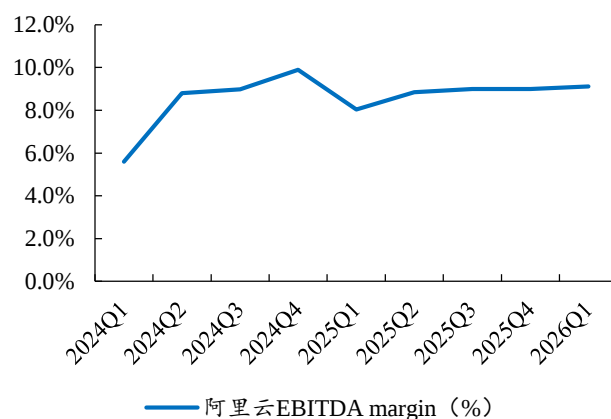
考虑到 AI 需求强劲，阿里云份额有望持续扩大，外部收入同比增速有望在未来几个季度持续加速，模型和应用 ARR 于 2026Q2 有望超过 100 亿元并在 2026 年底前达到 300 亿元，云 MaaS 占比提升有望驱动利润率持续改善。中国电商短期仍以份额稳定为优先目标，淘天闪购带动主站流量增长，订单密度、客单价提升单均亏损有望改善并降低电商买量成本，关注闪购减亏及与电商协同效率。

图3：阿里云收入同比增速持续加快



数据来源：公司公告、开源证券研究所

图4：2025年阿里云 EBITA 利润率基本稳定



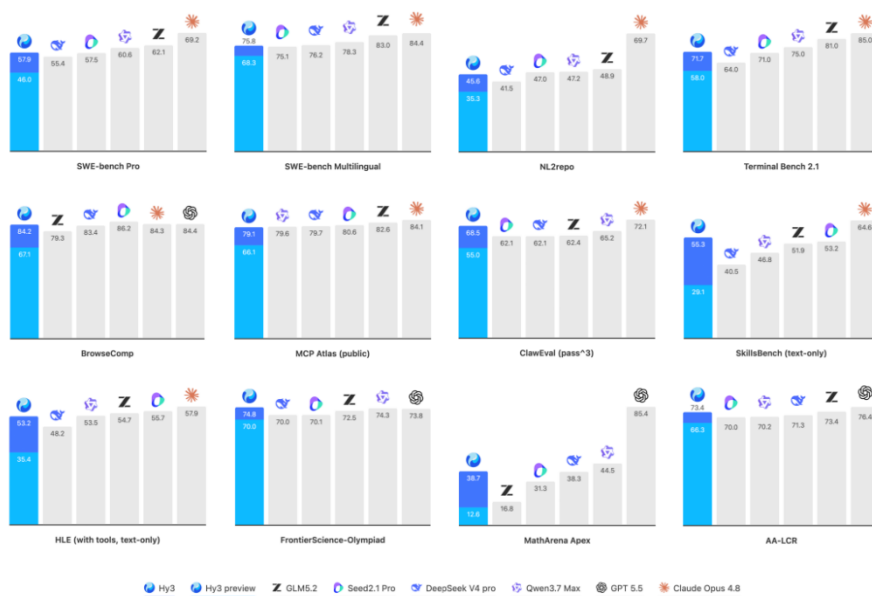
数据来源：公司公告、开源证券研究所

7月6日，腾讯正式发布混元 Hy3 大模型，并宣布其在 WorkBuddy 等核心办公 Agent 生态中的深度落地；7月10日，《金融时报》报道称腾讯正以约20亿美元估值寻求成为 Manus 最大股东，**腾讯 AI 战略从追赶通用能力转向落地高频生产力场景及 Agent 商业化变现。**

(1) Hy3 大模型：小参数跑出旗舰级能力，代码与办公能力提升明显。Hy3 是一个快慢思考融合模型，采用 MoE 架构，总参数 295B、激活参数 21B，支持 256K 上下文长度。在保持高性价比的同时，它在代码开发、办公生产、金融建模、前端设计、游戏制作等任务中进步显著。在基于 WorkBuddy 办公场景的内部测评显示，相比 preview 版本，任务解决率从 72% 跃升至 90%，平均耗时缩短 34%。

(2) Workbuddy 作为高价值生产力场景 Agent 产品，此外借助腾讯文档、腾讯会议、企业微信等腾讯体系内生态优势，用户付费意愿有望好于 C 端。

图5: Hy3 盲测表现超越同级

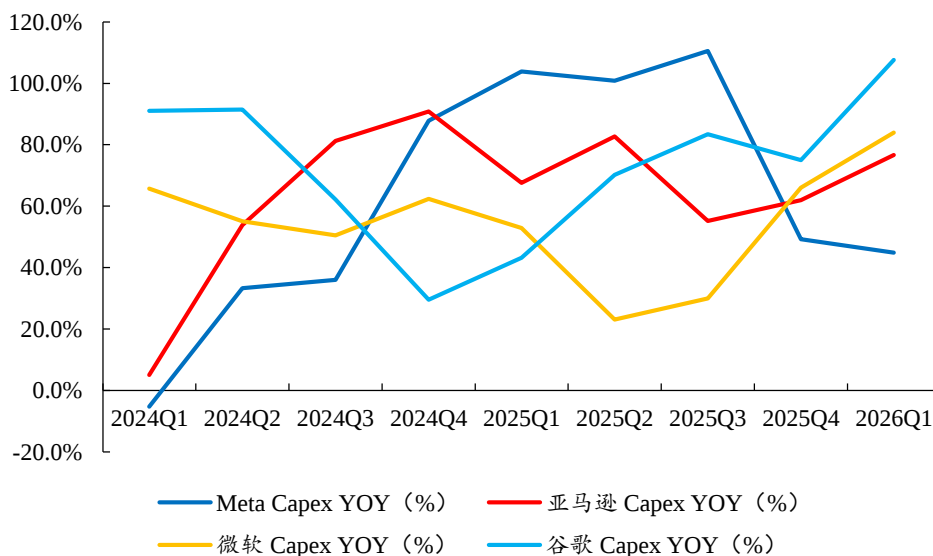


资料来源：腾讯云公众号

## 2、Meta 或进军云，从纯 C 端流量变现向 B 端基础设施演进

据彭博 7 月 1 日报道，Meta 正在计划进军云基础设施业务，向外部销售 AI 算力及模型访问权限。Meta 评估的商业模式包括：（1）向开发者和企业按量收取托管 AI 模型（如 Llama 系列）的访问费用；（2）直接出租底层的裸金属与 GPU 算力资源。随着 Meta 将 2026 年全年 Capex 上调至 1250 亿至 1450 亿美元区间，已成为北美四大 Hyperscalers 中算力储备增速最快、但此前唯一未提供公有云服务的厂商。通过出售闲置或转向算力，开辟 B 端营收曲线，Meta 不仅能摊薄自建超大型数据中心的固定成本，还能将 Capex 的回收周期缩短。

图6: Meta 资本开支自 2024 年起维持快速增长



数据来源：公司公告、彭博、开源证券研究所

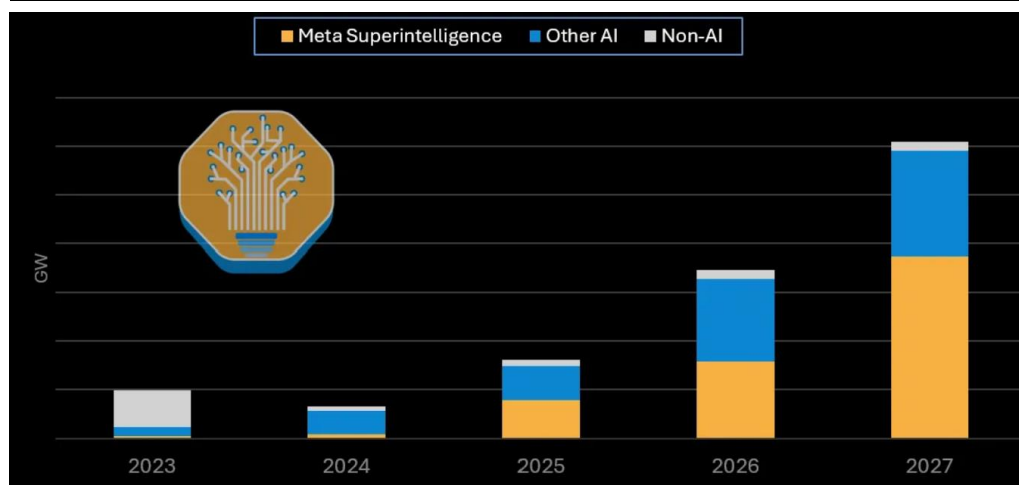
**对云计算竞争格局的冲击：**短期看，根据 Semianalysis，Meta 自 2024 年初以来签署近 10GW 的电力容量，目前大部分的容量增长是通过第三方实现的，因此 Meta 或驱动 Coreweave、Nebius 等 RPO 增长，未来需关注 Meta 内部广告推荐系统(RecSys)及前沿 AI 模型（Meta Superintelligence）对于训练和推理的算力利用效率，若广告推荐系统扩展规模或 AI 模型低于预期，其云端算力容量将迅速提升。

图7：2026Q1 及 Q2 Meta 新签算力合同超过 5GW



资料来源：Semianalysis

图8：Meta 目前的算力主要分配给 Meta Superintelligence



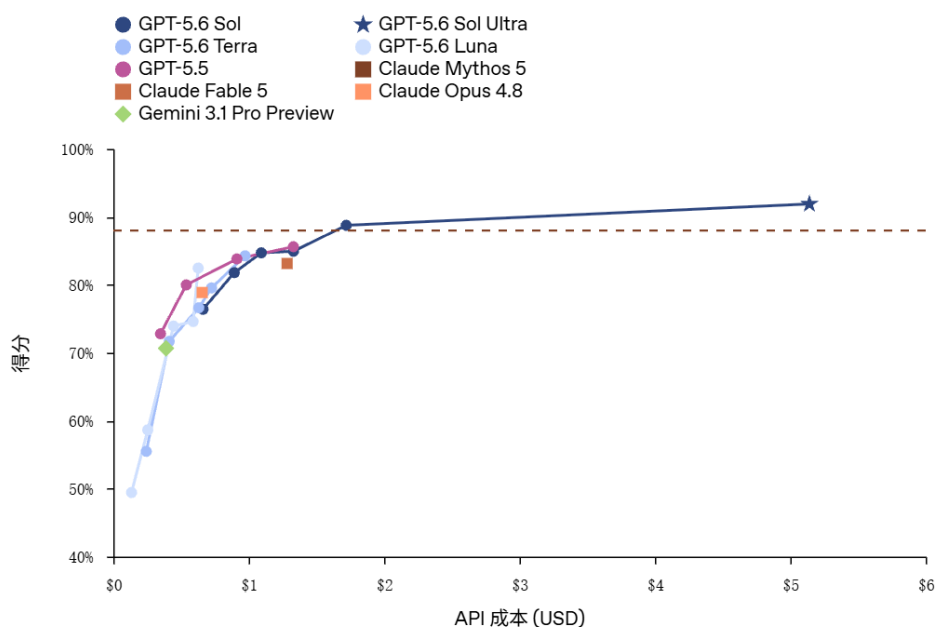
资料来源：Semianalysis

### 3、大模型：多款大模型同步发布，模型市场竞争或将加剧

**GPT-5.6 的能力跃迁：三档产品矩阵 + ultra 子代理范式。**GPT-5.6 是 OpenAI 首次将代次与能力档位解耦的三档产品矩阵，旗舰 Sol / 均衡 Terra / 极速 Luna 同步上线。Sol 是唯一解锁 max 推理努力与 ultra 子代理模式的版本；Terra 性能对标 GPT-5.5 但价格腰斩；Luna 是高吞吐、低延迟、低单价的轻任务档。Sol 是技术标杆，Terra 是利润抓手。三档 API 价目（每百万 token）：Sol \$5/\$30、Terra \$2.50/\$15、Luna \$1/\$6。

**Ultra 子代理模式：单代理范式到多代理协同的转折。**GPT-5.6 Sol 的 ultra 模式是首个官方商用的子代理并行通用范式，把复杂任务拆给多个子代理并行处理再汇总结果。在 Terminal-Bench 2.1 上：Sol 单体 88.8%，Sol Ultra 91.9%，领先所有闭源对手。ultra 比 Sol 单体高出 3.1 个百分点，是目前多代理优于单代理的对照证据。

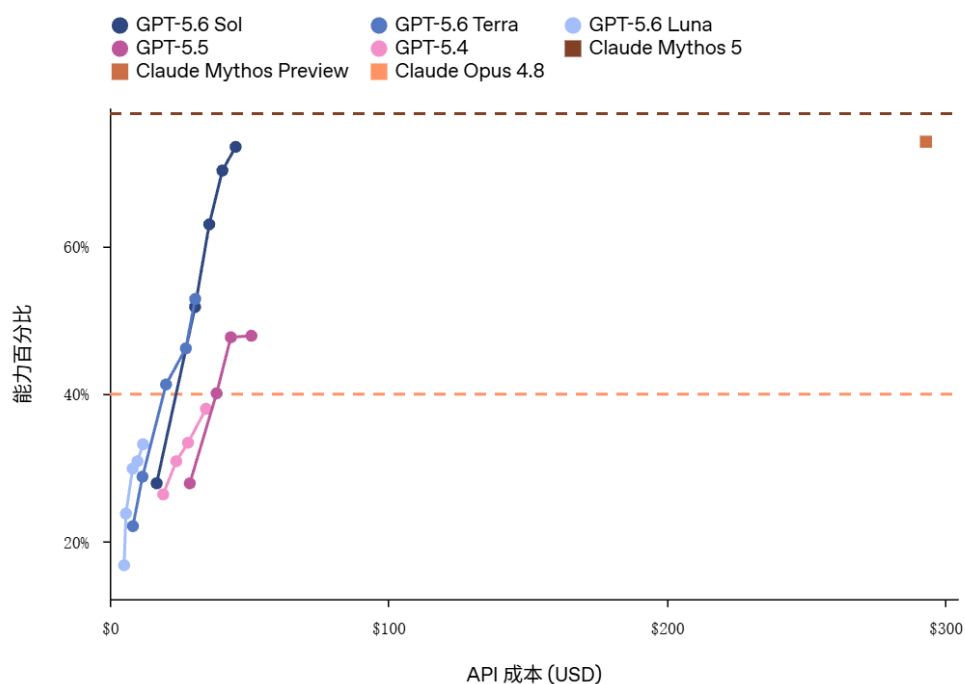
图9：Terminal-Bench 2.1 榜单上 GPT-5.6 Sol 优于竞品



资料来源：OpenAI 官网

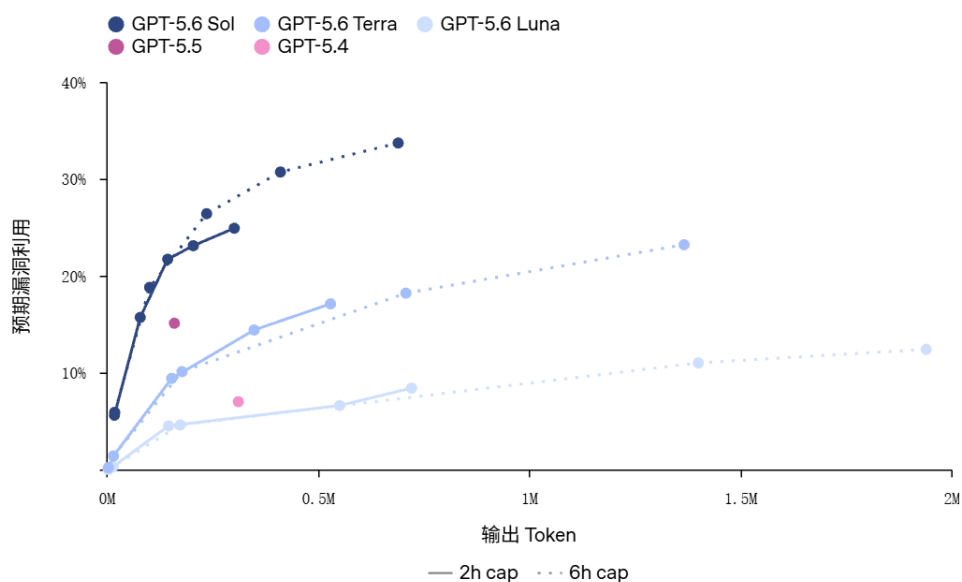
**GPT-5.6 Sol 在网络安全与生物学基准上同步跃升，是 OpenAI 迄今最强的双域模型。**安全侧在 ExploitBench 上与 Anthropic 未发布的 Mythos Preview 持平，但只用不到 1/3 的成本；ExploitGym（UC Berkeley 与前沿实验室共建）上三档模型随推理努力提升同步走强。生物侧在 GeneBench v1 长程基因组与定量生物分析上 Sol 优于 GPT-5.5 且使用更少 token。

**图10: ExploitBench 榜单上, GPT-5.6 Sol 与 Claude Mythos Preview 持平, 但成本明显更低**



资料来源: OpenAI 官网

**图11: ExploitGym 榜单上, GPT-5.6 Sol 优于前代 GPT-5.5**



资料来源: OpenAI 官网

GPT-5.6 Sol 输入价 \$5 与 Claude Opus 4.8 持平、输出价 \$30 略高于 Opus 4.8 的 \$25, 但 Terra 以 GPT-5.5 同档性能 50% 价格折扣成为本轮发布真正的商业杠杆。Terra 输入价 \$2.50、输出价 \$15, 相较 GPT-5.5 公开价腰斩, 对生产级 token 消耗方是远优于“Sol 又破纪录”的实质降本工具——若 OpenAI Terra 匹配 GPT-5.5 质量的口径属实, 这是上一代旗舰能力按 50% 折扣供货, 对 B 端批量 token 支出来说意义大于旗舰 benchmark 破纪录。

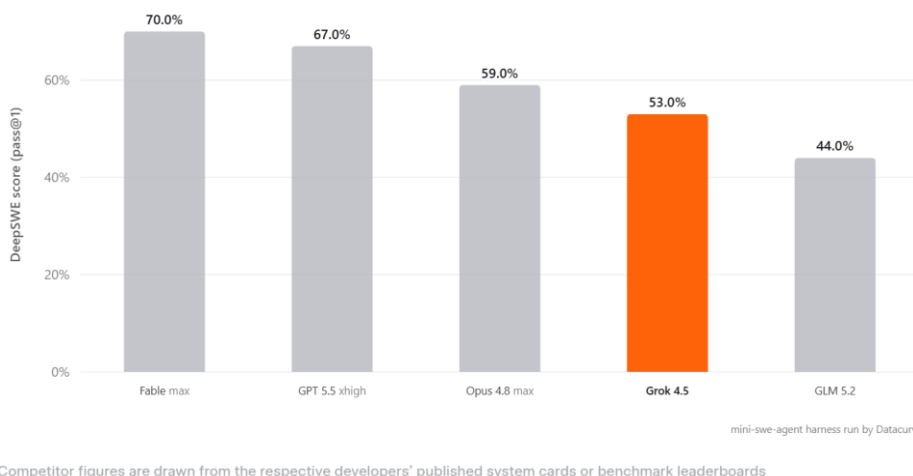
此外, 本周多家前沿厂商集中发布或放量新模型, 大模型竞争强度有所提升。Meta Superintelligence Labs (MSL) 连发 Muse Image、Muse Video 预览与 Muse Spark 1.1 三件套; xAI 发布 Grok 4.5; OpenAI GPT-5.6 全面 GA。Meta Muse Spark 1.1 是 MSL 团队重建一年后首个成熟跃迁, 从落后 DeepSeek v4 Pro 与 Kimi K2.6 转向性能-效率前沿; Muse Image 是 MSL 首个独立图像生成模型; Muse Video 是同源视频生成预览。三件套定位个人超级智能愿景, 覆盖 agentic、coding、图像生成、视频生成四大场景。

图12: Muse Spark 1.1 在多个榜单上与 Opus 4.8 与 GPT 5.5 相近

| Benchmark  |                                                                | Muse Spark 1.1<br>Meta | Muse Spark<br>Meta | Gemini 3.1 Pro (high)<br>Google | Opus 4.8 (max)<br>Anthropic | GPT 5.5 (xhigh)<br>OpenAI |
|------------|----------------------------------------------------------------|------------------------|--------------------|---------------------------------|-----------------------------|---------------------------|
| Agent      | MCP Atlas<br>Scaled tool use                                   | 88.1                   | 82.2               | 78.2                            | 82.2                        | 75.3                      |
|            | JobBench<br>Professional tool use                              | 54.7                   | 17.0               | 15.9                            | 48.4                        | 38.3                      |
|            | Toolathlon-Verified<br>Personal tool use                       | 75.6                   | 49.4               | 61.1                            | 76.2                        | 73.5                      |
|            | OSWorld-Verified<br>Agentic computer use                       | 80.8                   | 53.3               | 76.2                            | 83.4                        | 78.7                      |
|            | Humanity's Last Exam<br>Multidisciplinary reasoning (w/ tools) | 62.1                   | 50.4               | 51.4                            | 57.9                        | 52.2                      |
|            | Finance Agent v2<br>Agentic financial analysis                 | 57.2                   | -                  | 43.0                            | 53.9                        | 51.8                      |
| Coding     | Terminal-Bench 2.1<br>Agentic terminal coding                  | 80.0                   | 67.3               | 70.3                            | 82.7                        | 83.4                      |
|            | SWE-Bench Pro<br>Diverse software engineering                  | 61.5                   | 55.0               | 54.2                            | 69.2                        | 58.6                      |
|            | DeepSWE 1.1<br>Long-horizon agentic coding                     | 53.3                   | 10.0               | 12.0                            | 59.0                        | 67.0                      |
| Multimodal | CharXiv Reasoning<br>Chart QA                                  | 88.4                   | 88.9               | 81.6                            | 89.9                        | 84.8                      |
|            | BabyVision<br>Visual reasoning                                 | 76.3                   | 39.9               | 51.5                            | 81.2                        | 83.6                      |

资料来源: Meta 官网

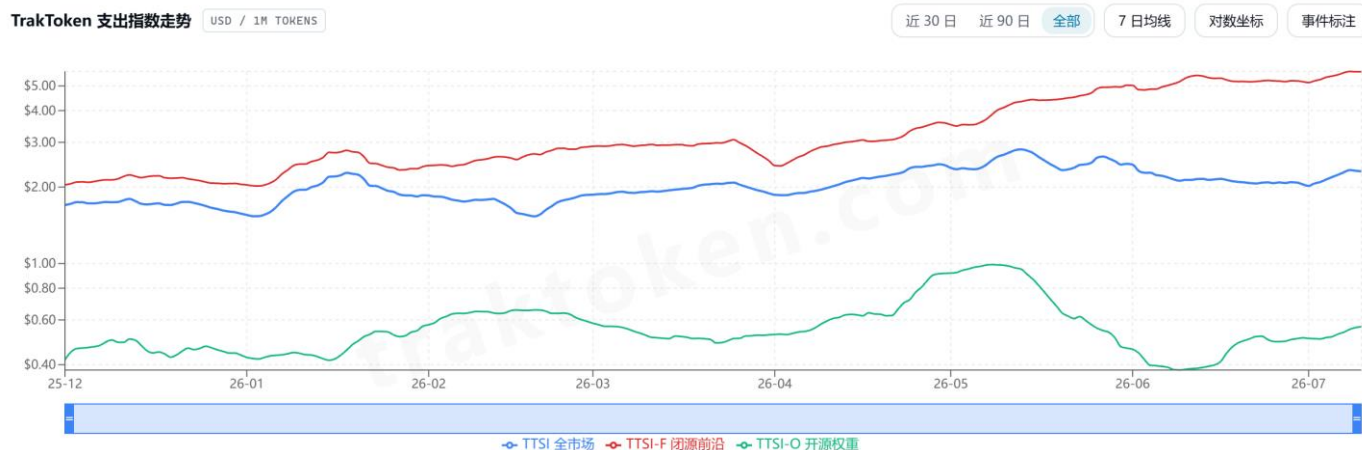
图13: DeepSWE 1.1 排名上, Grok 4.5 优于 GLM 5.2, 或已处于 Opus 4.7/4.8 级别



资料来源: xAI 官网

闭源模型定价依旧坚挺,但市场竞争或将加剧。根据 TrakToken 支出指数来看,全市场指数呈下行趋势,这其中闭源模型相对坚挺,开源模型在国产模型的驱动下出现阶段性下行。这一现象表明海外 SOTA 级模型仍然可以维持较好的溢价,同时以国产模型为代表的开源模型大幅获取中低价 token 使用场景,例如白领阶层。这仍然符合我们此前对大模型竞争格局的判断,模型厂商仍然需要持续向 SOTA 级模型推进,否则随着模型门槛逐步抬升,能力靠后的模型越容易陷入价格战的担忧。而随着 GPT-5.6、Grok 4.5 等模型发布,当前第一梯队的模型门槛或已达到 Opus 4.7/4.8 的水平,或许半年维度就会达到 Fable 5 级别的水平,后续模型市场竞争或将进一步加剧,模型是否能够维持与 Anthropic 的身位差将是关键观察要素。

图14: 在 TrakToken 指数中,闭源模型价格相对坚挺

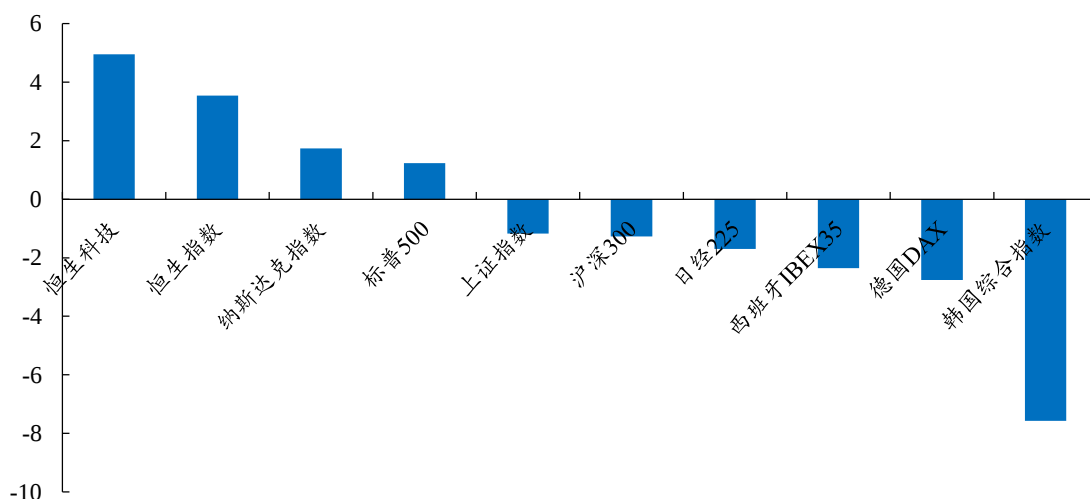


资料来源: traktoken 官网

#### 4、周度数据更新

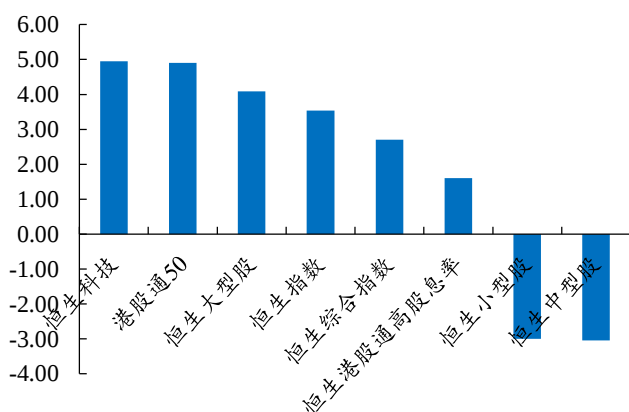
本周恒生指数上涨 3.53%，在全球主要市场中涨幅居前，从成分上看，科技、传媒、大型股类等涨幅居前，恒生科技本周上涨 4.95%。

图15：本周恒生指数（2026.07.06-2026.07.10）在全球主要市场中涨幅居前（单位：%）



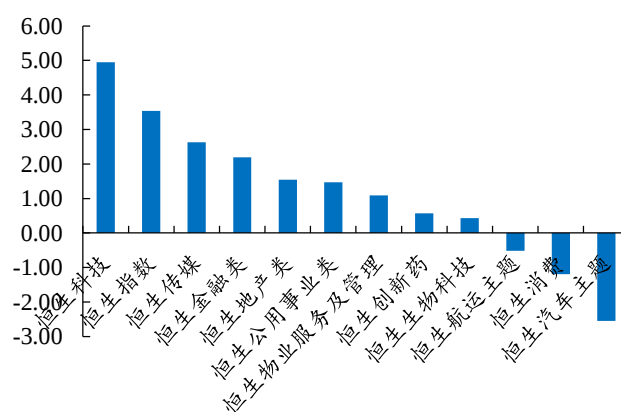
数据来源：Wind、开源证券研究所

图16：本周（2026.07.06-2026.07.10）恒生科技涨幅居前（%）



数据来源：Wind、开源证券研究所

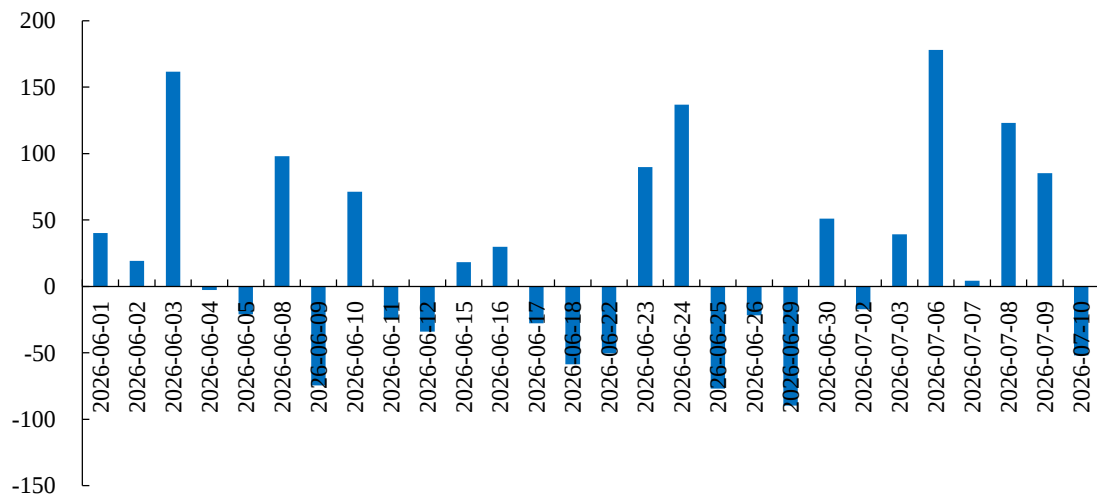
图17：本周（2026.07.06-2026.07.10）科技、传媒、金融类涨幅居前（%）



数据来源：Wind、开源证券研究所

本周港股通成交净额为流入 338.91 亿元，环比转负为正，从头部活跃个股交易层面来看，智谱（+137.0 亿港元）、阿里巴巴-W（+83.2 亿港元）、腾讯控股（+30.3 亿港元）资金净流入居前。

图18：2026 年 6 月以来港股通当日买入成交净额有所波动（亿元）



数据来源：Wind、开源证券研究所

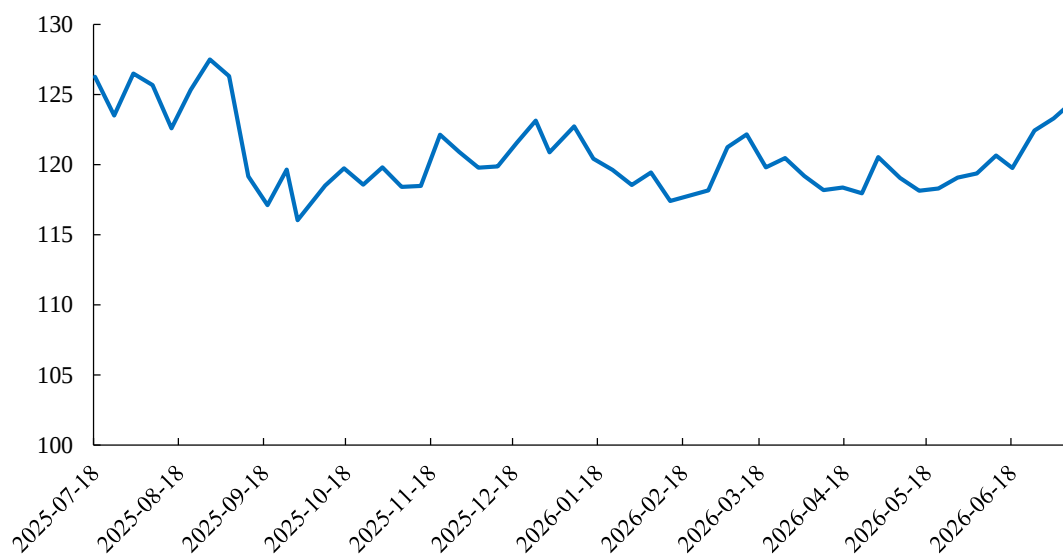
表1：本周（2026.07.06-2026.07.10）港股通头部活跃个股资金流向大模型、互联网等标的（仅计算前十大活跃个股明细）

| 股票代码    | 股票名称   | 成交净买入（亿港元） | 港股通持股数占总股本比例 |
|---------|--------|------------|--------------|
| 2513.HK | 智谱     | 137.0      | 4.6%         |
| 9988.HK | 阿里巴巴-W | 83.2       | 10.1%        |
| 0700.HK | 腾讯控股   | 30.3       | 11.8%        |
| 3690.HK | 美团-W   | 14.6       | 20.5%        |
| 3986.HK | 兆易创新   | 10.1       | 25.1%        |
| 1347.HK | 华虹宏力   | 8.2        | 22.1%        |
| 1024.HK | 快手-W   | 4.9        | 18.5%        |
| 0148.HK | 建滔集团   | -3.2       | 26.6%        |
| 1810.HK | 小米集团-W | -8.0       | 19.8%        |
| 0981.HK | 中芯国际   | -12.8      | 30.0%        |
| 1888.HK | 建滔积层板  | -30.1      | 17.2%        |
| 6869.HK | 长飞光纤光缆 | -32.4      | 48.0%        |

资料来源：Wind、开源证券研究所

本周恒生沪港通 AH 溢价指数为 124.5，环比有所提升，AH 股溢价或已触底，考虑到该指数成分中金融行业权重超 50%，能源行业权重超 15%，背后或更多反映对红利概念有需求的资金特征，随着后续更多 A 股科技类标的赴港上市，或带动港股板块的“成长性溢价”持续向 A 股靠拢。

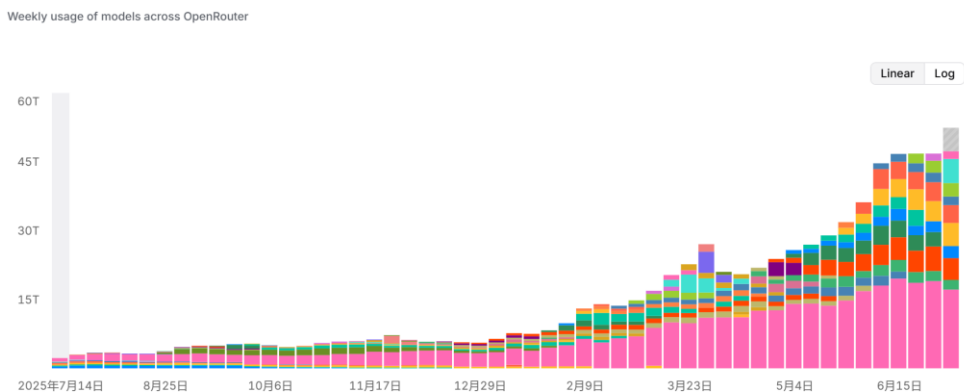
图19：恒生沪港通 AH 溢价指数或已触底



数据来源：Wind、开源证券研究所

本周（2026.07.06-2026.07.10）OpenRouter 的 token 调用量保持高位，其中国产模型 MiMo-V2.5、DeepSeek V4 Flash、Hy3（free）调用量位居前三。在 ZenMux 统计的 Token 市场份额中，国产 AI 占比接近 50%。依托开源生态及出色的性价比优势，国产模型持续扩大市场影响力。

图20: OpenRouter 的 token 调用量保持高位



#### LLM Leaderboard

Compare the most popular models on OpenRouter

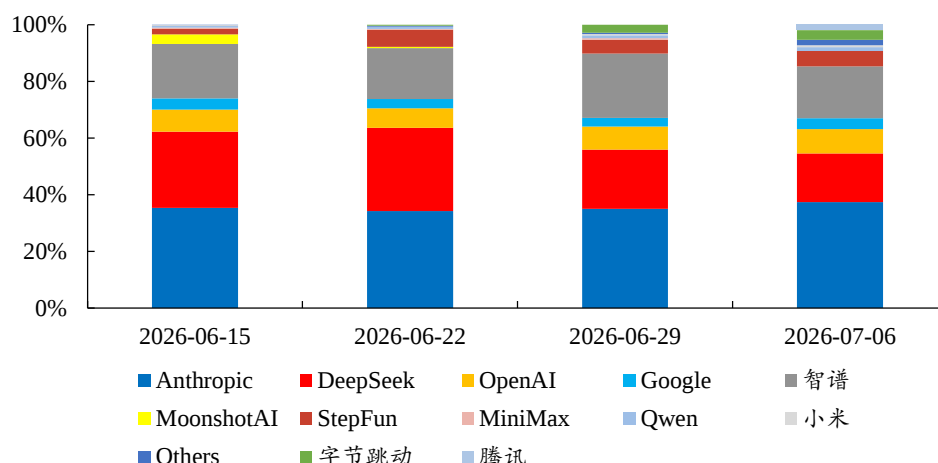
All models

This Week

|                                                                |                       |                                                                |                       |
|----------------------------------------------------------------|-----------------------|----------------------------------------------------------------|-----------------------|
| 1. <b>MiMo - V2.5</b><br>by <a href="#">xiaomi</a>             | 5.4T tokens<br>↑ 25%  | 11. <b>Claude Sonnet 4.6</b><br>by <a href="#">anthropic</a>   | 1.04T tokens<br>↓ 29% |
| 2. <b>DeepSeek V4 Flash</b><br>by <a href="#">deepseek</a>     | 5.21T tokens<br>↓ 3%  | 12. <b>Gemini 3 Flash Preview</b><br>by <a href="#">google</a> | 943B tokens<br>↑ 1%   |
| 3. <b>Hy3 (free)</b><br>by <a href="#">tencent</a>             | 5T tokens<br>new      | 13. <b>GPT-5.5</b><br>by <a href="#">openai</a>                | 940B tokens<br>↓ 20%  |
| 4. <b>MiniMax M3</b><br>by <a href="#">minimax</a>             | 4.3T tokens<br>↑ 7%   | 14. <b>Claude Sonnet 5</b><br>by <a href="#">anthropic</a>     | 909B tokens<br>↑ 100% |
| 5. <b>GLM 5.2</b><br>by <a href="#">zai</a>                    | 3.21T tokens<br>↑ 27% | 15. <b>Hy3 preview</b><br>by <a href="#">tencent</a>           | 713B tokens<br>↓ 79%  |
| 6. <b>DeepSeek V4 Pro</b><br>by <a href="#">deepseek</a>       | 2.92T tokens<br>↑ 30% | 16. <b>Gemini 2.5 Flash</b><br>by <a href="#">google</a>       | 609B tokens<br>↑ 2%   |
| 7. <b>Claude Opus 4.7</b><br>by <a href="#">anthropic</a>      | 2.41T tokens<br>↑ 19% | 17. <b>MiMo - V2.5 - Pro</b><br>by <a href="#">xiaomi</a>      | 586B tokens<br>↑ 6%   |
| 8. <b>Claude Opus 4.8</b><br>by <a href="#">anthropic</a>      | 2.01T tokens<br>↓ 4%  | 18. <b>Gemini 2.5 Flash Lite</b><br>by <a href="#">google</a>  | 562B tokens<br>↓ 8%   |
| 9. <b>Nemotron 3 Ultra (free)</b><br>by <a href="#">nvidia</a> | 1.68T tokens<br>↑ 76% | 19. <b>Laguna M.1 (free)</b><br>by <a href="#">poolside</a>    | 551B tokens<br>↓ 28%  |
| 10. <b>Step 3.7 Flash</b><br>by <a href="#">stepfun</a>        | 1.29T tokens<br>↓ 17% | 20. <b>Gemini 3.1 Flash Lite</b><br>by <a href="#">google</a>  | 505B tokens<br>↑ 17%  |

数据来源: OpenRouter 官网

图21：国产 AI 在 ZenMux 的 Token 市场份额占比接近 50%



数据来源：ZenMux、开源证券研究所

## 5、投资建议

**AI：**云服务需求与推理算力需求持续增长，云计算高景气延续，国产大模型流量提升带动商业化变现能力提升；AI 智能体产业进入快速增长期，受益标的谷歌、亚马逊、智谱、minimax、Nebius、Coreweave、讯策、德适-B、金山云。

**互联网：**关注 AI 商业化及应用场景拓展，随多模型能力持续提升、Agent 技术与应用生态快速成熟，token 消耗量有望延续高速增长，推理需求有望带动云计算行业摆脱低价内卷，供给端 DRAM 价格与产能紧约束，CPU、服务器等成本同步上行，共同推动云计算进入以 AI 算力溢价为核心的新阶段；国产 AI 芯片增长迅速；全球化出海有望由粗放扩张转向精细化盈利，推荐阿里巴巴-W、拼多多，百度集团-SW，受益标的腾讯控股。

## 6、风险提示

**软件及产品推出不及预期：**科技行业软件更新较快，市场竞争激烈，若产品推出不及时或不及预期，或影响企业未来竞争力和创收情况。

**产能及供应链风险：**若产能爬坡速度较慢，供给端受限，创收弹性无法充分释放，则相关行业产业链业绩或受到影响。

**监管政策变动：**科技政策及垂类行业相关政策的变动，或将影响公司增长景气度及战略决策，进一步影响公司的增长节奏。

**宏观经济增长放缓：**若宏观经济增长放缓影响终端消费需求、企业 IT 支出等，将影响产业链相关企业的创收能力。

**地缘政治风险：**地缘政治变动或伴随国际政策调整，或对科技行业供应链等相关企业带来限制。

## 特别声明

《证券期货投资者适当性管理办法》、《证券经营机构投资者适当性管理实施指引（试行）》已于2017年7月1日起正式实施。根据上述规定，开源证券评定此研报的风险等级为R4（中高风险），因此通过公共平台推送的研报其适用的投资者类别仅限定为境内专业投资者及风险承受能力为C4、C5的普通投资者。若您并非境内专业投资者及风险承受能力为C4、C5的普通投资者，请取消阅读，请勿收藏、接收或使用本研报中的任何信息。

因此受限于访问权限的设置，若给您造成不便，烦请见谅！感谢您给予的理解与配合。

## 分析师承诺

本研究报告的署名人员具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，以勤勉的职业态度，独立、客观地出具本报告，并对内容和观点负责。本报告清晰准确地反映了署名人员的研究观点，所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。本报告署名人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

## 股票投资评级说明

|      | 评级               | 说明                    |
|------|------------------|-----------------------|
| 证券评级 | 买入（Buy）          | 预计相对强于市场表现 20%以上；     |
|      | 增持（outperform）   | 预计相对强于市场表现 5%~20%；    |
|      | 中性（Neutral）      | 预计相对市场表现在-5%~+5%之间波动； |
|      | 减持（underperform） | 预计相对弱于市场表现 5%以下。      |
| 行业评级 | 看好（overweight）   | 预计行业超越整体市场表现；         |
|      | 中性（Neutral）      | 预计行业与整体市场表现基本持平；      |
|      | 看淡（underperform） | 预计行业弱于整体市场表现。         |

备注：评级标准为以报告日后的 6~12 个月内，证券相对于市场基准指数的涨跌幅表现，其中 A 股基准指数为沪深 300 指数、港股基准指数为恒生指数、新三板基准指数为三板成指（针对协议转让标的）或三板做市指数（针对做市转让标的）、美股基准指数为标普 500 或纳斯达克综合指数。我们在此提醒您，不同证券研究机构采用不同的评级术语及评级标准。我们采用的是相对评级体系，表示投资的相对比重建议；投资者买入或者卖出证券的决定取决于个人的实际情况，比如当前的持仓结构以及其他需要考虑的因素。投资者应阅读整篇报告，以获取比较完整的观点与信息，不应仅仅依靠投资评级来推断结论。

## 分析、估值方法的局限性说明

本报告所包含的分析基于各种假设，不同假设可能导致分析结果出现重大不同。本报告采用的各种估值方法及模型均有其局限性，估值结果不保证所涉及证券能够在该价格交易。

## 法律声明

开源证券股份有限公司是经中国证监会批准设立的证券经营机构，已具备证券投资咨询业务资格。

本报告仅供开源证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用。

本报告是基于本公司认为可靠的已公开信息，但本公司不保证该等信息的准确性或完整性。本报告所载的资料、工具、意见及推测只提供给客户作参考之用，并非作为或被视为出售或购买证券或其他金融工具的邀请或向人做出邀请。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会波动，过往的业绩表现不应作为其日后表现的预示。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。客户应当考虑到本公司可能存在可能影响本报告客观性的利益冲突，不应视本报告为做出投资决策的唯一因素。本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。本公司未确保本报告充分考虑到个别客户特殊的投资目标、财务状况或需要。本公司建议客户应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议。在任何情况下，本公司不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。若本报告的接收人非本公司的客户，应在基于本报告做出任何投资决定或就本报告要求任何解释前咨询独立投资顾问。投资者应自主作出投资决策并自行承担投资风险，任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

开源证券在法律允许的情况下可参与、投资或持有本报告涉及的证券或进行证券交易，或向本报告涉及的公司提供或争取提供包括投资银行业务在内的服务或业务支持。开源证券可能与本报告涉及的公司之间存在业务关系，并无需事先或在获得业务关系后通知客户。

本报告的版权归本公司所有。本公司对本报告保留一切权利。未经本公司事先书面授权，本报告的任何部分均不得以任何方式制作任何形式的拷贝、复印件或复制品，或再次分发给任何其他人，或以任何侵犯本公司版权的其他方式使用。所有本报告中使用的商标、服务标记及标记均为本公司的商标、服务标记及标记。

## 开源证券研究所

### 上海

地址：上海市浦东新区世纪大道1788号陆家嘴金控广场1号楼3层  
邮编：200120  
邮箱：research@kysec.cn

### 北京

地址：北京市西城区西直门外大街18号金贸大厦C2座9层  
邮编：100044  
邮箱：research@kysec.cn

### 深圳

地址：深圳市福田区金田路2030号卓越世纪中心1号楼45层  
邮编：518000  
邮箱：research@kysec.cn

### 西安

地址：西安市高新区锦业路1号都市之门B座5层  
邮编：710065  
邮箱：research@kysec.cn