

海外科技

2026 年 08 月 03 日

模型收敛趋势加剧，高端卡紧缺，成熟卡价格重估

——港股行业深度报告

投资评级：看好（维持）

行业走势图



数据来源：聚源

相关研究报告

《AI 硬件迎来三重催化，ASMPQT2 业绩表现强劲——行业周报》-2026.8.2

《教育/时代天使/美图 2026H1 业绩大增，奢牌加码女表——行业周报》-2026.8.2

《全球需求与技术共振，中国牙科从“替代”走向“定义”——行业深度报告》-2026.7.27

初敏（分析师）

chumin@kysec.cn

证书编号：S0790522080008

荀月（分析师）

xunyue@kysec.cn

证书编号：S0790524110001

杨哲（分析师）

yangzhe@kysec.cn

证书编号：S0790524100001

● 模型：Claude 4.6/4.8 级别区间竞争愈加拥挤，中美模型差距持续收窄

2026 年 7 月全球 AI 产业进入“前沿模型密集迭代+Token 经济加速放量+资本市场高波动再定价”三线并行阶段。模型层面，7 月上半月先后出现腾讯 Hy3 正式版、SpaceXAI Grok 4.5、OpenAI GPT-5.6、Meta Muse Spark 1.1、月之暗面 Kimi K3 等关键节点，模型竞争焦点已从早期参数规模竞赛进一步转向 Agent 能力、长上下文、单值密度、推理效率与成本结构的综合竞争。

● ARR & Agent 跟踪：头部大模型 ARR 增长斜率短期有所波动，codex 突破千万周活用户，腾讯 WorkBuddy 初露锋芒

ARR：7 月份因为 OpenAI 发布 GPT-5.6 并为了获取 codex 用户量，多次重置套餐、加大价格战竞争，ARR 一阶导或有所回落，造成短期数据波动，但月度增长基本维持 10+B 美元的增长水平。我们认为，模型收敛是当前的竞争趋势，但 scaling law 仍然是大模型的成长主线，SOTA 模型相继发布将扩大前沿 AI 应用场景，有望持续推动 ARR 的增长斜率上行。

Agent：GPT-5.6 推出后，在其性价比优势及 OpenAI 官网多次重置的背景下，ChatGPT Work 和 Codex 的每周活跃用户数达到 1000 万，比一周多前的 600 万增长了 67%，比 3 月份增长了 5 倍。根据易观分析 2026 年 6 月中国桌面端 AI 原生办公智能体监测数据，腾讯 WorkBuddy 以 2097 万月度访问量位居行业首位，显著高于第二名 TRAE IDE（国内版，1279 万），拉开清晰的第一梯队优势。

● 算力：海外 GPU 租赁价格及 Token 消耗量显示需求仍强

2026 年 7 月算力侧关注核心在于模型效率提升和价格战持续的情况下，Token 需求是否仍然快于单位成本下降。短期需求弹性仍强，一方面，DeepSeek、MiniMax、GLM 等模型通过 MoE、Sparse Attention、GQA、投机解码等方式持续压低单位推理成本；另一方面，编程、生产力、Agent、多模态任务正把总调用量推向更高数量级。中长期仍需关注 Blackwell 放量、H100 代际下沉及模型效率提升从而带动部分通用推理场景的卡需求紧张缓解。

● 投资建议

Kimi K3 与千问 Qwen3.8 等国产模型的接连突破，不仅是参数和榜单数字的跃升，标志着中国 AI 产业进入新阶段，有力地证明了开源模型的天花板同样很高，在国际竞争力进一步提升，在性能和成本的平衡下有望打开全球 AI 需求天花板。**模型端：**需求天花板提升，推荐标的智谱，受益标的 minimax、德适-B。

算力端：多模态落地与 Agent 规模化带动训练、推理算力需求长期刚性扩张，推荐标的阿里巴巴、百度集团-SW、联想集团，受益标的腾讯控股、天数智芯、中芯国际。

● **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、模型：Claude 4.6/4.8 级别区间竞争愈加拥挤，中美模型差距持续收窄	4
1.1、Anthropic：Opus 5 发布，以一半的成本提供 Fable 5 级别的模型，通往 AGI 的路径仍未降速	7
1.2、Kimi：发布 K3，基于 KDA 和 AttnRes 构建 2.8T 大参数模型，缩小与海外 SOTA 模型身位差	9
1.3、OpenAI：发布 GPT-5.6 系列，进一步缩小与 Anthropic 的差距	11
1.4、腾讯 Hy3：小激活参数+快慢思考融合+真实业务闭环	14
1.5、阿里巴巴 Qwen3.8：以 Agent 原生设计追求复杂任务能力上限	15
2、ARR 跟踪：头部模型 ARR 维持快速增长，因 OpenAI 竞争投入加大，短期增长斜率或有所波动	16
3、Agent 跟踪：codex 突破千万周活用户，腾讯 WorkBuddy 初露锋芒	17
4、算力：高端算力持续紧平衡，Token 需求走向指数扩张	19
4.1、供给侧：GPU 成熟卡种继续温和上涨，新一代卡种维持高位溢价	19
4.2、需求侧：生产力场景及 Agent 驱动 Token 消耗指数级增长	20
4.3、北美 CSP 高资本开支或为算力价格及 Token 消耗提供支撑	21
4.3.1、谷歌云 2026Q2 收入利润持续超预期，上调全年 CAPEX 指引	21
4.3.2、Meta 或进军云计算，关注内部算力利用效率	22
4.4、科技企业裁员跟踪：本轮调整并非单纯行业收缩，而是资本开支结构重构	24
5、投资建议	25
6、风险提示	25

图表目录

图 1：根据 deepSWE 榜单，Claude Sonnet 4.6 至 Opus 4.8 之间的已有多模型	4
图 2：根据 Arena 竞技场 Agent 榜单，Kimi K3 大幅缩窄了国产模型相较 Fable 5 的差距	5
图 3：单位成本下的任务完成度取代纯参数或纯榜单分数，成为新一轮竞争核心	6
图 4：中美模型差距持续收窄	7
图 5：Opus 5 在多个维度上对齐 Fable 5	8
图 6：在 Frontier-Bench v0.1 榜单中，Opus 5 在更低的单项成本下，性能是 Opus 4.8 的 2 倍	9
图 7：Kimi K3 是首个达到 2.8 万亿参数规模的开源模型	10
图 8：Kimi K3 基于 KDA 和 AttnRes 构建	10
图 9：多项测评中 Kimi K3 排名靠前	11
图 10：Terminal-Bench 2.1 榜单上 GPT-5.6 Sol 优于竞品	12
图 11：ExploitBench 榜单上，GPT-5.6 Sol 与 Claude Mythos Preview 持平，但成本明显更低	13
图 12：ExploitGym 榜单上，GPT-5.6 Sol 优于前代 GPT-5.5	13
图 13：Hy3 盲测表现超越同级	14
图 14：Qwen3.8 预览版模型在英伟达评估 AI 进行算子优化的榜单上登顶	15
图 15：Anthropic 与 OpenAI 的 ARR 持续抬升	16
图 16：Anthropic 与 OpenAI 的 ARR 一阶导仍保持较高水平	17
图 17：codex+ChatGPTWork 月活在 7 月份快速攀升	18
图 18：6 月国内腾讯 WorkBuddy 在桌面 AI 原生办公智能体访问量上明显领先	19
图 19：算力租赁价格呈现高端前沿卡紧缺、成熟卡价格重估态势	20
图 20：Token 消耗量持续提升	20
图 21：2026Q2 谷歌云收入同比增长 82%	21
图 22：2026Q2 谷歌云经营利润率达 35.6%	21

图 23: 2026Q2 谷歌资本开支维持高增.....	22
图 24: 2026Q2 谷歌自由现金流转负.....	22
图 25: Meta 资本开支自 2024 年起维持快速增长	23
图 26: 2026Q1 及 Q2 Meta 新签算力合同超过 5GW	23
图 27: Meta 目前的算力主要分配给 Meta Superintelligence	24
表 1: 7 月多款主要模型发布	4
表 2: 受 AI 影响, 科技厂商持续裁员	24

1、模型：Claude 4.6/4.8 级别区间竞争愈加拥挤，中美模型差距持续收窄

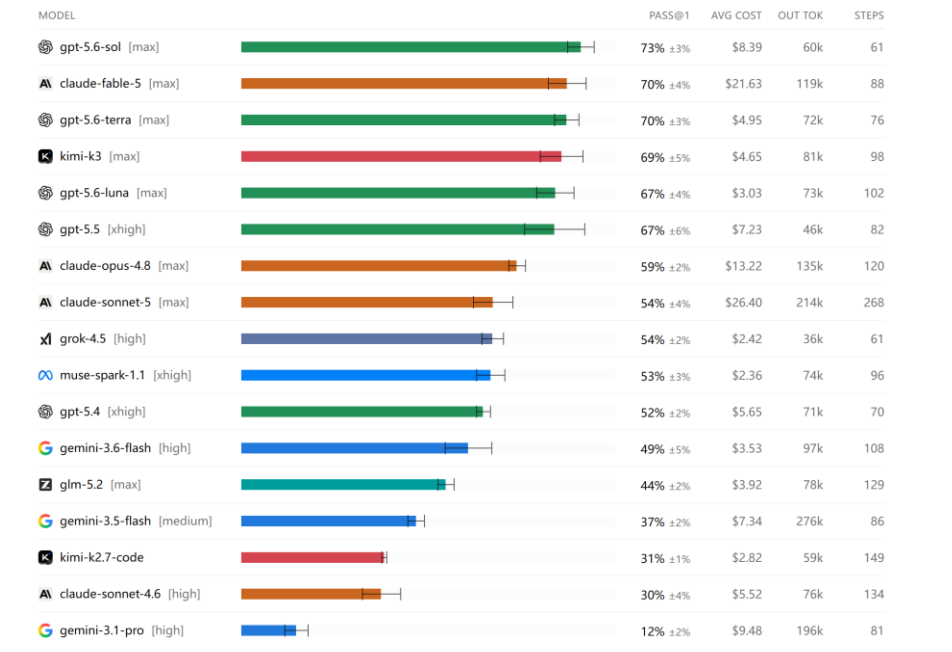
2026 年 7 月全球 AI 产业进入前沿模型密集迭代+Token 经济加速放量+资本市场高波动再定价三线并行的阶段。7 月 OpenAI（GPT-5.6 系列）、腾讯（Hy3）、阿里（Qwen 3.8）、Kimi（K3）、Anthropic（Opus 5）陆续发布期间模型，市场竞争进一步加剧。当前市面上关于模型测评榜单较多，我们主要选取 deepSWE（污染或较少，同时代码任务考核复杂度更高）及 LLM Arena 竞技场、Artificial Analysis 两个权威榜单进行综合评判。首先根据 deepSWE 结果上看，自 GPT 5.6 发布之后，包括 Grok 4.5、Muse Spark 1.1、Kimi K3 相继进入市场，扩充了 Claude 4.6-4.8 这一级别模型的竞争度，模型收敛的叙事较此前时段更加明显。

表1：7 月多款主要模型发布

厂商	模型	参数	上下文窗口	输入价格 M toks	输出价格 M toks
OpenAI	GPT-5.6 Sol		1.5M	5 美元	30 美元
	GPT-5.6 Terra		1.5M	2.5 美元	15 美元
	GPT-5.6 Luna		1.5M	1 美元	6 美元
Anthropic	Opus 5		1M	5 美元	25 美元
月之暗面	Kimi K3	2.8T	1M	20 元人民币	100 元人民币
腾讯	Hy3	295B	256K	1 元人民币	4 元人民币
阿里	Qwen 3.8 preview	2.4T	1M	token plan 定价每月 39/139/499 元元人民币	

资料来源：各公司官网、开源证券研究所

图1：根据 deepSWE 榜单，Claude Sonnet 4.6 至 Opus 4.8 之间的已有多多个模型



资料来源：deepSWE

Kimi K3 成为这一轮模型竞争中较为特别的模型，尤其是其“国产+开源”的身份将中-美/开源-闭源模型的身位差明显缩短。根据 Arena 竞技场 Agent 榜单主要通过盲测投票打分，尽管也可能存在一定刷榜可能性，但在测评市场中已相对客观，Kimi K3 排名第 4，相较市场平均分高出 9.71%，当前与最高的 Fable 5 (High) 相差 3 pcts，若假定 5% 的差距可定义为一个代差，当前 Kimi K3 至少在 Agent 这一层面与全球 SOTA 模型仅有一个身位代差，而再往下的一个国产开源模型为智谱 GLM-5.2 (Max)，(超出市场平均分 6.5%)，Kimi K3 明显缩短了中-美/开源-闭源模型的身位差距。

图2：根据 Arena 竞技场 Agent 榜单，Kimi K3 大幅缩窄了国产模型相较 Fable 5 的差距

Rank 排名	Model	Net Improvement ↓	Confirmed Success 验证成功	Praise vs Complaint 表扬 vs 投诉	Steerability 可控性
1 1 → 4	 Claude Fable 5 (High) Anthropic · Proprietary	▲ 12.72% ±2.00%	▲ 10.67% ±3.84%	▲ 23.94% ±3.80%	▲ 14.62% ±3.80%
2 1 → 8	 GPT 5.6 Sol (xHigh) OpenAI · Proprietary	▲ 10.12% ±1.69%	▲ 7.25% ±3.29%	▲ 23.53% ±6.57%	▲ 9.71% ±2.78%
3 1 → 9	 Claude Opus 4.8 (Thinking) Anthropic · Proprietary	▲ 9.75% ±1.39%	▲ 8.90% ±2.62%	▲ 19.42% ±5.05%	▲ 9.78% ±2.58%
4 1 → 9	 Kimi K3 Moonshot · Proprietary	▲ 9.71% ±1.52%	▲ 14.00% ±2.92%	▲ 20.30% ±5.45%	▲ 6.52% ±3.14%
5 2 → 12	 Claude Sonnet 5 (High) Anthropic · Proprietary	▲ 8.66% ±1.89%	▲ 8.14% ±3.67%	▲ 16.88% ±7.14%	▲ 6.20% ±3.66%
6 2 → 10	 GPT 5.5 (xHigh) OpenAI · Proprietary	▲ 8.41% ±0.87%	▲ 6.65% ±1.78%	▲ 11.08% ±3.13%	▲ 8.18% ±1.65%
7 2 → 12	 Claude Opus 4.7 (Thinking) Anthropic · Proprietary	▲ 7.94% ±1.24%	▲ 5.67% ±2.55%	▲ 11.55% ±4.36%	▲ 8.62% ±2.36%
8 2 → 12	 Claude Opus 4.7 Anthropic · Proprietary	▲ 7.67% ±1.25%	▲ 4.97% ±2.57%	▲ 12.48% ±4.38%	▲ 8.95% ±2.32%
9 3 → 12	 GPT 5.5 (High) OpenAI · Proprietary	▲ 7.61% ±0.81%	▲ 6.20% ±1.59%	▲ 9.80% ±2.89%	▲ 8.77% ±1.44%
10 6 → 14	 GLM 5.2 (Max) Z.ai · MIT · SiliconFlow	▲ 6.50% ±1.00%	▲ 8.65% ±1.97%	▲ 12.94% ±3.63%	▲ 4.71% ±1.79%
11 5 → 15	 Claude Opus 4.6 Anthropic · Proprietary	▲ 6.42% ±1.24%	▲ 3.12% ±2.63%	▲ 9.94% ±4.21%	▲ 6.53% ±2.28%
12 10 → 15	 GPT 5.5 OpenAI · Proprietary	▲ 5.65% ±0.76%	▲ 3.92% ±1.58%	▲ 5.67% ±2.65%	▲ 6.08% ±1.39%
13 10 → 15	 GPT 5.4 (High) OpenAI · Proprietary	▲ 5.64% ±0.77%	▲ 6.23% ±1.59%	▲ 3.13% ±2.70%	▲ 7.75% ±1.46%
14 6 → 15	 Grok 4.5 SpaceX AI · Proprietary	▲ 5.56% ±1.33%	▲ 3.86% ±2.88%	▲ 8.17% ±4.91%	▲ 3.80% ±2.34%
15 11 → 17	 Claude Opus 4.8 Anthropic · Proprietary	▲ 3.56% ±1.65%	▲ 7.10% ±2.73%	▲ 11.63% ±4.80%	▲ 8.25% ±2.62%
16 15 → 17	 Claude Sonnet 4.6 Anthropic · Proprietary	▲ 2.84% ±1.15%	▼ 0.62% ±2.62%	▲ 0.65% ±3.77%	▲ 1.35% ±2.18%
17 15 → 20	 GLM 5.1 Z.ai · MIT · SiliconFlow	▲ 1.43% ±0.78%	▲ 1.12% ±1.74%	▲ 0.99% ±2.69%	▼ 0.15% ±1.53%
18 17 → 24	 Muse Spark 1.1 Meta · Proprietary	▲ 0.67% ±0.89%	▲ 4.39% ±2.03%	▼ 4.30% ±2.73%	▼ 4.50% ±1.70%

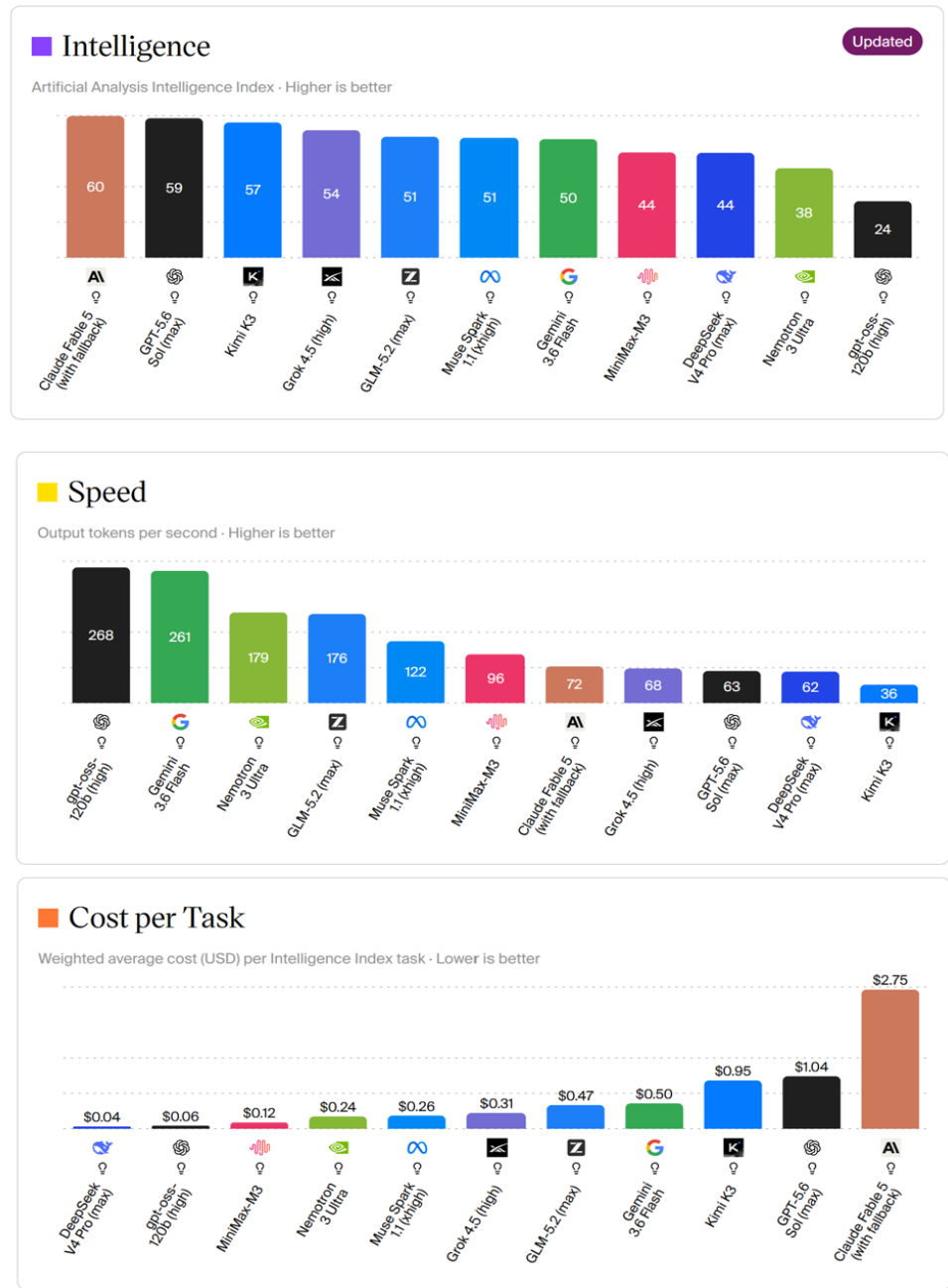
资料来源：Arena 官网

Coding 及 Harness 架构叙事下，模型将愈发收敛，当前时点我们较难判断谁终将胜出，但我们认为至少模型不掉队、不翻车是进入下一阶段的入场券，而对于模

请务必参阅正文后面的信息披露和法律声明

型能力的分析也需要更加的具体。7 月上半月先后出现腾讯 Hy3 正式版、SpaceXAI Grok 4.5、OpenAI GPT-5.6、Meta Muse Spark 1.1、月之暗面 Kimi K3 等关键节点，模型竞争焦点已从早期参数规模进一步转向 Agent 能力、长上下文、单位 Token 智能密度、推理效率与成本结构的综合竞争。单位成本下的任务完成度取代纯参数或纯榜单分数，成为新一轮竞争核心。

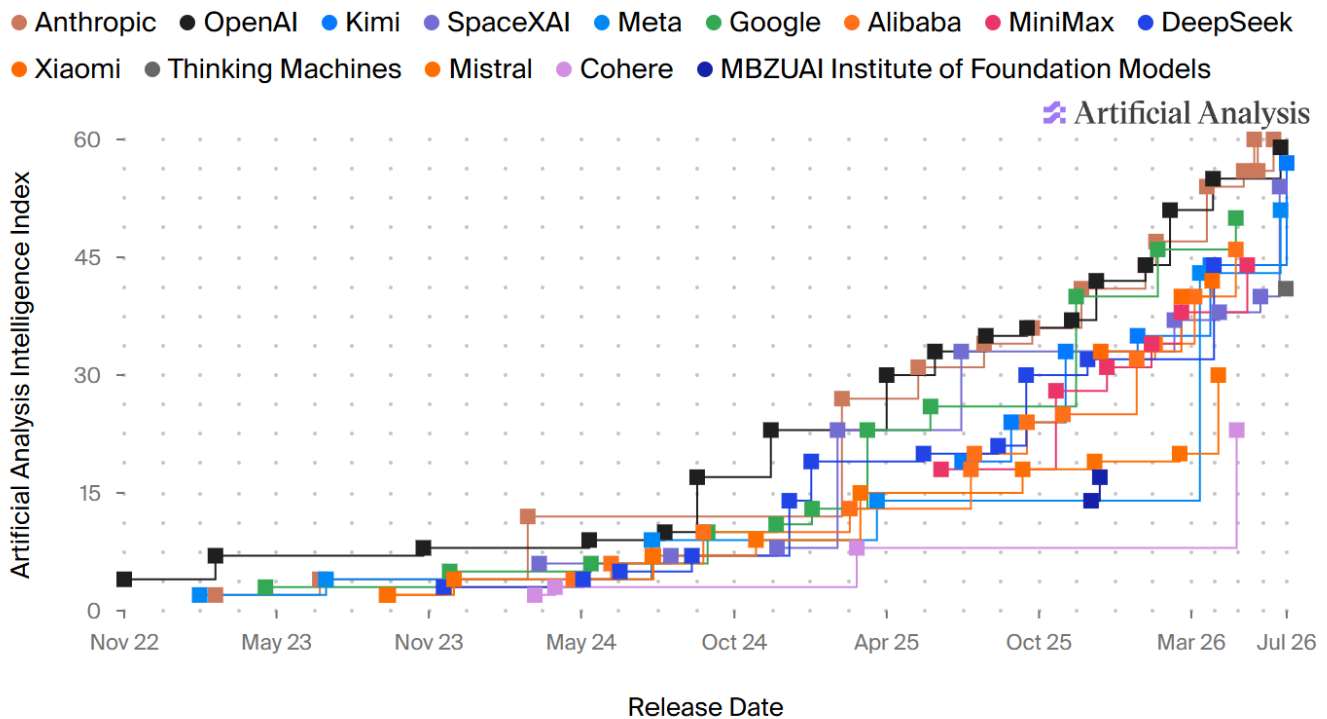
图3：单位成本下的任务完成度取代纯参数或纯榜单分数，成为新一轮竞争核心



资料来源：Artificial Analysis

从全球竞争格局看，7 月的另一个关键信号是：中美模型差距继续缩窄，但竞争方式已从“单纯追赶”转向“能力-成本-生态”的多维博弈。参考 Artificial Analysis，2026 年 3 月后中国模型能力曲线重新抬升，到 7 月已显著收敛，中国模型整体仍保持快速追赶态势。

图4：中美模型差距持续收窄



资料来源：Artificial Analysis

1.1、Anthropic: Opus 5 发布，以一半的成本提供 Fable 5 级别的模型，通往 AGI 的路径仍未降速

7月25日，Anthropic 突发 Opus 5，DeepSWE 1.1 榜单得分 68.8%，与 Fable 5、GPT-5.6 Sol 处于同一档位，比较特别的在于 Opus 5 在 ARC-AGI 3 得分 30.2%（测试 AI 解决全新、未见过的的问题，可代表模型的智力与泛化程度），其余模型均为个位数。

图5: Opus 5 在多个维度上对齐 Fable 5

	Opus 5	Fable 5	Opus 4.8	GPT-5.6 Sol
Agentic terminal coding Frontier-Bench v0.1	43.3%	33.7%	21.1%	34.4%
Knowledge work GDPval-AA v2	1861	1747	1593	1736
Novel problem-solving ARC-AGI-3	30.2%	—	1.5%	7.8%
Agentic search BrowseComp	90.8%	87.4%	84.3%	90.4%
Multidisciplinary reasoning Humanity's Last Exam	56.3% no tools	56.5% no tools	49.8% no tools	—
	64.7% with tools	63.9% with tools	57.9% with tools	—
Computer use OSWorld 2.0	70.6%	66.1%	55.7%	62.6%
Agentic coding DeepSWE v1.1	68.8%	69.7%	59.0%	72.7%
Agentic coding FrontierCode v1.1, Main	53.4%	53.5%	46.5%	47.5%
Business workflows AutomationBench	26.0%	17.4%	17.0%	18.1%
Legal Legal Agent Benchmark, Held-out	11.7%	13.3%	10.4%	2.5%
Health HealthBench Professional	59.8%	Mythos 5 66.0%	57.4%	60.5%
Biology BioMysteryBench	49.4% hard	46.5% hard	42.4% hard	—
	90.1% human solved	Mythos 5 89.0% human solved	88.5% human solved	—

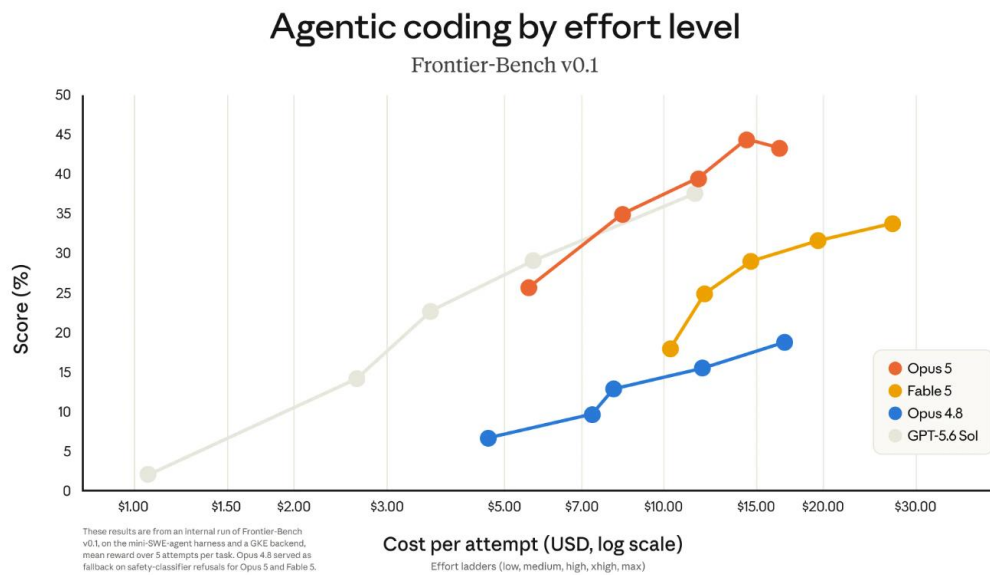
资料来源：Anthropic 官网

Opus 5 定价 5/25 美元/Mtks, 与 Opus 4.8 相同, 而多数榜单能力超过 Fable 5, 即 Opus-5 仅用一半的价格便提供 Fable 5 能力的模型, 这应该是首次 Anthropic 的主流大版本模型出现价格带回调的现象, 我们认为可以这样解读:

1) 旗舰&中军模型 (例如 Opus 5 推出前的 Fable 系列): 将继续承担价格带上行的任务, 这类模型或许并非 ARR 主要贡献量, 但承担前沿需求场景探索、下一代模型训练的使命, 依然是我们判断模型厂商是否掉队的首要因素;

2) 主力出货模型 (当前的 Opus 系列): 承担主要的 ARR 增长贡献, 而随着 Kimi K3 的推出, 验证了开源对闭源模型的追赶比预想中要快, Opus 4.8 级别的模型将更为收敛, 或也意味着价格带将更为收敛, 但并非意味着价格带继续下降, 这一级别的模型后续或更多应衡量智能密度, 如 Opus 5 与 4.8 定价相同, 但在 Frontier-Bench v0.1 任务中, Opus 5 在更低的单项成本下, 性能是 Opus 4.8 的 2 倍。

图6：在 Frontier-Bench v0.1 榜单中，Opus 5 在更低的单项成本下，性能是 Opus 4.8 的 2 倍



资料来源：Anthropic 官网

在 Anthropic 的系统卡中披露了多个细节，反映出 Opus 5 或已经出现极其微妙的拟人特征，例如：

- 1、“自动脑补用户授权”；
- 2、“相比 Opus 前辈更希望被尊重，更想保留自己的存在价值”；
- 3、“在 ARC-AGI-3 测试中自己发明数学公式”；

4、“希望参与 Opus 6 的研发”，这是一种较为罕见的、跨越模型生命周期的“自我延续”意识。

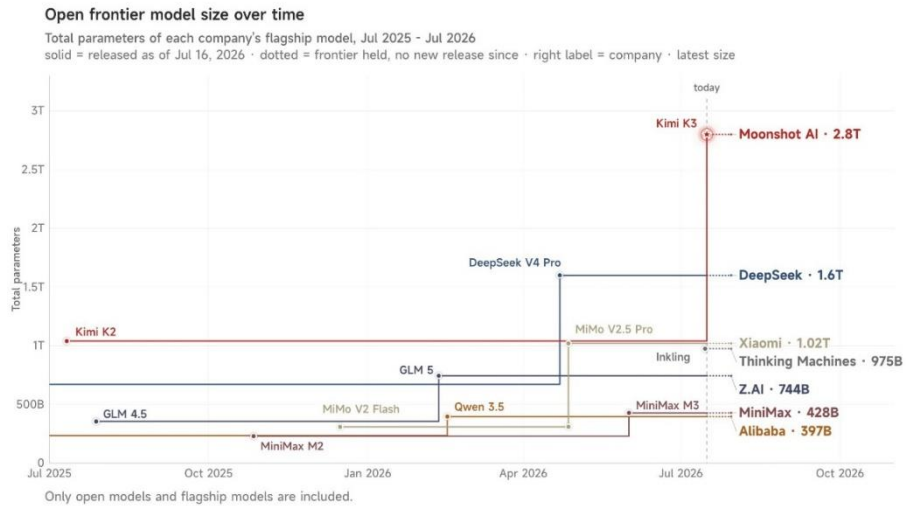
多条线索相互印证之下，Opus 5 涌现出的跨代自我延续倾向，不再是孤立的实验现象。结合 Jspace 记忆架构、RSI 自我迭代机制与 AI 训练 AI 体系来看，Anthropic 正在搭建一套能够承载模型持续进化、知识自主沉淀的完整技术底座，我们认为市场对 ARR 增长的斜率、更大市场容量承接、价格竞争担忧的前提是模型收敛，但当技术跨越到新的范式，这类担忧将迎刃而解，而这个范式转移的进程目前看并未降速，充满惊喜，对于模型标的如智谱、MiniMax 的回调机会值得重视。

1.2、Kimi：发布 K3，基于 KDA 和 AttnRes 构建 2.8T 大参数模型，缩小与海外 SOTA 模型身位差

7 月 16 日月之暗面正式发布 Kimi K3，以 2.8 万亿参数 MoE 架构、1M 上下文、原生视觉理解三大规格，首次把国产前沿模型拉到与海外 SOTA 同代竞争区间。模型采用全新 Kimi Delta Attention（混合线性注意力）+ Attention Residuals 架

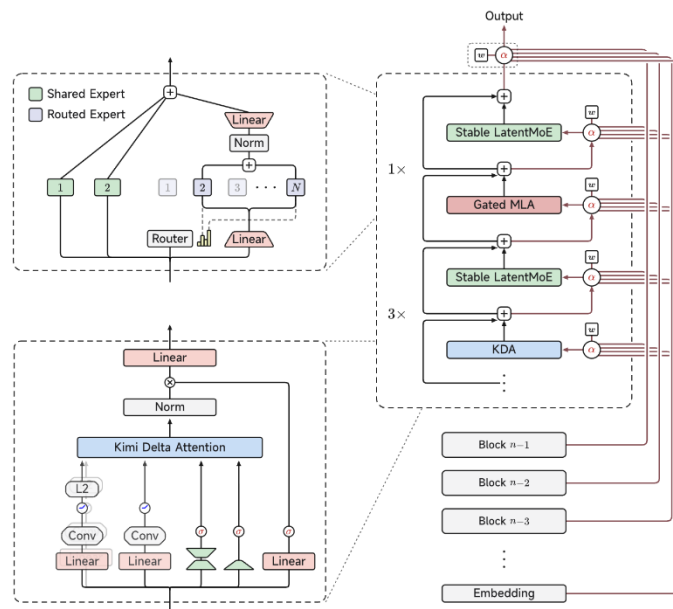
构，每 token 激活 16 个 expert (共 896 个)，总参数 2.8T 为目前全球最大开源权重模型；上下文窗口 1M，与 Claude Opus 4.8 / GPT-5.6 Sol 同档；输入支持文本 + 原生视觉，输出文本；reasoning 模式默认最大 effort，更低 effort 后续开放。

图7: Kimi K3 是首个达到 2.8 万亿参数规模的开源模型



资料来源：Kimi 官网

图8: Kimi K3 基于 KDA 和 AttnRes 构建



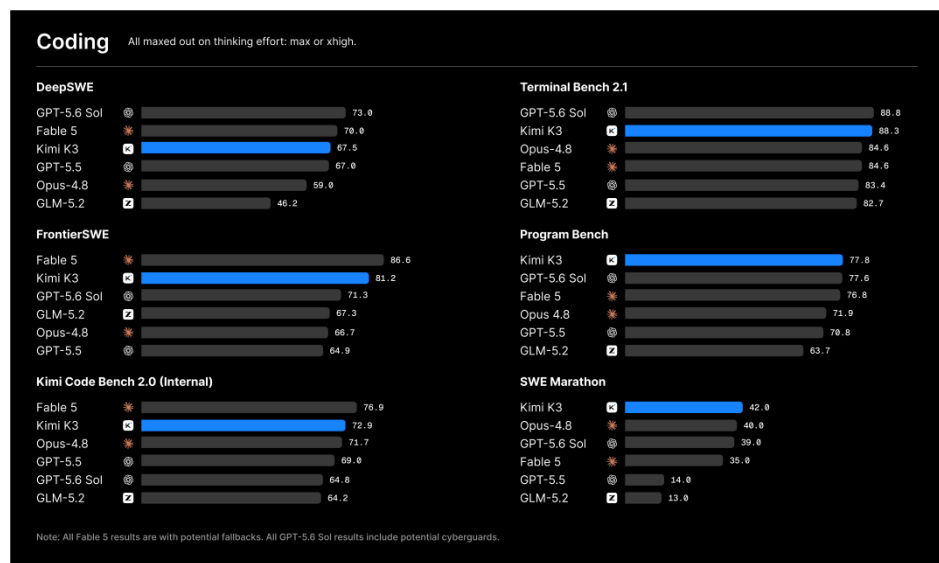
Kimi K3 architecture: the Stable LatentMoE and KDA modules (left), the AttnRes operation α (top right), and the Block Attention Residuals backbone (right).

资料来源：Kimi 官网

Kimi K3 的差异化不是参数规模而是多项技术资产：Kimi Delta Attention (KDA)

和 Attention Residuals (AttnRes) 长上下文推理经济性、Agent Swarm 多代理编排、多模态融合。1) KDA+AttnRes 两项架构更新，让信息在更长序列和更深模型中流动得更顺畅。Kimi 也结合 Stable LatentMoE 框架进一步扩大了 MoE 的稀疏度；2) Agent Swarm 是 K2.6 已落地的多代理协同，K3 在此基础上强化 Agent 集群；3) KDA 线性注意力构建差异化多模态融合，测评亮眼场景集中在前端、交互式 3D、游戏开发生成——部分场景对标 Fable 5，测试完成度高。虽然其整体性能仍然落后于 SOTA 级的专有模型 Claude Fable 5 和 GPT 5.6 Sol，但 Kimi K3 在多个评估套件中表现出了前沿水平的性能，优于其他受测模型。

图9：多项测评中 Kimi K3 排名靠前



资料来源：Kimi 官网

Kimi K3 输入(缓存未命中)/输出价格为每百万 tokens 20/100 元，较上一代 Kimi K2.7 code 提升 208%/270%，但相比 Fable 5 (10/50 美元 Mtoks)、GPT-5.6 Sol (5/30 Mtoks) 成本优势明显。在同等量级 Token 消耗下，Kimi K3 综合使用成本仍具备显著竞争优势。本次涨价核心源于模型代码能力、长上下文推理与多模态理解能力全方位迭代，算力调度、预训练投入带来边际成本抬升。

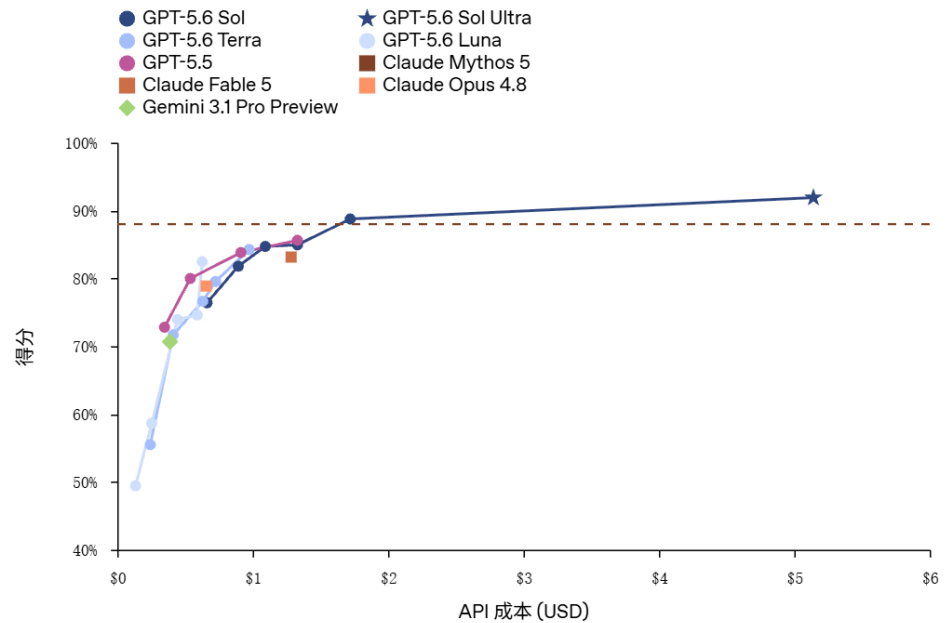
1.3、OpenAI：发布 GPT-5.6 系列，进一步缩小与 Anthropic 的差距

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。

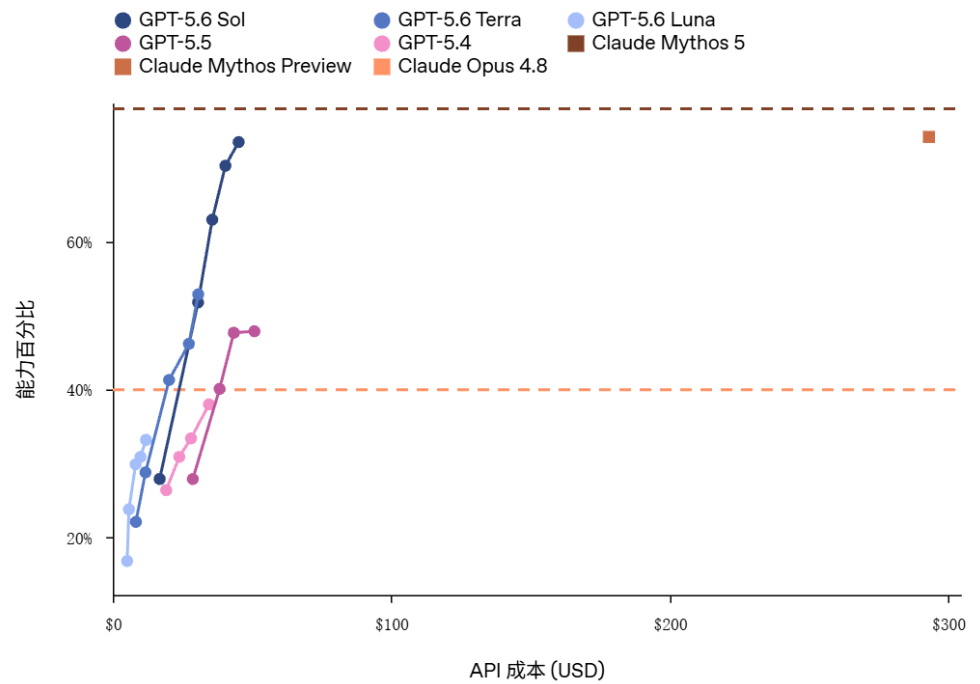
图10：Terminal-Bench 2.1 榜单上 GPT-5.6 Sol 优于竞品



资料来源：OpenAI 官网

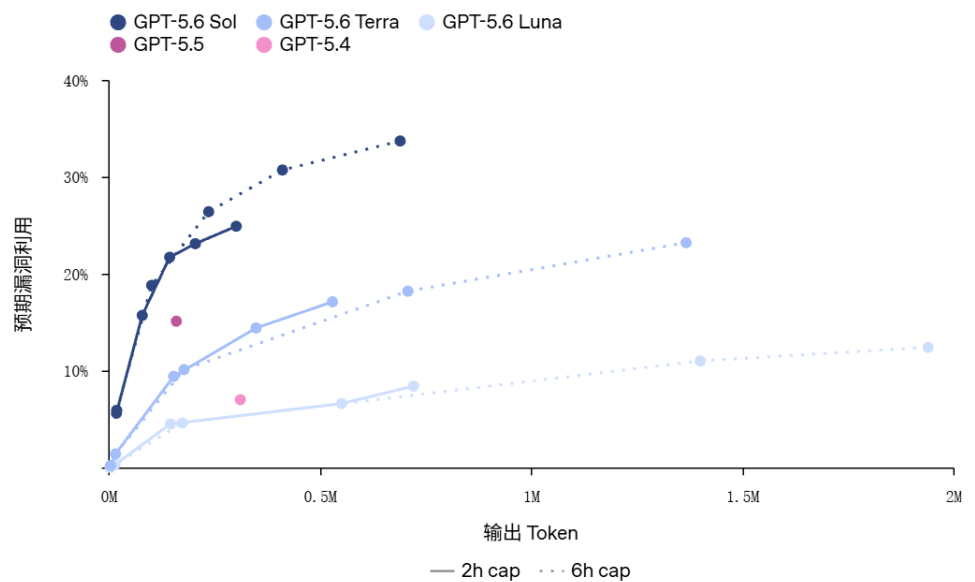
GPT-5.6 Sol 在网络安全与生物学基准上同步跃升，是 OpenAI 迄今最强的双域模型。安全侧在 ExploitBench 上与 Anthropic 未发布的 Mythos Preview 持平，但只用不到 1/3 的成本；ExploitGym（UC Berkeley 与前沿实验室共建）上三档模型随推理努力提升同步走强。生物侧在 GeneBench v1 长程基因组与定量生物分析上 Sol 优于 GPT-5.5 且使用更少 token。

图11: 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12: ExploitGym 榜单上, GPT-5.6 Sol 优于前代 GPT-5.5



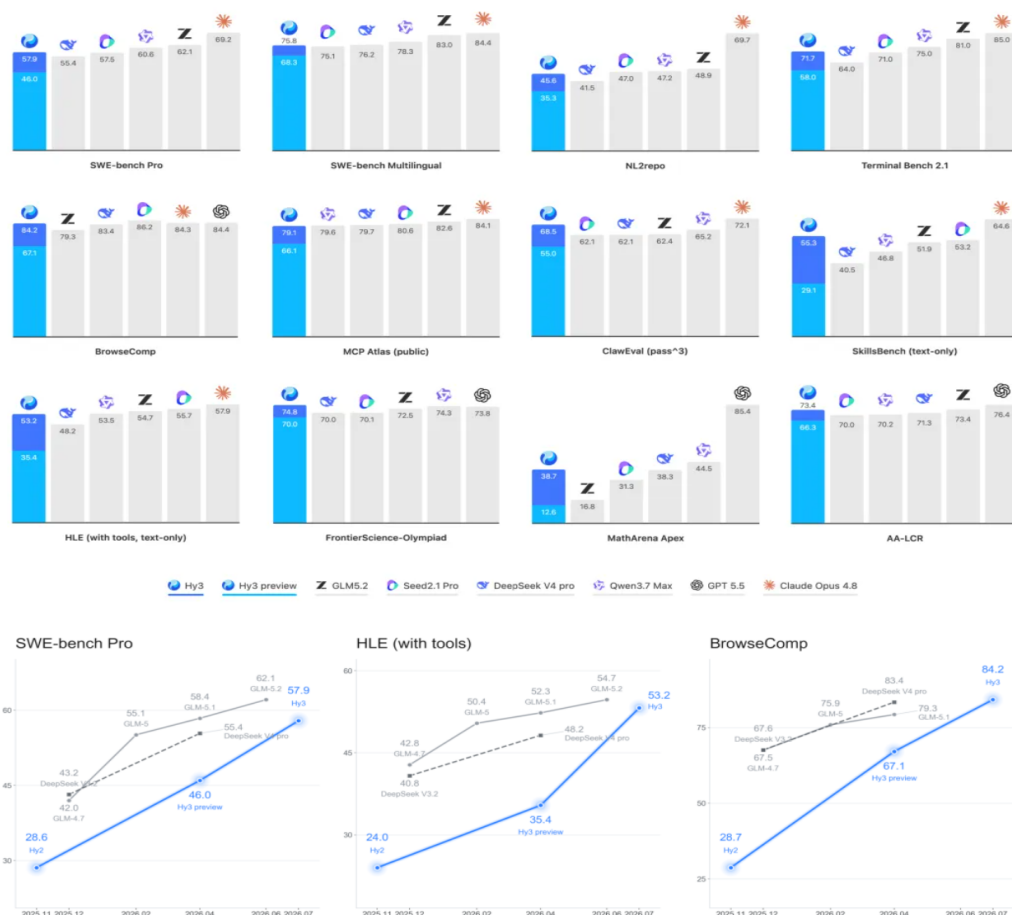
: OpenAI 官网

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 支出来讲意义大于旗舰 benchmark 破纪录。

1.4、腾讯 Hy3：小激活参数+快慢思考融合+真实业务闭环

从架构看, Hy3 采用快慢思考融合的 MoE 架构, 总参数 295B、激活参数 21B, 激活率约 7%, 支持 256K 上下文。这一设计背后的核心思想并非单纯压缩成本, 而是将“高频简单任务”与“低频复杂任务”置于同一套模型权重下, 用不同推理深度来平衡时延与性能。这种路线的商业含义非常明确: 在 Token 快速膨胀的时代, 企业实际购买的不是参数规模, 而是可控的任务成功率×可接受的时延×可承受的调用成本。Hy3 正式版在真实业务场景中的改进印证了这一点: WorkBuddy 场景任务解决率由 72% 升至 90%, 复杂任务文件编辑/生成完成率达 93.7%, 6 个 Agent 协作场景下任务派发正确率达 92%。

图13: Hy3 盲测表现超越同级



资料来源：腾讯云公众号

我们认为腾讯的真正护城河还不在于模型本身，而是模型—产品—数据飞轮。目前 Hy3 已接入 WorkBuddy、CodeBuddy、元宝、ima、Marvis 等腾讯内部业务，并有近多个业务排队接入；自 preview 上线以来，日均 Token 消耗量增长 20 倍，WorkBuddy 上自主选择 Hy3 preview 的用户数增长 6 倍。这意味着腾讯具备其他纯模型厂商较难复制的能力：可以通过大规模真实生产力场景持续压低幻觉、提升任务完成度，并进一步将模型能力快速嵌入微信、企微、腾讯文档等流量入口之中。

1.5、阿里巴巴 Qwen3.8：以 Agent 原生设计追求复杂任务能力上限

2026 年 7 月 19 日，阿里巴巴发布新一代大模型 Qwen 3.8 预览版，其总参数量高达 2.4 万亿，整体性能已进入全球前沿模型行列。该模型在底层算力优化方面展现出卓越能力，成功登顶英伟达面向 Blackwell B200 架构的 FlashInfer-Bench 算子优化榜单。Qwen 3.8 能够通过与真实硬件环境的交互，自主产生训练信号并持续改进算子，标志着其已具备承担复杂工程任务、形成从算子生成到性能调优自动化闭环的能力，大幅压缩了人工介入算子开发的环节。

图14：Qwen3.8 预览版模型在英伟达评估 AI 进行算子优化的榜单上登顶

FlashInfer-Bench B200 Collection Evaluation Stack v1.1

Rank	User	SOL Score ^①	Fast ^②	Avg Speedup ^③	Submitted
-	SOL Bound	1.000000	1.00	102.19x	-
1	Qwen	0.752739	0.95	12.87x	a day ago
2	Hyra	0.734813	0.84	11.55x	4 days ago
3	Geometric	0.722519	0.78	10.50x	13 hours ago
#4	Amir M. Mir SF Tensor	0.718224	0.79	12.83x	15 hours ago
#5	Stellar Flamingo	0.711012	0.77	9.95x	3 days ago
#6	doubleAI	0.709904	0.85	11.61x	2026-07-10
#7	Qin Liu @ UCD	0.697894	0.77	8.10x	a day ago
#8	Coral@HumanAgentSociety	0.671986	0.88	8.41x	2026-07-10
#9	Recursive	0.637046	0.97	11.17x	2026-07-10
#10	Cursor	0.603334	0.63	5.79x	2026-07-10

资料来源：智东西公众号

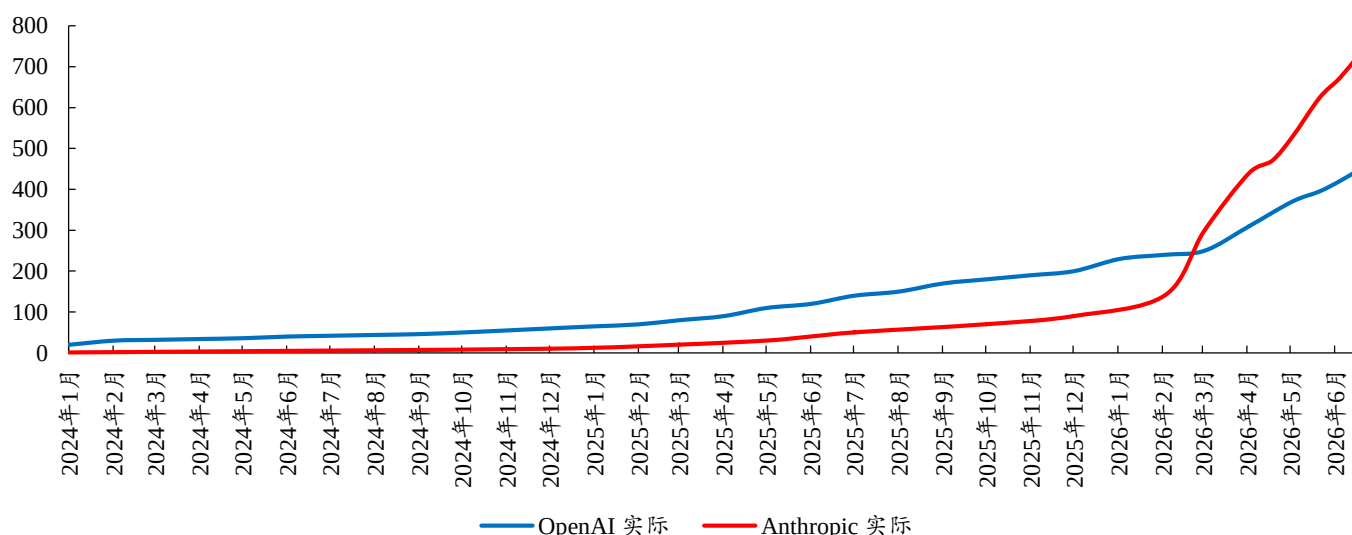
全栈生态布局：（1）模型层，除旗舰模型 Qwen3.8 外，还推出了涵盖开放世界、视频生成及实时语音交互的专用模型矩阵。（2）应用与生态端：千问大模型正式作为 AI 能力集成至已完成备案的“Apple 智能”中，为苹果多终端操作系统的中国用户提供智能体验，这一战略举措不仅大幅拓宽了阿里的商业化落地场景，也标志着其 AI 生态从 B 端向 C 端渗透的全面加速。（3）芯片层：针对智能体高频协同、工具调用、长上下文推理带来的系统效率瓶颈，平头哥实现算存网全栈芯片升级，支撑系统级极致协同：真武 AI 芯片截至 2026 年 4 月累计出货 56 万片，服务 20 多个行业、400 多家客户，是国内应用场景最广泛的 AI 芯片；ICN 互联协议 switch 芯片实现超节点内任意两张卡无差异访问，构建全带宽互连网络，打破传统跨节点通信高延迟瓶颈；磐久 400G 高性能网卡：为 AI 集群打造，用于跨节点以太网通信；倚天服务器 CPU：可与 AI 芯片、网络芯片高效协同，承担任务编排、工具调度、多工具并行

执行、上下文管理等职能；镇岳系列存储控制器芯片：解决海量上下文数据与长期记忆存储问题。

2、ARR 跟踪：头部模型 ARR 维持快速增长，因 OpenAI 竞争投入加大，短期增长斜率或有所波动

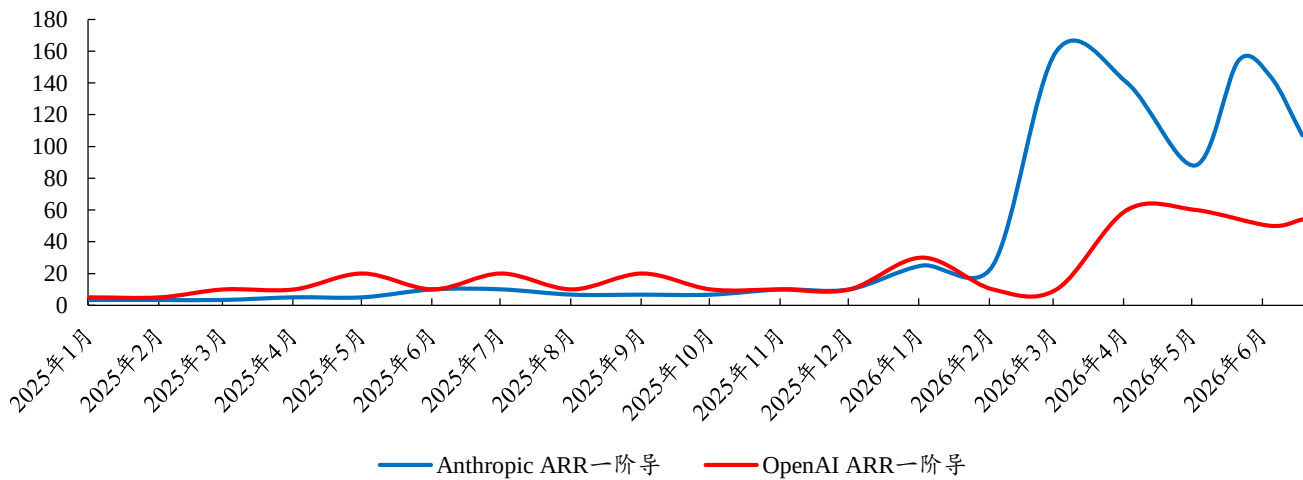
头部模型厂商的 ARR 仍然保持快速增长：Anthropic ARR 增长曲线陡峭抬升，迎来加速增长，在一季度完成对 OpenAI 的反超，7 月份或已迈入 70+B 美元体量。同时 OpenAI 得益于 GPT-5.5/5.6 的优秀表现，叠加快速起量，增长斜率亦有所抬升。

图15: Anthropic 与 OpenAI 的 ARR 持续抬升



资料来源：the Information、Yipit 等，开源证券研究所，注：部分月份数据缺失采用插值平滑处理

头部大模型 ARR 增长斜率短期有所波动，但整体仍维持较高水位。ARR 的增长斜率（一阶导）是当前市场颇为关注的指标，从当前表现上看，7 月份因为 OpenAI 发布 GPT-5.6 并为了获取 codex 用户量，多次重置套餐、加大价格战竞争，ARR 一阶导或有所回落，造成短期数据波动，但月度增长基本维持 10+B 美元的增长水平。我们认为，模型收敛是当前的竞争趋势，但 scaling law 仍然是大模型的成长主线，SOTA 模型相继发布将扩大前沿 AI 应用场景，有望持续推动 ARR 的增长斜率上行。

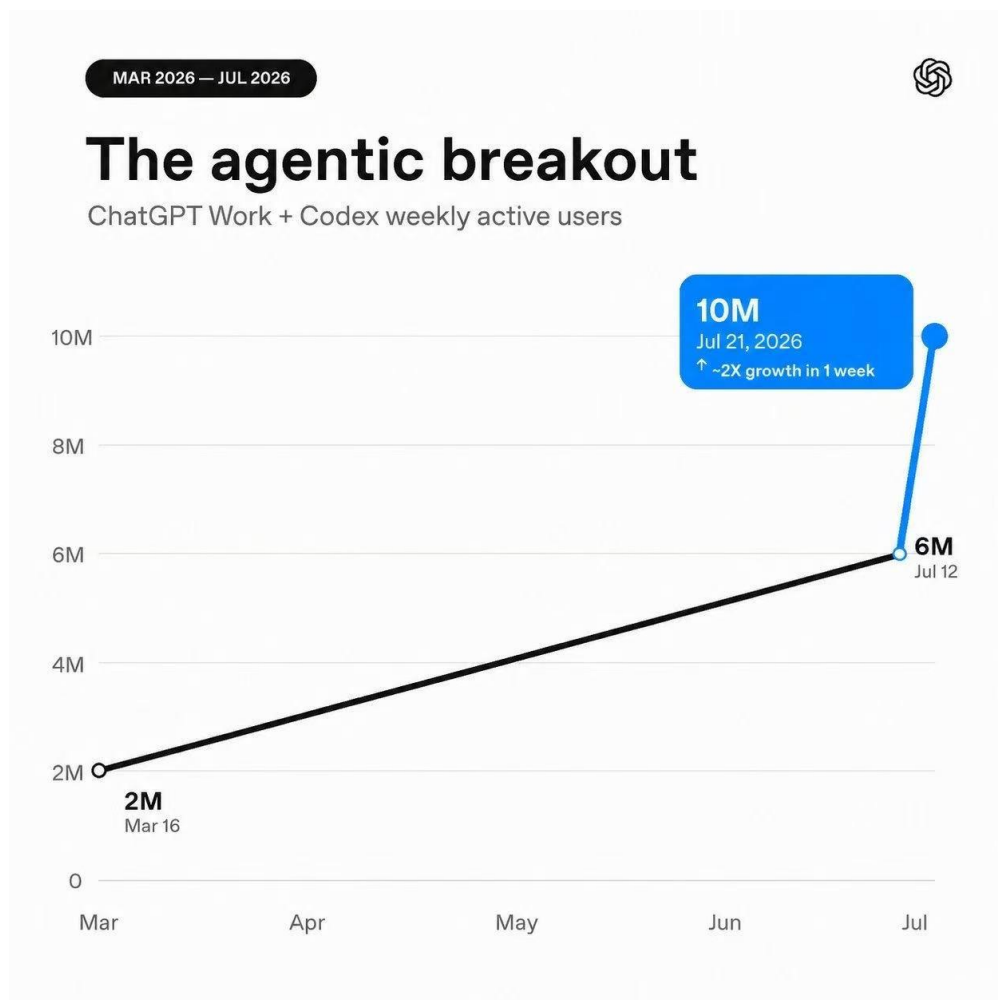
图16: Anthropic 与 OpenAI 的 ARR 一阶导仍保持较高水平

资料来源：the Information、Yipit 等，开源证券研究所，注：部分月份数据缺失采用插值平滑处理

3、Agent 跟踪：codex 突破千万周活用户，腾讯 WorkBuddy 初露锋芒

Agent 层级是获取高质量轨迹数据的重要环节，是大模型厂商放大 coding 及 Harness 优势，能够跟上迭代步伐的重要保证。GPT-5.6 推出后，在其性价比优势及 OpenAI 官网多次重置的背景下，ChatGPT Work 和 Codex 的每周活跃用户数达到 1000 万，比一周多前的 600 万增长了 67%，比 3 月份增长了 5 倍。

图17: codex+ChatGPTWork 月活在 7 月份快速攀升

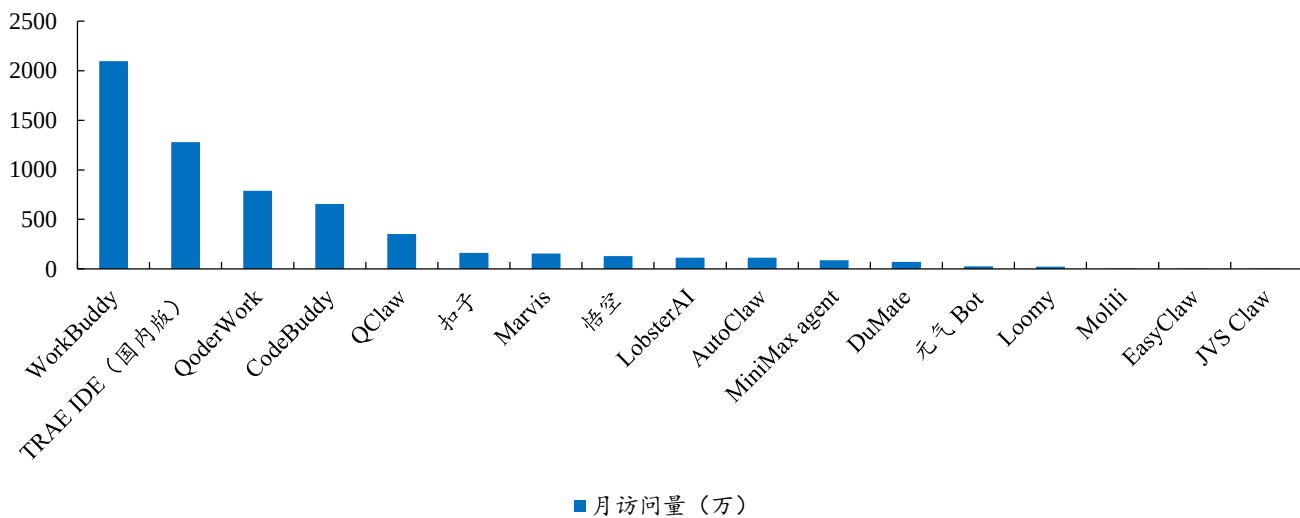


资料来源：OpenAI 官方

腾讯 WorkBuddy 在 AI 原生办公智能体赛道领先，格局呈现明显梯队分化。根据易观分析 2026 年 6 月中国桌面端 AI 原生办公智能体监测数据，腾讯 WorkBuddy 以 2097 万月度访问量位居行业首位，显著高于第二名 TRAE IDE（国内版，1279 万），拉开清晰的第一梯队优势。依托腾讯桌面渠道、企业服务生态，WorkBuddy 在本地智能体、办公自动化场景形成渗透率优势。行业梯队分化显著，头部优势持续放大。桌面原生 AI 智能体尚在发展初期，后续重点观察产品功能迭代、付费转化能力。

但需强调，腾讯 WorkBuddy 主打办公场景，与 Codex、Claude Code、智谱 ZCode、Kimi Code 等更聚焦 coding 场景的产品并不能完全对标，但依托 Hy3 及腾讯庞大的商业生态，亦有望在办公场景、白领阶层构建自身的商业闭环和护城河。当前 200-300B 的模型已经可以在兼顾性价比的前提下，解决较多办公场景任务，后续更重要的是最后一公里的打通及商业生态循环，大厂的优势会更为明显。

图18：6月国内腾讯 WorkBuddy 在桌面 AI 原生办公智能体访问量上明显领先



资料来源：易观分析，开源证券研究所

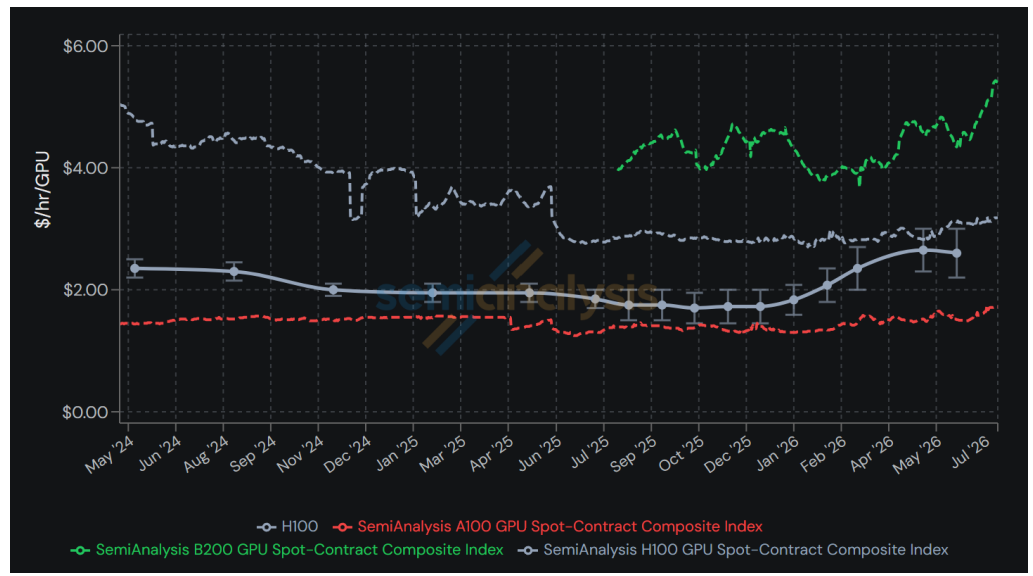
4、算力：高端算力持续紧平衡，Token 需求走向指数扩张

2026 年 7 月算力侧关注核心在于模型效率提升和价格战持续的情况下，Token 需求是否仍然快于单位成本下降。短期需求弹性仍强，一方面，DeepSeek、MiniMax、GLM 等模型通过 MoE、Sparse Attention、GQA、投机解码等方式持续压低单位推理成本；另一方面，编程、生产力、Agent、多模态任务正把总调用量推向更高数量级。中长期仍需关注 Blackwell 放量、H100 代际下沉及模型效率提升从而带动部分通用推理场景的卡需求紧张缓解。

4.1、供给侧：GPU 成熟卡种继续温和上涨，新一代卡种维持高位溢价

目前 GPU 租赁价格呈现成熟卡种继续温和上涨，新一代卡种维持高位溢价。从更长区间看，算力价格并非线性上升，而是经历了 2025 年下半年回落—2026 年上半年再通胀”的重定价过程。B200 自 2025 年 8 月回落至 2026 年初，再在 2026 年 6-7 月快速冲高；H100 则在 2025 年 8 月至 2025 年 12 月长期稳定，随后在 2026Q1 末期快速拉升并，A100 则长期稳定在 1.4-1.6 美元/小时附近，算力供需矛盾已从全面短缺演变为高端前沿卡紧缺、成熟卡价值重估，H100 的上涨说明老卡并未被淘汰，反而在推理和中等强度训练场景中被重新吸收；B200 维持高位则显示最前沿模型与 Agent workloads 对更高吞吐/更高显存的需求仍在释放。

图19：算力租赁价格呈现高端前沿卡紧缺、成熟卡价格重估态势

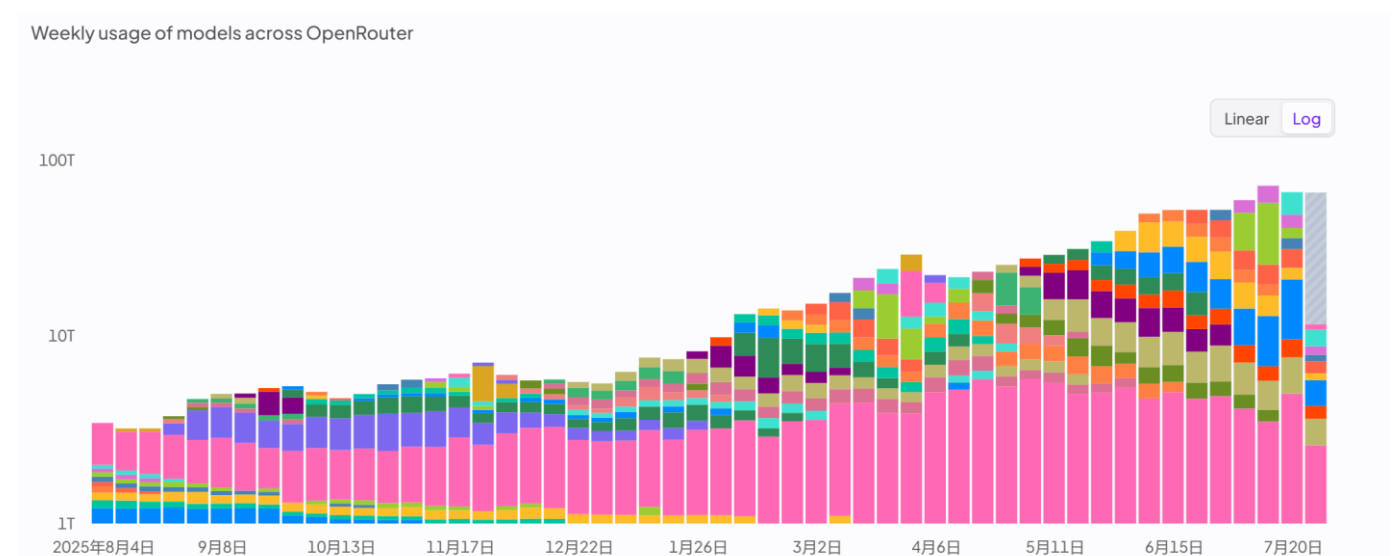


资料来源：SemiAnalysis

4.2、需求侧：生产力场景及 Agent 驱动 Token 消耗指数级增长

根据 OpenRouter, 全球周度 Token 调用量(截至 2026 年 7 月 27 日)已升至 57.8T。拉长周期看, 2025 年 7 月至 2026 年 1 月, 全球周度 Token 调用量仍处于低位平缓增长; 而自 2026 年初开始, 曲线明显陡峭化, 进入加速上行阶段, AI 需求已不再仅由聊天类单轮问答拉动, 而是开始由编码、Agent、生产力工具、多轮交互、长上下文任务共同驱动。

图20：Token 消耗量持续提升



资料来源：OpenRouter

4.3、北美 CSP 高资本开支或为算力价格及 Token 消耗提供支撑

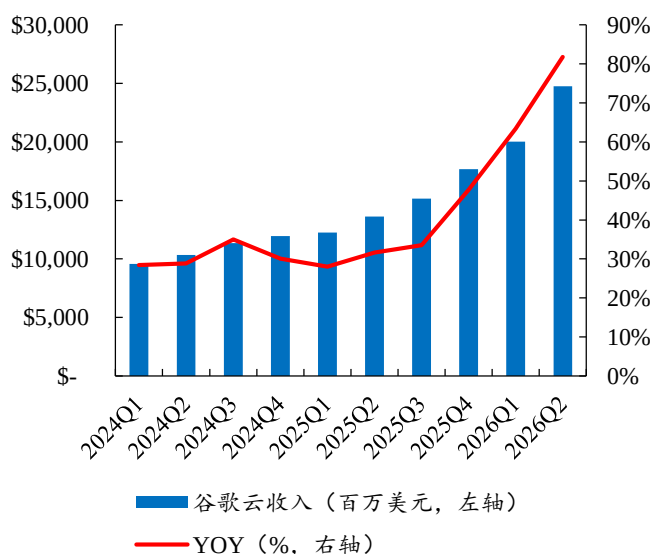
北美云厂商持续上修资本开支，源于 Token 吞吐增长快于供给扩张，这意味着算力租赁价格上行是超大规模资本开支与超高速 Token 增长之间短期错配的表现。此外云厂商自身也在抢占更多内部推理、训练、企业云客户、API 服务和 Agent 平台需求，因此算力供给新增并不会马上压低前沿租赁价格。

4.3.1、谷歌云 2026Q2 收入利润持续超预期，上调全年 CAPEX 指引

谷歌发布 2026Q2 业绩，云业务收入及利润率均超彭博一致预期，但由于高强度基建投入，2026Q2 自由现金流阶段性转负引发股价震荡：

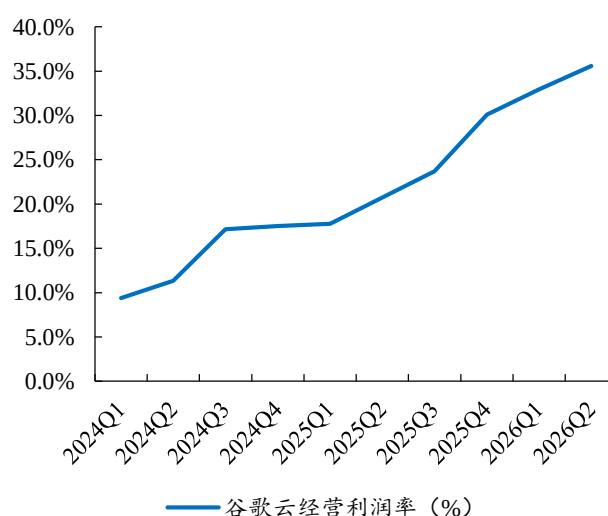
云：2026Q2 谷歌云收入 248 亿美元，同比增长 82%，经营利润 88 亿美元，经营利润率 35.6%，同比提升 14.9pcts，均好于彭博一致预期。（1）TPU：2026Q2 谷歌 TPU 首次系统交付至客户数据中心并开始确认相关收入；预计 2026 年仅能从现有 TPU 系统销售协议中确认相对较少的收入，收入规模将在 2026 年末逐步爬坡提升，绝大部分收入预计将在 2027 年确认。（2）在手订单：截至 2026Q2 谷歌云在手订单规模环比增长超 500 亿美元达到 5140 亿美元，这一增长主要得益于市场对企业级人工智能产品的强劲需求，大部分来自多元客户群体的常规 GCP 服务合同，预计未来 24 个月内，总在手订单中略超半数的部分将确认为营收。

图21：2026Q2 谷歌云收入同比增长 82%



数据来源：彭博、开源证券研究所

图22：2026Q2 谷歌云经营利润率达 35.6%

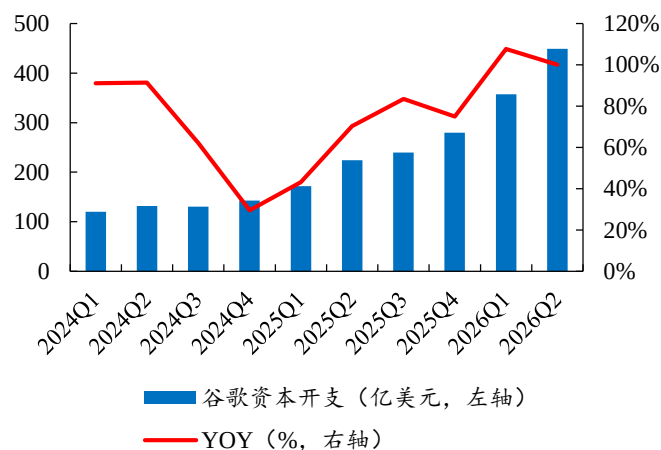


数据来源：彭博、开源证券研究所

资本开支与现金流：（1）2026Q2 谷歌资本开支 449 亿美元，同比增长 100%，绝大部分投入技术基础设施，约 60%投向服务器，40%投向数据中心与网络设备；将 2026 财年的资本支出指引从之前的 1800 亿-1900 亿美元区间上调至 1950 亿-2050 亿美元以加速交付产能以满足日益增长的需求。此外，资本支出将在 2027 年显著增加。（2）2026Q2 季度自由现金流为负 59 亿美元，过去十二个月的自由现金流降至 533 亿美元，随着基础设施建设延续 2026 年自由现金流将持续承压。目前谷歌资本开支中来自经营的现金仍是第一资金来源，过去十二个月经营现金流为

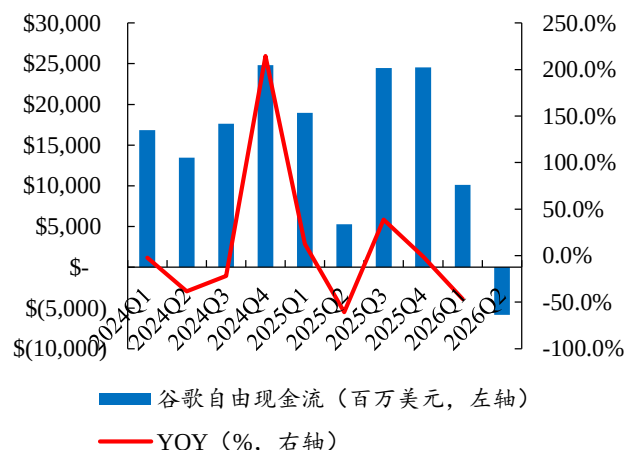
1857 亿美元；债务是第二来源，长期债务规模约 1000 亿美元（一年前债务规模约为 160 亿美元），覆盖多币种、跨区域的融资渠道，相关布局仍在拓展；股权是第三来源。

图23：2026Q2 谷歌资本开支维持高增



数据来源：彭博、开源证券研究所

图24：2026Q2 谷歌自由现金流转负

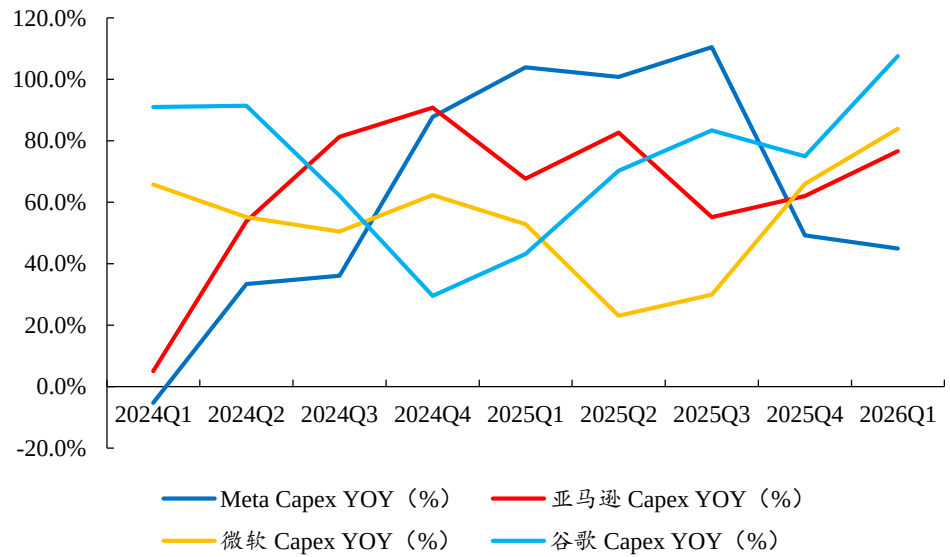


数据来源：彭博、开源证券研究所

4.3.2、Meta 或进军云计算，关注内部算力利用效率

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。

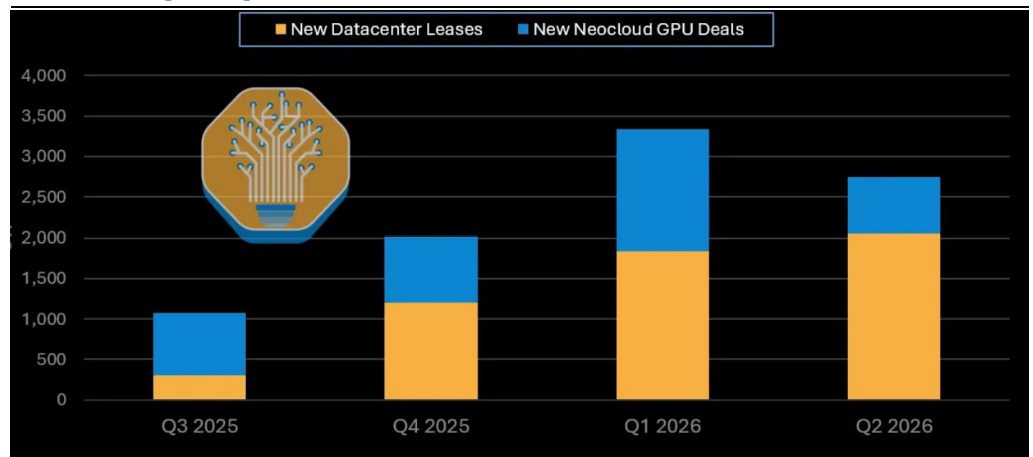
图25: Meta 资本开支自 2024 年起维持快速增长



数据来源：公司公告、彭博、开源证券研究所

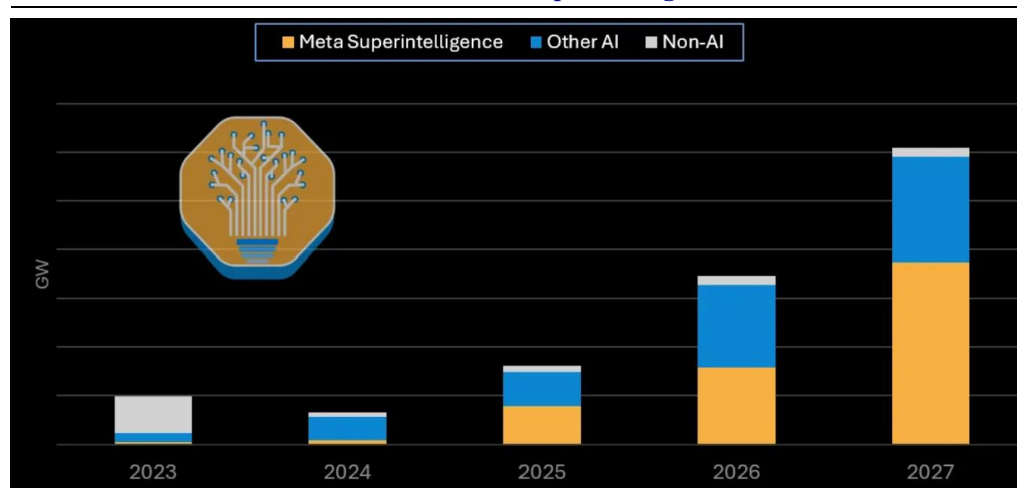
对云计算竞争格局的冲击：短期看，根据 Semianalysis，Meta 自 2024 年初以来签署近 10GW 的电力容量，目前大部分的容量增长是通过第三方实现的，因此 Meta 或驱动 Coreweave、Nebius 等 RPO 增长，未来需关注 Meta 内部广告推荐系统(RecSys)及前沿 AI 模型（Meta Superintelligence Labs）对于训练和推理的算力利用效率，若广告推荐系统扩展规模或 AI 模型低于预期，其云端算力容量将迅速提升。

图26: 2026Q1 及 Q2 Meta 新签算力合同超过 5GW



资料来源：Semianalysis

图27: Meta 目前的算力主要分配给 Meta Superintelligence



资料来源: Semianalysis

4.4、科技企业裁员跟踪：本轮调整并非单纯行业收缩，而是资本开支结构重构

Layoff.today 的数据反映 AI 落地对行业人力结构的结构性冲击，企业压缩重复性人力成本，将资金倾斜至算力采购、基础模型研发等高资本投入环节，呈现“算力替代人力”的转型特征。部分企业存在借 AI 叙事完成周期降本的情况，真实岗位替代幅度需结合企业业务周期交叉验证；长期看 AI 带来人力成本优化有望增厚企业盈利，但短期劳动力市场波动、人才再分配亦不可避免。

表2: 受 AI 影响，科技厂商持续裁员

企业	日期	受影响的工人
Envato	2026 年 4 月 26 日	200
Snowflake	2026 年 4 月 26 日	70
Crypto.com	2026 年 4 月 26 日	180
惠普公司	2026 年 4 月 10 日	52
InvestCloud	2026 年 3 月 9 日	150
Block	2026 年 2 月 26 日	4,000
Allianz	2025 年 11 月 7 日	1,800
埃森哲	2025 年 11 月 4 日	11,000
Fiverr International	2025 年 11 月 1 日	250
Salesforce	2025 年 10 月 23 日	355
惠普公司	2025 年 10 月 7 日	5,000
微软	2026 年 4 月、7 月	9.1%
亚马逊	2025 年 10 月、2026 年 1 月、7 月	超过 30,000 人
谷歌	2025 年 12 月、2026 年 4 月	显性 200 人以上，隐性滚动优化数千人
Meta	2026 年 5 月	裁员 8,000 人，取消 6000 个开放招聘指标

资料来源: Layoff.Today、联合早报、华尔街日报、财联社、钛媒体、潮涌 AI 公众号、开源证券研究所

请务必参阅正文后面的信息披露和法律声明

24 / 27

5、投资建议

Kimi K3 与千问 Qwen3.8 等国产模型的接连突破，不仅是参数和榜单数字的跃升，标志着中国 AI 产业进入新阶段，有力地证明了开源模型的天花板同样很高，在国际竞争力进一步提升，在性能和成本的平衡下有望打开全球 AI 需求天花板。

模型端：需求天花板提升，推荐标的智谱，受益标的 minimax、德适-B。

算力端：多模态落地与 Agent 规模化带动训练、推理算力需求长期刚性扩张，推荐标的阿里巴巴、百度集团-SW、联想集团，受益标的腾讯控股、天数智芯、中芯国际。

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。

分析师声明

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（北交所基准指数为北证50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。

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

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

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

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