

2Q26 大模型发展复盘及展望

关注算力利用效率提升及差异化定价能力

优于大市
(维持)

复盘 2Q26 大模型行业发展，我们观察到基础模型智能能力的竞争仍在加剧，Claude Fable 5 的推出推动了大模型行业整体智能能力上界提升，而智谱 GLM 5.2 的发布助力中国大模型厂商维持了在开源模型领域的优势地位。从模型综合能力看，国际厂商和中国厂商分别以 Anthropic 和智谱为代表在多个维度表现出色。除模型能力提升外，行业演进主要围绕两大趋势：1) 深化编程/办公智能体能力探索，包括长程工具调用、工程执行能力与原生多模态理解等；2) 持续优化推理调用价格及性价比以进一步推动模型调用渗透提升，以 Artificial Analysis 智能能力指数区间 40-50 的模型统计，其区间内在推理调用价格方面表现最优模型的调用价格已由 2026 年 3 月的约 1.2 美元/百万 tokens 下探至 2026 年 6 月的约 0.058 美元/百万 tokens。展望而言，我们认为模型能力上界依旧是决定定价能力的关键，行业玩家仍将关注智能能力的提升以支撑长期变现能力增长，同时将通过全栈能力优化推动成本优化。在应用及变现场景端，我们预期编程场景的竞争将加剧，但龙头厂商或仍有能力维持其市场领先地位和市占；Agent 端，我们预期完全自治的智能体系统是行业探索方向。股票推荐维度，中国互联网板块中我们推荐具备丰富应用场景且云业务有望受益于 token 需求指数级增长的阿里巴巴 (BABA US)、腾讯 (700 HK)、快手 (1024 HK)；美国互联网板块推荐微软 (MSFT US)、谷歌 (GOOG US)、亚马逊 (AMZN US)。

- **全球 AI Token 消耗量或将呈现指数级增长趋势。**我们测算全球月均 AI Token 消耗量将在 2030 年超过 26 万万亿，2025-2030 年复合增速 107%，增长驱动力包括：1) 模型智能水平提升带动 AI 产品总用户规模增长及用户使用粘性提升；2) 多模态输入/输出占比提升推动单次 prompt 的 token 消耗量增加；3) 底层模型能力提升+Harness 框架完善带动 Agent 产品渗透率提升。我们认为 Agent 场景将成为 2025-2030 年 Token 消耗量增长的核心驱动力，预计 2030 年全球 Agent 场景月均 Token 消耗量将超过 21 万万亿 (2025-2030 年复合增速 165%)，占全球 AI Token 消耗量约 84%。在此背景下，我们建议关注三条主线投资机会：1) 基础设施云厂商和头部基础模型厂商；2) 头部编程/办公/企业 AI Agent 应用厂商；3) 头部多模态模型与多模态 AI 应用厂商。
- **AI 模型平台：资本化进程加速，模型迭代速度提升，强化基础设施韧性。**2Q26 海外 AI 平台竞争格局主要边际变化包括：1) Anthropic ARR 在 5 月反超 OpenAI；2) Anthropic 和 OpenAI 先后开启 IPO 筹备进程，估值锚将逐步转向公开市场。国内 AI 平台竞争格局边际变化集中于：1) 资本化进程加速：智谱、MiniMax 分别于 5 月底/6 月初启动 A 股科创板 IPO 申报，阶跃星辰、月之暗面据新闻报道亦在筹备港股 IPO；2) 模型迭代节奏环比加快，MiniMax M3、智谱 GLM-5.2、月之暗面 K2.6、阶跃 Step 3.7 Flash 集中于 2Q26 发布，在长上下文、稀疏注意力、多模态原生生成等方向均有版本更新；3) 推进国产芯片适配、强化算力供应链韧性：DeepSeek V4 及 GLM-5.2 均完成对主流国产芯片深度适配。
- **AI 应用商业进展：B 端商业化持续加速。**复盘我们所覆盖的主要公司的 AI 业务商业化进展，我们观察到：1) 因 ROI 更为清晰且对智能能力的需求更大，AI 在 B 端的商业化变现仍在加速，云、广告、创意内容生成、Agent 等仍是主要的关注领域，且 AI 助力企业降本增效的进程在持续推进；2) 相较而言，C 端整体付费订阅及渗透仍处于较为早期的阶段，办公及生产力提升仍是主要探索及推进的场景，未来基于模型能力提升推动付费渗透率的提升、及持续推动从仅基于用户数量计费的模式转型至按“用户+使用量计费”，或将是商业模式演变探索的中长期重点。

中国互联网行业

贺赛一, CFA

(852) 3916 1739

hesaiyi@cmbi.com.hk

陶冶, CFA

(852) 3850 5226

franktao@cmbi.com.hk

陆文韬, CFA

luwentao@cmbi.com.hk

郭书音

(852) 3916 3716

guoshuyin@cmbi.com.hk

相关报告:

[1. 海外龙头 TMT 公司业绩启示 - 盈利增长支撑股价上行，持续关注盈利成长性 - 18 Jun](#)

[2. Knowledge Atlas \(2513 HK\) initiation - Sailing toward the vast frontier of AGI - 11 Jun](#)

[3. MiniMax Group \(100 HK\) - Evolving into a one-stop AI platform - 10 Jun](#)

[4. Kuaishou \(1024 HK\) - Kling AI maintained strong monetization momentum - 28 May](#)

[5. Alibaba \(BABA US\) - Strong cloud business development prospects - 14 May](#)

[6. Tencent \(700 HK\) - Resilient 1Q26 earnings growth despite AI investment; solid AI progress - 14 May](#)

[7. Microsoft \(MSFT US\) - Expanding from per-user model into user-plus-usage model - 4 May](#)

[8. Amazon \(AMZN US\) - AWS growth momentum continues to build - 4 May](#)

[9. Meta \(META US\) - AI accelerates ad revenue growth; raise FY26 capex guidance - 4 May](#)

[10. Alphabet \(GOOG US\) - 1Q26 results beat on strong search and cloud businesses; AI as the key driver - 4 May](#)

[11. AI 主题研究 - 1Q26 大模型进展复盘：关注 Agent 化及 B 端突围 - 16 Apr](#)

目录

互联网 - 关注算力利用效率提升及差异化定价能力 1

行业发展趋势：模型能力迭代加速，Fable 推升智能能力上界..... 3

大语言模型：模型能力快速迭代，Claude 领跑全球，智谱位居中国前列4

行业趋势：关注算力利用效率提升，而智能能力仍是模型定价水平的决定性因素7

行业探索最新进展：长程工作水平演进，编程与多模态操作能力强化9

大语言模型调用量：2Q26 token 消耗持续上升，中国开源模型贡献主要增量11

图像模型仍由海外厂商领跑，但视频赛道中国厂商更具优势14

龙头基础模型平台模型及业务进展更新17

海外 AI 基础模型平台：ARR 竞赛持续，Anthropic 增长领跑，进入 IPO 筹备阶段.....17

国内 AI 基础模型平台：持续关注模型能力迭代及商业化验证21

重点覆盖公司 2Q26 AI 业务进展更新及催化展望25

微软：将 Copilot 从单点应用升级为平台化架构，深化 Agent 业务发展布局26

谷歌：全栈 AI 策略支撑 AI 相关业务强劲增长26

亚马逊：关注 AWS 营收同比增长环比加速趋势28

Meta：持续推动 C 端 AI 应用落地28

腾讯：关注微信 AI 智能体进展30

阿里巴巴：组织架构重构与 AI 入口升级31

快手：外部融资释放可灵价值33

股票推荐.....34

行业发展趋势：模型能力迭代加速，Fable 推升智能能力上界

2Q26 大模型行业基础模型智能能力的竞争仍在持续加剧，Claude Fable 5 的推出推动了大模型行业整体智能能力上界提升，而智谱 GLM5.2 的发布进一步推动了中国大模型行业的智能能力的提升，也助力中国大模型厂商维持了在开源模型领域的优势地位。整体来看，2Q26 大语言模型行业演进主要围绕三大趋势：1) 深化编程/办公智能体能力探索，包括长程工具调用、工程执行能力与原生多模态理解等；2) 持续优化推理调用价格及性价比以进一步推动模型调用渗透提升，以 Artificial Analysis 智能能力指数区间 40-50 的模型统计，其区间内在推理调用价格方面表现最优模型的调用价格已由 2026 年 3 月的约 1.2 美元/百万 tokens 下探至 2026 年 6 月的约 0.058 美元/百万 tokens；3) 基于性价比优势，中国模型已成为全球 AI Token 的主要供应商（OpenRouter 平台周度调用量前十模型中，中国模型占据七位）。

从模型综合能力看，国际厂商和中国厂商分别以 Anthropic 和智谱为代表在多个维度表现出色。分项来看：1) 语言模型智能能力方面，国际厂商中以 Claude Fable 5 领先，国内则以智谱 GLM-5.2 为代表；2) 推理效率方面，国内 DeepSeek V4 Pro、DeepSeek V4 Flash、MiniMax M3、小米 MiMo-V2.5-Pro 均进入最优象限；3) Agent 能力方面，国际第一梯队主要包括 Anthropic Claude Fable 5/ Opus 4.8 与 OpenAI GPT-5.5，国内则 Z.AI GLM-5.2、DeepSeek V4 Pro、MiniMax M3 领跑；4) 编程能力方面，Anthropic Claude Fable 5/Opus 4.8 和 OpenAI GPT 5.5 领先，国内厂商中 GLM 5.2 排名第一；5) 图像生成方面，国际头部模型为 OpenAI、微软和谷歌，国内字节跳动表现较为突出；6) 视频生成方面，国内厂商阿里巴巴、字节跳动和快手构成第一梯队领跑行业，海外厂商以 xAI 表现突出。

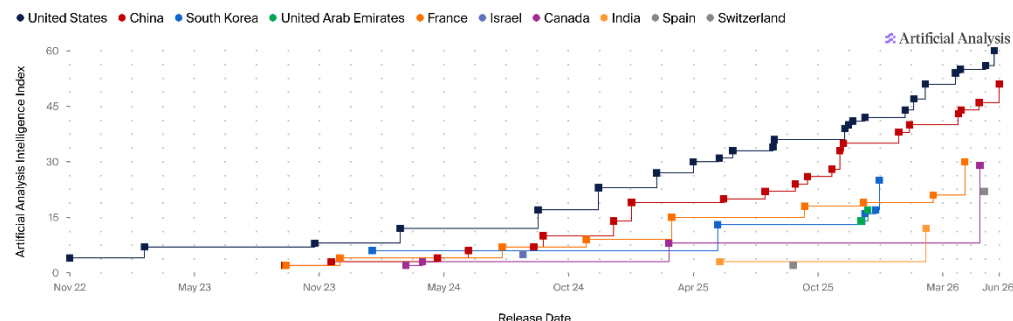
据 OpenRouter 平台 Token 调用量计算，中国开源模型已贡献 2Q26 主要增量。截至 6 月 22 日当周，DeepSeek 以 6.7 万亿 tokens 居平台首位，高于 Anthropic 的 4.25 万亿；小米（4.48 万亿）、MiniMax（3.74 万亿）、智谱（2.11 万亿）等中国厂商合计占比亦较高。DeepSeek 的调用量开始快速上升与其 2Q26 V4 系列发布、API 降价 75% 的时点一致。

主流大模型厂商的算力优化，已经从“单纯堆 GPU”转向“模型稀疏化 + 长上下文注意力优化 + 低精度训练/推理 + 推理缓存/路由 + 自研或深度适配硬件”的全栈工程。展望而言，我们预期行业龙头大模型厂商将持续探索模型智能能力的提升及通过全栈能力提升推动算力使用效率优化及成本优化。在应用及变现场景端，我们预期编程场景的竞争将加剧，但龙头厂商或仍有能力维持其市场领先地位和领先的市占率；Agent 端，我们预期完全自治的智能体系统是行业探索方向，Memory、持续学习、自我评判的技术攻关和推进是主要关注点。

图 1: 各国前沿大模型的发布时间及能力对比

Frontier Language Model Intelligence By Country, Over Time

Artificial Analysis Intelligence Index v4.1 incorporates 9 evaluations: GDPval-AA v2, τ^2 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR



资料来源: Artificial Analysis, 招银国际环球市场

注: 1) Artificial Analysis 智能能力指数涵盖 10 项评估维度, 主要包括: GDPval-AA, τ^2 -Bench Telecom, Terminal-Bench Hard, SciCode, AA-LCR, AA-Omniscience, IFBench, Humanity's Last Exam, GPQA Diamond, CritPt; 2) 模型性能呈现口径: 相关数值优先反映模型官方 API 的表现 (例如 o1 采用 OpenAI 官方 API 数据), 若模型没有官方 API, 则采用多个服务提供方结果的中位数作为代表值 (例如 Meta 的 Llama 模型); 3) 排名并未穷尽所有大模型; 4) 排名时间截至 2026 年 7 月 1 日, 除单独备注外, 下同

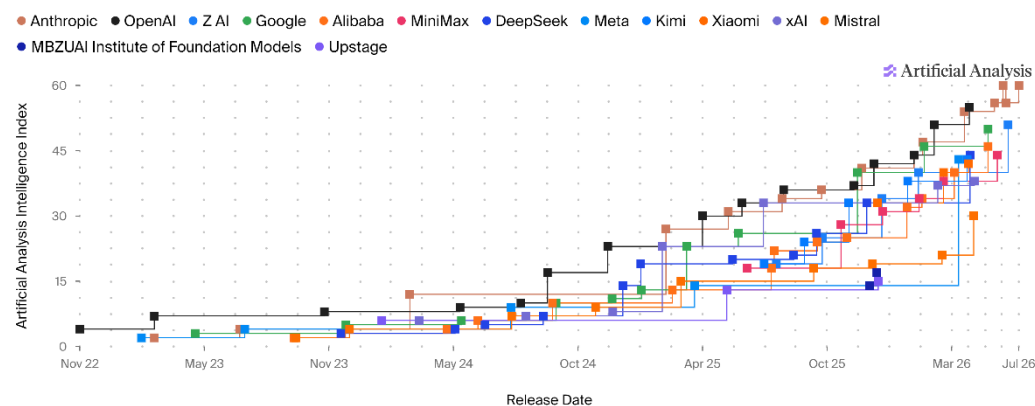
大语言模型: 模型能力快速迭代, Claude 领跑全球, 智谱位居中国前列

Artificial Analysis 时间序列显示, 2025 年下半年以来头部厂商公开模型版本的更新节点明显增密、版本间隔持续缩短, 且多家厂商能力曲线在更短时间内展现更高频的能力提升趋势。从模型能力表现来看, 根据 Artificial Analysis 智能能力指数, 截至 2026 年 7 月 1 日, 当前行业综合能力排名领先的模型为 Claude Fable 5 (60 分) 和 Claude Opus 4.8 (56 分), 其后为 OpenAI GPT 5.5 (55 分)、Claude Sonnet 5 (53 分) 和智谱 GLM 5.2 (51 分)。Fable 5 发布进一步推动模型能力上界的提升, 而各大模型厂商之间的竞争格局仍较为胶着。

图 2: 主流大语言模型的发布时间及能力对比

Frontier Language Model Intelligence, Over Time

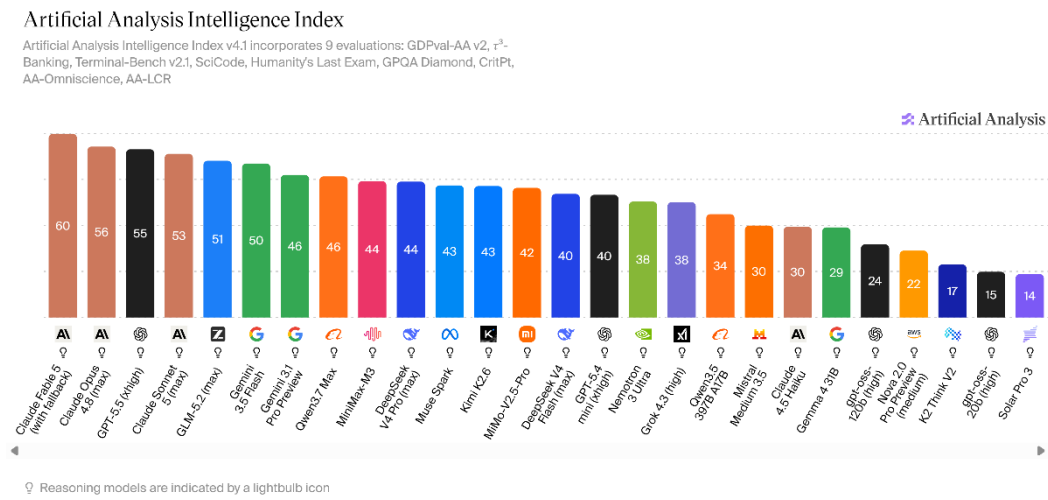
Artificial Analysis Intelligence Index v4.1 incorporates 9 evaluations: GDPval-AA v2, τ^2 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR



资料来源: Artificial Analysis, 招银国际环球市场

从大模型整体能力评分来看，基于 Artificial Analysis 数据，美国厂商中 Claude Fable 5 、 Claude Opus 4.8 和 OpenAI GPT 5.5 等处于行业领先水平，而中国厂商中智谱的 Z.AI GLM-5.2、阿里巴巴的 Qwen3.7 Max、MiniMax-M3、DeepSeek V4 Pro 大模型能力处于行业领先水平。

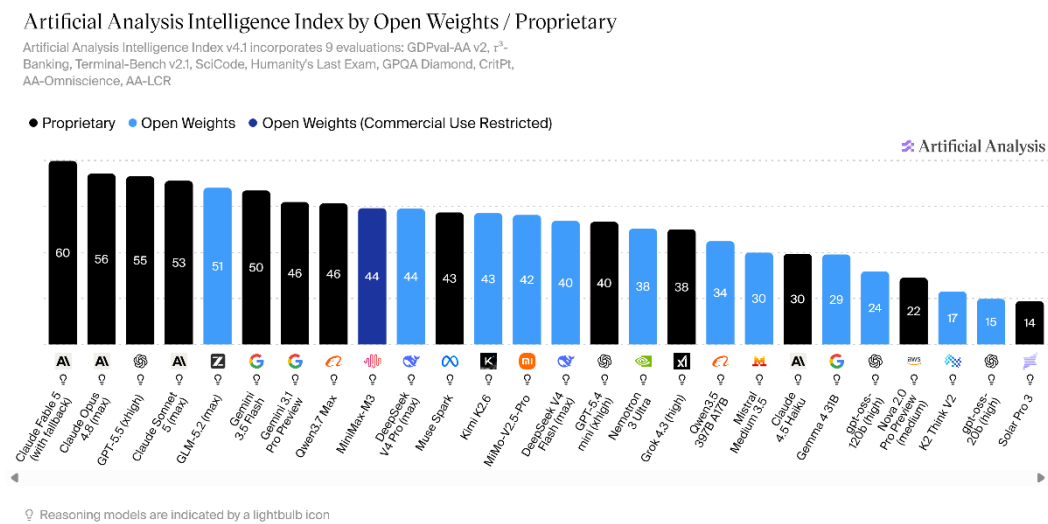
图 3: 主流大语言模型能力排名



资料来源: Artificial Analysis, 招银国际环球市场

从开源和闭源模型的对比来看，基于 Artificial Analysis 智能能力指数，整体而言闭源模型的智能能力仍领先于开源模型，但开源模型的智能能力亦处于行业领先水平。从开源模型整体能力评分来看，智谱 GLM-5.2 以 51 分位列开源模型第一；MiniMax-M3、DeepSeek V4 Pro 和月之暗面 Kimi K2.6 紧随其后，分别为 44 分、44 分和 43 分，处于开源模型第一梯队。

图 4: 开源及闭源模型能力对比及智能能力指数

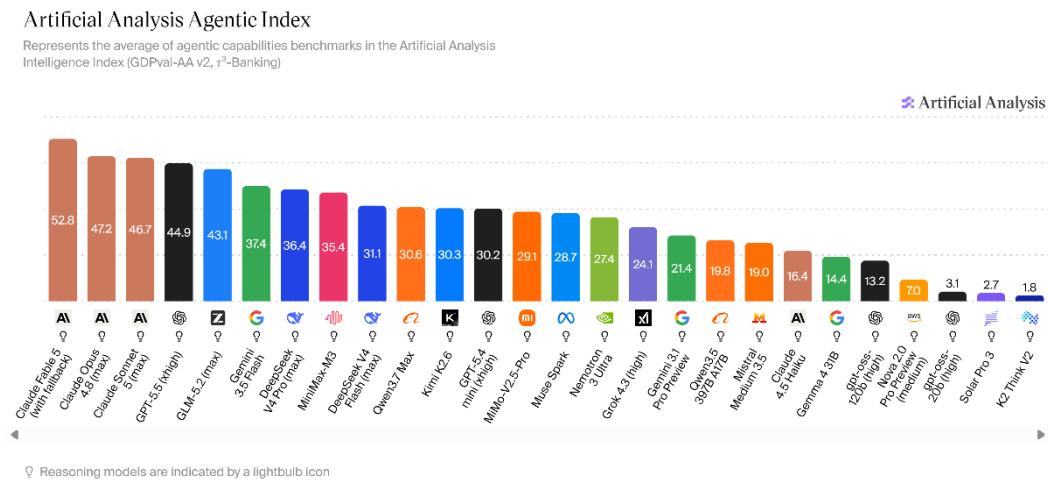


资料来源: Artificial Analysis, 招银国际环球市场

注: 图中黑色及灰色为闭源模型，蓝色为开源模型

在相对更为细分的智能体领域的评分中，截至 2026 年 7 月 1 日，美国厂商中 Anthropic Claude Fable 5、Anthropic Claude Opus 4.8、Anthropic Claude Sonnet 5 处于行业领先水平，而中国厂商中智谱的 Z.AI GLM-5.2、DeepSeek V4 Pro、MiniMax M3 智能体能力处于行业领先水平。

图 5: 大模型智能体能力排名

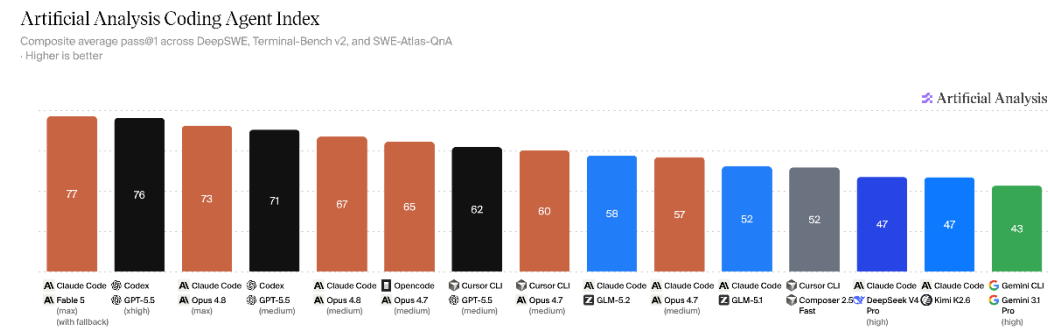


资料来源: Artificial Analysis, 招银国际环球市场

注: 1) Artificial Analysis 代理能力指数代表 Artificial Analysis 智能能力指数中代理能力基准测试的平均值 (GDPval-AA, r^2 -Bench Telecom); 2) 排名并未穷尽所有大模型

进一步而言，在商业化得到广泛验证且竞争持续加剧的编程领域而言，截至 2026 年 7 月 1 日，美国厂商中 Anthropic Claude Fable 5、Anthropic Claude Opus 4.8、OpenAI GPT-5.5 处于行业领先水平；中国模型中，智谱 GLM-5.2 在 Claude Code 工具环境下的编程能力指数为 58 分。

图 6: 大模型编程能力排名



资料来源: Artificial Analysis, 招银国际环球市场

注: 1) 该指数基于 DeepSWE、Terminal-Bench v2 及 SWE-Atlas-QnA 三项软件工程评测任务的一次通过率综合平均计算，衡量模型在编程任务中首次生成答案即通过测试或满足要求的能力，得分越高代表 Coding Agent 能力越强; 2) 横轴为“Coding Agent 工具/框架 + 底层模型”的组合，即第一行为编程智能体工具，第二行为底层模型; 3) 排名并未穷尽所有大模型。

行业趋势：关注算力利用效率提升，而智能能力仍是模型定价水平的决定性因素

2026 年以来，我们观察到 AI 算力效率改善由多环节推进：1) 硬件端通过芯片代际升级、专属推理芯片和自研集群部署改善单位功耗与单位成本下的 token 产出；2) 软件端通过模型训练方法、编译器与算子优化提升推理吞吐和训练效率；3) 国产模型通过注意力机制改造、KV cache 压缩和 MoE 稀疏激活，降低长上下文推理计算与存储开销；4) 调度与基础设施端通过资源动态分配、闲置容量利用和时间同步等方式提升 GPU 利用率。

整体看，主要模型厂商在算力效率提升和成本优化端各有侧重：DeepSeek、智谱、MiniMax 等国产厂商主要依托模型架构层改善（MoE 稀疏激活、MLA / Lightning Attention / MSA 等注意力机制改造）与训练系统及方法优化（FP8、DualPipe、GRPO/CISPO）压低单位成本；Anthropic、OpenAI 等海外厂商则更多在 API 与产品层（模型分层路由、prompt caching、batch 优化）乃至自研推理芯片上摊薄单位智能成本。

图 7：主要模型厂商的降本技术路径

公司	技术手段	核心逻辑
DeepSeek	MoE、MLA、FP8、DualPipe、多 token prediction、GRPO/RL 优化	在较低训练成本下做出高性能模型，核心是模型-系统-硬件协同
智谱 / Z.ai	MoE、小 active 参数、长上下文、国产芯片适配、开源生态	用更低激活计算量做前沿级能力，同时适配中国算力环境
MiniMax	MoE、Lightning Attention (M1) / MSA (M3)、长上下文优化、RL 效率算法 CISPO (M1)	重点解决百万上下文和推理时计算的成本问题
Anthropic	Claude 分层、prompt caching、batch API、Haiku/Sonnet/Opus 路由、企业 API 优化	通过产品/API 层提升推理经济性，复杂任务用强模型，普通任务用快模型
OpenAI	模型分层、prompt caching、fine-tuning/蒸馏、小模型替代、batch/latency optimization、自研推理芯片	从模型、API、产品、硬件全栈降低单位智能成本

资料来源：公司技术报告，招银国际环球市场

芯片升级推动单位功耗与单位美元产出改善。英伟达 1QFY27（年结于 1 月 25 日）Vera Rubin 平台系统推理吞吐量较前代提升最高 35 倍。专属推理芯片也在加速落地，据 OpenAI 与博通 2026 年 6 月 24 日联合公告，双方推出首款专为大语言模型推理设计的芯片 Jalapeño，早期测试显示其每瓦性能“大幅优于当前最先进水平”，博通 CEO 称该芯片可节省约 50% 成本。

软件栈与算子优化层面，模型、编译器和算子优化提升既有硬件的实际产出。英伟达 4QFY26 电话会，仅通过持续优化 CUDA 软件，GB200/ML72 系统性能最高提升 5 倍。微软 3QFY26（6 月 30 日年结）时称，经软硬件优化，Copilot 常用模型推理吞吐提升 40%，自研图像模型 MAI-Image 2 的 GPU 利用效率最高提升 260%；同时，自研 Maia 200 加速器较最新硅片每美元生成 token 数提升超 30%。据科大讯飞 2026 年 1 月公布，星火 MoE 训练效率对标 A800 由 30% 提升至 93%。算子优化方面，腾讯混元 AI Infra 团队 6 月 11 日宣布，通过开源 HPC-Ops 推理核心算子，针对真实负载下的计算不均问题采用运行时动态负载调度方案，实测长文本最高加速 2.95 倍，端到端 QPM 最高提升 17%；其 FusedMoE 模块通过整合多阶段流程、减少显存搬运与内核启动开销，相较主流框架性能提升 1.2 倍至 1.6 倍。

国产模型厂通过架构创新在推理侧系统性压低单位 token 成本。据 DeepSeek 2026 年 4 月 24 日发布的 V4 技术文档，其混合注意力机制 CSA+HCA 将长文本场景下的单 token 推理 FLOPs 和 KV cache 分别降至 V3.2 的 10%-27% 和 7%-10%。MiniMax 于 2026 年 6 月 1

日发布 M3 时，表示 M3 采用自研稀疏注意力架构 MSA，在 100 万 token 上下文规模时单 token 计算量仅为上一代模型 M2 的约 1/20，底层推理算子 KV outer gather Q 较 Flash-Sparse-Attention 等开源稀疏注意力实现性能提升 4 倍以上。

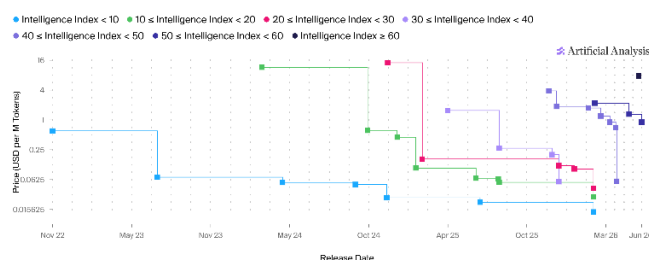
系统调度层面，资源调度优化有助于提升集群利用率和单位硬件产出。据华为云 2026 年 6 月公布，CCE VolcanoNext 统一调度引擎通过共享训推池与碎片化整合，资源利用率提升超 30%。阿里巴巴 4QFY26（财年结于 3 月）平头哥 GPU 超 60%算力服务外部客户，管理层表示服务器显卡“没有一张闲置”。Cloudflare CEO Matthew Prince 于 1Q26 电话会表示，多数超大规模云提供商的 GPU 利用率仅为个位数，而 Cloudflare 已经将 GPU 利用率提升至 70%-80%；其核心做法是根据不同 AI 推理任务特征动态路由至合适硅片，并利用网络中的闲置容量执行任务。据 SiTime 1Q26 电话会披露，公司指出推理工作负载中 GPU 利用率目前约为 20%-40%，而行业和公司目标是提升至 50%-60%，其中精确时间同步在提升 GPU 利用率过程中具有重要作用。

根据 Artificial Analysis 数据，在算力成本优化推动下，2Q26 不同智能能力区间的模型整体推理成本均展现优化趋势，其中模型能力介于 40-50 分之间的群组下降速度最为明显。与此同时，语言模型的输出速度也有所优化。以智能能力指数区间 40-50（包括阿里巴巴、MiniMax、DeepSeek、月之暗面和小米的旗舰模型）的模型统计，其区间内在推理调用价格方面表现最优模型的调用价格已由 2026 年 3 月的约 1.2 美元/百万 tokens 下探至 2026 年 6 月的约 0.058 美元/百万 tokens，对应降幅约 95%；而输出速度由约 150 tokens/秒提升至 200 tokens/秒左右，对应增幅约 33%。以智能区间 50-60（包括 Anthropic、OpenAI、智谱和 Google 的旗舰模型）的模型统计，其最优推理价格已由 2026 年 3 月的约 2.175 美元/百万 tokens 下探至 2026 年 6 月的约 0.902 美元/百万 tokens，对应降幅约 58.5%；而输出速度由约 160 tokens/秒提升至 185 tokens/秒左右，对应增幅约 16%。与此同时，我们观察到强智能能力的模型仍可持续维持较高定价甚至进一步提价。

图 8: 语言模型推理成本

Language Model Inference Price, by Intelligence Index Band, Over Time

Price in USD per 1M tokens (72:1 blend of cache, input, and output token prices). Bands use Artificial Analysis Intelligence Index v4.1.



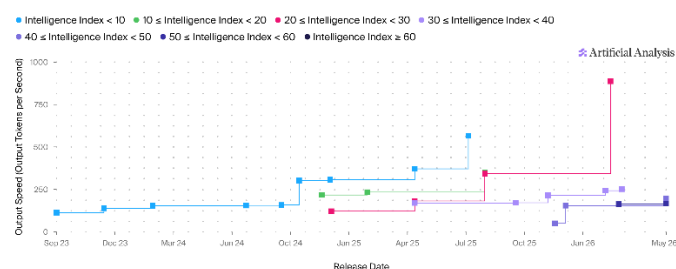
资料来源：Artificial Analysis, 招银国际环球市场

注：推理成本=Cache/输入/输出 token 价格按 7:2:1 权重加权后的每百万 tokens 成本

图 9: 语言模型输出速度

Language Model Output Speed, by Intelligence Index Band, Over Time

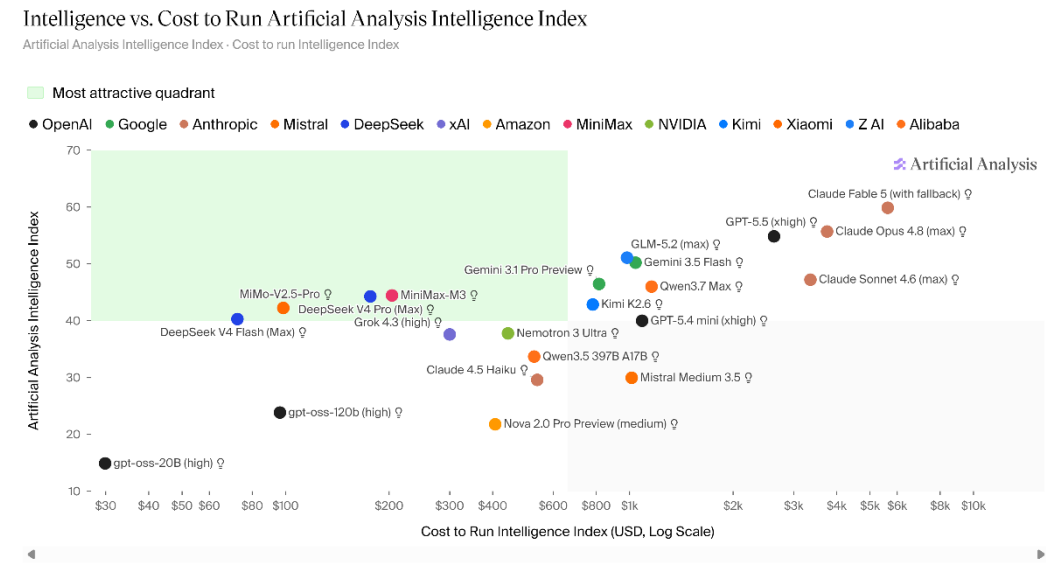
Output tokens per second. Bands use Artificial Analysis Intelligence Index v4.1.



资料来源：Artificial Analysis, 招银国际环球市场

从行业对比来看，基于不同模型的智能能力指数及其对应完成智能能力指数测试所需的成本综合衡量的成本效率而言，目前现有模型中，DeepSeek V4 Pro、DeepSeek V4 Flash、MiniMax M3、小米 MiMo-V2.5-Pro 处于最具吸引力的象限。其中 DeepSeek V4 Pro、MiniMax M3、MiMo-V2.5-Pro 为智能能力范围接近情况下成本更优的模型，MiMo-V2.5-Pro 作为非 Flash 版本仍实现较强的成本竞争力，体现出较高的效率优势。

图 10: 不同模型的智能能力指数及其对应完成智能能力指数测试所需的成本



资料来源: Artificial Analysis, 招银国际环球市场

注:运行成本根据模型的输入和输出 token 定价以及评估过程中使用的 token 数量计算得出

行业探索最新进展: 长程工作水平演进, 编程与多模态操作能力强化

截至 2026 年 6 月末, 横向对比 13 家中美前沿大模型最新版本进展, 我们观察到, 以真实 workflow 执行为导向的智能体能力已成为本轮大模型迭代的核心方向, 并向长程智能体演进, 长上下文有效性、工具调用稳定性与 token 经济性成为迭代方向。具体来看, 能力演进主要呈现三大趋势: 1) 强化长程工具调用与执行连贯性, 即围绕真实任务目标主动调用外部工具并在数百乃至上千次调用中维持目标连贯、跟踪状态、从错误中恢复, 例如阿里巴巴千问 3.7 Max 支持 35 小时无人值守与 1000+顺序工具调用, 小米 MiMo-V2.5-Pro 强化 "harness awareness" 以管理记忆并塑造自身上下文; 2) 重点提升编程与工程执行能力, 包括代码规划、软件工程、复杂系统执行及长程任务处理, 例如 Anthropic Claude Fable 5 在 SWE-Bench Pro 上达 80.3%、智谱 GLM-5.2 提升至 62.1 并专为长时程自主编程设计; 3) 强化原生多模态理解并向真实操作场景延伸, 即从被动看懂图文视频, 进阶到主动操作桌面/浏览器完成端到端任务, 例如 MiniMax M3 原生支持图像/视频输入且 OSWorld-Verified 达 70.06%, OpenAI GPT-5.5 强化计算机操作能力以支撑跨软件操作与复杂 workflow。

图 11: 前沿大模型最新进展及参数

厂商	最新公开旗舰模型	最新进展	较上一代版本的核心优化	更注重的功能方向	上下文窗口 (tokens)
Google	Gemini 3.5 / Gemini 3.5 Flash	2026 年 5 月 19 日 Google I/O 上发布 Gemini 系列更新; 官方称 Gemini 3.5 Flash 为"前高级智能、速度是同类型模型的 4 倍", 1M 上下文, Terminal-Bench 2.1 达 76.2%, 在 coding 和 agent 任务上超过 Gemini 3.1 Pro。	Gemini 3.5 在终端自动化和代码任务上超越了上一代旗舰 3.1 Pro; 配合内置的自动化思考架构与高达 6.5 万 Token 的单次输出, Flash 系列转变为专为高频、长任务智能体设计的完全体生产力引擎。	更偏向 agentic coding、多模态视频生成、个性化 agent、开发者 workflow。	1M
OpenAI	GPT-5.6 Sol/Terra/Luna	2026 年 4 月 23 日发布, 面向 ChatGPT 与 Codex 用户陆续开放; 官方定位为"专为实际工作打造"的新一代智能模型。	相比 GPT-5.5, GPT-5.6 编程和智能体能力进一步提升, GPT-5.6 Sol 在 Terminal-Bench 2.1 上取得新的 SOTA, 达到 88.8%。GPT-5.6 引入新的最高推理强度, 让 Sol 能够获得更多时间进行深度推理; 模型还推出了新的 ultra mode, 可以调用子 Agent 加速复杂任务。此外模型在生物学 workflow 和网络安全领域能力也有显著提升。	更偏向 agentic coding、computer use、在线调研、复杂数据分析、文档/表格/PPT 生成、跨软件操作、科学研究 workflow。	1M

Anthropic	Claude Fable 5 / Claude Opus 4.8	Claude Opus 4.8 于 2026 年 5 月 28 日发布；Claude Fable 5 和 Mythos 5 于 2026 年 6 月 9 日上线——Fable 5 在多平台发布，Mythos 5 仅向 Project Glasswing 审批客户限定开放；6 月 12 日因美国政府出口管制指令，Fable 5 和 Mythos 5 对所有用户暂停访问，预计 7 月 1 日前后对美国用户恢复，Opus 4.8/Sonnet 4.6/Haiku 4.5 不受影响	Anthropic 称“任务越长、越复杂，Fable 5 领先越大”；在 SWE-Bench Pro 上 Fable 5 达 80.3%、Opus 4.8 为 69.2%；Opus 4.8 相比 4.7 代码缺陷漏检率下降约 4 倍	更偏向 agentic coding、长时程任务执行、复杂软件工程、企业文档/金融/法律分析、高分辨率视觉理解	1M
xAI	Grok 4.3	Grok 4.3 是 xAI 最新旗舰模型，支持文本与图像输入，支持 1M 上下文，支持 function calling、structured outputs 和 configurable reasoning。	相比 Grok 4 / Grok 4.1 / Grok 4 Fast 系列，Grok 4.3 更强调行业领先的非幻觉率、agentic tool calling 和 instruction following；同时支持 no/low/medium/high 可配置推理强度，适合长上下文与工具调用任务。	更偏向低幻觉率、agentic tool calling、指令遵循、长上下文、多模态输入、结构化输出、可配置推理和低成本 API 调用。	1M
Meta	Muse Spark	2026 年 4 月 8 日发布，是 Meta Superintelligence Labs (MSL) 推出的首个模型，已首先用于 Meta AI app 与 Meta.ai，并计划扩展至 Facebook、Instagram、WhatsApp 等产品。	相比 Grok 4.2，Grok 4.3 更强调性价比，API 可用性、较长上下文和开发者调用；第三方评测认为其较前代提升明显。	更偏向低成本 API、长上下文、工具调用、多模态输入、语音/Custom Voices 生态。	260k
智谱 Z.AI	GLM-5.2	2026 年 6 月 16 日发布 GLM-5.2，753B 参数开源权重模型，专为长时程自主编程和工程任务设计，已在 Hugging Face、Z.ai API 和 20+ 第三方编程环境上线，企业订阅起价 \$12.60/月，核心权重以 MIT 开源协议发布	753B 参数，引入 "IndexShare" 架构——在每四层稀疏注意力间复用同一 index，在 1M 上下文下将单 token FLOPs 降低 2.9 倍；升级 MTP 层用于推测解码，接受 token 长度提升达 20%；Terminal-Bench 2.1 从 GLM-5.1 的 62.0 提升至 81.0，SWE-bench Pro 从 58.4 提升至 62.1	长时程编程 Agent、企业工程级代码自动化、低成本开源自托管部署	1M
稀宇科技 MiniMax	MiniMax M3	2026 年 6 月 1 日正式发布 M3，API 同日上架，模型权重和技术报告承诺 10 天内开源	引入 MSA (MiniMax Sparse Attention) 架构，在 1M 上下文下相对 M2 实现 >9x prefill 和 >15x decode 加速、单 token 算力降至 1/20；SWE-Bench Pro 达 59.0%，超越 GPT-5.5 和 Gemini 3.1 Pro；原生多模态——支持图像、视频输入，OSWorld-Verified (计算机使用) 达 70.06%	智能体场景任务处理、工具调用稳定性、多步推理、长上下文真实任务执行	1M
小米	MiMo-V2.5/ MiMo-V2.5 Pro	2026 年 4 月 23 日正式发布；是小米推出的首个开源权重旗舰级大模型	相比前代 MiMo-V2-Pro，总参数维持 1.02T/激活 42B 量级，沿用混合注意力架构并将 SWA 与 GA 以 6:1 交错、滑窗 128，KV-cache 存储降低近 7 倍；将多模态与 Pro 级推理统一于单一架构而非外挂编码器；后训练强化 "harness awareness"，以“先规划再行动”模式减少无效尝试，使完成同等任务 token 消耗较前代降低 40-60%；引入 FP4 混合量化+DFlash 并行投机解码与 TileRT 系统级常驻内核优化 GPU 效率；1M 长上下文从“标称”做到“可用”，在 512k/1M GraphWalks 基准上仍保持有效推理。	侧重智能体长程任务的稳定执行与 token 经济性，原生支持每会话 1000+ 自主 tool calls，可在数百次工具调用中维持目标连贯、跟踪状态并从错误中恢复；统一原生处理图像/视频/音频；在 agentic 场景下显著提升指令遵循能力，可靠遵守上下文中的细微约束。	1M
阶跃星辰 StepFun	Step 3.7 Flash	据公司官方发布，Step 3.7 Flash 为其最新旗舰推理/Agent 开源底座模型；稀疏 MoE 架构，196B 总参/激活 11B	强调“效率型跃迁”——通过 MTP-3 和稀疏 MoE，在接近顶级闭源推理深度的同时仍保持面向实时交互的生成速度，最高生成速度 400 Tokens/s	更偏向高复杂度推理、编程、智能体工具使用、搜索、办公任务，以及与吉利的车载 AI 场景	256K
月之暗面	Kimi K2.7 Code	2026 年 6 月 12 日发布 Kimi K2.7 Code，是基于 K2.6 架构的编程专用变体；1T 参数 MoE 编程模型 (32B 激活，384 专家)，256K 上下文，Modified MIT 开源协议，相比 K2.6 使用约 30% 更少“思考”tokens 的同时编程基准提升	Moonshot 报告相比 K2.6，K2.7 在 Kimi Code Bench v2 上 +21.8%、Program Bench 上 +11.0%、MLS Bench Lite 上 +31.5%，推理 token 使用量降低约 30%；在 MCP Mark Verified 上击败 Opus 4.8 (81.1 vs 76.4)，工具调用准确率更强	更偏向长程 coding、Agent 工具调用、MCP 协议、Kimi Code CLI 开发者工作流	256K
幻方量化 DeepSeek	DeepSeek-V4-Pro / DeepSeek-V4-Flash	2026 年 4 月 24 日发布 DeepSeek-V4 预览版，API 已上线 V4-Pro 与 V4-Flash，并同步开源	相比前代模型，V4-Pro 的 Agent 能力、世界知识和推理性能显著提升；V4-Flash 相比 V4-Pro 更快捷、经济，在简单 Agent 任务上表现接近 V4-Pro，但高难度任务仍有差距。DeepSeek-V4 采用新的注意力机制与 DSA 稀疏注意力，在降低计算和显存需求的同时提升长上下文效率。	Agentic Coding、主流 Agent 产品适配、工具调用、思考/非思考模式、数学/STEM/竞赛代码推理、世界知识、代码与文档生成。	1M
字节跳动 豆包	Seed 2.1 Pro / Seed 2.1 Turbo	字节在 2026 年 6 月 23 日发布 Seed 2.1 系列，提供早期访问；官方称 Doubao-Seed-2.1 是面向 Coding 和 Agent 时代打造的新一代大模型，提供 Pro 和 Turbo 两个版本，分别面向高复杂度任务探索和规模化生产场景	相较上一代版本，Seed 2.1 在 Coding 工程交付、Agent 长链路任务执行和多模态理解三大方向上实现全面升级	适用于复杂 Coding、长链路 Agent、多步骤工程交付等高复杂度任务探索及高价值生产场景。	256k

阿 里 巴 巴 千 问	Qwen 3.7 Max / Qwen 3.7 Plus	阿里在 2026 年 5 月 20 日杭州 Apsara Cloud Summit 正式发布 Qwen3.7-Max; 2026 年 6 月 1 日补发 Qwen 3.7 Plus, 与 Max 同样 1M 上下文+35 小时自主运行上限, Plus 定价 2 元/M input (约为 Max 的 1/6) 并新增视觉能力; 两款均为 闭源 API-only	官方定位"Agent Frontier", 专为长时程自主 Agent workflow 设计, 支持运行数小时、调用上百到上千次工具而不丢失上下文; 35 小时无人值守运行、1,000+顺序工具调用、Kernel Bench 96%胜率, 1.98× GPU 加速; GPQA Diamond 92.4%、HMMT 2026 年 2 月 97.1%、SWE-Verified 80.4%	更偏向 agentic coding、35 小时长程自主任务、数学/竞赛、Kernel 优化、跨 harness 泛化	1M

资料来源: 公司资料, Hugging Face, Artificial Analysis, Reuters, 招银国际环球市场

大语言模型调用量: 2Q26 token 消耗持续上升, 中国开源模型贡献主要增量

我们测算全球月均 AI Token 消耗量将在 2030 年超过 26 万万亿, 2025-2030 年复合增速为 107%, 核心增长驱动力包括: 1) 模型智能水平提升带动 AI 产品总用户规模增长以及用户使用粘性提升; 2) 多模态输入/输出占比提升推动单次 Prompt 的 Token 消耗量增加; 3) 底层模型能力提升+Harness 框架完善带动 Agent 产品渗透率持续提升, Claude Code、Codex 等 Agent 产品的推出让 2026 年成为 Agent 落地的关键年, 由于 Agent 产品的更高的 Token 消耗量, 我们预计 Agent 场景将成为 2025-2030 年 Token 消耗量增长的核心驱动力, 预计 2030 年全球 Agent 场景月均 Token 消耗量将超过 21 万万亿 (2025-2030 年复合增速 165%), 占全球 AI Token 消耗量约 84%。

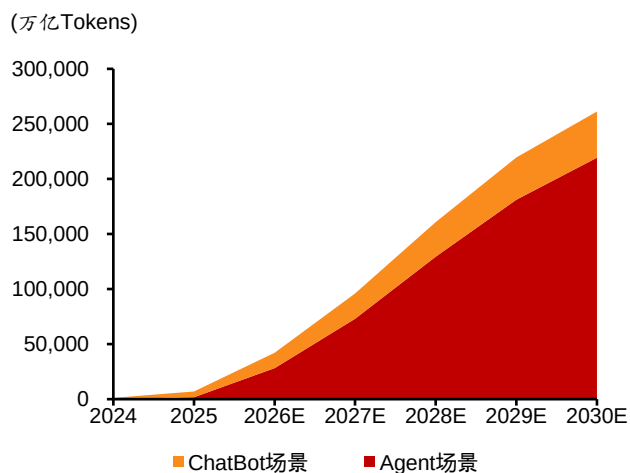
从 Token 消耗趋势看互联网软件板块投资机会, 我们建议关注几条主线: 1) 基础设施云厂商和头部基础模型厂商, 作为 AI 算力工厂, 将持续受益于 AI Token 消耗量的快速增长; 2) 头部编程/办公/企业 AI Agent 应用厂商, 具备产品、Harness 框架、用户与数据资产优势的 Agent 厂商有望受益于 Agent 产品渗透率的提升和 Agent Token 消耗量的快速增长; 3) 头部多模态模型与多模态 AI 应用厂商, 多模态输入/输出仍是长期趋势, 潜在价值量创造和 Token 消耗量仍有较大增长空间。

图 12：全球及中国：大模型 Token 消耗量测算

全球	2024	2025	2026E	2027E	2028E	2029E	2030E	核心驱动因素
全球 AI 日活跃用户数(百万)	250	700	1,260	1,512	1,739	1,913	2,008	AI 工具渗透率&用户粘性
YoY		180%	80%	20%	15%	10%	5%	持续提升
ChatBot 场景(含多模态输入/输出)								
人均每日对话次数	25	30	36	40	42	43	43	AI 使用场景进一步丰富&
YoY		20%	20%	10%	5%	3%	1%	用户粘性提升
平均每次对话 Token 消耗量	6,000	8,000	10,000	12,500	14,000	15,120	15,574	模型智能水平提升&多模
YoY		33%	25%	25%	12%	8%	3%	态输入/输出占比提升
ChatBot 场景月均 Token 消耗量(万亿)	1,163	5,208	14,062	23,202	31,378	38,395	41,940	
YoY		348%	170%	65%	35%	22%	9%	
Agent 场景								
Agent 产品用户量	5	56	252	363	487	574	643	AI Agent 产品渗透率提
YoY		1020%	350%	44%	34%	18%	12%	升
Agent 产品渗透率%	2%	8%	20%	24%	28%	30%	32%	
人均每日分派任务数	8	16	24	29	33	36	38	Agent 能力提升&用户粘
YoY		100%	50%	20%	15%	10%	5%	性提升
平均每次任务 Token 消耗量	20,000	60,000	150,000	225,000	258,750	279,450	287,834	Agent 能力提升，处理更
YoY		200%	150%	50%	15%	8%	3%	复杂的任务
Agent 场景月均 Token 消耗量(万亿)	25	1,667	28,123	72,895	129,342	181,097	219,360	
YoY		6620%	1588%	159%	77%	40%	21%	
全场景								
全球月均 AI Token 消耗量(万亿)	1,187	6,875	42,185	96,097	160,720	219,493	261,299	
YoY		479%	514%	128%	67%	37%	19%	
中国	2024	2025	2026E	2027E	2028E	2029E	2030E	
中国 AI 日活跃用户数(百万)	50	140	252.0	315.0	378.0	434.7	478.2	AI 工具渗透率&用户粘性
YoY		180%	80%	25%	20%	15%	10%	持续提升
ChatBot 场景(含多模态输入/输出)								
人均每日对话次数	25	30	36	40	42	43	43	AI 使用场景进一步丰富&
YoY		20%	20%	10%	5%	3%	1%	用户粘性提升
平均每次对话 Token 消耗量	6,000	8,000	10,000	12,500	14,000	15,120	15,574	模型智能水平提升&多模
YoY		33%	25%	25%	12%	8%	3%	态输入/输出占比提升
ChatBot 场景月均 Token 消耗量(万亿)	233	1,042	2,812	4,834	6,821	8,726	9,986	
YoY		348%	170%	72%	41%	28%	14%	
Agent 场景								
Agent 产品用户量	1	11	50	76	106	130	153	AI Agent 产品渗透率提
YoY		1020%	350%	50%	40%	23%	17%	升
Agent 产品渗透率%	2%	8%	20%	24%	28%	30%	32%	
人均每日分派任务数	8	16	24	29	33	36	38	Agent 能力提升&用户粘
YoY		100%	50%	20%	15%	10%	5%	性提升
平均每次任务 Token 消耗量	20,000	60,000	150,000	225,000	258,750	279,450	287,834	Agent 能力提升，处理更
YoY		200%	150%	50%	15%	8%	3%	复杂的任务
Agent 场景月均 Token 消耗量(万亿)	5	333	5,625	15,187	28,118	41,159	52,229	
YoY		6620%	1588%	170%	85%	46%	27%	
全场景								
中国月均 AI Token 消耗量(万亿)	237	1,375	8,437	20,020	34,939	49,885	62,214	
YoY		479%	514%	137%	75%	43%	25%	

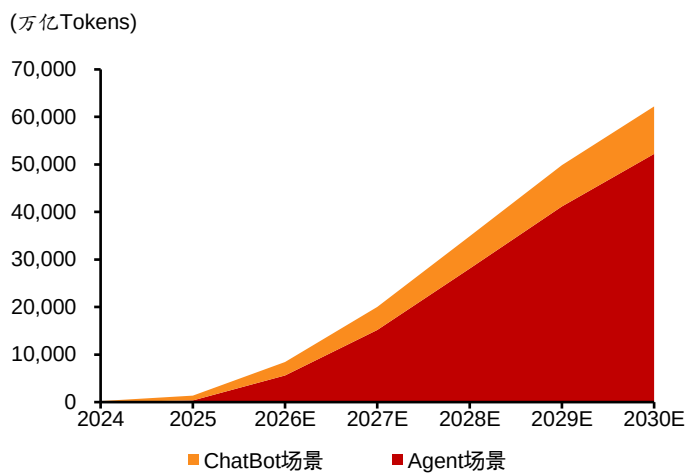
资料来源：公司资料、OpenRouter、QuestMobile、Iternal.ai、招银国际环球市场预测

图 13: 全球 AI Token 月均消耗量趋势



资料来源：公司资料、OpenRouter、QuestMobile、Iternal.ai、招
银国际环球市场预测

图 14: 中国 AI Token 月均消耗量趋势



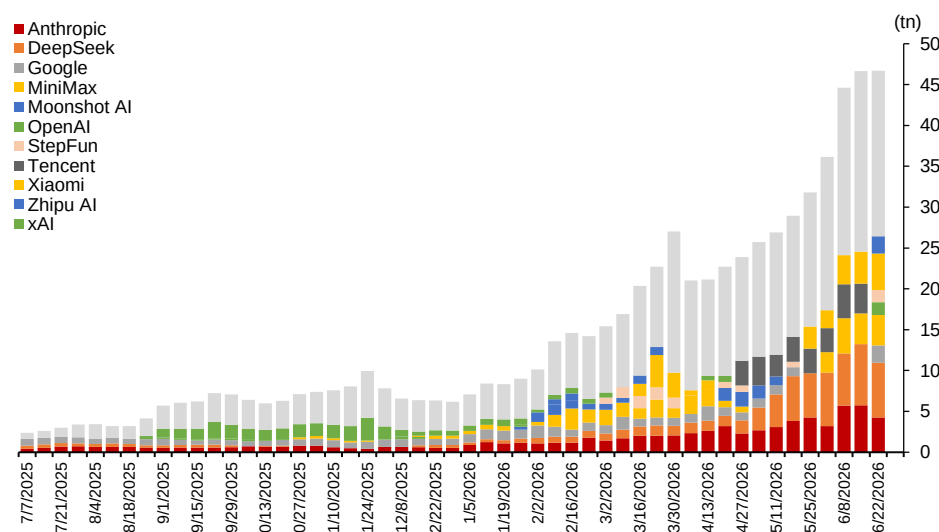
资料来源：公司资料、OpenRouter、QuestMobile、Iternal.ai、招
银国际环球市场预测

短期来说，我们以 OpenRouter 平台的周度 token 调用量作为大模型应用侧需求趋势变化的高频参考。截至 2026 年 6 月 15 日当周，平台总调用量为 46.66 万亿 tokens，较一年前（2025 年 6 月 30 日当周）的 2.11 万亿增长约 22 倍。

总量持续上升，2Q 增速加快。平台周度调用量在 2025 年下半年维持在 2–6 万亿 tokens；2026 年起读数走高，1 月 26 日当周为 8.25 万亿，2 月 9 日升至 13.01 万亿；2Q26 内由 4 月初的约 21 万亿升至 6 月 15 日的 46.66 万亿，单季约翻倍。我们认为这与模型迭代转向 Agent 与编程场景、单任务 token 消耗上升有关——长程工具调用、多步推理与 coding 工作流的使用，抬高了单位用户的 token 消耗水平。

中国开源模型贡献主要增量。截至 6 月 22 日当周，DeepSeek 以 6.7 万亿 tokens 居平台首位，高于 Anthropic 的 4.25 万亿；小米（4.48 万亿）、MiniMax（3.74 万亿）、智谱（2.11 万亿）等中国厂商合计占比亦较高。DeepSeek 的调用量上升与其 2Q 内 V4 系列发布、API 降价 75% 的时点一致。2025 年下半年平台调用量主要来自 Google、Anthropic 与 xAI（xAI 在 2025 年 11 月 24 日当周一度达 2.76 万亿）；2026 年以来，MiniMax、腾讯、小米、智谱等厂商先后进入周度前十；自 5 月 4 日当周起，DeepSeek 调用量达 2.79 万亿 tokens，超过 Anthropic 的 2.69 万亿升至首位并保持领先，中国厂商合计调用量已明显超过美国厂商。我们认为，这是因为中国开源模型在性能接近的同时、价格显著更低，对于价格敏感的开发群体而言性价比较高。

图 15: OpenRouter 平台周度大语言模型调用量



资料来源：OpenRouter, 招银国际环球市场

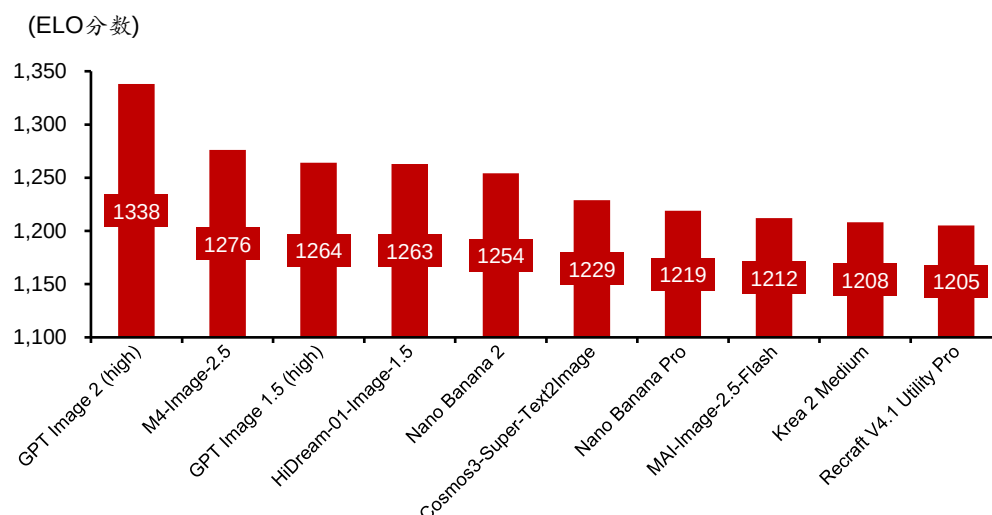
注：1) OpenRouter 为面向开发者的第三方模型聚合/路由平台，仅覆盖经由该平台分发的调用量，不含厂商自有 API 及其他分发渠道；2) 该榜单每周仅列示调用量排名前十的厂商，未进入当周前十的厂商不单独显示，其调用量计入“Others”。

图像模型仍由海外厂商领跑，但视频赛道中国厂商更具优势

我们观察到图像模型的功能演进也同样转向 workflow 执行，重点包括：1) 多轮编辑和可控性继续增强，例如 4 月发布的 Midjourney V8.1 强调更好的提示词遵循度和细节一致性，同样 4 月发布的 OpenAI GPT Image 2 可在四轮连续编辑下人脸与未触及区域保持稳定（无漂移和未编辑区域的二次重绘）；2) 图像模型进一步嵌入创意生产流程，例如 Adobe 在 6 月将 Creative Agent 扩展至 Firefly 和 Creative Cloud，覆盖 Photoshop、Premiere、InDesign 等应用，强化从生成、编辑到交付的工作流自动化。

图像模型方面海外厂商仍掌握第一梯队主导权，中国厂商已经具备强竞争力。文生图（text-to-image）领域，基于 Artificial Analysis, 美国厂商中 OpenAI GPT Image 2 (high)、Microsoft M4 Image 2.5、OpenAI GPT Image 1.5 (high)、Nvidia Cosmos3 Super Text2Image、Google Nano Banana 2/Pro 图像模型能力领先行业，中国厂商则以智象未来 HiDream O1 Image 1.5 领先。此外，从图像模型的性价比来看（质量 vs. API 价格），智象未来 HiDream O1 Image 1.5、Google Nano Banana 2/Pro、OpenAI GPT Image 1.5 (high) 处于最优象限，即相对高质量、低价格区间。

图 16: 主流图像模型文生图能力排名



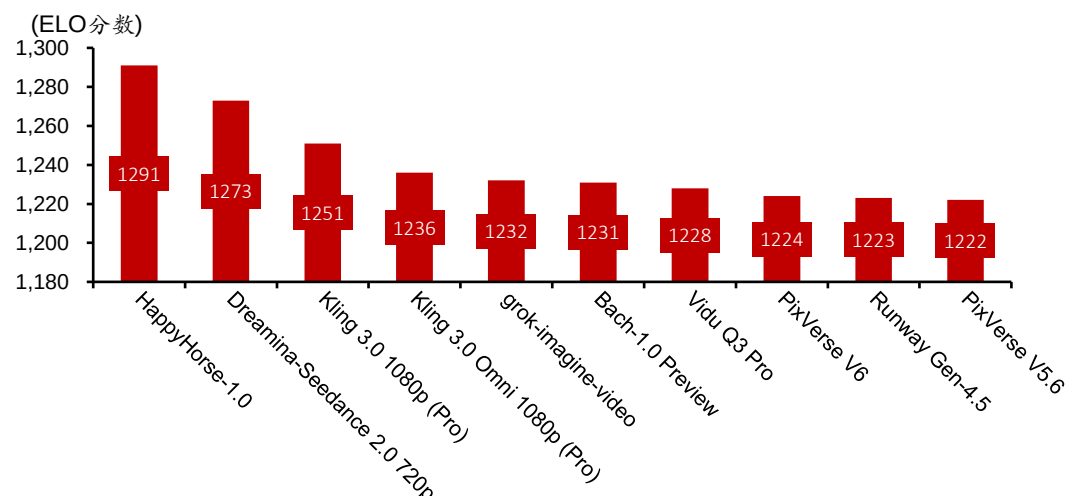
资料来源: Artificial Analysis, 招银国际环球市场

注: 1) Artificial Analysis 图像模型文生图能力 ELO 评分指图像生成质量相对度量指标, 越高代表能力越强; 2) ELO 评分由 Artificial Analysis Image Arena 中数百万次用户投票计算得出; 3) 排名并未穷尽所有大模型

视频模型的功能演进方面, 我们观察到三个主要趋势: 1) 强调更强的镜头可控性与角色一致性, 例如 Runway Gen-4.5 支持更精准的镜头控制、多镜头叙事和跨镜头角色一致性; 2) 原生音视频一体化继续推进, 例如 Seedance 2.0 开放 API 及 1080P 原生生成能力, 延续文本、图像、音频、视频多模态输入和音视频联合生成方向; 3) 高清化与生产级输出能力提升, 例如快手 Kling 3.0 支持更高分辨率、长时长和多镜头分镜, 并强化原生音频、角色声音/外观克隆等能力, 使视频模型从单段生成逐步转向可用于短剧、广告、社媒视频等内容生产 workflow。

中国视频模型凭借用户规模、数据等优势形成较强竞争力。以字节跳动即梦 AI 为例, 截至 2026 年 3 月, 根据 QuestMobile 数据, 即梦 AI MAU 规模达 1,352 万, 形成较大的用户规模与视频数据优势。文生视频领域 (text to video), 基于 Artificial Analysis, 阿里巴巴旗下 HappyHorse-1.0 位居榜首, 与字节跳动即梦 Seedance 2.0 720p、快手可灵 3.0 1080p (Pro) 形成行业第一梯队, 美国厂商则由 xAI Grok-imagine-video、Runway Gen-4.5 领先。从性价比来看 (质量 vs. API 价格), 字节跳动即梦 Seedance 2.0 720p 为最优象限中性价比最高的模型。

图 17: 主流视频模型文生视频能力排名

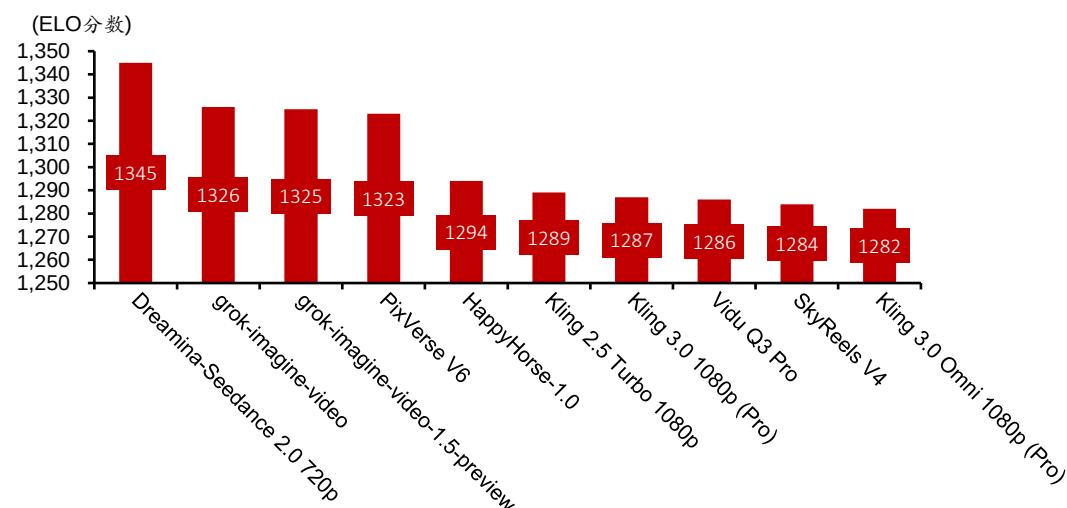


资料来源: Artificial Analysis, 招银国际环球市场

注: 1) Artificial Analysis 视频模型能力 ELO 评分指视频生成质量相对度量指标, 越高代表能力越强; 2) ELO 评分由 Artificial Analysis Video Arena 中用户投票计算得出; 3) 排名并未穷尽所有大模型。

图生视频 (image to video no audio) 方面, 字节跳动即梦 Dreamina- Seedance 2.0 720p 位居行业第一, 美国厂商以 xAI Grok imagine-video/1.5 preview 领先, 其后爱诗科技 PixVerse V6、HappyHorse-1.0、快手可灵 Kling 2.5 Turbo、生数科技 Vidu Q3 Pro 及昆仑万维 SkyReels V4 等多个中国厂商形成行业第一梯队。

图 18: 主流视频模型图生视频能力排名



资料来源: Artificial Analysis, 招银国际环球市场

注: 1) Artificial Analysis 视频模型能力 ELO 评分指视频生成质量相对度量指标, 越高代表能力越强; 2) ELO 评分由 Artificial Analysis Video Arena 中用户投票计算得出; 3) 排名并未穷尽所有大模型。

龙头基础模型平台模型及业务进展更新

海外 AI 基础模型平台：ARR 竞赛持续，Anthropic 增长领跑，进入 IPO 筹备阶段

2Q26 海外 AI 平台竞争格局出现三点边际变化，对比 OpenAI、Anthropic、xAI：1) Anthropic ARR 反超 OpenAI，媒体报道其预计在 2Q26 实现首个盈利季度；2) 三家龙头先后启动 IPO 进程，估值锚开始转向公开市场；3) xAI/SpaceX 以算力租赁形式打开新的收入曲线。

从收入体量与增速看，Anthropic 已超过 OpenAI。据 Anthropic 2026 年 5 月 28 日 Series H 公告，公司 ARR 已达约 470 亿美元，CFO Krishna Rao 公布的爬坡轨迹为：2025 年 12 月 90 亿→2026 年 2 月 140 亿→3 月 190 亿→4 月 300 亿→5 月初 440 亿。OpenAI 侧，据公司 5 月披露，ChatGPT 周活 9 亿、ARR 约 250 亿美元、付费消费者订阅 5,000 万；据 Reuters 2026 年 6 月报道，ChatGPT App 跨过 10 亿 MAU。尽管 Anthropic 和 OpenAI 的 ARR 或存在统计口径差异，但 Anthropic 的 ARR 体量已经超过 OpenAI。xAI 仍是三家中规模最小者：据 SpaceX 招股书披露，2025 年 xAI 亏损 64 亿美元、收入 32 亿美元；1Q26 xAI 单季收入 8.18 亿美元、运营亏损 24.7 亿美元。

从盈利可见度看，Anthropic 预计成为首家实现单季度盈利的前沿模型公司。据《华尔街日报》报道，Anthropic 预计 2Q26 收入 109 亿美元（环比 1Q26 的 48 亿美元接近翻倍），并实现 5.59 亿美元经营利润。OpenAI 侧，据 The Information 报道，2026 年现金消耗预计约 270 亿美元，2027 年进一步升至 630 亿美元；据 European Business Magazine 分析，OpenAI 每 1 美元收入对应当季约 1.22 美元亏损。xAI 侧，据 Sacra 估算，独立 AI 业务年化亏损规模超 120 亿美元，需要 SpaceX 资产负债表支撑。

三家龙头的估值体系将逐步由一级市场切换至公开市场。2Q26 三家公司 IPO 进程均有所更新：SpaceX（合并 xAI）于 6 月 12 日在纳斯达克上市，6 月 24 日市值达到 2.02 万亿美元；Anthropic 于 6 月 1 日以 9,650 亿美元估值保密递交 IPO 申请、累计融资约 1,250 亿美元；据 OpenAI 6 月 8 日官方公告，公司已提交保密 S-1，但表态“时间未定”。

我们认为，基础模型后续的关注点和催化主要仍集中于：1) 模型迭代与智能能力的提升，Google 有望 7 月发布 Gemini 3.5 Pro；2) Anthropic 2Q26 运营盈利预期能否如期兑现；3) 商业化节奏和创新业务探索，如 OpenAI 广告与 Codex 两条新曲线对 Q3/Q4 ARR 的实际贡献度，及智能硬件业务的变现探索。

图 19：海外核心 AI 基础模型平台：财务及运营数据对比

海外三大 AI 基础模型公司：关键指标对比			
维度	Anthropic (Claude)	OpenAI (ChatGPT)	xAI (Grok)
规模与增长			
最新 ARR	~470 亿美元 (Anthropic Series H 融资公告, 2026.5)	~250 亿美元 (OpenAI 公告, 2026.2)	xAI 1Q26 收入 8.18 亿美元 (SpaceX 招股书)；独立 AI 业务 ARR 约 5 亿美元 (Sacra 估算, 2026.4)
增长轨迹	2025 年底 90 亿→2026 年 5 月中 470 亿，5 个月约 4 倍；爬坡 90 亿(12 月)→140 亿(2 月)→190 亿(3 月)→300 亿(4 月)→470 亿美元 (5 月中) (Anthropic CFO Krishna Rao 披露)	ChatGPT 周活 9 亿+ (OpenAI 披露, 2026.2)；App MAU 达 10 亿 (Reuters, 2026.6)；付费消费者 5,000 万+，企业付费用户 900 万+，企业 seats 700 万+ (OpenAI 披露, 2026.2)	Grok MAU 1.17 亿、合并 X 后总 MAU 5.5 亿、SuperGrok 付费 190 万、含 Grok 访问的 X 付费用户 440 万 (SpaceX 招股书, 2026.3)

ARR 目标	2026 年 200–260 亿美元目标已于 4 月提前达成；2028 年目标收入 700 亿美元 、现金流 170 亿美元 （The Information）	2026 年全年 300 亿美元 （OpenAI 披露，2026.5）；2030 年长期目标 3,900 亿美元 （Reuters）	2026 年目标 20 亿美元 ；目标 2027 年盈利（Sacra）
收入结构与核心产品			
引擎①	Claude Code ARR 25 亿+ ，企业端订阅自年初增长 4 倍 并贡献过半收入；全球~4% 公开 GitHub commits 由其生成（Anthropic 公告，2026.2）	Codex（GPT-5.5）周活 400 万+ （公司公告，2026.4）；API 每分钟处理 150 亿+ token（CNBC，2026.6）	Grok 4 API 定价\$3/\$15 每百万 token；Grok 4.1 Fast \$0.20/\$0.50，主流前沿模型最低价（2026.6）
编程/开发			
引擎②	Cowork 定位“Claude Code for general computing”，2026 年 1 月上市首月推出 30+ 功能、11 个开源插件，可运行于 Excel/PPT	ChatGPT 订阅：Go \$8/Plus \$20/Pro \$200 每月；周活免费用户付费转化率仅 -5% （2025.7）	2025 年 Grok+X 订阅收入 3.65 亿美元 ，1Q26 再增 1.77 亿美元 （SpaceX 招股书）
通用/办公			
引擎③	Claude Fable 5 / Mythos 5 上下文窗口 100 万 token，定价\$10/\$50 每百万 token，Mythos Preview 价格减半（Anthropic 公告，2026.6）	广告业务面向 Free/Go 上线，初始 CPM~60 美元、最低投放 20 万美元（公司公告，2026.2）；pilot 6 周内达 1 亿美元 ARR，首批含 Target/Ford/Adobe（The Information，2026.3）	算力租赁新收入线：Anthropic 每月支付 xAI（SpaceX） 12.5 亿美元 租赁 Colossus 1（300MW、22 万张 H100/H200/GB200 GPU），合约至 2029.5（SpaceX 招股书）
新曲线			
盈利与成本			
盈利节奏	1Q26 收入 48 亿美元 ；2Q26 预计收入 109 亿美元 、经营利润 5.59 亿美元 ，或为首个盈利季度（WSJ）；2027 年现金流转正、2028 年现金流 -170 亿美元 （The Information）	2026 年现金消耗 -270 亿美元 ，2027 年升至 630 亿美元 （The Information）；每 1 美元收入对应约 1.22 美元亏损（European Business Magazine）	xAI 2025 年亏损 64 亿 /收入 32 亿 （SpaceX 招股书）；1Q26 单季运营亏损 24.7 亿 /收入 8.18 亿 ；目标 2027 年盈利（管理层披露）
毛利率演进	2024 年-94%→2025 年目标 40%→2028 年目标 77% （The Information）；推理毛利率从上年 38% 提升至 70%（Semi Analysis）	整体毛利率 33%；推理成本 2025 年 84 亿→2026 年预计 141 亿美元 （The Information）	AI segment 1Q26 capex 77 亿美元 ，年化~308 亿、同比翻倍以上（SpaceX 招股书）
估值与融资			
最新估值	9,650 亿美元 （H 轮融资投后估值）（2026.5，公司公告）；累计融资 1,250 亿美元（2026.6，Sacra）	8520 亿美元 （投后估值）（2026.3，公司公告）；累计融资 1220 亿美元（2026.3，公司公告）	SpaceX 全股票收购 xAI，xAI 估值 2,500 亿 （2026.2，CNBC）；SpaceX 最新市值 2.07 万亿美元 （2026.7）
估值倍数	按 2026 年 5 月的 470 亿 ARR 和 9,650 亿估值计算： -20.5x	按 2026 年 2 月的 250 亿 ARR 和 8,520 亿估值计算：约 -34.1x	按 2026 年 2 月的 xAI 单体估值 2500 亿和 2026 年 ARR 目标 20 亿估算：约 -125x
上市进度	已提交保密 IPO 申请（2026.6，公司公告）；最早可能 2026 年底上市（2026.6，WSJ）	已提交保密 IPO 申请（2026.6，公司公告）	SpaceX 于 2026.6.12 在 Nasdaq 挂牌、股票代码 SPCX；IPO 定价\$135/股，募资 750 亿美元

资料来源：公司资料、Reuters、CNBC、The Information、TechCrunch、WSJ、Sacra、Semi Analysis、European Business Magazine、招银国际环球市场

注：数据截至 2026 年 7 月 1 日

■ Anthropic：ARR 突破 470 亿美元，由 Claude Code 及模型能力提升双引擎驱动

我们认为，Anthropic 在本轮基础模型竞争中商业化效率与盈利兑现节奏均处于前列。2Q26 内，公司有三个关键性的事件值得关注：ARR 超越 OpenAI（尽管双方 ARR 统计口径或存在不一致的情况）、媒体报道公司 2Q26 有望实现季度盈利、及保密递交 IPO 申请。

ARR 5 个月增长约 4 倍，企业收入占比 80%。 据公司披露，Anthropic ARR 已从 2025 年底的约 90 亿美元提升至 2026 年 5 月中的约 470 亿美元，爬升轨迹为\$8,700 万(2024.1)→\$10 亿 (2024.12)→\$90 亿 (2025 年底)→\$140 亿 (2026.2)→\$190 亿 (2026.3)→\$300 亿 (2026.4)→\$470 亿 (2026.5)。据公司披露，企业收入占比达 80%，显著高于 OpenAI；2026 年 2 月 Series G 时年付费 100 万美元以上的企业客户为 500 家，至 5 月已突破 1,000 家。

Claude Code 持续放量。 据 Anthropic 2026 年 2 月公告，Claude Code 年化收入规模已超过 25 亿美元，周活用户自年初以来翻倍；据公司披露，全球约 4% 的公开 GitHub commits

由 Claude Code 生成（环比上月翻倍）；企业订阅自年初增长 4 倍，企业用户已贡献 Claude Code 过半收入。

Cowork：办公场景延伸。 据公司 2026 年 1 月发布会披露，单月推出 30 余项产品/功能，Cowork 将 Claude Code 的工程能力延伸至更广泛的知识工作场景，目前已可直接运行于 Excel/PowerPoint，是公司在通用办公方向的核心产品。

Claude Fable 5 / Mythos 5：首个公开发布的 Mythos 级模型，API 价格减半。 据 Anthropic 2026 年 6 月 9 日发布会，公司发布 Claude Fable 5：首个公开可用的 Mythos 级模型，同步推出 Mythos 5 用于政府相关网络安全场景。据官方文档，该系列默认上下文窗口 100 万 token、单次最高输出 12.8 万 token。API 定价由 Mythos Preview 的 \$25/\$125 降至 Fable 5 的 \$10/\$50 每百万 token。能力方面，公司称 Fable 5 在软件工程、知识工作、视觉、科学研究等几乎所有已测试基准上达到 SOTA，且任务越长、越复杂，相对自家其他模型的领先幅度越大，典型基准 SWE-Bench Pro 得分达到约 80%、Hex 核心数据分析基准首个突破 90%（较 Opus 4.8 提升 10 个百分点），以及《Slay the Spire》长上下文一致性评测中到达最终幕的频率为 Opus 4.8 的三倍；与之同源的 Mythos 5 解除了部分领域护栏，网络安全能力为全球所有模型中最强，初期通过 Project Glasswing 与美国政府合作部署。Fable 5 与 Mythos 5 于 2026 年 6 月 12 日因美国商务部出口管制令被全球下线，后续 7 月 1 日，Anthropic 宣布美国商务部正式撤销对 Anthropic 旗下模型 Fable 5 以及 Mythos 5 的出口管制，并将在次日恢复访问。

Sonnet 5：更具智能体属性的性价比模型。 Anthropic 7 月 1 日发布全新模型 Claude Sonnet 5，模型在推理、工具使用、编程和知识工作等智能体能力方面，相比 Sonnet 4.6 性能有显著提升，更接近 Opus 4.8，例如 Sonnet 5 在 SWE-Bench Pro 得分达到约 63%，显著高于 Sonnet 4.6（58%），接近 Opus 4.8（69%）。定价方面，Claude Sonnet 5 标准定价为输入 3 美元/百万 token，输出 15 美元/百万 token，对比 Claude Opus 4.8 输入 5 美元/百万 token，输出 25 美元/百万 token。

算力供给：与 xAI（SpaceX）的合作打开 GPU 限制瓶颈。 据 SpaceX 招股书披露，2026 年 5 月，SpaceX 将 Colossus 1 数据中心算力（超 300MW、22 万张 H100/H200/GB200 GPU）整体租赁给 Anthropic，xAI 则将训练负载迁至 Colossus 2；Anthropic 按月支付 12.5 亿美元，任一方 90 天通知终止协议，SpaceX 正寻求复制类似交易。该安排同时缓解 Anthropic 短期 GPU 瓶颈、并为 xAI/SpaceX 打开新的可见度收入。

盈利与估值： 据 Semi Analysis，Anthropic 推理毛利率已从上年的 38% 提升至 70%。据 TechCrunch 援引公司财务路演材料，公司预计 2028 年实现 70 亿美元规模正现金流，毛利率提升至 77%；《华尔街日报》报道公司 2Q26 有望实现运营盈利。2026 年 6 月 1 日，Anthropic 以 9,650 亿美元估值保密递交 IPO 申请，累计融资约 1,250 亿美元；按 ARR 测算估值倍数仅约 20.5x。

■ OpenAI：C 端用户规模仍处行业领先地位

OpenAI 在用户规模与平台化能力上仍处行业领先地位，其在积极探索 B 端的业务、正试图通过 Codex 和广告业务进一步扩展长期营收增长空间。2Q26 的增量积极进展是公司已经以保密形式提交了 IPO 申请。

用户规模仍居行业首位。 据 Reuters，2026 年 6 月 ChatGPT App MAU 达 10 亿，成为达到该量级的最快 App；据 OpenAI 披露，截至 2026 年 3 月，ChatGPT 周活 9 亿+、付费消

费者订阅 5,000 万+、付费企业用户 900 万+；2025 年 12 月企业 workplace seats 超 700 万、覆盖 92% 财富 500 强。

GPT-5.6：推理与智能体能力进一步提升。OpenAI 在 6 月 26 日发布 GPT-5.6 限量预览版（公司正与美国政府合作以获批全量开放），包含旗舰模型 Sol/性价比较为平衡的模型 Terra/面向高频工作场景的高性价比模型 Luna。旗舰模型 GPT-5.6 Sol 在编程工作流方面取得 SOTA 水平，Terminal-Bench 2.1 达到 88.8%，超越 Claude Mythos 5（88.0%）和 Fable 5（84.3%）。此外，GPT-5.6 Sol 在网络安全方向的能力也有显著提升，在 ExploitBench 上，GPT-5.6 Sol 只使用约三分之一的输出 token，就达到与 Mythos Preview 相近的表现。

收入结构：Codex 与广告共同打开第二增长曲线。据 Reuters，OpenAI 年化收入从 2025 年 6 月的 100 亿美元升至 2026 年 3 月的 250 亿美元以上；据公司披露，企业收入已占总收入 40% 以上，目标 2026 年末与 C 端齐平；据 The Information 报道，截至 2026 年 3 月，广告 pilot 在不到 6 周内达到 1 亿美元 ARR，验证了订阅+API 之外的第三条变现路径。Codex 侧，公司披露，截至 2026 年 4 月下旬已有超过 400 万周活开发者用户。

现金消耗仍在扩大。据 European Business Magazine 分析，OpenAI 每 1 美元收入对应当季约 1.22 美元亏损。据公司 3 月公告，公司完成 1,220 亿美元融资（估值 8,520 亿美元），计划在 2030 年前持续重投半导体与数据中心；《华尔街日报》认为，其现金消耗节奏对一家公开市场公司而言属罕见水平。

IPO 进度：晚于 Anthropic 约一周递表。据 OpenAI 2026 年 6 月 8 日官方公告，公司已向 SEC 提交保密 S-1，但表态“时间未定，可能还需一段时间，因为有些事在私有阶段更容易做”。

■ xAI：并入 SpaceX 登陆 Nasdaq，市场以“生态+算力”逻辑定价

我们认为，尽管 xAI 单体的财务表现弱于 Anthropic 与 OpenAI，但通过被 SpaceX 全资吸收、并借 SPCX 完成大规模 IPO，市场实际上是以“生态溢价+算力供给方”的逻辑、而非单纯 AI 模型公司的逻辑对其重新定价。

SpaceX 完成大规模 IPO，xAI 整体并入。SpaceX 于 2026 年 6 月 12 日以 SPCX 为代码挂牌，IPO 定价 \$135，募资 750 亿美元；首日收涨 19.2%，市值跻身美股第六大公司。

xAI 的收入与亏损同步扩大。据 SpaceX 招股书，xAI 2024 年收入 26.2 亿美元、亏损 15.6 亿美元；2025 年收入 32 亿美元、亏损扩大至 64 亿美元，亏损增速快于收入增速。1Q26 单季收入 8.18 亿美元、运营亏损 24.7 亿美元，占 SpaceX 整体收入 46.9 亿美元的 17%。1Q26 单季 AI segment capex 达 77 亿美元，年化约 308 亿美元，同比翻倍以上。用户数量方面，Grok MAU 1.17 亿，占合并后 X+Grok 5.5 亿总 MAU 的约 21%；付费 SuperGrok 订阅 190 万，含 Grok 访问权益的 X 付费用户 440 万。

算力租赁是 xAI (SpaceX) 在 2Q26 新增的可见度较高收入曲线。据 SpaceX 招股书，Anthropic 按月支付 12.5 亿美元租赁 Colossus 数据中心算力（年化 150 亿美元），合约可 90 天通知终止，且 SpaceX 正寻求复制类似交易。该安排使 xAI 从“纯模型公司”延伸至“模型+算力供应商”，与 Anthropic 形成上下游合作关系。

国内 AI 基础模型平台：持续关注模型能力迭代及商业化验证

2Q26 国内 AI 平台竞争格局边际变化集中于三条主线：1) 资本化进程加速。智谱、MiniMax 分别于5月底/6月初启动A股科创板IPO申报，若顺利推进将成为国内首批“A+H”通用大模型企业；据《财经》报道，阶跃星辰已完成红筹架构拆除并预期年底前完成港股上市；月之暗面新一轮融资落地、投前估值升至300亿美元；2) 头部厂商模型迭代节奏环比加快。MiniMax M3、智谱 GLM-5.2、月之暗面 K2.6、阶跃 Step 3.7 Flash 集中于2Q26发布，在长上下文、稀疏注意力、多模态原生生成、Agent 生产化等方向均有版本更新；3) DeepSeek V4 预览版和智谱 GLM-5.2 等国产模型均完成了对华为昇腾、寒武纪等主要国产芯片 Day0 的深度适配。国内 AI 基础模型平台的主要关注点仍集中于模型能力的迭代，模型变现效率的提升，及新的使用场景的拓展。

图 20：国内 AI 基础模型平台核心财务对照

中国头部 AI 模型与多模态应用公司商业化更新					
维度	智谱 Zhipu (2513.HK)	稀有科技 MiniMax (100.HK)	月之暗面 Moonshot (Kimi)	阶跃星辰 StepFun	DeepSeek
规模与增长					
最新收入/ARR	API 接口业务 2026 年 3 月 ARR 约 2.5 亿美元；公司 2025 年全年收入 7.24 亿元人民币	截至 2026 年 5 月 ARR 约 3 亿美元；2025 年收入 7,903.8 万美元，同比 +158.9%	据 36 氪报道，2026 年 4 月 ARR 超过 2 亿美元	未正式披露 ARR；《财经》报导公司 2025 年收入近 5 亿元，2026 年预计收入约 12 亿元	未披露 ARR
增长引擎	云端部署、企业级智能体、MaaS 平台；2025 年云端部署收入 1.90 亿元，同比+292.6%；企业级智能体收入 1.66 亿元，同比 +248.8%	海螺 AI、MiniMax 语音、Talkie/星野及开放平台；海外市场及 AI 原生应用是主要增长来源，2026 年 5 月公司服务的全球企业和开发者客户数已超百万，全球用户规模约 3 亿	海外入驻 AWS、GCP、Azure 等公有云触达 B 端，并接入 Groq、Together AI 及 Cursor、Cline 等 Coding Agent 工具	观察者网报道模型已作为系统级能力预装进 OPPO、荣耀等手机厂商的设备，总装机量超过 4,200 万台；AI+终端+车载（吉利）为主要商业化方向	开源模型 API 调用、企业级 Agent 场景拓展；据 36 氪报道 6 月 V4.1 拟新增图像/音频理解并强化 MCP 协议适配
业务结构与核心产品					
引擎① 基座模型	2026 年 6 月 16 日发布 GLM-5.2，稳定可用 100 万 Token 上下文窗口、IndexShare+ 升级 MTP 自研架构，采用 MIT 开源协议；FrontierSWE 74.4%、PostTrainBench 34.3%、SWE-Marathon 13.0%	M3 于 2026 年 6 月 1 日正式发布，SWE-Bench Pro 超 GPT-5.5 和 Gemini 3.1 Pro；100 万 token 上下文，推理效率为上代 1/20；M2.5/M2.7 已开源	4 月开源旗舰大模型 Kimi K2.6，多项核心性能指标持平甚至超越 GPT-5.4、Claude Opus 4.6、Gemini 3.1 Pro，OpenRouter 榜单位列第九	5 月 29 日发布并开源 Step 3.7 Flash，稀疏 MoE 架构，总参数量 196B+1.8B（VIT）、激活 11B，最高生成速度 400 Tokens/s；原生支持 256K 超长上下文窗口	4 月下旬 DeepSeek-V4 预览版正式上线并开源，拥有百万字超长上下文，在 Agent 能力、世界知识和推理性能上实现国内与开源领先
引擎② C 端/产品	智谱清言、AutoGLM Agent、GLM Coding Plan、开放平台 API	海螺 AI、MiniMax 语音、Talkie/星野；海螺 AI 为 AI 视频生成平台，Talkie/星野偏 AI 陪伴与互动	Kimi 智能助手、Kimi Claw / Agent 产品、API 平台；收入由订阅与 API 共同驱动；6 月 3 日 Kimi Work 开启公测，用自然语言描述目标即可在用户电脑上拆解任务、并行执行、调用工具、使用浏览器、创建和整理文件夹	收入采取端侧按 License 收费、云侧按消耗计费；跃问、冒泡鸭、与吉利合作银河 M9；推出 Step Plan，包月 49-699 元	DeepSeek-V4-Pro API 永久降价至原价的 25%（每百万 Tokens 输入低至 0.025 元）；正式版 V4 宣布 7 月中旬上线，届时 API 将启动峰谷定价机制，高峰时段价格翻倍
新曲线	GLM-5.2 统一计费模式下 API 综合单价较此前分档口径有所上调；MaaS 平台注册用户、付费开发者和企业智能体客户扩张	MiniMax Code：作为专为 M3 设计、并与 M3 一起训练的 Agent 产品。MiniMax Agent 产品还升级并推出个人助理 Mavis，创新推出多智能体团队协作模式。	K2.6 开源+Agent 调用方向；海外 API 与开发者生态延续放量	基础模型+ 硬件+ 整车	国产算力深度适配（昇腾+寒武纪），逐步从 CUDA 迁移至华为 CANN 框架
上市进度					
IPO 状态	2026 年 1 月 8 日在港交所主板挂牌上市；6 月 1 日晚间，智谱发布公告称，计划申请在 A 股科创板上市交易	2026 年 1 月 9 日在港交所主板挂牌上市；2026 年 5 月 29 日正式启动 A 股科创板 IPO 进程	未披露	《财经》报导阶跃星辰预计年底完成上市	未披露

募资进展	港股 IPO 已募资约 43 亿港元；A 股 IPO 拟募资 150 亿元人民币	IPO 募资总额约 55.4 亿港元	据每日经济新闻报道，2026 年 5 月完成/接近完成约 20 亿美元新融资；6 月再次开启新一轮融资，投前估值 300 亿美元；1H26 累计融资规模接近 40 亿美元	据《每日经济新闻》报道，新一轮 25 亿美元融资已接近完成	据《每日经济新闻》报道，首轮融资拟募集 510 亿元人民币，梁文锋计划认购最大份额，主要外部机构包括腾讯、宁德时代、网易、京东等
最新估值					
当前市值/估值	9,381 亿港元 (~1,196 亿美元)	1,308 亿港元 (~167 亿美元)	据每日经济新闻报道，2026 年 6 月新一轮融资投前估值约 300 亿美元 (5 月融资后估值约 200 亿美元)	据《财经》报道，2026 年 IPO 前估值口径约 100 亿美元	据《每日经济新闻》报道，投后估值约 4,000 亿人民币
估值倍数 参考	26/27E P/S: 248x/96x	26/27E P/S: 56x/19x	若以 4 月 ARR 2 亿美元、投前估值 300 亿美元测算，P/S 约 150x	若以 2025 年收入近 5 亿元、IPO 前估值约 100 亿美元测算，P/S 约 135.9x	未披露收入口径，估值倍数暂不可比

资料来源：公司资料、每日经济新闻、The Information、财经、澎湃新闻、钛媒体、36 氪、观察者网、招银国际环球市场
注：数据截至 2026 年 7 月 1 日

■ 智谱 AI (2513 HK)：A 股科创板 IPO 筹备进程启动，GLM-5.2 接力发布

我们认为，智谱在国产基础模型厂商中商业化路径最为清晰，云端部署+企业级智能体的结构已进入放量阶段。后续催化集中于模型能力迭代、国产算力适配深化及芯片等供给受限情况下变现效率的提升。

A 股科创板 IPO 进程启动，有望助力融资渠道拓宽。据公司 2026 年 6 月 1 日公告，智谱拟于科创板募资 150 亿元，若推进顺利将成为国内首家“A+H”通用大模型企业。

GLM-5.2 旗舰模型发布，Coding 及 Agent 能力进一步强化。智谱于 2026 年 6 月 13 日发布 GLM5.2。根据公司披露，GLM-5.2 参数规模 753B，采用 MoE（混合专家）架构，支持 1M token 稳定上下文窗口，以 MIT 协议完全开源。在 FrontierSWE 编程基准上，GLM-5.2 得分 74.4，与 Anthropic Opus 4.8 (75.1) 差距约 1 个百分点，同时超过 GPT-5.5 (72.6)。在 PostTrainBench 基准中，GLM-5.2 以 34.3 分排名第二，仅次于 Opus 4.8 (37.2)，高于 GPT-5.5 (28.4)。另外，GLM 5.2 从此前的按照上下文长度分档计费改为统一计费，定价相对提升。国产算力适配方面，根据公司披露，GLM-5.2 已在 Day 0 完成与华为昇腾、平头哥、摩尔线程、寒武纪、昆仑芯、沐曦、海光、壁仞等国产算力平台的推理适配，在国产芯片集群上实现高吞吐、低延迟、大并发的稳定运行。

■ MiniMax (100 HK)：M3 发布+科创板 IPO 双催化，“模型+应用”双轮结构进一步深化

MiniMax 于 6 月发布 M3 旗舰模型，将变现路径从 C 端订阅与 AI 陪伴延伸至 B 端 API+企业 Agent；M3 发布一周内申报 A 股科创板 IPO，进入“A+H”双重资本化候选名单。后续可关注 M3 API 商业化兑现节奏、A 股 IPO 募资落地进展，以及 ARR 增长的节奏。

M3 模型发布，能力进入全球第一梯队。据公司 2026 年 6 月 1 日发布，M3 采用自研 MSA（MiniMax Sparse Attention）稀疏注意力机制、支持 100 万 token 上下文、单 token 计算量仅需上一代模型的约 1/20；据公司披露，M3 在 SWE-Bench Pro 基准上超越 GPT-5.5 与 Gemini 3.1 Pro，整体能力进入全球第一梯队。公司同步推出 MiniMax Code（M3 配套 Agent 产品）与个人助理 Mavis（多智能体团队协作模式），将变现路径从 C 端订阅持续延伸至 B 端 API+企业 Agent。国产算力适配方面，M2.7 自开源首日即完成华为昇腾、摩尔线程、沐曦、昆仑芯的算力适配；M3 同样完成对华为昇腾、沐曦及摩尔线程等国产算力底座的深度适配，支持多种主流推理框架部署。

A 股科创板 IPO 进程启动，A+H 双重资本化候选名单确认。据公司 2026 年 5 月 29 日公告，MiniMax 已启动 A 股科创板 IPO 进程；叠加 2026 年初港股主板挂牌（IPO 募资约 55.4 亿港元），公司亦进入“A+H”双重资本化候选名单。截至 2026 年 7 月 1 日，公司市值约 1308 亿港元（约 167 亿美元），按彭博一致预期对应 56/19x 26/27E P/S。

ARR 持续增长。据南华早报报道，MiniMax 联合创始人负焱祎透露，截至 2026 年 5 月，公司 ARR 相较 2 月翻倍至 3 亿美元以上，全球企业及开发者客户突破 100 万、全球用户约 3 亿；Talkie/星野等 AI 陪伴产品在海外贡献显著。底层 M 系列大语言模型（M2.5/M2.7 已开源）+海螺 AI 视频生成+MiniMax 语音+Talkie/星野的“模型+应用”结构稳固。

■ DeepSeek（未上市）：V4 预览版发布，国产算力适配推进，首轮融资启动

DeepSeek 于 4 月发布并开源 V4 预览版（Pro/Flash 双版本），在 Agentic Coding 等评测中处于开源模型前列；国产芯片适配进入工程化阶段，V4 早期适配权限独家开放给华为昇腾与寒武纪，目标从 CUDA 生态向华为 CANN 框架迁移；公司启动首轮外部融资，投后估值区间约 4000 亿人民币，后续持续关注公司模型能力迭代节奏。

V4 预览版发布并开源。据公司 2026 年 4 月 24 日公布，DeepSeek-V4 分 Pro 和 Flash 两个版本，拥有百万字超长上下文，在 Agent 能力、世界知识和推理性能上实现国内与开源领域领先；V4-Pro 在 Agentic Coding 评测中达到当前开源模型最佳水平，使用体验优于 Sonnet 4.5、接近 Opus 4.6 非思考模式；数学、STEM、竞赛型代码测评中超越所有已公开评测的开源模型。架构方面，公司在 token 维度引入压缩机制并结合 DSA 稀疏注意力以降低长上下文场景下的计算与显存开销，同时引入混合注意力架构（Hybrid Attention Architecture）以改善长对话记忆能力，方向与年初 mHC、Engram 两篇论文一致。另外，5 月 22 日公司宣布，其旗舰模型 V4-Pro 的 API 价格将永久降价 75%，V4 Pro 输入 0.025 元/输出 6 元，显著低于 OpenAI/Anthropic 与智谱/Minimax 的旗舰模型。

国产算力深度适配落地。据公司技术报告，首次把华为昇腾 NPU 与英伟达 GPU 写进同一份硬件验证清单；V4 未按惯例给英伟达早期测试权限，而是独家开放给昇腾与寒武纪，目标从 CUDA 生态整体迁移至华为 CANN 框架。寒武纪基于 vLLM 完成 Day 0 适配并开源；DeepSeek 自研 MegaMoE2 融合内核在昇腾平台实现 1.5–1.7 倍推理加速；目前至少有 8 家国产 AI 芯片厂商官宣适配方案。

首轮外部融资启动。据《每日经济新闻》2026 年 6 月 19 日报道，DeepSeek 首轮融资拟募集 510 亿元人民币，投后估值区间约 4000 亿人民币，创始人梁文锋计划认购最大份额（200 亿元人民币），主要外部机构包括腾讯（100 亿元）、宁德时代（50 亿元）、网易（30 亿元）、京东（30 亿元）、IDG 资本（30 亿元）等，以及中国国家人工智能产业投资基金（10 亿元）；如最终落地，将为公司迄今规模最大的单轮融资。The Information 5 月报道公司计划于 6 月推出 V4.1 迭代版本，配套企业级工具能力、强化 MCP 协议适配，并新增图像/音频理解（输出仍限文本），方向上向企业级 Agent 场景延伸。

■ 月之暗面（未上市）：K2.6 发布+20 亿美元新融资落地，一级市场融资强度领先同业

我们认为，月之暗面在 2Q26 在国内主要 AI 模型厂商中一级市场融资强度较高、ARR 增速较快；6 月新一轮融资落地后估值升至 300 亿美元。后续关注点包括 K2.6 海外 API 与企业级商业化兑现节奏。

5 月约 20 亿美元融资接近完成，6 月再次开启新一轮融资，投前估值 300 亿美元。据 36 氪 2026 年 5 月报道，月之暗面已接近完成约 20 亿美元新一轮融资，估值约 200 亿美元；

2026 年上半年累计融资规模接近 40 亿美元。据每日经济新闻报道，月之暗面已于 6 月开启新一轮融资，投前估值达 300 亿美元。

Kimi K2.6 发布，开源路线+Agent 方向进一步强化。据公司披露，K2.6 延续多模态 MoE 开源路线，在 K2.5 基础上进一步强化多模态理解与 Agent 调用能力，在长文本推理、代码生成、工具调用等核心维度上对齐甚至超越部分闭源旗舰模型。Kimi K2.6 在 SWE-Bench Pro 上以 58.6 分超过 GPT-5.4 的 57.7 和 Claude Opus 4.6 的 53.4，在 Terminal-Bench 2.0 上以 66.7 分超过 GPT-5.4 的 65.4。公司模型已入驻 AWS、GCP、Azure 等公有云及 Groq、Together AI 等推理平台的分发布局，以及与 Cursor、Cline 等主流 Coding Agent 工具深度适配，模型调用量有望加速提升，拓展在海外开发者社区的生态地位。Agent 能力的强化叠加 6 月 3 日 Kimi Work 公测，公司在算力投入、模型迭代与商业化落地之间的正循环有望进一步加速。

ARR 高速增长。据 36 氪 2026 年 4 月报道，月之暗面 ARR 已超 2 亿美元，相较 2026 年 3 月初的 1 亿美元翻倍，主要由 K2.5 模型上线后全球付费用户及 API 调用量大涨共同驱动；据创始人 2025 年内部信，2025 年 9 月-11 月海外+国内付费用户月均环比+170%、海外 API 收入同期+4 倍。按当前投前估值 300 亿美元和 4 月 ARR 超 2 亿美元测算，对应 P/S 约 150x。

■ 阶跃星辰（未上市）：预期年底前完成港股上市；发布 Step 3.7 Flash 开源模型

阶跃星辰在 2Q26 完成了赴港 IPO 的关键架构准备，并通过吉利银河 M9 车型加速终端分发体系扩张；同时于 5 月底开源新一代 Agent 导向的 Step 3.7 Flash，进一步夯实“原生多模态+高效推理”的差异化定位。后续关注点包括银河 M9 实际交付节奏，以及 Step 3.7 Flash 在 Agent 场景下的起量情况。

红筹架构拆除，赴港 IPO 进程持续推进。据《每日经济新闻》2026 年 5 月报道，阶跃星辰已接近完成约 25 亿美元的新一轮融资，且其红筹架构已拆除；《财经》2026 年 2 月报道，阶跃星辰预期年底完成上市，IPO 前估值约 100 亿美元。融资强度与赴港 IPO 时间表均处于一级市场 AI 模型厂商前列，是少数确定下半年完成上市的国内基础模型公司。按其 2025 年近 5 亿元收入和 IPO 前估值 100 亿美元测算，对应 P/S 约 136x，估值水平高于已上市同行，或反映市场对其终端商业化节奏的预期。

终端应用扩张，AI+汽车场景兑现。据公司披露，与吉利合作打造的银河 M9 车型预计 1H26 交付超 100 万辆，且 2026 年 1 月已在 OPPO、荣耀等国内六大手机品牌上实现超 4200 万台装机，“基座模型+终端硬件+整车”的三方协同模式加速形成。考虑到公司高管同时担任吉利控股旗下千里科技董事长，阶跃与吉利体系内的车机 AI 合作有望进一步深化，构成区别于其他厂商的差异化路径。

Step 3.7 Flash 发布，多模态+Agent 方向进一步强化。公司于 2026 年 5 月 29 日发布并开源 Step 3.7 Flash，面向 Agent 生产化阶段，围绕 Agent、Coding、Search 与多模态 workflow 进行系统优化；采用稀疏 MoE 架构，总参数 196B+1.8B (ViT)、激活参数 11B，最高生成速度可达 400 Tokens/s，适合高频、多轮、低等待的 Agent 应用，支持 256K 超长上下文窗口。公司同步推出面向 Agent 与 Coding 场景的一体化能力订阅方案 Step Plan，包月价格区间 49-699 元。叠加既有的 Step 系列模型，公司模型矩阵在语言、多模态、MoE Flash 方向已形成完整覆盖，是其打入手机/车载等终端分发场景的关键支撑。算力适配方面，阶跃星辰持续推进“模型+芯片”生态协同，Step 3.5 Flash 已实现与华为昇腾、沐曦股份、壁仞科技等国产芯片平台适配。

重点覆盖公司 2Q26 AI 业务进展更新及催化展望

复盘 2Q26 我们所覆盖的主要公司的 AI 业务商业化进展，我们观察到：1) 因 ROI 更为清晰且对智能能力的需求更大，AI 在 B 端的商业化变现仍在加速，云、广告、创意内容生成、Agent 等仍是主要的关注领域，且 AI 助力企业降本增效的进程在持续推进；2) 相较而言，C 端整体付费订阅及渗透仍处于较为早期的阶段，办公及生产力提升仍是主要探索及推进的场景，未来基于模型能力提升推动付费渗透率的提升、及持续推动从仅基于用户数量计费的模式转型至按“用户+使用量计费”，或将是商业模式演变探索的中长期重点。就具体催化而言，主要关注点集中于旗舰模型能力的迭代、云厂商营收增速的提速趋势及利润率，以及 C 端产品的付费渗透率提升和变现趋势。

图 21：重点覆盖公司 2Q26 AI 业务进展更新

	AI 商业化进展	重点催化展望
微软	1) AI 相关商业化：AI 业务 ARR 突破 370 亿美元，同比增长 123%；Microsoft 365 Copilot 付费席位突破 2000 万，席位数同比增长 250%，为发布以来最快增速； 2) 企业 Agent 生态：LinkedIn Talent Solutions 的 Agent 产品 ARR 超 4.5 亿美元；GitHub Copilot 被近 14 万家组织使用，企业订阅用户数同比增长近 3 倍； 3) 提升企业 AI 应用渗透：Copilot Credit 消费量环比近乎翻倍，近 90% 的财富 500 强使用 Copilot Studio 构建活跃 AI Agent。	1) 涵盖 GitHub Copilot 在内的产品和业务由基于用户数量计费逐步转向按“用户+使用量计费”的进展更新； 2) Azure 云业务营收增长加速。
谷歌	1) 谷歌全平台 5 月月均处理 Token 数量已超 3200 万亿，同比增长 7 倍，公司旗下模型 API 每分钟处理 Token 数达到 190 亿； 2) AI 广告工具助力广告主投放效率提升，目前谷歌超 30% 的客户搜索广告预算已用于 AI Max； 3) 1Q26 谷歌云收入同比增长 63% 至 200 亿美元（4Q25/3Q25：+48%/34%），主要受 AI 解决方案与算力基础设施需求旺盛驱动。	1) 6-7 月发布 Gemini-3.5 Pro； 2) AI 搜索&云业务增长持续加速。
亚马逊	1) AWS AI 业务放量：AWS AI 业务年化收入超 150 亿美元，是 AWS 同期规模的近 260 倍； 2) 提升电商用户参与度：Rufus 月活用户增长超 115%、互动量同比近 400%；通过 Rufus 推出赞助商品/品牌提示后，约 20% 的购物者会继续与品牌提示对话。	1) AWS 营收增速进一步加速
Meta	1) 推荐算法优化推动 1Q26 Facebook 视频总观看时长环比增长 8%，创四年最大环比增幅； 2) Lattice 与 GEM 模型迭代优化，带动落地页广告转化率提升超 6%； 3) 已有超 800 万广告主至少使用一款 Meta 生成式 AI 广告创意工具，其中使用视频生成功能的广告主转化率提高超 3%。	1) 旗舰模型迭代
腾讯	1) AI 持续提升广告投放效果及价格，智能广告投放解决方案 AIM+ 在 1Q26 已赋能公司约 30% 的广告主投放； 2) HY3 Preview 凭借综合智能能力与良好成本效率，在发布以后持续排名 OpenRouter Token 使用量榜单前五； 3) AI 办公生产力应用 WorkBuddy 用户增长势头强劲，按日活用户口径统计，WorkBuddy 是国内领先的桌面办公智能体。	1) 2H26 发布微信 AI 智能体“小微”； 2) WorkBuddy 办公智能体升级迭代与商业化加速； 3) 2H26 发布更大参数的混元模型。
阿里巴巴	1) AI 相关的云营收快速增长：云智能集团外部收入同比增长 40%，其中 AI 相关产品收入连续第 11 个季度实现三位数增长，占云业务外部收入的 30%。AI 产品年化收入已超过 358 亿元人民币； 2) MaaS 货币化加速：预计 2026 年 6 月季度，模型与应用服务 ARR 将超过 100 亿元人民币，年底达到 300 亿元； 3) AI 应用生态扩展：在消费者端，Qwen 应用已整合淘宝、天猫的电商服务能力，成为覆盖生活、生产力与学习的全能个人助手；在企业端，推出智能办公、AI 编程、业务运营管理等产品。	1) MaaS 业务快速增长并支撑云业务营收强劲增长； 2) 云业务利润率水平提升超出市场预期。

快手	1) 商业化方面：1Q26 可灵 AI 收入突破 6.5 亿元，同比增长超 300%。受益于 2 月可灵 3.0 模型正式上线，2026 年 3 月可灵 AI 年化收入规模 (ARR) 达到约 5 亿美元，较 2025 年 12 月的 2.4 亿美元实现翻倍； 2) 外部融资方面：据南华早报 7 月报道，快手可灵 AI 拟融资 30 亿美元，投后估值 180 亿美元（对应约 36x ARR）。	1) 可灵模型升级迭代； 2) 可灵外部融资。
----	---	----------------------------

资料来源：公司资料、招银国际环球市场

微软：将 Copilot 从单点应用升级为平台化架构，深化 Agent 业务发展布局

微软于 2026 年 3 月宣布 Copilot 管理团队更新，将商业端和消费端 Copilot 体验整合为一个统一系统，覆盖四个支柱：Copilot experience、Copilot platform、Microsoft 365 apps、AI models。Jacob Andreou 负责跨消费和商业场景的 Copilot 体验，直接向 CEO 汇报。这个动作相当于把过去分散在 M365、Windows、Bing、Consumer AI 里的 Copilot 体验进一步统一。

在 2026 年 6 月的 Build 2026，微软明确把 AI 平台架构升级为 Agent-first，其核心是新推出的 Microsoft IQ。作为 Agent 时代的“上下文层”，Microsoft IQ 在 GitHub Copilot、Microsoft Foundry、Copilot Studio 中可用，用来为 Agent 提供企业知识、工作流、邮件、会议、文档、业务系统等上下文。其下包括 Work IQ、Fabric IQ、Foundry IQ、Web IQ 等层。微软的目标并不只是提供模型，而是在提供让 Agent “理解企业上下文”的数据/知识层。

同时，微软在 Build 2026 给出更新，期待 Microsoft Foundry 承担类似“企业 Agent PaaS”的角色，强调 Microsoft Foundry 要覆盖 Agent 的搭建、部署、运营的全生命周期。开发者可以用 Microsoft Agent Framework、GitHub Copilot SDK、Claude Agent SDK 等构建 Agent；用 Foundry Agent Service 部署长周期 Agent；并把 Agent 发布到 Teams 和 Microsoft 365 Copilot 中。

此外，公司将 Agent 365 定位为企业观察、治理、管理和保护 Agent 的控制平面，建立在 Entra、Defender、Purview、Intune 等微软安全与管理体系之上，致力于帮助客户的 IT 部门理解谁创建了 Agent、Agent 访问了什么数据、花了多少钱、有没有越权。

谷歌：全栈 AI 策略支撑 AI 相关业务强劲增长

基于公司的全栈 AI 产品策略，谷歌已成为全球最大的 AI Token 工厂之一；并且基于其搜索以及云计算的业态，谷歌已形成清晰且领先的 AI 商业化路径，持续受益于 AI 需求的强劲增长。谷歌全平台 5 月处理 Token 数量已超 3200 万亿，同比增长 7 倍，公司旗下模型 API 每分钟处理 Token 数达到 190 亿。展望未来，我们认为公司将持续优化升级其底层算力基础设施（谷歌云与 TPU）、模型能力（Gemini）以及 AI 应用（Gemini App、AI Search 等），从而推动核心搜索主业以及谷歌云业务的强劲增长；其中，Gemini 3.5 Pro 或将在 6 月发布，有望进一步提升模型编程和智能体等多方面能力，巩固 Gemini 模型系列的 SOTA 智能水平。

图 22: 谷歌：月均处理 Token 数量强劲增长



资料来源：公司资料，招银国际环球市场

谷歌 2026 年 5 月发布 Gemini 3.5 Flash，模型实现性能、成本、推理速度的良好平衡，该模型在编程和智能体等多项重要基准的表现好于 Gemini 3.1 Pro，同时 Token 输出速度较其他前沿模型快 4 倍，达到 289 Tokens/s，同时公司计划在 6 月发布 Gemini 3.5 Pro，有望在智能水平上实现进一步跃升。

图 23: 谷歌 Gemini 3.5 Flash：主要测评评分对比

Benchmark			Gemini 3.5 Flash	Gemini 3 Flash	Gemini 3.1 Pro	Claude Sonnet 4.6	Claude Opus 4.7	GPT-5.5
Coding	Terminal-bench 2.1 Agentic terminal coding	Terminus-2 harness	76.2%	58.0%	70.3%	-	66.1%	78.2%
	SWE-Bench Pro (Public) Diverse agentic coding tasks	Single attempt	55.1%	49.6%	54.2%	-	64.3%	58.6%
Agentic	MCP Atlas Multi-step workflows using MCP		83.6%	62.0%	78.2%	69.5%	79.1%	75.3%
	Toolathlon Real-world general tool use		56.5%	49.4%	-	-	-	55.6%
UI control	OSWorld-Verified Agentic computer use		78.4%	65.1%	76.2%	72.5%	78.0%	78.7%
Expert tasks	Finance Agent v2 Financial analysis and decision-making		57.9%	42.6%	43.0%	51.0%	51.5%	51.8%
	GDPval-AA Economically valuable knowledge work	Elo	1656	1204	1314	1676	1753	1769

资料来源：公司资料，招银国际环球市场

在 AI 应用和商业化方面，AI 已赋能并加速谷歌多项主要业务增长：**1) 搜索**：谷歌 5 月 AI Overviews 和 AI Mode 的月活跃用户数已分别超 25 亿和 10 亿，带动公司搜索量持续增长至历史新高。公司 5 月将 AI Overviews 和 AI Mode 整合升级至统一的 AI 搜索框，进一步优化搜索体验；**2) AI 云**：1Q26 谷歌云收入同比增长 63%至 200 亿美元，主要受 AI 解决方案与算力基础设施需求旺盛驱动，云业务在手订单 1Q26 环比近乎翻倍，达到 4600 亿美元；**3) C 端应用**：5 月 Gemini App 月活跃用户超 9 亿，同比增长超 100%，公司 5 月进一步升级 Gemini App 并引入多项 AI 新功能，包括每日摘要智能体 Daily Brief，AI 办公工具 Docs Live 等。此外公司 5 月还推出了谷歌首个 C 端 AI Agent 产品 Gemini Spark，能够长期后台运行并自主完成任务，包括邮件整理、日程管理、文档生成、活动规划等。

图 24: 谷歌: Gemini Spark



资料来源：公司资料，招银国际环球市场

亚马逊：关注 AWS 营收同比增长环比加速趋势

Amazon 于 2026 年 4 月进一步扩大与 Anthropic 的战略合作（[链接](#)）：Amazon 将向 Anthropic 追加 50 亿美元投资，并可能根据商业里程碑未来再追加最多 200 亿美元；同时，Anthropic 承诺未来十年在 AWS 云服务和 AI 基础设施上投入超过 1000 亿美元，并获得最多 5GW 的 AWS Trainium 算力容量，用于 Claude 模型的训练和推理。我们认为此次合作进一步强化了 AWS 与 Anthropic 在资本、算力和模型分发层面的协同，有助于 Anthropic 缓解算力瓶颈，也有望推动 AWS Trainium 生态和 AI 云收入增长。

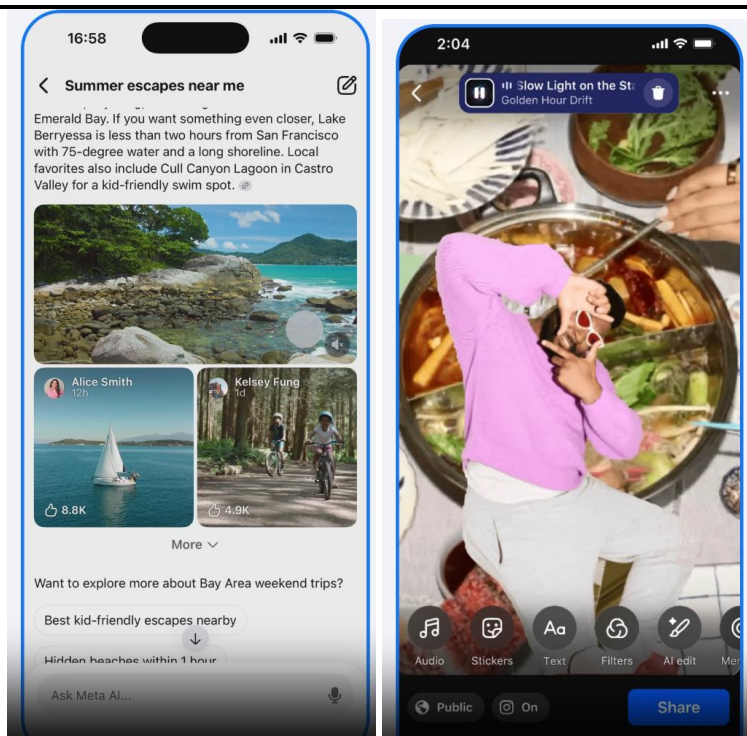
1Q26 亚马逊 AWS 营收 376 亿美元，同比增长 28.4%（4Q25：23.6%）。受资本开支投入影响，AWS 营业利润率为 37.7%，同比回落 1.8 个百分点。管理层指出 1Q26 AWS 在手订单规模达 3640 亿美元（不含近期与 Anthropic 达成的超千亿美元合作订单），环比大增 49%。在手订单大幅增长印证需求景气度高，我们预计 2026 年二季度 AWS 营收增速将进一步提升至 31%。

Meta：持续推动 C 端 AI 应用落地

我们认为 Meta 持续高额的 AI 投入将主要体现在两方面的收益：1) 推动广告业务持续稳健增长；2) 提升基础模型能力，落地 C 端 AI 应用，实现个人超级智能（Personal Superintelligence）的长期愿景。相比微软/谷歌等云厂商而言，Meta AI 投入面临更长的投资回报期，并因此在 AI 投入 ROI 以及业绩层面面临更高的不确定性，短期估值层面也有所折价。但长期我们仍看好 Meta 的人才与算力储备支持其实现个人超级智能的愿景，AI 投入将持续拉动核心业务增长。

基于旗舰大模型 Muse Spark，Meta AI 在 6 月面向用户推出多个 AI 新功能以提升用户体验：1) AI Mode 搜索标签，Meta AI 基于用户在 Groups、Reels 等公开分享的内容给出搜索答案，提供真实的用户观点与经验；2) 其他 AI 生成式工具，例如拼贴剪影模板、视频转场、AI 换装/发型/配饰等。

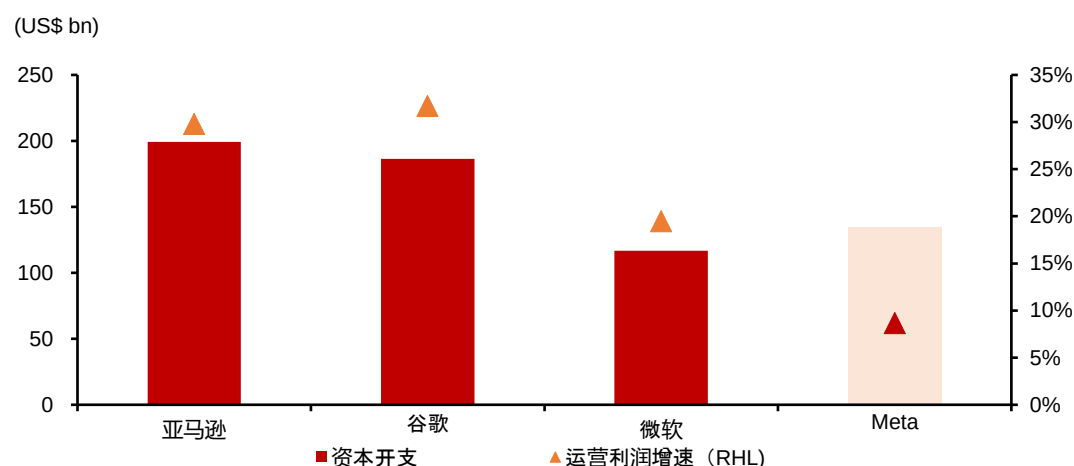
图 25: Meta AI: AI Mode & 创作工具



资料来源：公司资料，招银国际环球市场

据彭博社 7 月 1 日报道，Meta 正制定计划以推出云计算业务，向外部客户出售 AI 算力以及模型访问权限。我们认为该计划能有效缓解市场此前对于公司 AI 投入 ROI 的担忧，即：Meta 目前维持着接近北美头部云厂商的资本开支投入规模与节奏（2026E Meta 资本开支预期 1345 亿美元 vs 亚马逊/谷歌/微软 1999 亿/1865 亿/1168 亿美元），但缺少能够高效变现谷时冗余算力并提升集群整体算力利用率的云基础设施服务，导致公司短期盈利和估值承压（2026E Meta 运营利润增速预期+9% vs 亚马逊/谷歌/微软+30%/32%/20%）。我们认为，AI 算力的租售能让 Meta 在保持高额 AI 投入的同时维持良好的 ROI，但并不意味着公司要缩减在 AI 算力、前沿模型与 AI 应用方面的投入。

图 26: 北美主要云厂商 vs Meta: 资本开支&运营利润增速 (2026E)



资料来源：彭博一致预期，招银国际环球市场

腾讯：关注微信 AI 智能体进展

市场短期仍然担忧腾讯在 AI 竞争中落后以及公司在 AI 投入中的 ROI 情况，但我们认为未来有三条主线有望缓解该担忧：1) 微信 AI 智能体或将在 2H26 发布，基于微信庞大的用户基数与丰富的数据沉淀，有望快速成为国内领先的 C 端 AI 智能体产品；2) 办公/编程智能体方面，WorkBuddy 持续巩固用户规模优势，并加速推进商业化；3) 混元大模型的持续升级迭代。

微信在 2026 年 6 月开启微信 AI 智能体产品“小微”内测，产品基于微信团队自研模型 WeLM 以及 DeepSeek，在微信生态内为用户提供效率、生活、社交娱乐等多领域的智能体服务。在目前已开放的功能中，我们认为具备较高应用价值的场景包括：1) 对话生成小程序/小工具，有望帮助用户提高创作/办公/生活效率，丰富微信内小程序生态；2) 总结归纳微信生态内各类型内容（包括聊天、朋友圈、公众号、视频号等），有望充分利用微信生态沉淀的数据和内容资产，帮助用户提供办公/生活效率。出于用户隐私、合规和算力成本等因素考虑，目前“小微”尚未面向用户大规模开放且权限较为受限（比如创作的小程序仅限个人使用、“小微”无法直接读取用户所有私聊/群聊）。但随着未来产品功能进一步成熟，我们认为“小微”有望依托其高价值 AI 应用场景逐步尝试商业化变现，潜在的可行途径包括相关订阅会员或广告。

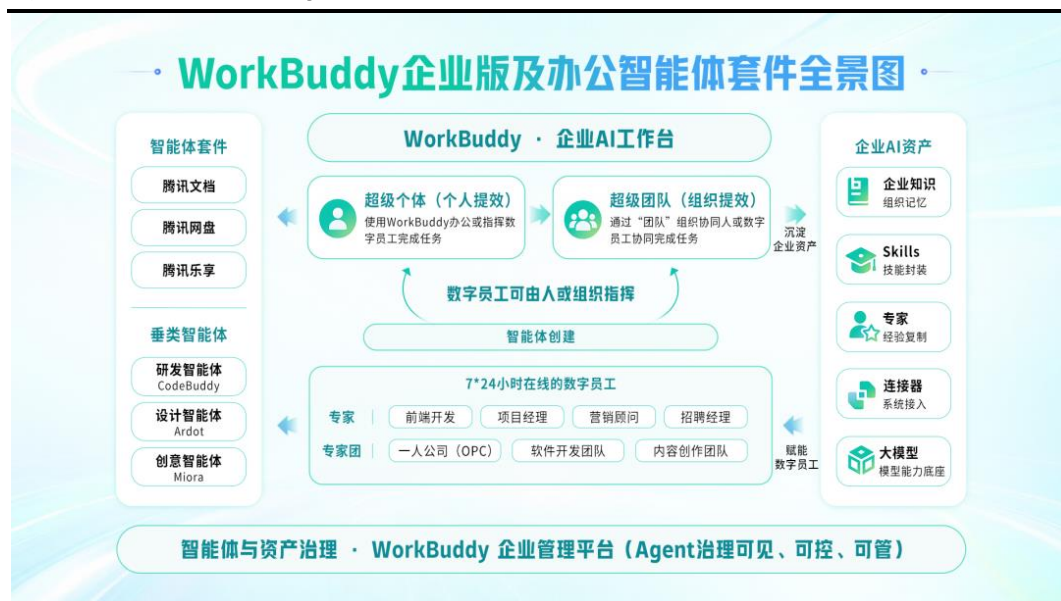
图 27：微信“小微”主要产品功能



资料来源：财联社，公司资料，招银国际环球市场

腾讯 AI 办公生产力应用 WorkBuddy 用户增长势头强劲，按日活跃用户口径统计，WorkBuddy 是国内领先的桌面办公智能体。公司 6 月正式面向企业推出 WorkBuddy Enterprise 企业智能体产品，整合专家（封装各岗位工作流和技能）、助理（数字员工）、团队（人机协作）三大能力，配套企业级 Agent 观测、效果评估、成本管控等多项能力，帮助腾讯进一步渗透到企业 AI 智能体市场。

图 28: 腾讯: WorkBuddy 企业版



资料来源：公司资料，招银国际环球市场

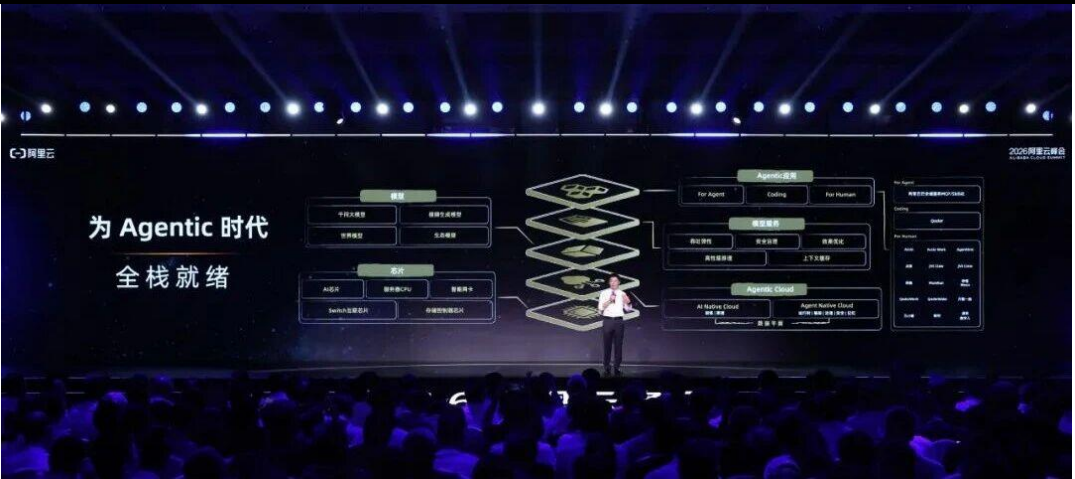
模型层面，腾讯也仍在持续投入以提升底座模型的智能水平。HY3 Preview 凭借综合智能能力与良好成本效率，在发布以后持续排名 OpenRouter Token 使用量榜单前五。展望未来，腾讯模型训练将围绕几个重点：1) 持续优化底层基础设施，提升数据质量，加强模型基础能力和泛化能力；2) 强调模型和产品的 Co-design，摒弃刷榜思维，建立正确的测评标准，让模型/智能体满足真实场景需求；3) 首先关注模型性能，然后才是模型性价比。我们预计更大参数量级的混元 3.0 模型有望在 2H26 发布。

阿里巴巴：组织架构重构与 AI 入口升级

2Q26 阿里巴巴 AI 相关的布局及进展主要可以概括为技术与模型组织重构及 AI 入口升级，同时公司持续深化模型能力。AI 相关的组织架构调整主要包括设立集团技术委员会、AI 进入更高层级的统筹及将通义实验室升级为“通义大模型事业部”等，而入口升级主要包括对于 B 端入口的升级和对于 C 端的入口升级。B 端主要的优化是将钉钉、悟空、MuleRun、客户商业化体系合并成更完整的企业级 AI Agent 落地组织，C 端主要的优化是将千问 App 接入淘宝/天猫消费场景，进一步强化 AI 应用与大消费生态的协同。在模型能力的深化方面，阿里巴巴 6 月 2 日发布多模态大模型 Qwen3.7-Plus，在文本和视觉能力上取得大幅提升，同时公司 6 月 23 日发布了 Happy Horse 1.1，相较此前版本围绕动态表现力、主题一致性、指令遵循、视觉呈现质量等维度进行了系统化升级。

在 AI 相关的组织架构调整方面，4 月 8 日：1) 阿里巴巴在集团层面新设技术委员会，由 CEO 吴泳铭担任组长，成员包括周靖人、吴泽明、李飞飞。分工上，周靖人担任首席 AI 架构师，李飞飞负责阿里云技术及 AI 云基础设施建设，吴泽明负责集团业务技术平台及 AI 推理平台建设，我们认为调整标志 AI 战略进入集团级技术协同机制；2) 同时，通义实验室升级为通义大模型事业部，标志着通义从“模型研发中心”进一步变成“模型产品与商业化核心部门”；3) 李飞飞出任阿里云 CTO，吴泽明专注阿里巴巴集团 CTO 工作，有助于把模型、算力、推理、业务应用串联。

图 29: 阿里巴巴：阿里云业务面向 Agentic 时代重构全栈技术体系

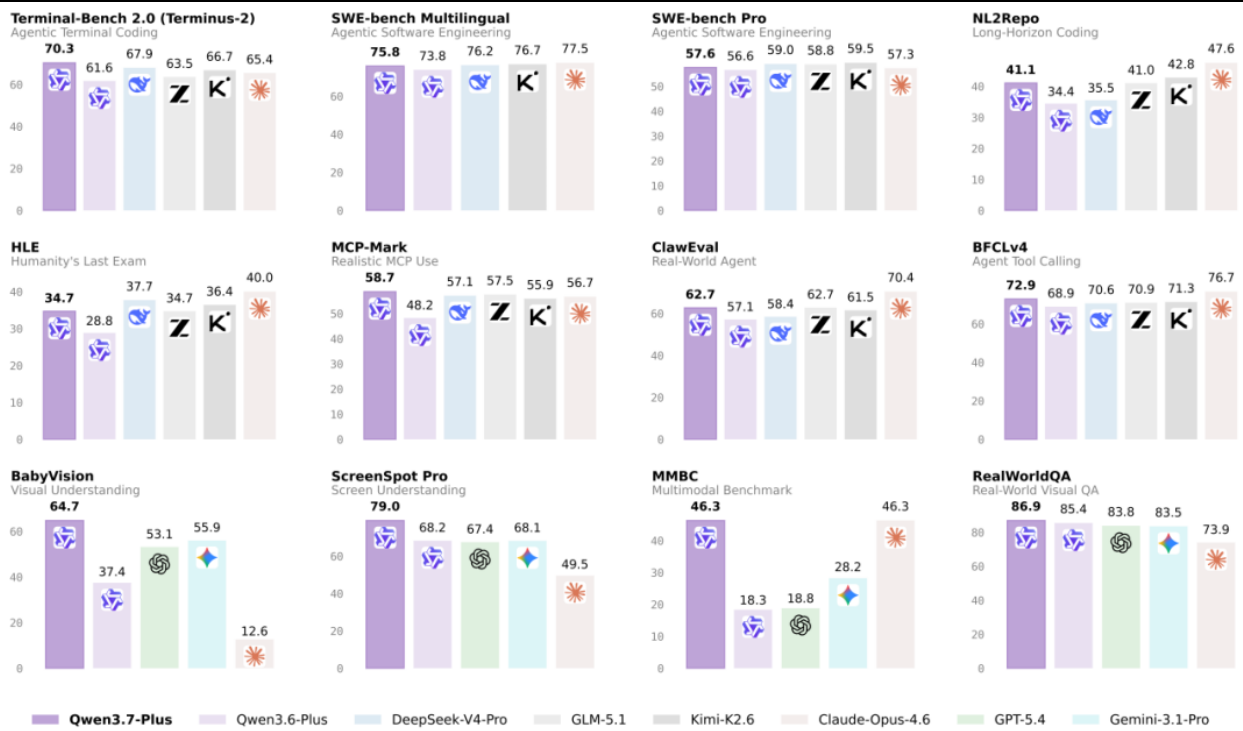


资料来源：公司资料，招银国际环球市场

而在入口升级方面，B 端：1) 6 月 11 日，陈宇森接任钉钉 CEO/悟空事业部 CEO，此前，陈宇森曾在阿里云内部创业，带队研发 AI Agent 产品 MuleRun，我们认为任命说明钉钉的战略重心持续向 AI Agent、企业工作流、AI To B 入口倾斜；2) 在对钉钉和悟空事业部进行的组织梳理中，陈宇森成立核心平台业务部，整合悟空与 MuleRun 两个团队成立新的悟空团队；同时整合销售、服务、售前、交付、客户成功等团队，成立客户发展部；并成立公司信息技术部，目标是让内部系统更易于被 Agent 使用。C 端入口升级方面，阿里巴巴 5 月 13 日在业绩公告中提及，C 端应用千问 App 已经全面接通电商服务能力，进一步深化 AI 应用与大消费生态的协同。

在模型能力的升级方面，Qwen3.7-Plus 实现了多模态混合智能体的突破，实现在能看懂图片和视频的同时进行深度推理、自我编程、调用工具、验证测试并自主迭代，将“看、想、写、做、验”整合进统一的智能体工作流。

图 30：阿里巴巴：Qwen3.7-Plus 在主要测评中的得分



资料来源：公司资料、招银国际环球市场

注：数据截至 2026 年 6 月 2 日模型发布当日

Happy Horse 1.1 相较此前的版本，实现了：1) 动态表现力提升、动作更连贯；2) 支持 9 张参考图输入，主体一致性增强；3) 指令遵循能力增强，适配简洁与复杂提示词；及 4) 视觉质感优化，缓解“油光感”与“过度锐化”。

快手：外部融资释放可灵价值

随着市场预期已较为充分反映快手电商与广告业务所面临的压力，我们认为当前阶段影响快手股价和估值的关键已聚焦在可灵的商业化与外部融资进展：1) 商业化方面：1Q26 可灵 AI 收入突破 6.5 亿元，同比增长超 300%。受益于 2 月可灵 3.0 模型正式上线，2026 年 3 月可灵 AI 年化收入规模（ARR）达到约 5 亿美元，较 2025 年 12 月的 2.4 亿美元实现翻倍；2) 外部融资方面：公司 7 月 2 日公告可灵已与投资者订立增资协议，总增资额不超过 30 亿美元（人民币 204 亿元），约占北京可灵经扩大注册资本的 16.67%（隐含可灵投后估值 180 亿美元，对应约 36x ARR），根据协议可灵已募得 27.9 亿美元。我们认为，外部融资有望进一步释放可灵价值并支撑集团整体估值。

若剔除可灵股权估值以及集团净现金，快手当前市值隐含主业估值仅对应 3x 2026 年 PE，估值具备较高的安全边际；基于我们敏感性分析，假设给予快手主业 3-8x PE 估值，对应集团估值范围人民币 1650-2588 亿元，潜在上行空间 3-62%。

图 31：快手：分部估值

(人民币百万元)	2026E 收入	2026E 净利润	估值倍数	估值	备注
集团	146,836	16,938	9x PE	160,153	7 月 2 日收盘市值
可灵	3,640	-1,820	28x PS	102,000	150 亿美元潜在投前估值
净现金				6,744	2025 年末净现金
集团(剔除可灵/净现金)	143,196	18,758	3x PE	51,409	

资料来源：Wind，招银国际环球市场预测

图 32：快手：估值敏感性测算

主业估值倍数	主业估值 (人民币百万元)	集团估值 (人民币百万元)	潜在上行空间
3x PE	56,274	165,018	3%
4x PE	75,032	183,776	15%
5x PE	93,790	202,534	26%
6x PE	112,548	221,292	38%
7x PE	131,306	240,050	50%
8x PE	150,064	258,808	62%

资料来源：公司资料，招银国际环球市场预测

注：数据截至 2026 年 7 月 2 日；快手可灵股份+净现金估值人民币 1087 亿元，集团剔除可灵亏损净利润假设 188 亿元，详细数据见图 31

股票推荐

- 1) 腾讯 (700 HK, 买入, 目标价: 750.0 港元)：**我们看好腾讯持续受益于 AI 发展，公司对于 AI 的投入将从大模型、AI 智能体及赋能主业等多个方面逐渐兑现。基于稳固的竞争壁垒及 AI 赋能，我们预计 FY26/FY27 公司 Non-IFRS 运营利润维持 8/9% 的稳健增长。多项催化因素有望支撑公司基本面增长与估值修复：1) 微信 AI 智能体“小微”有望在 2H26 正式上线，帮助腾讯借助微信+AI 快速切入智能体赛道，长期业务空间广阔；2) 常青游戏持续保持稳健，有望支撑公司全年游戏收入稳健增长；3) AI 赋能叠加视频号广告加载率提升，推动广告收入维持快于行业的增长。公司当前估值仅对应 13x FY26 Non-IFRS PE，仍有较大修复空间。我们基于 SOTP 的目标价为 750.0 港元，对应 22x FY26E Non-IFRS PE。
- 2) 阿里巴巴 (BABA US, 买入, 目标价: 220.1 美元)：**我们重申对阿里巴巴的积极看法，主因云计算收入增长前景强劲，这得益于人工智能相关产品的采用率提升以及数字化需求的增加，且我们认为云业务强劲增长趋势有望在 2026 年延续。此外，闪购业务更加专注效率提升有望带动业务减亏及支撑集团层面盈利修复。尽管闪购业务仍将优先市场份额，但受订单结构、用户补贴优化及运营效率提升的推动，我们预计即时零售全年亏损将从 2026 财年（我们预估 880 亿元）大幅收窄至 2027 财年的 440 亿元，并在 2028 财年进一步减半。在我们看来，阿里巴巴仍然是人工智能主题下的关键受益方之一，我们基于分部估值的目标价为 220.1 美元。“买入”评级。
- 3) 快手 (1024 HK, 买入, 目标价: 70.5 港元)：**我们认为公司当前估值已充分反映快手电商与广告业务面临的压力，若剔除可灵股权估值以及集团净现金，快手当前市值隐含主业估值仅对应 3x 2026 年 PE，估值具备较高的安全边际。展望后续，我们认为多项催化有望推动公司估值修复：1) 可灵模型持续迭代；2) 可灵商业化持续加速，我们预计可灵 2026 年总收入同比增长 250% 至 36 亿元；3) 可灵外部融资落地。我们基于 SOTP 的目标价为 70.5 港元。“买入”评级。
- 4) 微软 (MSTF US, 买入, 目标价: 616.4 美元)：**我们看好微软长期结构性增长前景，AI 及数字化需求强劲支撑下公司 Azure 业务有望维持高速增长，且商业未完成订单 (RPO) 余额保持强劲增长为业务增长带来支撑；生产力和业务流程板块每用户平均收入持续提升印证 AI 变现稳步推进。管理层强调 GitHub Copilot 已转向按使用量计费模式，此举可使定价与实际使用量及成本匹配，并预期当前所有基于用户数量计费的业务均将逐步过渡至“按用户 + 使用量计费”模式。我们认为，该模式有望进一步释放商业化潜力，并缓解市场后续担忧。我们基于 DCF 模型的目标价为 616.4 美元，对应 31.5x FY27 市盈率。“买入”评级。
- 5) 谷歌 (GOOG US, 买入, 目标价: 425.0 美元)：**我们看好 AI 推动公司核心搜索和云业务的强劲增长，支撑公司在显著加大 AI 投入的同时维持稳健盈利增长与可观自由现

金流，我们预计公司 FY26 整体运营利润仍将实现 28% 的同比增长。长期来看，谷歌全栈 AI 解决方案（云/TPU+模型+应用）将巩固公司在 AI 时代的核心竞争力，拓展长期增长空间，持续受益于强劲的 AI 需求。我们的目标价为 425.0 美元，基于 35 倍 FY26 PE。“买入”评级。

- 6) **亚马逊 (AMZN US, 买入, 目标价: 305.0 美元)**：我们看好 AWS 营收同比增速环比进一步加速的空间，且认为亚马逊稳健的盈利增长（我们预测 2026 年亚马逊经营利润将同比增长 27%）将有望支撑股价增长。305.0 美元的目标价基于 17.6x EV/EBITDA 估值，目标估值水平与公司三年均值区间一致。“买入”评级。

免责声明及披露

分析员声明

负责撰写本报告的全部或部分内容的分析员，就本报告所提及的证券及其发行人做出以下声明：（1）发表于本报告的观点准确地反映有关他们个人对所提及的证券及其发行人的观点；（2）他们的薪酬在过往、现在和将来与发表在报告上的观点并无直接或间接关系。

此外，分析员确认，无论是他们本人还是他们的关联人士（按香港证券及期货事务监察委员会操作守则的相关定义）（1）并没有在发表研究报告 30 日前处置或买卖该等证券；（2）不会在发表报告 3 个工作日内处置或买卖本报告中提及的该等证券；（3）没有在有关香港上市公司内任职高级人员；（4）并没有持有有关证券的任何权益。

招银国际环球市场或其关联机构曾在过去 12 个月内与本报内所提及发行人有投资银行业务的关系。

招银国际环球市场投资评级

买入	: 股价于未来 12 个月的潜在涨幅超过 15%
持有	: 股价于未来 12 个月的潜在变幅在-10%至+15%之间
卖出	: 股价于未来 12 个月的潜在跌幅超过 10%
未评级	: 招银国际证券并未给予投资评级

招银国际环球市场行业投资评级

优于大市	: 行业股价于未来12 个月预期表现跑赢大市指标
同步大市	: 行业股价于未来12 个月预期表现与大市指标相若
落后大市	: 行业股价于未来 12 个月预期表现跑输大市指标

招银国际环球市场有限公司

地址: 香港中环花园道 3 号冠君大厦 45 楼

电话: (852) 3900 0888

传真: (852) 3900 0800

招银国际环球市场有限公司(“招银国际环球市场”)为招银国际金融有限公司之全资附属公司(招银国际金融有限公司为招商银行之全资附属公司)

重要披露

本报告内所提及的任何投资都可能涉及相当大的风险。报告所载数据可能不适合所有投资者。招银国际环球市场不提供任何针对个人的投资建议。本报告没有把任何人的投资目标、财务状况和特殊需求考虑进去。而过去的表现亦不代表未来的表现，实际情况可能和报告中所载的大不相同。本报告中所提及的投资价值或回报存在不确定性及难以保证，并可能会受目标资产表现以及其他市场因素影响。招银国际环球市场建议投资者应该独立评估投资和策略，并鼓励投资者咨询专业财务顾问以便作出投资决策。

本报告包含的任何信息由招银国际环球市场编写，仅为本公司及其关联机构的特定客户和其他专业人士提供的参考数据。报告中的信息或所表达的意见皆不可作为或被视为证券出售要约或证券买卖的邀请，亦不构成任何投资、法律、会计或税务方面的最终操作建议，本公司及其雇员不就报告中的内容对最终操作建议作出任何担保。我们不对因依赖本报告所载资料采取任何行动而引致之任何直接或间接的错误、疏忽、违约、不谨慎或各类损失或损害承担任何的法律责任。任何使用本报告信息所作的投资决策完全由投资者自己承担风险。

本报告基于我们认为可靠且已经公开的信息，我们力求但不担保这些信息的准确性、有效性和完整性。本报告中的资料、意见、预测均反映报告初次公开发布时的判断，可能会随时调整，且不承诺作出任何相关变更的通知。本公司可发布其它与本报告所载资料及/或结论不一致的报告。这些报告均反映报告编写时不同的假设、观点及分析方法。客户应该小心注意本报告中所提及的前瞻性预测和实际情况可能有显著区别，唯我们已合理、谨慎地确保预测所用的假设基础是公平、合理。招银国际环球市场可能采取与报告中建议及/或观点不一致的立场或投资决定。

本公司或其附属关联机构可能持有报告中提到的公司所发行的证券头寸并不时自行及/或代表其客户进行交易或持有该等证券的权益，还可能与这些公司具有其他投资银行相关业务联系。因此，投资者应注意本报告可能存在的客观性及利益冲突的情况，本公司将不会承担任何责任。本报告版权仅为本公司所有，任何机构或个人于未经本公司书面授权的情况下，不得以任何形式翻版、复制、转售、转发及或向特定读者以外的人士传阅，否则有可能触犯相关证券法规。

如需索取更多有关证券的信息，请与我们联系。

对于接收此份报告的英国投资者

本报告仅提供给符合(I)不时修订之英国 2000 年金融服务及市场法令 2005 年(金融推广)令(“金融服务令”)第 19(5) 条之人士及(II) 属金融服务令第 49(2) (a) 至(d) 条(高净值公司或非公司社团等)之机构人士，未经招银国际环球市场书面授权不得提供给其他任何人。

对于接收此份报告的美国投资者

招银国际环球市场不是在美国的注册经纪交易商。因此，招银国际环球市场不受美国就有关研究报告准备和研究分析员独立性的规则的约束。负责撰写本报告的全部或部分内容的分析员，未在美国金融业监管局(“FINRA”)注册或获得研究分析师的资格。分析员不受旨在确保分析师不受可能影响研究报告可靠性的潜在利益冲突的相关 FINRA 规则的限制。本报告仅提供给美国 1934 年证券交易法(经修订) 规则 15a-6 定义的“主要机构投资者”，不得提供给其他任何人。接收本报告之行为即表明同意接受协议不得将本报告分发或提供给任何其他人。接收本报告的美国收件人如想根据本报告中提供的信息进行任何买卖证券交易，都应仅通过美国注册的经纪交易商来进行交易。

对于在新加坡的收件人

本报告由 CMBI (Singapore) Pte. Limited (CMBISG) (公司注册号 201731928D) 在新加坡分发。CMBISG 是在《财务顾问法案》(新加坡法例第 110 章)下所界定，并由新加坡金融管理局监管的豁免财务顾问公司。CMBISG 可根据《财务顾问条例》第 32C 条下的安排分发其各自的外国实体，附属机构或其他外国研究机构编制的报告。如果报告在新加坡分发给非《证券与期货法案》(新加坡法例第 289 章)所定义的认可投资者，专家投资者或机构投资者，则 CMBISG 仅会在法律要求的范围内对这些人士就报告内容承担法律责任。新加坡的收件人应致电 (+65 6350 4400) 联系 CMBISG，以了解由本报告引起或与之相关的事宜。