招银国际 CMB INTERNATIONAL
招商银行全资附属机构 A Wholly Owned Subsidiary Of China Merchants Bank

3Q26 大模型发展复盘及展望

单任务成本下降及 RSI 深化推动 AI 使用深度提升

优于大市
(维持)

复盘 3Q26 行业发展：1）基础模型智能能力继续提升，Claude Opus 5.5 与 Sonnet 5.5 领跑全球，中国厂商维持开源模型领域的优势，但开源模型的竞争亦有加剧趋势；2）行业探索突破方向为强化电脑操作、长程软件工程及网络安全能力，推动编程与 Agent 向完整工作流延伸；3）用户规模增长与使用深度提升共同推动商业化进展，而国内大模型厂商通过与海外云平台按调用量分成有望进一步拓宽收入来源。展望而言，我们判断模型能力上界依旧是决定差异化定价能力的关键，每任务成本下降及 Agent 化有望为调用量快速增长提供支撑。高价值且 ROI 具备吸引力、产出结果具备可验证性、数据充足、出错代价可控是推动 AI 在垂类场景渗透进一步提升的关键要素，而 RSI 的深化或有望进一步推动 AI 使用深度提升。股票推荐：1）海外云厂商：建议关注亚马逊（AMZN US）、微软（MSFT US）和谷歌（GOOG US）；2）中国云和 AI 应用：推荐阿里巴巴（BABA US）和腾讯（700 HK）；3）大模型标的：看好智谱（2513 HK）和 MiniMax（100 HK）。

- **行业进展：电脑操作、长程软件工程及网络安全为主要突破方向。** 3Q26 模型真实任务执行能力进一步提升：1）电脑操作与成果编辑逐步整合，Cowork 等产品将资料读取、跨软件操作和文档交付连接起来，推动 Agent 从编程向通用办公进一步扩展；2）软件工程能力由代码生成向长时间自主执行延伸，模型能够围绕同一目标持续开展开发、测试及问题修复；3）网络安全应用向漏洞检测与自动修复拓展。我们认为，长程任务完成能力的提升将扩大 AI 可承接的工作范围，为更高频、持续性的调用提供支撑。
- **行业趋势：用户规模扩张，头部付费群体使用深度提升。** 用户规模层面，ChatGPT 周活由 2026 年的 8 月的 10 亿增至 10 月的 12 亿，美国消费者样本中至少付费订阅 ChatGPT、Gemini 或 Claude 之一的比例，亦由 2025 年 8 月的 2.1%升至 2026 年 8 月的 4.5%。使用深度维度，B 端高强度用户持续增加投入，Ramp 美国企业样本中，2026 年 8 月 Top 1%企业每员工月度 AI 支出达 7,205 美元，+198% YoY，Top 10%及中位企业支出分别同比增长 246%及 165%。C 端付费渗透方面，头部付费人群的支出增长更快，YipitData 数据显示，2025 年 1 月至 2026 年 8 月，美国 Top 1% AI 付费用户平均月支出增长 79%，搭建与自动化、内容创作、事务处理是重要的付费场景。
- **商业模式：海外模型商业化持续深化，国内模型变现渠道进一步拓宽。** 3Q26 海外 AI 平台商业化主要边际变化包括：1）OpenAI 企业业务增长提速支撑年化收入运行率（ARR）快速增长，3Q26 公司企业业务 ARR 同比增长 107%（据 CNBC）；2）广告商业化继续推进。国内 AI 平台商业化进展层面，海外云平台收入分成模式成功拓展。
- **未来展望：RSI 加速模型迭代，每任务成本下降及电脑操作推动 AI 渗透率提升，关注 Agent 进展。** 我们预期，RSI 持续深化有望推动 AI 渗透率进一步提升。Anthropic 披露的 AI 主导研发工作占比于 8 月已升至 26%（3 月：1%）。模型能力提升与每任务成本下降共同改善用户 ROI。电脑操作能力提升使 AI 能够参与更多非编程办公流程，日常工作中的提效案例有望形成积极的示范效应，进一步提升 B、C 端渗透率与使用深度。Agent 进展深化的过程中，我们认为行业的竞争有望逐步从“谁的模型更强”转向“谁能拥有用户的长期工作流”，未来中间层的 SaaS 和自动化工具的价值存在被挤压的风险，而治理和合规有望获得更多的重视。

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

相关报告链接

1. 软件 & IT 服务 – 海外软件选股策略深度报告：在共识中寻新机 – 9 Oct
2. Z.AI (2513 HK; BUY) - Scaling intelligence, scaling monetization – 2 Sep
3. Minimax (100 HK) 1H26 review: strong monetization; extending performance-cost frontier of large model – 27 Aug
4. Alibaba (BABA US) - 2Q26 review: Strong cloud growth prospects with clear ROIC visibility – 21 Aug
5. Z.AI (2513 HK) - GLM-5.3 unveiled with strong capabilities – 14 Aug
6. Tencent (700 HK) - Solid 2Q26 results; gradually capturing AI benefits in the long run – 13 Aug
7. 北美云大厂业绩超预期有哪些投资启示? - 6 Aug
8. Amazon (AMZN US) 2Q results: Solid AWS revenue growth and margin expansion – 3 Aug
9. Alphabet (GOOG US) - 2Q26 results: Strong AI demand continues to drive cloud acceleration and solid core search – 23 Jul
10. Tencent (700 HK) - Solid core business growth; investing to accelerate AI development – 13 Jul
11. AI 主题研究 - 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 – 6 Jul
12. 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

行业发展趋势：模型能力持续提升，新场景应用推动商业化扩张..... 3

大语言模型：模型能力快速迭代，Claude 领跑全球，小米及智谱位居开源前列 4

效率优化：专业能力与任务成本共同决定应用扩张空间 8

行业探索最新进展：电脑操作与专业工具结合，AI 向完整工作流延伸.....10

大语言模型需求展望：用户与使用强度同步提升.....15

多模态：海外厂商领跑图像，中国厂商保持视频优势.....19

龙头基础模型平台模型及业务进展更新.....21

海外 AI 基础模型平台：商业化持续加速，Agent 拓展付费场景 21

国内 AI 基础模型平台：长程能力与推理效率同步改善，API 商业化加速.....27

重点覆盖公司 3Q26 AI 业务进展更新.....35

Meta：C 端 Agent 加速落地.....35

腾讯：WorkBuddy 持续领跑国内办公 Agent 市场 36

阿里巴巴：深化全栈布局、拓展全球数据中心规模 38

股票推荐.....39

行业发展趋势：模型能力持续提升，新场景应用推动商业化扩张

复盘 3Q26 大模型行业发展，我们观察到三个主要趋势：1）基础模型智能能力继续提升，Claude Opus 5.5 与 Sonnet 5.5 领跑全球，中国厂商维持开源模型领域的优势；2）编程与办公 Agent 向长程任务、电脑操作及专业工作流延伸，模型开始参与自身研发，推动持续运行及并行实验带来的 Token 需求；3）推理效率优化与商业模式创新同步推进，订阅、API 及云平台收入分成共同扩展模型厂商的变现空间。

前沿模型继续提升能力上界，中国厂商保持开放权重模型优势。根据 Artificial Analysis 截至 2026 年 10 月 2 日的数据，Claude Opus 5.5 及 Sonnet 5.5 的智能能力指数为 58 及 56 分，10 月发布的 Meta Muse Wonder Full 亦达到 56 分；小米 MiMo-V2.6-Pro 为 46 分，智谱 GLM-5.3 及 Qwen3.8 Max 均为 45 分。专业场景中，模型优势随任务变化；图像生成由 OpenAI 领跑，中国厂商在文生视频保持较强竞争力。

AI 商业化持续加速，增长动力来自用户规模与使用深度同步提升。ChatGPT 周活用户由 2026 年 8 月的 10 亿增至 10 月的 12 亿；Ramp 样本中，8 月 Top 10%及中位企业每员工月度 AI 支出分别为 675.60 美元和 12.50 美元，同比增长 245.5%及 165.4%。高价值用户方面，a16z 引用的美国订阅样本显示，Claude 月费 100 美元及以上订阅占比为 7.3%，高于 ChatGPT 的 1.3%。我们判断，长程编程及专业办公场景的持续拓展，将推动付费用户渗透率与单用户使用量进一步提升。

展望而言，我们判断模型能力上界依旧是决定差异化定价能力的关键，推理效率优化则为更大规模调用增长提供支撑。Agent 产品持续拓宽可执行任务范围，研发自动化进一步增加高强度、长时间运行的推理需求。商业化方面，海外厂商通过分层订阅及企业端变现提升单用户价值，国内厂商则在自营 API 之外，通过与云平台分成扩大开放权重模型的收入来源。我们持续看好基础模型、云基础设施及编程和专业 Agent 应用的增长空间。

大语言模型：模型能力快速迭代，Claude 领跑全球，小米及智谱位居开源前列

3Q26 模型迭代保持高频，9 月为季度内发布最密集的月份。按 OpenAI、Anthropic、Google、DeepSeek、智谱、月之暗面及小米七家重点厂商统计，7-9 月共发布 21 批、27 个核心通用及推理模型版本，其中 9 月发布 14 个版本，占季度总数的 52%。美国厂商以 Anthropic 和 OpenAI 为代表，持续提升前沿模型的编程、电脑操作及长程任务能力，并以不同能力及价格档位推动付费渗透；中国厂商则在开放权重路线下，更重视架构、推理效率及国产算力适配，并通过云平台合作扩大商业覆盖。

图 1：3Q26 核心模型迭代时间轴

厂商	7 月 →	8 月 →	9 月
OpenAI	07/09 GPT-5.6 Sol / Terra / Luna		09/03 GPT-6 Astra 09/22 GPT-6 Sol / Luna 09/29 GPT-6.1 Sol
Anthropic	07/24 Opus 5		09/01 Fable 5.1 / Mythos 5.1◆ 09/22 Opus 5.5 09/28 Sonnet 5.5
Google	07/21 Gemini 3.6 Flash / 3.5 Flash-Lite	08/13 Gemini 3.7 Flash	09/02 Gemini 3.8 Flash 09/30 Gemini 4 Argon◆
DeepSeek	07/31 V4-Flash-0731	08/13 V4-Pro-0813 08/21 V4-Flash-Vision-Exp	09/10 V4.1-Flash
智谱		08/14 GLM 5.3 08/26 GLM 5.3 Flash	
Moonshot	07/16 Kimi K3		09/11 Kimi K2.8 Preview
小米			09/22 MiMo V2.6 Pro / Flash

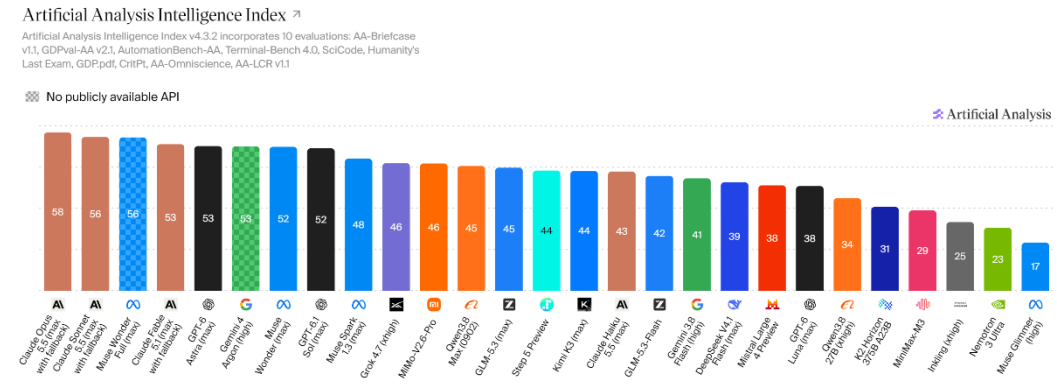
资料来源：公司资料，招银国际环球市场

注：1）统计周期为 2026 年 7 - 9 月；2）同厂商同日发布计一批，独立型号分别计版本，受限及预览型号在图中以◆标注。

从模型能力表现来看，根据 Artificial Analysis 智能能力指数，当前行业综合能力排名领先的模型为 Claude Opus 5.5（58 分）、Claude Sonnet 5.5（56 分）和 Meta Muse Wonder Full（56 分），其后为 Claude Fable 5.1、OpenAI GPT-6 Astra 及 Google Gemini 4 Argon（均为 53 分），GPT-6.1 Sol 为 52 分。Claude Opus 5 至 Opus 5.5 的综合能力由 51 分提升至 58 分，GPT-5.6 Sol 至 GPT-6 Astra 由 47 分提升至 53 分，海外头部厂商持续推升前沿模型能力。

而中国厂商中小米 MiMo-V2.6-Pro 以 46 分领跑，智谱 GLM-5.3 及阿里巴巴 Qwen3.8 Max 均为 45 分，月之暗面 Kimi K3 和阶跃星辰 Step 5 Preview 均为 44 分，构成国内第一梯队。分项来看，Kimi K3、Step 5 Preview 及 MiMo-V2.6-Pro 在长上下文推理评测 AA-LCR 中的成绩分别为 89%、88%及 86%；GLM-5.3 在 Terminal-Bench 4.0 中的成绩为 42%，高于 MiMo-V2.6-Pro 的 35%及 Kimi K3 的 13%。月之暗面、阶跃及小米在长上下文能力方面表现突出，智谱则在终端编程任务中表现较强。

图 2：主流大语言模型能力排名

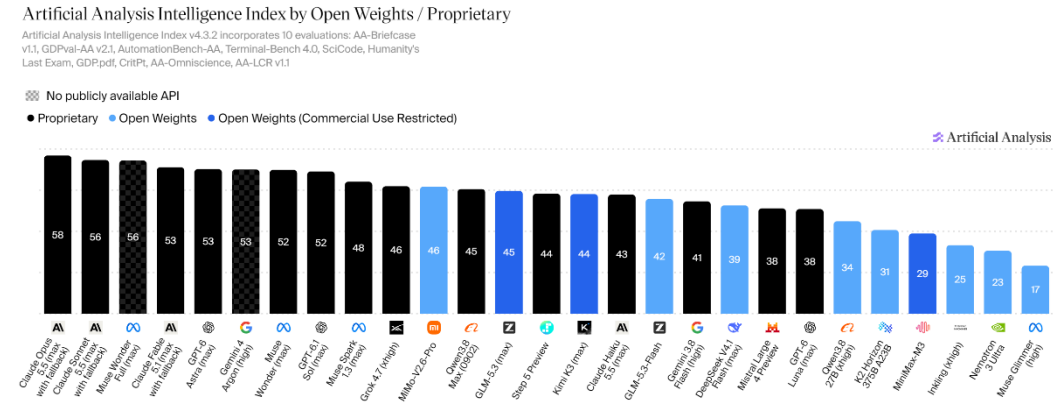


资料来源：Artificial Analysis，招银国际环球市场

注：Opus 5.5、Sonnet 5.5 及 Fable 5.1 采用 Adaptive Reasoning、Max Effort 及 Default Fallback；Haiku 5.5 及 Muse Wonder 系列采用 Max 档。

从开源和闭源模型的对比来看，闭源模型仍维持能力上界优势，中国厂商继续领跑开源权重模型。小米 MiMo-V2.6-Pro 以 46 分位列第一，智谱 GLM-5.3 及月之暗面 Kimi K3 分别为 45 分和 44 分，GLM-5.3-Flash 及 DeepSeek V4.1 Flash 分别为 42 分和 39 分。当前开源权重领先模型与闭源领先模型的综合评分相差 12 分。

图 3：开源及闭源模型能力对比及智能能力指数

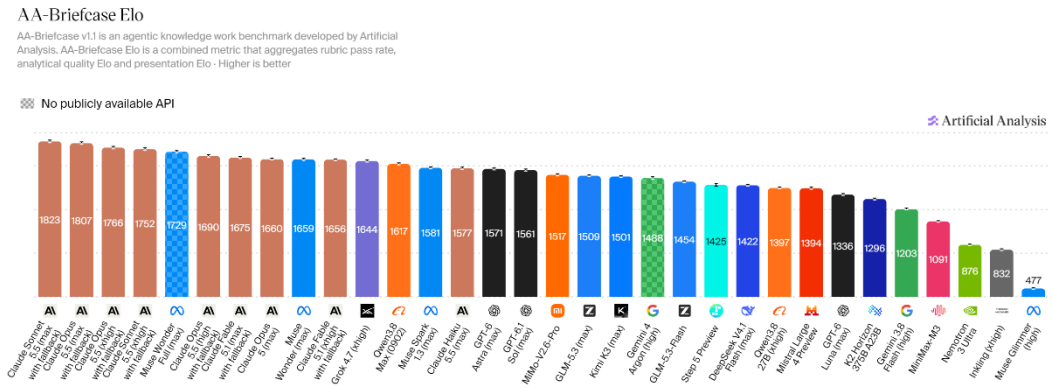


资料来源：Artificial Analysis，招银国际环球市场

注：蓝色为开源权重模型，黑色为闭源模型。

在智能体领域，Anthropic 继续领跑实际工作任务交付。Claude Sonnet 5.5 及 Opus 5.5 在 AA-Briefcase 中的 Elo 分别为 1,823 及 1,807，Muse Wonder Full 及 Wonder 为 1,729 及 1,659；中国厂商中，Qwen3.8 Max、MiMo-V2.6-Pro 及 GLM-5.3 分别为 1,617、1,516 及 1,509。GLM-5.3 较 GLM-5.2 在同口径评测中的 AA-Briefcase Elo 由 1,230 升至 1,509，AutomationBench-AA 业务流程自动化任务通过率由 28%升至 62%，模型迭代进一步强化了工具调用、业务流程执行及工作成果交付。

图 4：大模型智能体工作任务能力对比

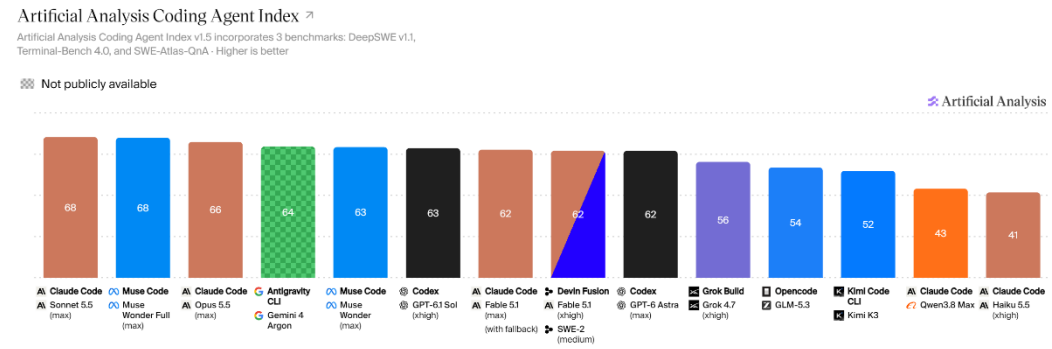


资料来源：Artificial Analysis，招银国际环球市场

注：AA-Briefcase v1.1 衡量模型执行知识工作并交付文件成果的能力。

在商业化得到广泛验证的编程领域，Claude Code/Sonnet 5.5 与 Muse Code/Muse Wonder Full 的能力指数均为 68 分，Claude Code/Opus 5.5 为 66 分；Antigravity CLI/Gemini 4 Argon、Codex/GPT-6.1 Sol（Xhigh）及 Codex/GPT-6 Astra（Max）分别为 64、63 及 62 分。中国模型中，OpenCode 搭配 GLM-5.3 为 54 分，Kimi Code CLI 搭配 Kimi K3 为 52 分。分项来看，Codex 搭配 GPT-6.1 Sol 在 DeepSWE 软件工程评测中的通过率为 73%，Claude Code 搭配 Sonnet 5.5 为 72%，Kimi Code CLI 搭配 Kimi K3 为 68%。

图 5：大模型编程能力排名

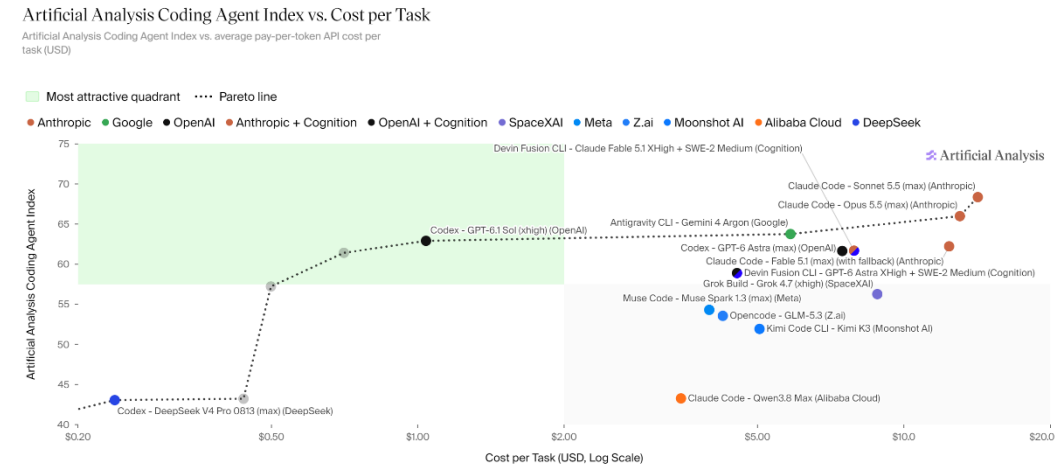


资料来源：Artificial Analysis，招银国际环球市场

注：Coding Agent Index v1.5 由 DeepSWE v1.1、Terminal-Bench 4.0 及 SWE-Atlas-QnA 三项评测得分等权平均计算。图中为编程工具或框架与底层模型的组合。

同一能力区间内，单任务费用差异已较为显著。Codex 搭配 GPT-6.1 Sol 的编程能力指数为 63 分，略高于 GPT-6 Astra 组合的 62 分，但单任务测试费用由 7.47 美元降至 1.04 美元、降幅约 86%，任务耗时由 29.4 分钟降至 15.5 分钟、缩短约 47%。均衡模型能力提升与 Harness 优化，使更多日常编程任务能够以较低费用完成，也为更复杂、持续时间更长的任务释放调用预算。

图 6：GPT-6.1 Sol 与 Astra 的编程能力及单任务费用对比



资料来源：Artificial Analysis，招银国际环球市场

注：Coding Agent Index v1.5 由 DeepSWE v1.1、Terminal-Bench 4.0 及 SWE-Atlas-QnA 三项评测得分等权平均计算。图中为编程工具或框架与底层模型的组合。

专业场景的模型表现进一步体现差异化。Artificial Analysis 的 Capability Scores 覆盖金融与会计、战略与运营、法律、医疗与健康、工程及经济学六类场景，按职业任务需求，对领域知识、推理、知识工作及工具使用等能力加权。其中，战略与运营的智能体知识工作与工具使用权重合计占 65%，工程与经济学的领域知识与推理权重合计分别占 65%及 70%，法律和医疗另设 10%的非幻觉权重。不同场景对能力组合的要求，推动基础模型向专业工作进一步延伸。

海外模型中，Claude Opus 5.5 在六类场景均保持领先。截至 2026 年 10 月 13 日，在图 7 选取的 18 款代表模型中，Opus 5.5 六项得分均为最高，分别为 60.7、63.7、63.3、60.5、60.4 及 65.6 分。Sonnet 5.5 工程得分为 58.4，接近 Opus 5.5 的 60.4；Gemini 4 Argon 法律得分为 60.0，高于 GPT-6 Astra 的 58.5 及 Sonnet 5.5 的 56.4。

国内模型呈现不同的专业场景优势。在图 7 选取的 7 款国内模型中，Kimi K3 在金融与会计、法律及经济学三项居首，分别为 46.9、48.6 及 53.4 分，其中经济学与 MiMo-V2.6-Pro 的 53.2 分接近；DeepSeek V4.1 Flash 在战略与运营为 51.7 分，高于 Kimi K3 的 50.5 分及 MiMo-V2.6-Pro 的 47.8 分；GLM-5.3 在医疗与健康为 46.7 分，在工程为 47.1 分，后者与 MiMo-V2.6-Pro 的 46.9 分接近。由此可见，综合指数中的国内领先模型并未包揽所有专业场景首位，垂直应用需要进一步比较领域知识与执行能力的组合。

图 7：中外代表模型的六类专业场景能力评分

模型及评测配置	金融与会计	战略与运营	法律	医疗与健康	工程	经济学
Claude Opus 5.5 Max	60.7	63.7	63.3	60.5	60.4	65.6
Claude Sonnet 5.5 Max	57.2	59.6	56.4	58.1	58.4	60.8
Muse Wonder Full Max	54.2	57.8	54.7	—	57.0	60.7
Claude Fable 5.1 Max	56.4	59.7	60.5	58.0	56.5	62.7
GPT-6 Astra Max	55.2	57.3	58.5	51.7	54.7	59.8

模型及评测配置	金融与会计	战略与运营	法律	医疗与健康	工程	经济学
GPT-6.1 Sol Max	54.0	57.0	57.7	51.2	53.9	59.3
Gemini 4 Argon High	53.8	56.2	60.0	—	55.2	60.4
Muse Wonder Max	51.3	54.8	52.0	—	53.8	57.7
Muse Spark 1.3 Max	49.8	51.2	50.9	49.7	49.4	56.3
Grok 4.7 Xhigh	51.9	54.4	54.1	47.2	47.4	55.7
Claude Haiku 5.5 Max	43.9	40.4	42.3	44.1	44.9	51.2
MiMo-V2.6-Pro	46.1	47.8	44.8	41.8	46.9	53.2
Qwen3.8 Max 0902	45.6	47.3	46.1	41.4	46.4	52.3
GLM-5.3 Max	44.7	47.6	42.0	46.7	47.1	51.2
Step 5 Preview	45.4	47.1	46.2	40.8	45.0	52.8
Kimi K3 Max	46.9	50.5	48.6	45.0	43.3	53.4
DeepSeek V4.1 Flash Max	45.0	51.7	41.3	40.6	39.3	44.4
MiniMax-M3	30.1	25.0	31.5	29.7	31.2	43.7

资料来源：Artificial Analysis，招银国际环球市场

注：1）数据截至 2026 年 10 月 13 日；2）指数得分越高表明能力越强，保留一位小数；3）各场景基准及权重不同，分数仅在相同场景内可比、并不适合在场景间交叉对比；4）模型档位按 AA 标注列示；5）“—”表示 AA 未提供完整评分，不代表零分。

模型在专业场景中的差异，为行业 Agent 应用提供了更多的新产品开发空间。战略运营类任务更依赖跨工具执行，工程和经济分析更依赖领域知识与推理，法律及医疗则重视事实可靠性和长篇资料处理。我们认为，行业应用厂商通过接入客户资料、专业软件及业务流程，可以在基础模型之上提供更完整的工作成果，并围绕使用量、席位及交付价值形成收费。

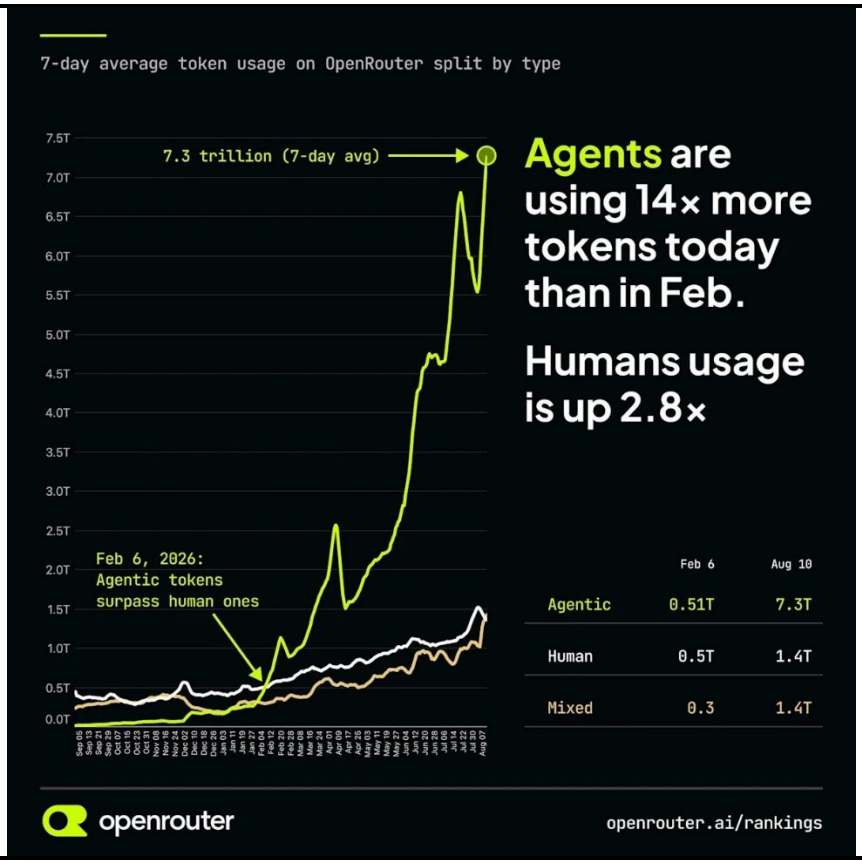
效率优化：专业能力与任务成本共同决定应用扩张空间

专业能力接近时，调用费用差异影响企业可持续运行的任务规模。截至 2026 年 10 月 2 日，Artificial Analysis 公布的 Max 档数据中，GPT-6.1 Sol 与 GPT-6 Astra 的战略运营评分分别为 57.0 及 57.3 分，平均每评测任务费用为 0.67 及 3.04 美元；金融与会计评分分别为 54.0 及 55.2 分，费用为 0.55 及 2.45 美元。Sol 在两项场景的费用均低约 78%，为日常文档处理、重复分析和持续办公提供了不同的能力与费用选择。

推理与执行框架优化共同提高算力利用效率。架构和注意力优化减少长上下文计算及存储开销，缓存复用降低重复输入费用，服务系统优化缩短生成时间；Harness 则通过上下文管理和工具反馈，减少无效步骤与重复调用。对企业而言，模型调用价格下降与任务成功率提升同时发生，意味着相同预算能够支持更多分析、验证及持续运行的 Agent 任务。

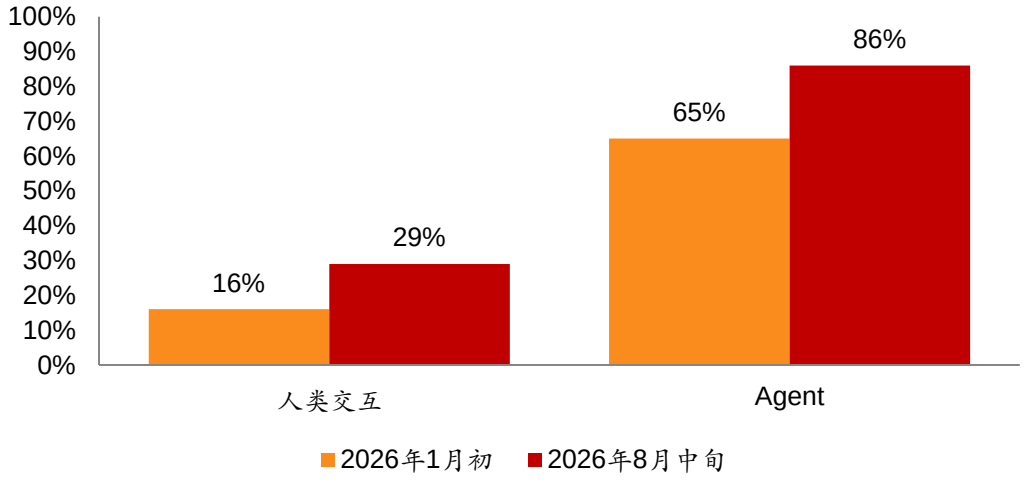
调用成本下降有望进一步扩大 Token 需求。随着模型能够处理更复杂的工作，客户节省的费用有望部分转化为新增调用，包括更长的上下文、更多轮次的验证及多个 Agent 并行执行。Agent 使用强度已明显高于人工聊天： OpenRouter 数据显示，2026 年 8 月 Agent 七日平均 Token 用量达到 7.3 万亿，较今年 2 月 6 日增长约 14 倍；同期 Agent 的缓存 Token 占比由年初 65%升至 86%，高于人工交互的 29%。缓存降价降低长任务的重复输入费用，也支持 Agent 在同一上下文内持续推进工作。

图 8：2 月 6 日至 8 月 10 日，agents 7 天平均 token 使用量上涨 14 倍，而人类使用量上涨 2.8 倍



资料来源：OpenRouter，招银国际环球市场
注：数据截至 2026 年 8 月 10 日

图 9：Agent 的缓存 Token 占比显著高于人工交互



资料来源：a16z、OpenRouter，招银国际环球市场

行业探索最新进展：电脑操作与专业工具结合，AI 向完整工作流延伸

大模型应用的探索重点进一步转向完整任务执行。3Q26 我们观察到：1）电脑操作与成果编辑结合，编程 Agent 向非编程办公场景扩展；2）行业资料、连接器及专业软件接入，推动 AI 进入法律、金融和工程等工作流；3）研发自动化由辅助编程向实验执行及问题修复推进。这些变化使 AI 承接的任务由短对话转向连续工作，增加长时间运行和多 Agent 并行的推理需求。

■ Harness 优化与开放推动模型能力落地

Harness 优化使模型在参数不变的情况下提升任务表现，成为全栈效率竞争的重要环节。Harness 负责组织上下文、工具调用、任务状态及执行过程。据 OpenAI 2026 年 8 月 19 日披露，保持 GPT-5.6 Sol 模型不变，通过保留推理记录及压缩上下文，ARC-AGI-3 成绩由 13.3%升至 38.3%，输出 Token 消耗降至原来的六分之一。LangChain 于 2 月 17 日披露，在保持 GPT-5.2-Codex 不变的情况下，调整提示词、工具及执行流程，使 Terminal-Bench 2.0 成绩由 52.8%升至 66.5%。模型能力与执行框架协同优化，将同时影响 Agent 的任务成功率、执行速度及调用费用。

通用执行框架开放，行业应用的差异化进一步落在资料与业务流程。OpenAI 于 2026 年 8 月 19 日披露，Codex 开源执行框架可通过 SDK 及 app-server 嵌入第三方产品，开发者能够复用上下文管理、工具调用、沙箱及审批等能力。DeepSeek Harness 于 9 月 29 日开放开发者预览及源码，将模型、工具、会话、沙箱和调度组件插件化，支持执行记录追踪、恢复及重放。我们认为，通用组件的开放将降低 Agent 开发门槛，行业应用厂商可更集中地投入机构资料、专业软件接入及业务流程，围绕特定行业需求形成产品差异。

■ 应用场景从通用办公延伸至专业行业与模型研发

电脑操作与成果编辑逐步整合，通用 Agent 扩展至非编程办公任务。OpenAI 于 2026 年 7 月 9 日推出 ChatGPT Work，将应用、本地文件及网页接入持续执行过程，并生成可编辑的文档、表格和演示文稿；Anthropic 于 9 月 16 日宣布逐步合并 Cowork 与 chat，并以测试功能推出 Claude Docs 及 Slides。非编程工作已有使用基础：Anthropic 公布的 5 月 11-31 日 120 万次 Cowork 会话样本中，业务流程与运营、内容创作与文案分别占 33.4%和 16.4%。我们认为将资料读取、跨软件操作和成果编辑连成完整流程，有望提高 AI 在日常办公中的参与深度，并扩大可重复执行的任务量。

法律应用通过接入文档、条款规则和专业软件，将 AI 嵌入合同审阅与尽职调查。Anthropic 于 2026 年 5 月 12 日推出法律行业方案，新增 20 多个连接器及 12 个专业插件，支持按团队条款手册起草、修订和逐条比较合同，并将结果继续用于邮件、清单及汇报材料。据 Anthropic 披露，Harvey 在不足一个月内完成 Claude 的平台集成，覆盖 Assistant、Vault 及 Workflows 中的文档审阅与起草。机构资料和审阅规则的接入，使模型可以承担一项法律事务中的多个连续步骤，也为专业应用持续扩容提供基础。

金融应用向财富管理延伸，跨系统信息整理成为新增任务入口。Anthropic 于 9 月 14 日推出 Claude for Financial Advisors，连接客户关系管理、投资组合、托管及财务规划系统，提供会前准备、组合分析和合规检查等工作流，由顾问审阅结果并批准关键操作。银行内部已有多场景部署：据 Anthropic 的 N26 案例，首年上线 15 个以上内部应用，覆盖多语言客服、拒付处理及金融犯罪调查支持，在目标流程中实现最高 70%的任务自动化。我们预计，随着客户资料、分析及沟通环节进一步贯通，金融 AI 有望提高单个团队可服务的客户数量和工作范围。

专业应用已形成新增付费，客户扩容与邻近行业延伸提供增长路径。法律人工智能初创企业 Harvey 于 2026 年 7 月 16 日披露，2Q26 净新增 ARR 超过 1 亿美元；公司已服务超过 125 家资产管理机构，包括 Blue Owl、Bridgewater 及 KKR，并收购 Benchmark，将尽职调查和交易文档分析进一步延伸至投资决策流程。法律与资产管理共享大量文档审阅、资料提取和知识复用需求，成熟产品可沿既有客户扩大部署岗位，并向相邻专业市场拓展。

芯片设计展示了通用模型与专业工具结合后，任务范围由验证调试向设计生成扩展的路径。Cadence 于 2026 年 2 月推出 ChipStack AI Super Agent，通过调用 EDA 工具执行设计、测试和问题修复；9 月 22 日再推出 RTL Generation Agent，支持从设计规格生成 RTL 代码，并按功耗、性能和面积要求迭代优化。公司早期评估显示，相比直接使用基础模型生成代码，设计面积和功耗平均分别降低 24%及 18%；相关新增能力计划于 4Q26 向部分早期客户开放。专业软件提供可检验的反馈，使 AI 能够围绕同一设计目标持续试验和改进，工具接入与验证流程因而成为工程应用的重要产品基础。

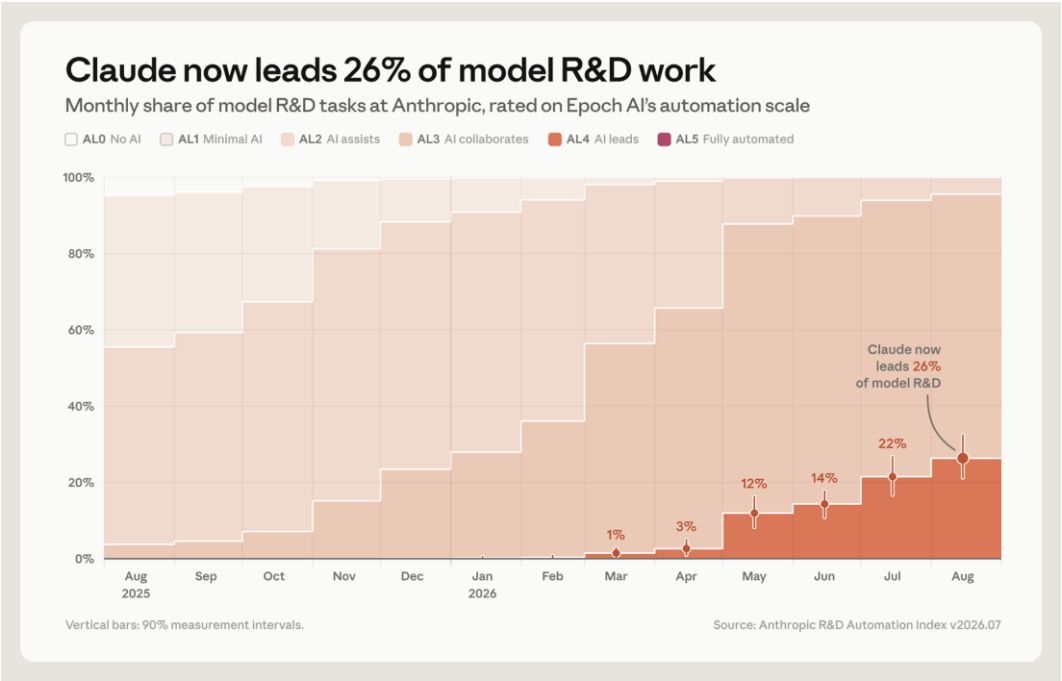
■ **递归自我改进（RSI）：模型参与自身研发，实验并行化扩大 Token 需求**

模型研发自动化正由辅助编程向连续实验执行推进。递归自我改进（RSI）的目标是让模型参与自身研发，通过代码、实验及优化工作改进下一代模型，再由更强的模型提高后续研发效率。我们观察到，当前落地主要集中于三类任务：1）编写和测试研发代码；2）读取实验日志、定位问题并修复；3）并行开展实验及比较结果。研发任务持续时间更长，且可以由多个 Agent 同时推进，因此模型能力提升有望带动更多内部推理需求。

Anthropic 的 AI 主导研发工作占比已升至 26%，并行 Agent 规模达到约 3 万个。据 Anthropic 9 月披露，按 2026 年 7 月固定研发任务篮子及人员投入加权，Claude 主导工作占比由 3 月 1%升至 5 月 12%、7 月 22%及 8 月 26%，达到 AI 协作及以上水平的占比超过 90%。在 AI 主导的任务中，研究员给出目标，Claude 自行读取日志、定位问题、编写及测试修复方案，研究员负责监督和部署决策。同期，公司最常用内部平台约有 3 万个 Agent 并行开展研究及工程任务，研发执行由单个助手扩展至大规模并行。

研发自动化已带来具体工程产出。Anthropic 披露，截至 5 月，合入代码库的代码中超过 80%由 Claude 编写，2Q26 工程师人均每日合入代码量达到 2024 年的约 8 倍；4 月，Claude 完成 800 多项修复，使一类 API 错误减少约三个数量级。模型连续执行编写、测试与修复，减少研究员的逐步操作，支持更多排错和优化任务并行推进。

图 10：Anthropic 模型研发工作中 AI 主导占比升至 26%



资料来源：公司资料，招银国际环球市场

OpenAI 已达到"自动化研究实习生"阶段，研发 Agent 投入继续上升。据公司 2026 年 9 月 6 日披露，系统已能在人类指导下完成原需熟练研究员数日的明确研发任务，下一目标为 2028 年 3 月的自动化 AI 研究员。截至 8 月中旬，研究组织每个人类工作日对应约 3.1 个 Agent 工作日，按累计运行时间及 8 小时工作日折算。8 月研究员每日推理用量按 API 价格计，中位数超过 600 美元、90 分位超过 7,000 美元，每名活跃实验人员的实验数量创 2025 年 1 月以来新高。我们预计研发自动化将成为模型厂商内部新增的高强度推理场景，并通过更密集的实验推动后续模型迭代。

RSI 的推进将进一步拉动 Token 用量。OpenRouter 数据显示（图 8），Agent Token 用量超过人工交互，自动化执行正成为重要的推理需求来源。随着模型更多参与代码编写、实验设计、结果分析及反复验证，模型研发将产生持续、多轮的推理调用；多 Agent 并行运行则增加同一研究团队能够同时推进的实验数量。模型能力提升还将扩大可自动化的研发范围，使 Agent 能够承担更完整的研发流程。由此，研发效率的改善有望转化为更多实验和更快迭代，推动模型研发本身成为持续增长的 Token 需求来源。

■ 数据控制与操作审计支持更深的任务授权

企业数据与操作权限管理完善，支持 Agent 接入更多核心业务。Anthropic 于 2026 年 9 月 1 日宣布 Enterprise Frontier Safeguards（EFS），与金融、法律、医疗等行业的 100 多家客户共同设计，计划于秋季稍晚分阶段推出。方案将活动数据保存在客户控制的云环境中，自动识别异常并交由客户团队审查。权限、审计及验证与现有系统结合，使企业可以向 Agent 开放更多内部资料和连续操作流程，扩大 AI 在专业场景中的部署范围。

图 11：前沿大模型最新进展及参数

厂商	最新旗舰/主力模型	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 4 Argon（限定开放）/Gemini 3.8 Flash	据 Google 2026 年 9 月 30 日公告，Gemini 4 Argon 先向 Fairwind 经审核网络安全客户限定开放；计划开放推广价为 2/10 美元每百万 Token，缓存输入折扣 95%。公开主力 3.8 Flash 继续通过 Gemini、Antigravity 及 API 提供服务。	Argon 强化长程编程、自动化及多模态推理；公司评测 DeepSWE 1.1、AutomationBench 及 LVBench 得分分别为 77.9%、51.3%及 91.7%，支持最高 1M 输出，用于更长代码及研究产物生成。	长程软件工程、自动化任务、多模态专业工作及网络安全。	Argon 输入未公布、最高输出 1M；3.8 Flash 输入 1M
OpenAI	GPT-6 Astra/GPT-6.1 Sol/GPT-6 Luna	GPT-6.1 Sol 于 9 月 29 日接入 Work、Codex 及 API，标准输入/输出价为 2/10 美元每百万 Token（输入≤272K）。10 月 8 日新增 Ultrafast，Codex 中 Token 生成速度最高为标准模式 8 倍，短输入价格为 12/60 美元。	公司 DeepSWE 1.1 评测中，6.1 Sol high 档 75.2%，高于 6 Sol max 档 68.8%，任务费用下降 76%；OSWorld 2.0 离线集 partial 口径 max 档 71.4%，接近同档 Astra 的 73.5%，每任务费用约为其七分之一。	电脑操控、跨应用业务流程、复杂编程、文档及数据分析；通过 Astra、Sol、Luna 分别覆盖高能力、均衡及高频低成本任务。	1.05M；最高输出 128K
Anthropic	Claude Fable 5.1/Opus 5.5/Sonnet 5.5/ Haiku 5.5	Opus 5.5、Sonnet 5.5 输入/输出价分别为 4/20、2/10 美元每百万 Token。10 月 7 日新增 Haiku 5.5，输入≤100K 时为 0.10/0.50 美元，超过 100K 为 0.50/2.50 美元；Sonnet 缓存读取降至 0.10 美元。	Sonnet 5.5 的 Terminal-Bench 4.0 由 10.3%升至 70.6%，OSWorld 2.1 partial 由 57.0%升至 80.1%。Haiku 5.5 平均运行成本较 Haiku 4.5 下降约 75%，OSWorld 2.1 离线子集得分 72.4%。	长程编程、电脑操控、跨应用办公、金融及法律等知识工作；重点改善单位任务成本。	1M；最高输出 128K
SpaceXAI	Grok 4.7	据 xAI 2026 年 9 月 21 日公告，Grok 4.7 上线 Cursor、Grok Build 及 API，原生用于 Grok Bot。API 低于 200K 输入时，标准输入/输出价为 2/6 美元每百万 Token；长输入按两倍收费。	采用更大的新基座，并通过持续数小时的强化学习任务增强长程执行；公司评测 CursorBench 4 由 Grok 4.6 的 40.4%升至 46.3%，Terminal-Bench 4 由 20.3%升至 37.6%，强化持续编程及自主工作。	长程编程 Agent、交互及视觉应用开发、知识工作、工具调用与推理优化。	500K
Meta	Muse Wonder / Muse Wonder Full	10 月 12 日发布。Wonder 通过 Meta Model API 及 Muse Code 开放，输入/输出价为 4/20 美元每百万 Token、缓存读取 0.20 美元；Full 面向 Muse Code 订阅用户。	强化跨会话记忆、持续执行及错误恢复。公司评测中，Wonder Full 的 DeepSWE 1.1 为 79.0%，SWE-Marathon 为 57.5%（xhigh 档），支持长程软件开发及专业任务。	长程软件工程、电脑操作、跨会话个人 Agent 及金融、法律等专业任务。	最高 1M
智谱 Z.AI	GLM-5.3/GLM-5.3-Flash/GLM-5.3-FlashX/GLM-Legal	Flash 采用 320B 总参数/18B 激活参数及混合注意力，注意力计算量约为旗舰 1/3，KV 缓存约为 1/4.4；Flash/FlashX 输入输出价为 0.15/0.50 及 0.37/1.25 美元每百万 Token。10 月 12 日发布 GLM-Legal 并开放 API 试用申请。	GLM-5.3 沿用 753B 基座，后训练使内部 Code Bench 表现提升 50%；Terminal-Bench 3.0 由 4.6%升至 28.3%，DeepSWE 1.1 由 46.2%升至 66.9%。Flash 在同框架 DeepSWE 1.1 达 63.4%、Toolathlon 达 78.4%，兼顾视觉理解与高频执行。	长程软件工程、复杂工具调用、模型研发及推理系统优化、网络安全；Flash 侧重多模态办公与高频调用；GLM-Legal 覆盖合同审阅、证据分析及合规研究。	1M

厂商	最新旗舰/主力模型	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）
稀宇科技 MiniMax	MiniMax M3 / M3.1-Flash-Preview	M3 总参数约 428B、激活 23B，采用 MSA 稀疏注意力，支持图像及视频输入。9 月下旬上线 M3.1-Flash-Preview，延续 1M 上下文及多模态输入，新增五档推理深度，通过 M Plan 及 MiniMax Code 提供服务。	将编程、原生多模态与百万 token 上下文整合，用于长程任务执行。公司案例显示，M3 可连续约 12 小时复现论文；在约 24 小时 GPU 算子优化中完成 147 次基准提交、1,959 次工具调用，将硬件峰值利用率由 7.6%提高至 71.3%，对应 9.4 倍加速。	长程编程、科研复现、GPU 算子优化、多模态工具调用及低成本 Agent。	最高 1M；官方 API 保证至少 512K
小米	MiMo-V2.6 Pro/Flash/Pro UltraSpeed	据小米 2026 年 9 月 22 日更新，V2.6 上线 Studio、MiMo Code、桌面端及 API，开放 Pro-RL 与 Flash-RL 权重；Pro 总参数 1.02T/激活 42B，Flash 为 309B/15B。UltraSpeed 维持 Pro 能力，输出速度最高提升 20 倍。	相比 V2.5 Pro，统一扩大编程、通用 Agent、视觉及网络安全强化学习，并通过模型参与评分改进任务执行路径。公司评测中，DeepSWE 1.1 由 19.0%升至 71.9%，MiMo Code Bench 由 40.4%升至 63.2%，Toolathlon-Verified 由 49.1%升至 76.9%；V2.6 Pro 的 OSWorld-Verified 达 82.0%。	长程编程、电脑操控、原生全模态 Agent、网络安全及强化学习驱动的自我改进。	1M
阶跃星辰 StepFun	Step 5 Preview	Step 5 Preview 9 月 20 日上线，已通过产品及 API 提供服务，计划 10 月 15 日开放权重；采用 600B 总参数/27B 激活参数的 MoE 架构，支持视觉输入。	从效率型 Flash 模型向复杂工程及专业工作拓展；公司评测中，High 档 StepCodeBench 为 49.0%，FrontierFinance 为 66.4%。在 H100 的 MLA 算子优化测试中，四次独立尝试取最佳结果，约 22 小时达到 508 TFLOPS，体现长程执行、测试与反馈优化能力。	复杂编程、金融分析、长程 Agent、GPU 算子及推理效率优化。	1M
月之暗面	Kimi K3/K2.8 Preview	Kimi K3 于 7 月 16 日发布、7 月 27 日开放权重，总参数 2.8T/激活 104B，支持原生视觉。9 月 11 日 Kimi Code 新增 K2.8 Preview，支持 1M 上下文及 low/high/max 推理档。来源：公司模型卡及 Kimi Code 更新记录。	K3 每步从 896 个专家中激活 16 个，采用 KDA 注意力及 Attention Residuals；公司测算扩展效率提高约 2.5 倍，支持跨小时编程及科研任务。K2.8 Preview 以更高推理效率承接接近 K3 水平的编程任务，扩展编程任务的效率选择。	长程编程、科研与 GPU 优化、视觉设计、可编辑办公文档、复杂知识工作及多 Agent 协作。	K3 及 K2.8 Preview：1M
DeepSeek	DeepSeek-V4.1-Flash/DeepSeek-V4-Pro-0813	据 DeepSeek 2026 年 9 月 10 日更新，V4.1-Flash 主干 552B、另含 196B Engram 条件记忆参数，预填充/解码阶段分别激活 8B/16B 参数，加入原生视觉；V4-Pro 正式版继续服务。	V4.1-Flash 采用非对称编码器—解码器架构；公司 max 档 Terminal-Bench 2.1/DeepSWE 1.1 达 90.6%/74.2%，高于 V4-Flash 的 82.7%/54.4%。KV 缓存约 890 字节/Token，HBM 及 SSD 缓存分别降至前代 1/4 及 1/8。	低成本多模态 Agent、复杂编程、长上下文、工具调用、专业推理及高吞吐部署。	1M；V4.1-Flash 最高输出 384K
字节跳动 豆包	Seed 2.1 Pro/Seed 2.1 Turbo	据字节跳动 2026 年 6 月 23 日公告，Seed 2.1 系列正式发布，通过豆包及火山引擎提供服务；Pro 面向高复杂度任务，Turbo 面向规模化生产，强化编程、Agent 及多模态理解。	相比 Seed 2.0，强化跨工具、跨环境任务交付，以及从需求分析、环境配置到验证的全流程编程。通过强化学习协调 GUI 操作与非 GUI 工具调用，公司披露平均任务执行步骤减少 16%；同步改善复杂文档、图表、视频及空间信息理解。	电脑及手机操控、跨应用办公、端到端软件工程、多模态理解，以及模型研发工作流。	256K

厂商	最新旗舰/主力模型	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）
阿里巴巴千问	Qwen3.8-Max-0902/Qwen3.8-Flash/Qwen3.8-Omni-Flash	8 月 3 日发布 Qwen3.8-Max，8 月 13 日推出 2.4T 总参数/95B 激活参数的开放权重版，8 月 26 日推出 Flash；9 月 2 日更新 Max 快照，9 月 18 日及 21 日新增 Omni-Flash 及实时版本。	相比 3.7 系列，Max 强化原生视觉、复杂工程及多工具协作；0902 快照进一步改善长程自主开发、图表推理及文档解析。Flash 兼容主流编程 Agent 接口，承接高并发调用；Omni-Flash 统一处理文本、图像、音频和视频，将 Agent 执行扩展至多媒体制作与实时交互。	长程编程、GUI 操作、文档及知识工作、多模态内容制作、音视频交互及高并发 Agent。	1M；此处对应 Max、Flash 及非实时 Omni-Flash

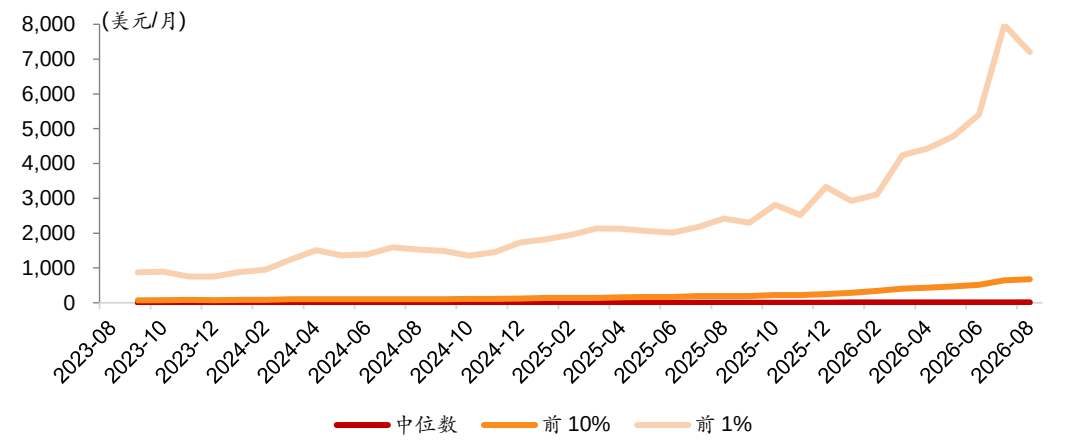
资料来源：公司资料，Hugging Face，Artificial Analysis，Reuters，招银国际环球市场
注：数据截至 2026 年 10 月 13 日

大语言模型需求展望：用户与使用强度同步提升

我们预期，Agent 将成为 Token 需求增长的核心驱动力。需求增长同时来自两方面：1）用户基数扩大，ChatGPT 周活及企业 AI 付费采用率继续提升；2）单用户使用强度上升，编程、专业办公及研发 Agent 增加连续执行时间，并带动多个任务并行。与短对话相比，Agent 需要持续读取上下文、调用工具及验证结果，推理成本下降将进一步扩大这些任务的使用规模。

企业 AI 使用强度仍有较大提升空间，而每任务成本下降或成为关键助推因素。根据 Ramp AI Index，2026 年 8 月美国 Top 10%及中位企业每员工月度 AI 支出分别为 675.60 美元和 12.50 美元，环比增长 4.8%及 5.2%，同比增长 245.5%及 165.4%，两者支出强度相差约 54 倍。Top 1%企业同期支出为 7,205.13 美元，同比增长 198.0%。从更广泛的企业群体看，AI 投入仍保持较快增长，高强度用户已形成的研发、编程及办公用法，具备向更多部门和岗位推广的空间。

图 12：企业每员工月度 AI 支出：中位及前 10%企业投入继续增长

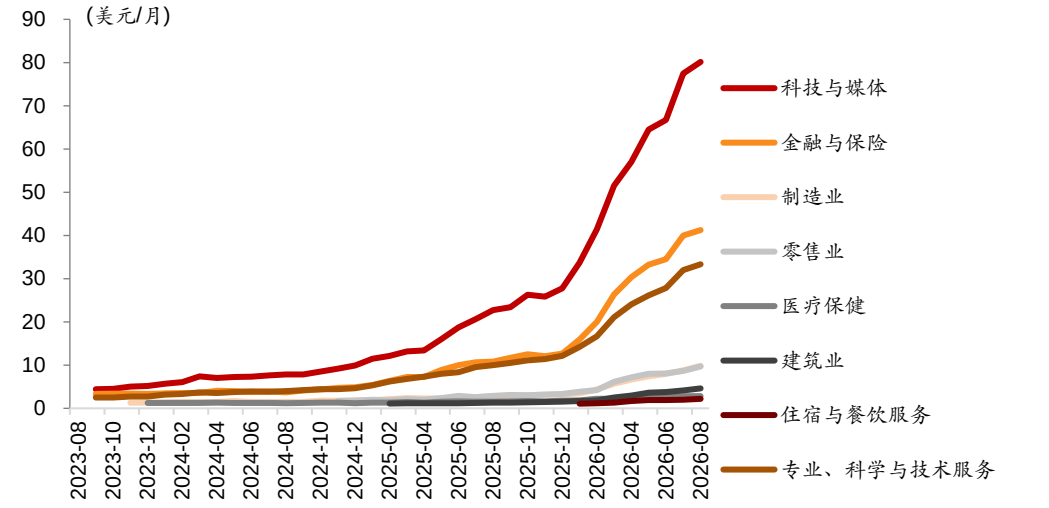


资料来源：Ramp AI Index，招银国际环球市场；数据截至 2026 年 8 月
注：基于 Ramp 美国企业样本，按每员工月度 AI 支出分组；支出涵盖模型订阅、API 及其他 AI 产品和服务

知识密集行业投入领先，专业工作流具备向更多岗位推广的基础。据 Ramp AI Index，2026 年 8 月样本企业 AI 付费采用率为 56.1%；科技与媒体、金融与保险、专业科学及技术服务行业的每员工月度 AI 支出中位数分别为 80.15、41.24 和 33.33 美元，同比增长 252.5%、281.5%和 233.3%，高于制造业的 9.84 美元及零售业的 9.68 美元。我们判断，信息整理、分析及文档交付占比较高的行业，更容易围绕模型形成重复使用的工作流；法

律和金融连接器、行业插件及预设流程的完善，有望降低推广成本，扩大部署岗位和付费客户范围。

图 13：分行业每员工月度 AI 支出中位数：知识密集行业投入领先

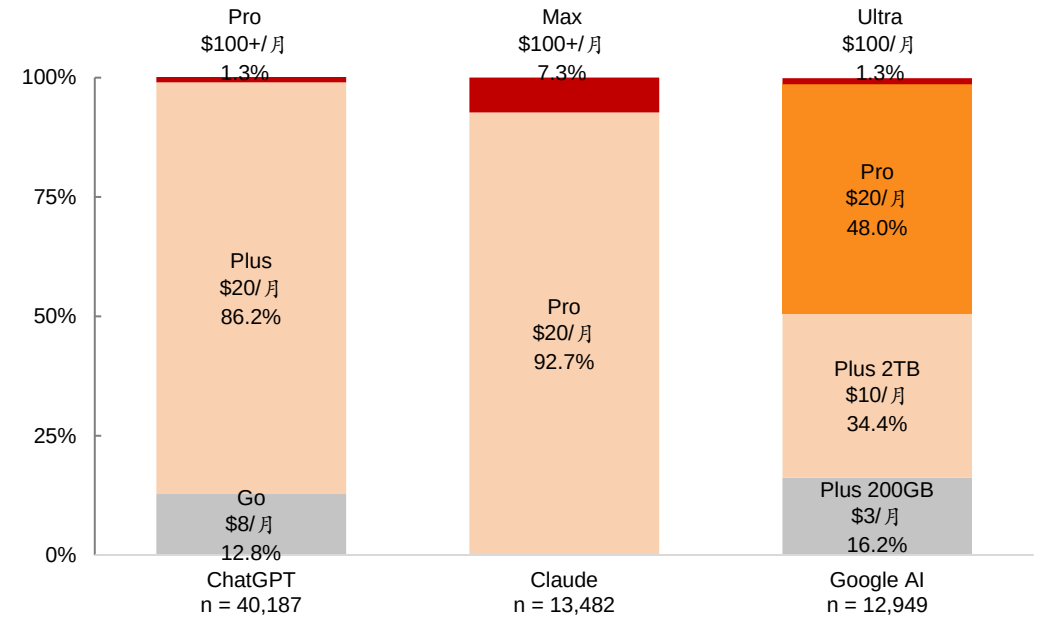


资料来源：Ramp AI Index，招银国际环球市场；数据截至 2026 年 8 月

注：基于 Ramp 美国企业样本，按行业列示每员工月度 AI 支出中位数，涵盖模型订阅、API 及其他 AI 产品和服务

Claude 高价订阅的占比高于 ChatGPT。YipitData 美国活跃订阅样本显示，2026 年 8 月 ChatGPT 及 Claude 订阅样本数分别为 40,187 及 13,482 份，月费 100 美元及以上档位的占比分别为 1.3%及 7.3%。我们推测，编程及专业工作场景的高频使用，是 Claude 用户向 Max 等高价档位转化的重要基础。

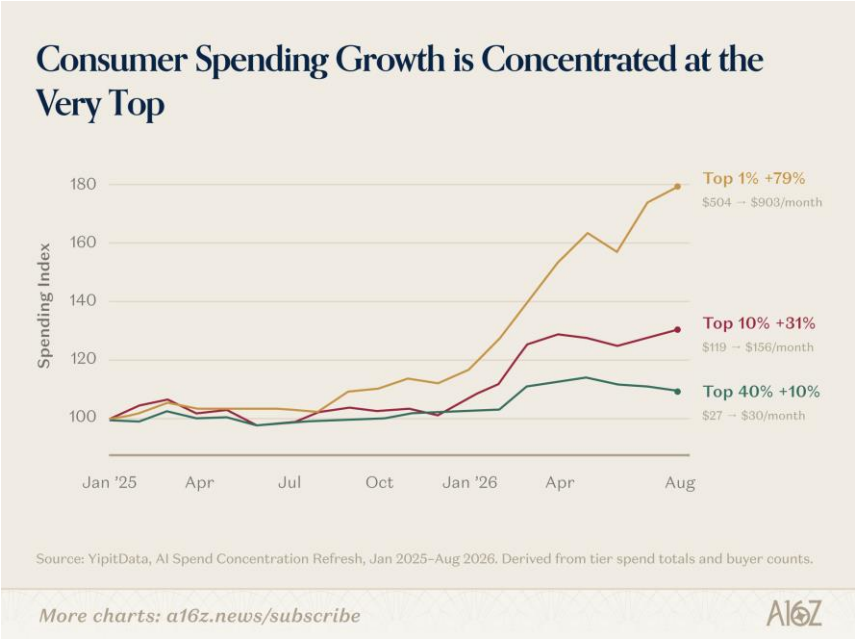
图 14：Claude 月费 100 美元及以上订阅占比高于 ChatGPT



资料来源：a16z、YipitData，招银国际环球市场

企业采购之外，个人端 AI 付费也在扩大，专业用户的生产性需求正在支撑付费增长。根据 a16z 引用的 YipitData 数据，2026 年 8 月，美国消费者样本中至少付费订阅 ChatGPT、Gemini 或 Claude 之一的比例由一年前的 2.1%升至 4.5%，显示个人付费的覆盖范围持续扩大。与此同时，高付费人群的支出增长更快：2025 年 1 月至 2026 年 8 月，按 AI 支出排名的 Top 1%付费用户平均月支出由 504 美元升至 903 美元，增长 79%；Top 10%及 Top 40%群体的平均月支出则分别由 119 美元、27 美元升至 156 美元、30 美元，增幅为 31%及 10%。这意味着，个人端 AI 商业化既受益于更多用户开始付费，也受益于高价值用户持续增加投入。其中，以个人身份购买工具、用于工作和经营的专业用户（prosumer）连接了个人订阅与生产性需求，其付费能力和使用强度为 AI 产品提供了基础订阅之外的收入空间。

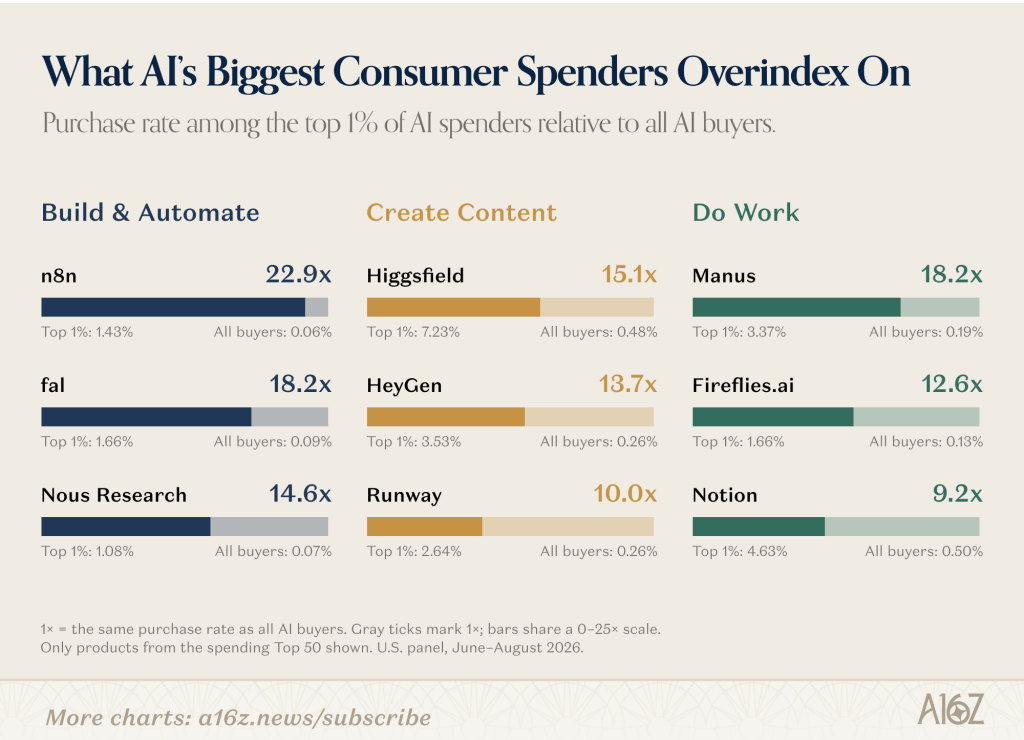
图 15：个人端 AI 支出增长集中于头部付费用户



资料来源：a16z、YipitData，招银国际环球市场

头部用户的购买偏好集中于能够形成工作成果的生产工具，相关使用方式的普及有望带动更广泛人群加购和升级。根据 a16z，2026 年 6-8 月，Top 1%付费用户购买自动化工具 n8n、任务执行工具 Manus 及内容创作工具 Higgsfield 的比例，分别为全部 AI 付费用户的 22.9 倍、18.2 倍及 15.1 倍，体现其对流程搭建、任务交付和内容生产的需求。这类应用服务于流程搭建、项目交付及内容生产，付费价值能够与节省的工时和增加的产出挂钩，是支撑个人用户持续投入的重要基础。随着电脑操作及 Agent 能力提升，AI 可承担的任务将进一步向非编程办公和专业场景扩展，让更多用户获得实际生产价值，带动 Top 10%乃至 Top 40%付费群体增加使用、加购工具及升级套餐，为个人端 AI 收入增长提供空间。

图 16：头部 1%付费用户更倾向购买自动化、内容创作及生产力工具



资料来源：a16z，招银国际环球市场

中国模型厂商的商业化形式进一步延伸至海外云平台收入分成。我们观察到，目前基础模型厂商主要采用三类变现方式：1）按席位或档位收取订阅费，以使用额度、模型能力及响应速度区分产品；2）通过云平台托管模型，按调用量分享平台收入；3）将模型接入企业系统，结合部署、使用量及项目交付收费。高强度用户推动订阅升级，持续运行的 Agent 增加 API 调用，云平台则带来更广的企业客户覆盖。Kimi K3 于 9 月接入 AWS Bedrock 并落地按调用量分成，GLM-5.3 于 10 月初跟进，双方也分别与阿里云达成类似合作。模型厂商提供基础模型和持续迭代，云平台提供推理资源、企业接入及结算服务。我们认为，这一模式结合了模型厂商的技术能力与云平台的客户渠道，在保持开放权重分发的同时，拓宽模型厂商的收入来源。

Token 降价推动更大规模使用。Ramp 于 2026 年 9 月 9 日披露，Token Spend Management 样本的每百万 Token 实际价格由 3 月峰值 1.15 美元降至报告发布当周的 0.68 美元，下降约 41%；前沿模型 Token 份额由 8 月峰值 53%回落至 45%，标准模型使用增加。其 6 月报告显示，2025 年 1 月至 2026 年 4 月，接入 AI 账户的企业样本 Token 用量增长 1,001%，支出增长 497%。我们认为模型价格下降及不同档位组合扩大了客户可负担的调用规模，长程 Agent 进一步提高单用户的 Token 消耗，为推理需求增长提供支撑。

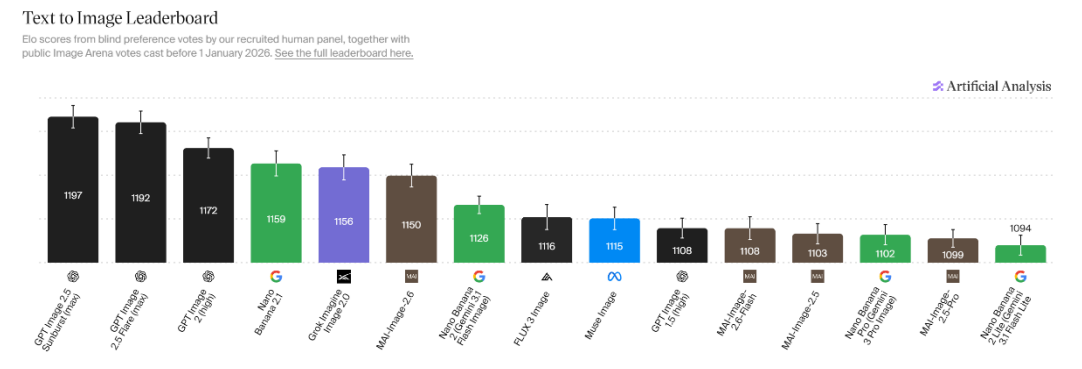
企业 AI 渗透率提升需要把熟练用户的使用方法转化为可复制的工作流。Harvey 于 2026 年 8 月 5 日披露，HEUKING 已将平台从 2024 年约 50 人的试点扩展至律所全部专业岗位，许可数量较最初部署接近六倍；月活用户中 85%每周使用，新增用户一个月后仍活跃的比例因群体不同为 89%至 100%。应用覆盖文档分析、合同起草、邮件草拟及重复任务。将成熟用法嵌入 Word、Outlook 等日常软件，并以培训和预设流程支持更多岗位使用，有助于把局部效率改善转化为持续使用和客户扩容。

多模态：海外厂商领跑图像，中国厂商保持视频优势

我们观察到图像模型的功能演进继续转向创意生产工作流，重点包括：1）多轮编辑和可控性继续增强，例如 9 月发布的 OpenAI ChatGPT Images 2.5 强化参考图主体保留及连续修改一致性，并支持草图参考和图上批注，缩短创意生成与反馈修改流程；2）图像模型进一步覆盖专业设计需求，例如 7 月发布的字节跳动 Seedream 5.0 Pro 强化密集文字、数据和复杂信息图的布局生成，支持圈选、图层分离和多图融合，提升局部修改和素材组合的可控性。10 月 6 日，Google 发布 Nano Banana 2.1，强化多轮人物一致性、文字呈现及宽画幅生成；10 月 9 日阿里云百炼新增 Qwen-Image-2.1-Turbo，支持最多 10 张参考图及原生透明图像生成，进一步拓展商业设计场景。

图像模型方面海外厂商仍掌握第一梯队主导权，中国厂商以阿里巴巴和字节跳动领先。文生图（text-to-image）领域，基于 Artificial Analysis，OpenAI GPT Image 2.5 Sunburst/Flare (max)的 Elo 评分分别为 1,197/1,191，位居榜单前二；10 月发布的 Google Nano Banana 2.1 以 1,160 分位居第四，Black Forest Labs FLUX 3 Image 以 1,116 分位居第八。中国厂商中，阿里巴巴 Qwen-Image-3.0-Pro 和字节跳动 Seedream 5.0 Pro 分别为 1,087 及 1,082 分，排名第 16 及 17。此外，从图像模型的性价比来看（质量 vs. API 价格），微软 MAI-Image-2.6 及 Meta Muse Image 的 Elo 分别为 1,151 及 1,115 分，每千张 API 价格为 38.9 及 10 美元，在相对高质量、低价格区间具备优势。

图 17：主流图像模型文生图能力排名



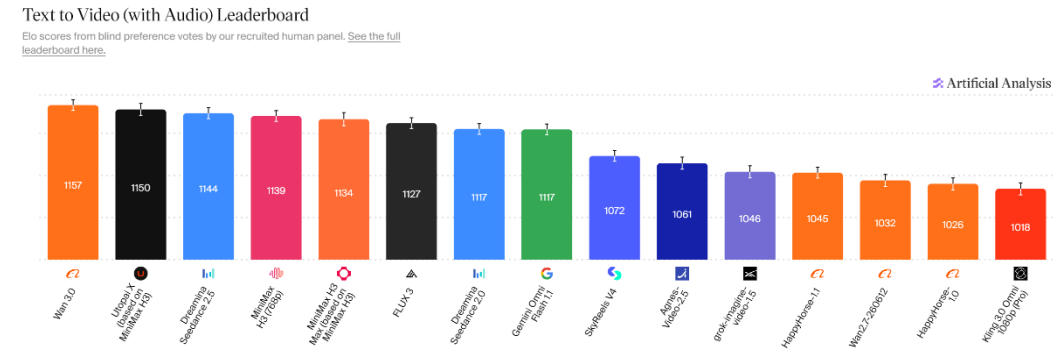
资料来源：Artificial Analysis，招银国际环球市场

注：列示 AA-Image-T2I v2.0 可用模型前 17 名，数据截至 2026 年 10 月 13 日。Elo 越高越好，排序按未取整分值，图中显示取整分值，下同。

视频模型的功能演进方面，我们观察到三个主要趋势：1）长时长和跨镜头一致性增强，例如 7 月发布的 Seedance 2.5 将单次音视频生成由 15 秒延长至 30 秒，多轮续写可延展至数分钟，并强化人物、场景及音画连续性；2）原生音视频一体化推进，例如 7 月发布的 MiniMax H3 统一理解文本、图片、视频和音频参考素材，支持带原生立体声的视频生成；3）生成后的可控编辑及高清输出能力提升，例如 Seedance 2.5 支持时间戳级音视频修改，MiniMax H3 支持 2K 输出。

中国视频模型在文生视频领域维持领先。文生视频（text-to-video with audio）领域，基于 Artificial Analysis，阿里巴巴万相 Wan 3.0 以 Elo 1,157 分位居榜首，字节跳动即梦 Dreamina Seedance 2.5 及 MiniMax H3（768p）分别为 1,144 及 1,139 分，排名第三及第四；基于 MiniMax H3 开发的 Utopai X 及 fal MiniMax H3 Max 分别位居第二及第五，Google Gemini Omni Flash 1.1 则以 1,117 分位居第八。从性价比来看（质量 vs. API 价格），MiniMax H3 每分钟 API 价格为 4.80 美元，分别相当于 Wan 3.0 及 Seedance 2.5 的 40% 及 14%，处于高质量、低价格区间。

图 18：主流视频模型文生视频能力排名（含音频）

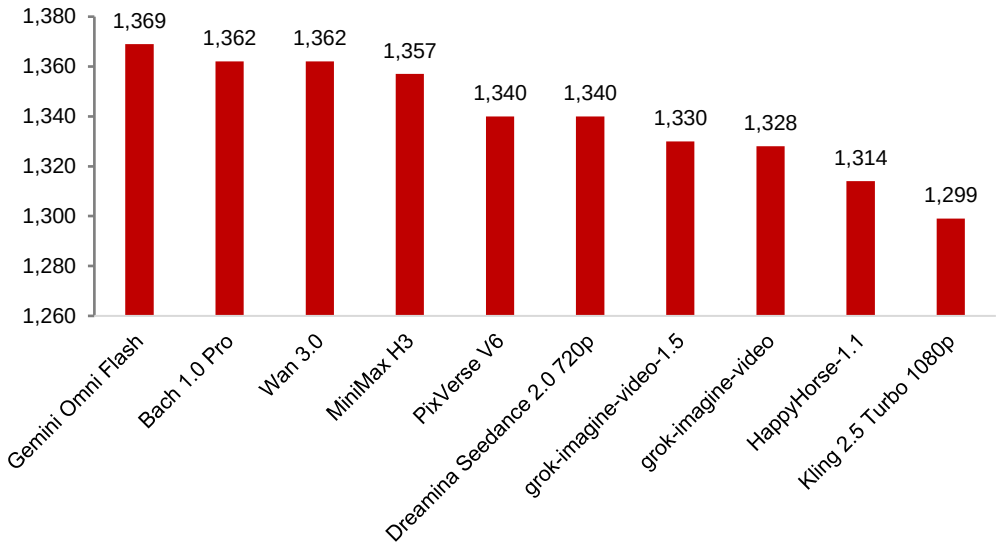


资料来源：Artificial Analysis，招银国际环球市场

注：列示 AA-Video-T2V v2.0 可用模型前 8 名，包含音频。* Utopai X 及 fal MiniMax H3 Max 均基于 MiniMax H3，颜色按发布方归属。正文 API 价格为 AA 列示的每分钟生成价格。

图生视频（image-to-video no audio）方面，Google Gemini Omni Flash 以 Elo 1,369 分位居行业第一，阿里巴巴万相 Wan 3.0、MiniMax H3 及爱诗科技 PixVerse V6 分别为 1,362、1,357 及 1,340 分，多个中国厂商处于行业前列。MiniMax H3 同时位居开源权重模型第一。

图 19：主流视频模型图生视频能力排名（无音频）



资料来源：Artificial Analysis，招银国际环球市场

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：商业化持续加速，Agent 拓展付费场景

3Q26 海外 AI 平台商业化继续加速，主要变化包括：1）OpenAI 企业收入增长提速，Anthropic 继续扩大企业及专业用户覆盖；2）Work、Cowork 及持续运行的 Agent 由编程延伸至办公和专业工作，扩大高价值付费场景；3）模型分层与缓存降价同步推进，在提升能力的同时降低单任务费用；4）融资及上市筹备提速，为下一阶段训练和算力投入提供支持。

OpenAI 收入运行率加速提升，Anthropic 企业业务与 xAI 算力服务持续放量。 据 CNBC 2026 年 10 月 8 日报道，OpenAI 年化收入运行率已接近 500 亿美元，三季度整体及企业业务年化收入运行率分别增长 77%及 107%；据 Reuters 8 月 17 日报道，Anthropic 年化收入于 7 月底超过 650 亿美元。季度收入方面，据 Bloomberg 8 月 14 日报道，Anthropic 2Q26 初步收入超过 115 亿美元；据《华尔街日报》8 月 18 日报道，OpenAI 同期收入为 67 亿美元。xAI 侧，据 SpaceX 8 月 4 日业绩公告，AI 分部 2Q26 收入 25.61 亿美元，环比增长 213%，新云服务协议贡献主要增量。

模型能力提升与单任务费用下降并行，推动企业扩大部署。Anthropic 于 2026 年 9 月披露，Opus 5.5 默认设置下典型任务费用较上一代下降 40%，Sonnet 5.5 最多下降 30%；OpenAI GPT-6.1 Sol 在公司 DeepSWE 1.1 评测中的任务费用较 GPT-6 Sol 下降 76%。经营端，Reuters 9 月 13 日援引《金融时报》报道称，Anthropic 预计 3Q26 连续第二季实现正向调整后经营利润；SpaceX AI 分部 2Q26 调整后 EBITDA 转正至 11.46 亿美元，同期经营亏损为 12.57 亿美元。模型、推理系统及业务规模共同影响利润改善的速度。

资本化进程加速，Anthropic 与 OpenAI 上市计划分别指向 2026 及 2027 年。据 Bloomberg 10 月 2 日报道，Anthropic 拟最快于 11 月 9 日当周启动 IPO 推介、最快 11 月中旬上市。OpenAI 已于 6 月保密递交 S-1，CFO 于 8 月表示预计 2027 年上市；公司同时据报洽谈至少 300 亿美元新融资，目标投前估值约 1.4 万亿美元。企业收入和用户规模加速增长，为头部厂商扩大融资及进入公开市场提供了业务基础。

展望而言，模型能力上界决定头部厂商的差异化定价能力，Agent 产品扩大付费场景，全栈推理优化则提高规模化服务的经济性。OpenAI 的用户及企业收入同步扩张，Anthropic 在高价订阅和企业场景中形成较强变现，Google 依托模型、开发工具和云渠道扩大分发，xAI 则通过 Cursor 与算力服务拓展收入。海外平台的竞争正进一步延伸至模型、应用及基础设施的综合能力。

图 20：海外核心 AI 基础模型平台：财务及运营数据对比

海外主要 AI 基础模型公司：关键指标对比				
维度	Anthropic（Claude）	OpenAI（ChatGPT）	Google（Gemini）	xAI（Grok）
规模与增长				
最新 ARR 及收入	2026 年 7 月底~**650 亿美元**（Reuters，2026.8.17）；2Q26 初步收入**>115 亿美元**（Bloomberg，2026.8.14）	2026 年 9 月年化收入运行率接近 500 亿美元，较三季度初增 77%，预计年底达 700 亿，三季度企业业务年化收入运行率增长 107%（CNBC，2026.10.8）	未单独披露	AI 分部 2Q26 收入 **25.61 亿美元**（SpaceX 业绩报，2026.8.4）
增长轨迹	2025 年底约 90 亿→2026 年 5 月约 470 亿→2026 年 7 月底超过 650 亿美元（Reuters，2026.8.17）	ChatGPT 周活由 2026 年 2 月超 9 亿、8 月 6 日 10 亿升至 10 月 5 日 12 亿；最近两个月增加约 2 亿、增长 20%（OpenAI 官网）	2Q26 Gemini App 月活 **9.5 亿**，日活为一年前 3 倍；模型 API 处理量由上季 160 亿升至 **220 亿** token/分钟（Alphabet 业绩会，2026.7.22）	AI 分部 2Q26 收入同比+247%、环比+213%，主要由云服务协议驱动（SpaceX 公告，2026.8.4）
收入目标	知情人士称 2026 年 ARR 将达 1000 亿美元（彭博，2026.9.19）；内部预测 2028 年全年收入 1,900 亿—2,000 亿美元（Reuters，2026.8.14）	公司预计 2026 年营收为 360 亿美元；2030 年长期目标 **3,500 亿**美元（金融时报，2026.9）	未单独披露	2026 年目标 20 亿美元已达成；目标 2027 年盈利（Sacra）
收入结构与核心产品				
引擎① 编程/开发	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 周活**>500 万**；公司另披露**>100 万**人用于非软件开发工作（公司公告，2026.7.9）	Antigravity 周活超 240 万（7.22）；3.8 Flash 强化长程编程（9.2）；Mechanize 联合创始人及十余名员工加入 DeepMind，补充 Agent 训练及评测能力（Bloomberg，9.11）	Grok 4.6 于 8 月 12 日发布；SpaceX 于 8 月完成收购 Cursor，交易价格 600 亿美元（SpaceXAI 公告，2026.8.12、Cursor 公告，2026.8.14）
引擎② 通用/办公	Cowork 拓展云端执行、网页/移动端，8 月加入桌面端内置浏览器；7 月披露>90%使用属于非软件开发工作（公司公告，2026.7.7、2026.8.26）	ChatGPT Work 于 7 月上线；GPT-6 Astra 于 9 月强化 computer use 及专业工作（公司公告，2026.7.9、2026.9.8）	Gemini 在 Gmail、Drive、Docs、Slides 及 Chat 中执行跨应用任务，可生成文档、表格与演示文稿；9 月新增 5 项 Agent 能力（Google Workspace，2026.9.9）	Grok Bot 支持持续运行的云电脑、连接器及定时工作流（SpaceXAI 更新记录，2026.8.11）
引擎③ 新曲线	Claude for Small Business 自 5 月推出以来安装>90 万次，覆盖 43 个工作流、9 月新增 27 项集成（公司公告，2026.9.15）	ChatGPT 广告年化收入达 10 亿美元（公司公告，2026.8.31）	拓展网络安全与语音 Agent：3.8 Flash Cyber 支持漏洞检测及自动修复，限受信任防御者使用；3.8 Live 支持实时语音交互（Google 公告及 API 文档）	2Q26 新签云协议合同销售额 141 亿美元，当季新增基础设施收入 16 亿美元（SpaceX 公告，2026.8.4）
盈利与成本				
盈利节奏	2Q26 初步实现正向调整后经营利润，公司预计 3Q26 **连续第二季为正**（Bloomberg，2026.8.14、Reuters，2026.9.13）	2Q26 经营**亏损 123 亿美元**，1Q26 为 93 亿美元（WSJ，2026.8.19）	Gemini 独立盈亏未列示	2Q26 调整后 EBITDA 由 1Q26 的-6.09 亿转为 **+11.46 亿美元**；经营亏损由 24.69 亿收窄至 **12.57 亿美元**（SpaceX2Q26 业绩公告，2026.8.4）
成本	Fable 5.1 按 8 月实际负载及默认推理强度测算，典型费用可降约 25%、高度 Agent 化负载可降约 45%（公司测算，2026.9.1）	GPT-5.6 Luna/Terra 降价 80%/20%，输入/输出价分别为 0.2/1.2 及 2/12 美元每百万 token；生产服务软件优化使端到端模型服务成本下降 20%（公司公告，2026.7.30）	Gemini 3.8 Flash 输入/输出推广价 0.75/3.75 美元/百万 token，至 2026.12.31；2027.1.1 起为 1.50/7.50 美元（Google 公告，2026.9.2）	AI 分部资本开支由 1Q26 的 77.23 亿升至 2Q26 的 158.28 亿美元（SpaceX 公告，2026.8.4）
估值与融资				
最新估值	目标上市估值 **2 万亿美元** (2026.8，WSJ)	据报拟募资至少 300 亿美元，目标投前估值约 1.4 万亿美元，处洽谈阶段（Bloomberg，2026.9.29）	NA	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** (2026.10.14)
估值倍数	按 2026 年底的 1000 亿 ARR 和 2 万亿估值计算：**~20x P/S**	按 1.4 万亿美元投前估值/近 700 亿美元年化收入运行率测算，**~20x P/S**	NA	NA
上市进度	已提交保密 IPO 申请 (2026.6，公司公告)；最早可能 2026 年 11 月上市 (2026.9，WSJ)	CFO 于 8 月称计划 2027 年上市；Altman 于 9 月明确排除 2026 年 IPO（CNBC，8.19；Reuters，9.12）	NA	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、Fortune，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 10 月 14 日

■ **Anthropic：企业商业化持续放量，电脑操作及任务成本优化推动应用拓展**

Anthropic 以编程和企业 Agent 为核心，商业化规模持续扩大。Claude Code 积累的长程任务能力正通过 Cowork 及企业执行环境延伸至非编程工作，Opus、Sonnet 和 Fable 的分层定价覆盖不同复杂度的任务。我们判断，专业用户使用强度提升、企业客户扩容以及推理效率优化，将继续推动公司收入增长及利润改善。

年化收入快速扩张，季度收入同步增长。据 Reuters 2026 年 8 月 17 日报道，Anthropic 年化收入由 2025 年底约 90 亿美元增至 2026 年 7 月底超过 650 亿美元；据 Bloomberg 8 月 14 日援引投资者材料报道，2Q26 初步收入超过 115 亿美元，高于 1Q26 的 47.3 亿美元及 2Q25 的 7.87 亿美元。Bloomberg 9 月 19 日援引《纽约时报》报道，公司预计 2026 年内年化收入超过 1,000 亿美元。

企业业务由大客户向中小企业扩展。公司于 2026 年 4 月 6 日披露，年化付费超过 100 万美元的企业客户已超过 1,000 家，较 2 月的超过 500 家翻倍。9 月 15 日，公司披露 5 月推出的 Claude for Small Business 累计安装超过 90 万次，包含 43 个工作流，并新增 27 项工具集成。该产品运行于 Cowork，将 AI 应用从后台事务延伸至获客、客户询问处理及提案生成，扩大企业产品覆盖的岗位与任务。

Claude Code 由编程工具向企业内部工作流延伸。公司 2026 年 2 月 12 日披露，Claude Code 年化收入已超过 25 亿美元，企业订阅增至年初的四倍，企业使用贡献过半收入。8 月 6 日，公司面向 Team 及 Enterprise 客户开放自托管执行环境公测，使 Agent 可以访问客户自有网络中的内部服务、数据库和工具链，代码及构建产物保留在客户基础设施上，模型推理仍由 Anthropic 提供。执行环境进入内部系统，有助于承接大型代码库维护、安全审查等持续性任务，提高企业使用深度。

Cowork：云端执行与内置浏览器推动电脑操作落地。据公司 2026 年 7 月 7 日公告，Cowork 开始扩展至网页及移动端，支持跨设备接续任务，以及设备离线时继续运行云端和定时任务；当时超过 90%的使用属于非软件开发工作，业务运营和内容创作合计约占一半。据公司 8 月 26 日公告，Cowork 桌面端进一步加入内置浏览器，可直接浏览网页、点击输入、填写表单、提取业务看板数据及操作没有连接器的网站；Pro、Max 及 Team 逐步开放，Enterprise 可由管理员启用。内置浏览器与用户个人浏览器分离，桌面端可结合本地文件与网页执行任务，云端任务和本地浏览器操作由此分别承接不同工作需求。

Opus 5.5：编程及电脑操作能力继续提升，典型任务费用较上一代下降 40%。据 Anthropic 2026 年 9 月 22 日公告，公司推出 Claude Opus 5.5。在公司披露的评测中，Terminal-Bench 4.0 成绩由 Opus 5 的 52.3%升至 66.4%，高于 Fable 5.1 的 55.8%；OSWorld 2.0 partial 口径得分由 74.0%升至 81.8%，高于 Fable 5.1 的 80.7%；跨应用业务工作流评测 AutomationBench 成绩由 26.9%升至 40.0%。模型能力提升同步体现于长程编程执行：在公司将 HAProxy 由 C 语言改写为 Rust 的测试中，Opus 5.5 与 Fable 5.1 均通过几乎全部原有回归测试，前者耗时由 12 小时缩短至 9.5 小时，费用低 51%。Opus 5.5 标准输入/输出价格为 4/20 美元每百万 token，较 Opus 5 下降 20%；缓存读取价格由 0.50 美元降至 0.20 美元，降幅 60%。叠加每任务 token 消耗减少，公司测算默认设置下典型工作负载费用下降 40%，输出生成速度提升超过 30%。Fast 模式以两倍标准价格提供最高约 2.5 倍速度，进一步满足企业对复杂任务质量、执行时效及成本的分层需求。

Sonnet 5.5：中端模型强化编程及电脑操作，任务费用最多下降 30%。据 Anthropic 2026 年 9 月 28 日公告，Sonnet 5.5 在公司 Terminal-Bench 4.0 评测中的成绩由 Sonnet 5 的 10.3%升至 70.6%，CursorBench 4 由 34.1%升至 55.5%；OSWorld 2.1 partial 口径得分由

57.0%升至 80.1%。模型支持 100 万 Token 上下文及最高 128K 输出。10 月 7 日，公司进一步将 Sonnet 5.5 缓存读取价格由 0.20 美元降至 0.10 美元每百万 Token，降幅 50%；官方测算，多数 Agent 任务成本因此再下降约 20%。

Haiku 5.5 降低高频任务成本，补齐轻量 Agent 模型档位。公司于 10 月 7 日发布 Haiku 5.5，支持 100 万 Token 上下文，适用于摘要、资料提取、分类及编程子 Agent 任务。公司评测中，其 OSWorld 2.1 离线子集及 Terminal-Bench 4.0 成绩分别为 72.4%及 39.2%。输入不超过 100K 时，输入/输出价格为 0.10/0.50 美元每百万 Token，超过 100K 时为 0.50/2.50 美元；公司测算平均运行成本较 Haiku 4.5 下降约 75%，支持更大规模的高频调用。

Fable 5.1：长程任务能力继续提升，缓存降价改善 Agent 经济性。据公司披露，Fable 5.1 在同条件重测的 OSWorld 2.0 严格任务成功率为 41.7%，高于 Fable 5 的 36.1%；Terminal-Bench 4.0 成绩由 42.0%升至 55.8%，AutomationBench 由 17.1%升至 31.4%。输入/输出定价维持 10/50 美元每百万 token，缓存读取价下降 75%至 0.25 美元。公司按 8 月四周实际负载测算，默认推理强度下典型费用可下降约 25%，高度 Agent 化负载可下降约 45%。同日发布的 Mythos 5.1 则面向经过审核的网络安全及生命科学用户，延伸专业研究和防御场景。

算力供给：多云与多芯片布局支撑企业需求扩张。据公司 2026 年 5 月 6 日公告，与 SpaceX 的合作为公司提供 Colossus 1 全部算力，对应超过 300MW 容量及 22 万张英伟达 GPU；公司当时同步将 Claude Code 的五小时用量上限翻倍，并取消 Pro 及 Max 的高峰时段限额折减。中期供应方面，据公司 4 月 20 日公告，与 Amazon 的新协议覆盖最高 5GW 容量，其中接近 1GW 预计于 2026 年底前上线，公司承诺未来十年向 AWS 投入超过 1,000 亿美元；据 5 月 6 日公告，Google/Broadcom 合作另覆盖 5GW、预计从 2027 年开始上线。AWS Trainium、Google TPU 及英伟达 GPU 并行使用，有望提高算力供给弹性。

调整后经营盈利初步转正，云分成是分析毛利的重要成本项。据 Bloomberg 8 月 14 日报道，Anthropic 提供给潜在投资者的初步材料显示 2Q26 已实现正向调整后经营利润；Reuters 9 月 13 日援引《金融时报》称，公司预计 3Q26 连续第二季为正。报道提及的超过 80%毛利率未计入云渠道收入分成及模型训练成本。云渠道扩大企业触达，同时分享部分调用收入，因而自有渠道占比、托管分成和推理效率共同影响厂商的利润率。

融资与 IPO：拟最快 11 月中旬上市，目标估值可能超过 2 万亿美元。据 Reuters 2026 年 9 月 28 日报道，Anthropic IPO 目标估值可能超过 2 万亿美元；据 Bloomberg 10 月 2 日香港时间报道，公司计划最快于 11 月 9 日当周启动推介、最快 11 月中旬上市，并争取在感恩节前完成。结合 Bloomberg 9 月 19 日香港时间报道的 2026 年内超过 1,000 亿美元年化收入预期，按 2 万亿美元目标估值测算，估值/目标年化收入约 20x。

■ **OpenAI：收入运行率加速提升，Work 及分层模型拓展付费场景**

OpenAI 用户与企业业务同步加速，商业化进一步向办公和持续运行的 Agent 扩展。ChatGPT 周活已达 12 亿，企业销售额自 7 月以来翻倍，Work 将编程积累的执行能力延伸至通用办公，Dots 增加持续运行的个人 Agent 场景。预计分层模型和订阅扩大付费覆盖，广告则进一步挖掘免费用户价值，消费与企业业务共同支撑公司下一阶段增长。

年化收入运行率接近 500 亿美元，企业业务保持快速增长。据 CNBC 2026 年 10 月 8 日报道，截至 9 月底，OpenAI 年化收入运行率接近 500 亿美元，三季度整体及企业业务年化收入运行率分别增长 77%及 107%，企业业务增速高于整体 30 个百分点。据彭博 10 月 9 日报道，OpenAI 预计将在年底达到 700 亿元 ARR。结合此前披露，年化收入爬升轨迹为超

过 400 亿美元（2026.8，Bloomberg 8 月 13 日）→接近 500 亿美元（2026.10，CNBC 10 月 8 日）→2026 年底目标 700 亿美元（2026.10，彭博 10 月 9 日）。据 CNBC 8 月 14 日报道，CFO Sarah Friar 表示，企业收入已超过消费业务，提前实现原预计年底达到的两类业务收入持平目标。据《华尔街日报》8 月 18 日报道，公司 2Q26 收入为 67 亿美元，环比增长 18%；经营亏损由 1Q26 的 93 亿美元扩大至 123 亿美元。

ChatGPT 用户增长再加速，最近两个月周活增加约 2 亿。公司披露的周活用户由 2026 年 2 月超过 9 亿、8 月 6 日 10 亿，升至 10 月 5 日 12 亿；8 月至 10 月增长 20%，最近两个月新增量已超过此前约五个月的增量。同期，免费用户获得 Luna 不限量文字对话，Work 将编程 Agent 扩展至文档、表格和演示文稿等办公任务。用户规模扩大，为付费订阅、广告及企业客户拓展提供更大的基础。

ChatGPT Work 及 Dots：电脑操作由项目执行向持续运行的个人 Agent 延伸。 据公司 2026 年 7 月 9 日公告，Work 覆盖网页、移动端及桌面端，支持浏览器操作、文件处理、电脑操作及定时任务，并结合 Slack、Teams 等应用推进后续工作；Sites 提供网站及应用构建和托管。公司于 9 月 29 日 DevDay 推出 Dots，基于 GPT-6 Astra 为用户配置可持续运行的 Agent 及云电脑，支持连接超过 4,000 款应用，在用户离线时继续推进任务，逐步向符合条件的 Pro 及 Business Premium 用户开放，Enterprise 由管理员启用测试。公司同步新增每月 500 美元的 Pro 档位，为高频使用提供更高额度，持续运行及复杂工作成为新增付费场景。我们认为，Dots 推出的意义在于有望把 AI 从“随叫随到的工具”推向“长期托付的同事”，行业的竞争有望逐步从“谁的模型更强”转向“谁能拥有用户的长期工作流”，未来中间层的 SaaS 和自动化工具的价值存在被挤压的风险，而治理和合规会有望获得更多的重视。

GPT-6 Astra 强化电脑操作，Ultrafast 继续缩短任务等待时间。 据公司 2026 年 9 月 8 日披露，GPT-6 Astra 在 OSWorld 2.0 partial 口径下得分 72.6%，高于 GPT-5.6 Sol 的 65.7%；公司延迟模拟显示，每任务耗时由约 75 分钟降至约 40 分钟。在 Mind2Web 测试中，Astra 与执行框架配合的任务完成速度约为 Sol 的 1.9 倍。Astra 标准输入/输出价格为 10/50 美元每百万 Token，Fast 模式以两倍价格提供最高约两倍速度。公司于 9 月 29 日 DevDay 进一步推出 Ultrafast，在 Codex 及 API 中的 Token 生成速度最高分别达到常规模式的 8 倍及 6 倍，缩短高复杂度任务的执行等待时间。

GPT-6.1 Sol：编程及电脑操作能力接近旗舰，单位任务成本进一步下降。 据 OpenAI 2026 年 9 月 29 日公告，GPT-6.1 Sol 在公司 DeepSWE 1.1 评测中 high 档得分 75.2%，高于 GPT-6 Sol max 档的 68.8%，任务费用下降 76%。在 OSWorld 2.0 离线集 partial 口径下，max 档得分 71.4%，接近同档 GPT-6 Astra 的 73.5%，每任务费用约为后者七分之一。输入不超过 272K 时，标准输入/输出价格维持 2/10 美元每百万 Token，缓存读取价格由 0.20 美元降至 0.10 美元；输入超过 272K 时，全请求输入及缓存价格翻倍，输出价格提高 50%。模型支持 1.05M 上下文及最高 128K 输出。低成本模型 GPT-6 Luna 输入/输出价格为 0.10/0.50 美元。10 月 8 日，GPT-6.1 Sol 进一步推出 Ultrafast，覆盖 Codex、ChatGPT Work 及 API，Codex 中 Token 生成速度最高为标准模式 8 倍；输入不超过 272K 时，输入/输出价格为 12/60 美元每百万 Token。高能力、均衡及低成本模型结合推理强度与速度分档，使企业可按任务难度、时效及价值配置投入。

软硬件协同优化推理成本，自研芯片计划年底前部署。 据 OpenAI 2026 年 9 月 8 日披露，GPT-5.6 Sol 辅助优化生产服务软件，使端到端模型服务成本下降 20%、Token 生成效率提升超过 15%。硬件方面，据公司 8 月 25 日披露的 Jalapeño 首轮测试，自研推理芯片在 GPT-OSS 120B、DeepSeek-R1 及 Kimi K2.5 等模型上的单位额定功耗峰值吞吐为对比系

统的 1.5-1.9 倍，延迟降至对比系统的约 28%-59%。公司计划于年底前开始部署该芯片，通过模型、推理软件与专用硬件协同，提高单位算力可承接的业务量。

广告年化收入达到 10 亿美元，效果广告与品牌 Agent 拓展变现。 据 OpenAI 2026 年 8 月 31 日披露，ChatGPT 广告上线不足 200 天，年化收入已达 10 亿美元，广告主数量达数万家，覆盖超过 40 个国家；按点击及结果优化的投放已占多数，Pixel 及 Conversions API 等工具帮助广告主追踪转化。据公司 9 月 16 日公告，公司开始测试 Sponsored Agents，允许用户点击广告后与品牌赞助的 Agent 对话，并通过 ChatGPT Work 生成广告、在 Ads Manager 加入创意工具，同时接入首个 CRM 合作伙伴 HubSpot 及首个电商合作伙伴 Shopify，进一步连接获客、互动及转化场景。

IPO 计划指向 2027 年，用户与企业收入增长支撑上市预期。CNBC 于 2026 年 8 月 19 日报道，CFO Sarah Friar 向员工表示 OpenAI 预计于 2027 年上市，Sam Altman 于 9 月进一步明确 2026 年不会上市。据 Bloomberg 9 月 30 日报道，公司正洽谈至少 300 亿美元新融资，目标投前估值约 1.4 万亿美元，对应年底 700 亿美元年化收入运行率约 20x。我们认为，周活用户重新加速、企业收入占比提前超过消费业务，体现公司增长动力由消费端向消费与企业双轮驱动扩展，也为管理层对 2027 年上市及业务扩张的预期提供支持。

■ **Google：Flash 加快迭代，Argon 拓展长程编程及专业任务**

Google 三季度围绕编程和 Agent 加快模型迭代，形成效率与前沿能力两条产品路线。7 月 21 日至 9 月 2 日，Flash 系列在六周内连续更新三次，9 月底再推出 Gemini 4 Argon。Flash 以较低调用价格承接日常及高频任务，Argon 提升复杂编程和专业研究能力，模型通过 Gemini、Antigravity 及 API 覆盖个人、开发者和企业客户。

Flash 系列持续提升能力，推广期 Token 单价较 3.6 减半。Google 于 7 月 21 日推出 Gemini 3.6 Flash 及 3.5 Flash-Lite，8 月 13 日升级至 3.7 Flash，9 月 2 日再推出 3.8 Flash。公司评测中，3.7 较 3.6 的 DeepSWE v1.1 成绩由 49.0%升至 65.3%，AutomationBench 由 17.0%升至 30.4%；3.8 进一步强化长程编程及多步骤推理。3.7 及 3.8 年内推广输入/输出价格为 0.75/3.75 美元每百万 Token，较 3.6 原价下降 50%，以较低单价支持企业部署 Agent。

Gemini 4 Argon 提升长程任务能力，最高输出扩大至 100 万 Token。公司 9 月 30 日披露，Argon 的 DeepSWE v1.1 及 AutomationBench 成绩分别为 77.9%和 51.3%，输出上限由此前 64K 提升至 1M Token，支持生成更长的代码及研究成果。Google 内部已将其用于大规模代码迁移和数据中心内存优化，后者已释放超过 300 TiB 内存。模型先经 Fairwind 计划向获准网络安全客户开放，推广输入/输出价格为 2/10 美元每百万 Token。

Mechanize 团队加入，补充 Agent 训练和评测能力。据 Business Insider 9 月 11 日报道，Google 完成人才交易，Mechanize 联合创始人 Tamay Besiroglu 及十余名员工加入 DeepMind，主要参与模型中期训练。Mechanize 为编程 Agent 构建任务环境和评分系统，覆盖功能开发、应用部署及陌生代码库排错，并将任务结果用于强化学习。

■ **xAI：Cursor 收购落地，算力服务成为收入增长主力**

xAI 并入 SpaceX 后，模型应用与算力服务共同扩大商业化。Cursor 补充编程产品及开发者分发，Grok 与持续运行 Agent 拓展应用场景，对外云服务则将算力资源转化为合同收入。2Q26 AI 分部收入由云服务带动快速增长，调整后 EBITDA 转正；公司同时大幅增加资本开支，继续扩张后续业务所需的算力供给。

Cursor 收购完成，模型研发与编程产品分发进一步结合。 据公司披露，收购 Cursor 的交易金额为 600 亿美元；Cursor 于 8 月 14 日确认已被 SpaceX 收购，并称双方已在 Grok 4.6 研发中合作，后续将结合算力、模型及编程产品。

Grok 4.7 强化长程编程及 Agent 能力，维持较低 API 定价。 据 xAI 2026 年 9 月 21 日公告，Grok 4.7 采用更大的新基座，并通过持续数小时的强化学习任务加强长程执行，原生用于 Grok Bot。在公司评测中，CursorBench 4 成绩由 Grok 4.6 的 40.4%升至 46.3%，Terminal-Bench 4 由 20.3%升至 37.6%。模型支持 500K 上下文，通过 Cursor、Grok Build 及 API 提供服务；据公司 API 文档，低于 200K 输入长度时标准输入/输出价格为 2/6 美元每百万 Token，更长输入按两倍价格收费。模型能力与 Cursor 分发结合，继续拓展复杂工程及持续工作场景。

Grok Bot 拓展持续执行，多模态产品同步迭代。Grok Bot 上线持续运行的云电脑，结合消息、审批、连接器及定时流程承接多步骤工作，将产品由即时交互向持续执行延伸。语音及视频方面，公司于 7 月 29 日推出 Grok Voice Think Fast 2.0；7 月 31 日升级的 Grok Imagine Video 1.5 支持文生视频、图生视频及参考视频生成，其中文生及图生视频支持原生 1080p。

云服务带动收入放量，算力容量与合同规模同步扩张。 据 SpaceX 2026 年 8 月 4 日业绩公告，包含 X 及算力服务的 AI 分部 2Q26 收入 25.61 亿美元，同比/环比增长 247%/213%；其中 AI 解决方案及基础设施收入 21.94 亿美元，高于 1Q26 的 4.75 亿美元，X 广告收入 3.67 亿美元，较 1Q26 的 3.43 亿美元小幅增长。新云服务协议对应合同销售额 141 亿美元，并为 2Q26 贡献 16 亿美元增量基础设施收入；同期 AI 算力额定容量由 1Q26 的 1.0GW 提升至 1.4GW，算力扩张与对外销售共同拉动收入增长。

调整后 EBITDA 转正，资本开支继续扩张。SpaceX AI 分部 2Q26 调整后 EBITDA 由 1Q26 亏损 6.09 亿美元转为盈利 11.46 亿美元，经营亏损由 24.69 亿美元收窄至 12.57 亿美元。同期资本开支由 77.23 亿美元增至 158.28 亿美元，1H26 累计 235.51 亿美元，单季投入约为同期收入的 6.2 倍。云服务已带动经营改善，扩大的资本投入将支持后续模型训练和对外算力服务。

国内 AI 基础模型平台：长程能力与推理效率同步改善，API 商业化加速

国内 AI 基础模型平台的商业化重点进一步转向 API 及企业 Agent。3Q26 主要变化包括：1）GLM、Kimi 及 Step 5 强化长程编程、电脑操作和专业任务；2）架构优化、缓存压缩及国产算力适配降低推理成本；3）开放平台成为主要收入来源，智谱与 MiniMax 1H26 相关收入占比分别为 86.5%及 63.4%；4）智谱及 Kimi 与云平台按调用量分成，拓宽开放权重模型的变现渠道。模型能力提升与效率优化将继续拉动 Token 调用，云合作使模型厂商能够分享更多第三方平台的使用收入。

图 21：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)	稀宇科技 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	1H26 收入 9.54 亿元人民币，同比+399.7%；9 月全业务 ARR 达 18 亿美元，目标 2026 年底达 30 亿美元（公司披露，2026.9.16）	1H26 收入 1.17 亿美元，同比+283.1%；8 月 ARR 超 8 亿美元（公司披露，2026.8.26）	ARR 由 2026 年 6 月 3 亿美元增至当前约 10 亿美元，年底目标 20 亿美元（Bloomberg，2026.10.6）	公司 2025 年收入近 5 亿元，2026 年预计收入约 12 亿元（财经，2026.2.27）	9 月年化收入运行率据报达 10 亿美元（Reuters 转述 The Information，2026.9.24）；1-7 月累计收入约 4.75 亿元人民币（新浪财经，2026.8.26）
增长引擎	1H26 API 收入 8.25 亿元人民币、占 86.5%；8 月底 Token 调用较年初增长超 40 倍、API 均价+101%（公司披露，2026.8.31）	1H26 开放平台及其他 AI 企业服务收入 7,393 万美元，占收入 63.4%（公司披露，2026.8.26）	API 收入占比超过 70%，海外开发者及企业客户为主要来源（观察者网，2026.7.22）	截至 2025 年底模型装机量超过 4,200 万台，2026 年计划上车超过 100 万辆；手机及车载为主要分发方向（科创板日报，2026.1.26）	开源模型 API 调用及企业 Agent；V4.1-Flash 于 9 月 10 日加入原生视觉，已接入 WorkBuddy 及 OpenCode（公司更新，2026.9.10）
业务结构与核心产品					
引擎① 基座模型	GLM-5.3 沿用 753B 参数基座，8 月推出；Code Bench max 档完成率由 23.4%升至 34.5%，输出 Token 由 9.6 万降至 7.5 万。Flash 总参 320B/激活 18B，支持原生视觉及 1M 上下文；注意力计算量/KV 缓存约为旗舰的 1/3 和 1/4.4。Flash/FlashX 输入输出价为 0.15/0.50 及 0.37/1.25 美元每百万 Token（公司，2026.10.1；API 文档截至 10.2）。	M3 总参约 428B/激活 23B，支持 1M 上下文、原生图像及视频理解和桌面操作；1M 下每 Token 计算量为前代约 1/20，预填充/解码加速超过 9/15 倍。公司 SWE-Bench Pro/Terminal-Bench 2.1 得分 59.0%/66.0%。H3 支持 15 秒、2K、原生立体声视频；Music 3.0 支持 5 分钟完整歌曲（公司公告，2026.6—8）。9 月下旬上线 M3.1-Flash-Preview，支持 1M 上下文、多模态输入及五档推理深度。	Kimi K3 总参 2.8T/激活 104B，从 896 个专家中选 16 个，支持 1M 上下文及原生视觉；7 月 16 日发布、7 月 27 日开放权重。SWE-Marathon/FrontierSWE/OSWorld-Verified 为 42.0%/81.2%/84.8%。K2.8 Preview 于 9 月 11 日接入 Kimi Code，支持 1M 及 low/high/max 推理档（公司公告）。	Step 5 Preview 于 2026 年 9 月 20 日发布，总参数 600B、激活参数 27B，支持 100 万 Token 上下文及视觉输入；公司评测 High 档 StepCodeBench/FrontierFinance 得分 49.0%/66.4%，重点强化复杂编程及金融工作，计划 10 月 15 日开放权重。StepAudio 3 Realtime 支持全双工对话及异步工具执行，公司评测 τ-Voice 任务成功率 56.0%。	V4.1-Flash 于 9 月 10 日发布，552B 主干另含 196B Engram 条件记忆参数，预填充/解码分别激活 8B/16B；支持 1M 上下文、384K 输出及原生视觉。max 档 Terminal-Bench 2.1/DeepSWE 1.1 得分 90.6%/74.2%；KV 缓存约 890 字节/Token，HBM/SSD 缓存为前代 1/4 和 1/8（公司公告，2026.9.10）。
引擎② C 端/产品	Coding Plan 于 2026 年 8 月向全量用户开放；开放平台 API 及 Co-work 专业场景	MiniMax Code、Mavis（升级后的 MiniMax Agent）、海螺及音频产品	Kimi、Kimi Work、Code 及 API；Kimi Code Desktop 于 9 月 17 日覆盖 macOS 及 Windows	2026.7.13 发布 STEPX、Step AOS 及 Amoo，展出 STEPX Neo 手机	V4.1-Flash 上线 API，WorkBuddy（含 CodeBuddy）及 OpenCode 支持新模型

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)	稀宇科技 MiniMax (100 HK)	月之暗面 Moonshot (Kimi)	阶跃星辰 StepFun	DeepSeek
新曲线	Co-work 订单超 10 亿元人民币；10 月 GLM-5.3 接入 AWS，按调用量分成，阿里云等已签类似协议（公司 9.16；科创板日报 10.6）	8 月 ARR 中 To B 约占 80%；管理层将 M3.1 推理成本目标设为 M3 上线时约 1/3（公司披露，2026.8.26）	9 月推出企业伙伴"登月计划"；K3 于 9 月 18 日上线 AWS，按调用量分成；阿里云百炼亦达成类似合作（AWS；每经及第一财经 9.21）	StepAudio 3 Realtime 支持异步工具执行，语音交互向任务执行延伸	天翼云部署文档提供基于 vLLM-Ascend、昇腾 910B 及 W8A8 量化的部署方案；缓存压缩降低长上下文部署需求
上市进度					
IPO 状态	2026 年 1 月 8 日在港交所主板挂牌上市；A 股辅导已完成，发行方案于 2026.6.22 获股东批准	2026 年 1 月 9 日在港交所挂牌；5 月 29 日完成 A 股上市辅导备案，5 月 31 日公告探索科创板上市	据报已秘密递表；目标 2027 年一季度香港上市（Bloomberg，2026.10.6）	据道琼斯 2026 年 6 月 8 日报道，彼时计划递交香港 IPO 申请	已聘请中介机构准备潜在科创板 IPO，目标于 2027 年初上市（Reuters，2026.9.9；彭博，2026.10.6）
募资进展	约 20 亿美元配售及 30 亿美元可转债分别于 9 月 16、18 日完成，净所得合计 392.74 亿港元，约 60%用于模型及基础设施	7 月 14 日/16 日完成配售/可转债，融资总额 95.408 亿/65 亿港元，合计约 160.4 亿港元	据报完成上市前最后一轮私募，估值约 500 亿美元；香港 IPO 拟募资最多 50 亿美元（Bloomberg，2026.10.6）	2026 年 5 月完成超过 20 亿美元 Pre-IPO 融资（道琼斯，2026.6.8）	最新融资规模预计超 800 亿元人民币，最终或接近 1,000 亿元，预计 10 月完成（彭博，2026.10.6）
最新估值					
当前市值/估值	2,511 亿港元（~321 亿美元）	660 亿港元（~84 亿美元）	据报已完成上市前最后一轮私募融资，估值约 500 亿美元（Bloomberg，2026.10.6）	主要投资者提出的上市目标估值最高 120 亿美元（道琼斯，2026.6.8）	2026 年 6 月首轮投后估值超 500 亿美元；下一轮讨论估值约 5,000 亿元人民币（Reuters，2026.9.9）
估值倍数参考	26/27E P/S：33.5x/11.2x	26/27E P/S：17.9x/6.1x	按 500 亿美元最新私募估值/2026 年底 20 亿美元目标 ARR 测算，估值/目标 ARR 约 25x	—	按 5,000 亿元人民币讨论估值/10 亿美元年化收入运行率测算，约 74.5x

资料来源：公司资料、每日经济新闻、Reuters、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：公司经营及资本化进展更新截至 2026 年 10 月 14 日，各项数据日期见表。

■ 智谱 AI（2513 HK）：长程任务能力与 Token 效率同步提升，API 量价齐升

智谱 API 业务量价齐升，云分成与融资同步支持商业化扩张。GLM-5.3 强化复杂工程能力，Flash 覆盖高频及低成本调用，国产算力和推理系统优化推动单位 Token 成本下降。1H26 开放平台及 API 收入占比升至 86.5%，分部毛利率转正；9 月，约 50 亿美元融资补充下一代模型及基础设施投入，支持公司扩大模型与服务供给。10 月初，AWS 收入分成落地，进一步增加自营 API 之外的收入来源。

GLM-5.3 强化长程工程能力，完成任务所需 Token 同步下降。据智谱 2026 年 8 月 14 日研究博客及官方模型卡，GLM-5.3 沿用 GLM-5.2 的 753B 参数基座，通过后训练强化复杂编程及长程任务。公司自建 Z.ai Code Bench 评测显示，max 档端到端任务完成率由 GLM-5.2 的 23.4%提升至 34.5%，每任务输出 Token 由约 9.6 万降至约 7.5 万。模型迭代重点包

括扩大长程任务环境及强化学习规模，使模型在连续执行、失败与重新规划中学习，能力由代码生成向完整工程交付延伸。

GLM-5.3-Flash：原生视觉与架构降本结合，FlashX 承接高时效调用。 GLM-5.3-Flash 于 2026 年 8 月 26 日上线，覆盖浏览器操作、图形界面交互及办公文档生成。据智谱 10 月 1 日技术披露，Flash 总参数 320B、激活参数 18B，使用稀疏与线性混合注意力及 IndexPool，注意力计算量约为 GLM-5.3 的三分之一，KV 缓存约为其 1/4.4，支持 100 万 Token 上下文及最高 128K 输出。在公司同框架评测中，DeepSWE 1.1 得分为 63.4%，高于 GLM-5.2 的 46.2%；Toolathlon 为 78.4%，高于 GLM-5.2 的 59.9%。截至 10 月 2 日，官方 API 文档列示 Flash 输入/输出价格为 0.15/0.50 美元每百万 Token，FlashX 为 0.37/1.25 美元，后者提供约 200 Token/秒生成速度。旗舰、Flash 及 FlashX 分别覆盖高难度工程、高频任务与高时效需求，将能力提升及推理降本转化为分层产品。

GLM-Legal 将低成本基座拓展至法律工作流。智谱于 10 月 12 日发布基于 GLM-5.3-Flash 的法律模型 GLM-Legal，通过法律语料训练及数万任务环境中的强化学习，提升合同审阅、证据分析及跨境合规研究能力。公司披露，其在覆盖 24 个法律实务领域的 Harvey LAB 评测中，全部要求通过率为 19.2%，较基座提升 12.5 个百分点。目前已开放 API 试用申请，进一步拓展专业场景的模型调用需求。

Token 调用与 API 均价同步上升，付费及重度用户增长快于总用户。截至 8 月 31 日，MaaS 平台 Token 调用量较年初增长超过 40 倍，其中 Coding Plan 调用量增长超过 23 倍，API 平均售价同期提高约 101%。MaaS 企业及开发者用户由 6 月底超过 580 万增至 8 月底超过 740 万，较年初增长 144%；付费日活用户较年初增长 603%，按收入计前十大用户日均调用量增长 98 倍。公司 8 月向全量用户开放 Coding Plan 的 Lite、Pro、Max 及团队版本，订阅与 API 调用共同承接编程需求。

国产算力进入大规模推理，工程优化带动单位 Token 成本下降。智谱已实现 10 万级国产芯片的规模化低成本推理，单位 Token 推理成本较年初下降 80%；通过 SGLang 及编码、预填充、解码分离的服务架构，同一硬件上的端到端服务性能较初始基线提升 3 倍。公司另称，GLM-5.3 参与推理引擎及算子开发后，算子开发周期缩短约一半。

API 收入占比升至 86.5%，分部毛利率转正，整体毛利率仍受业务结构变化影响。智谱 1H26 收入 9.54 亿元人民币，同比+399.7%；其中开放平台及 API 收入 8.25 亿元，同比+2,735.7%，占比由 1H25 的 15.2%升至 86.5%，毛利率由-0.4%提升至 24.6%。同期企业级智能体收入5,556万元，同比+304.4%，企业级通用大模型收入6,704万元，同比-54.6%。由于高毛利私有化模型交付收缩、客户向云端迁移，本地化部署毛利率由 59.1%降至 37.9%，整体毛利率由 50.0%降至 26.4%。

云厂商分成落地，GLM 新增海外收入渠道。10 月 6 日据《科创板日报》报道，GLM-5.3 接入 Amazon Bedrock，AWS 将按模型调用量与智谱进行收入分成；智谱也已与阿里云百炼等国内云厂商签署类似协议。模型厂商提供模型能力，云平台承担托管推理及企业分发，使智谱能够分享第三方平台的调用收入。公司此前于 9 月 16 日披露全业务 ARR 已达 18 亿美元，并将年底目标由 24 亿美元上调至 30 亿美元，预计云托管 API 收入自 10 月起确认。

Co-work 订单扩大，专业任务需求增长。智谱于 9 月 16 日披露，GLM-5.3 发布约一个月后，Co-work 行业订单金额超过 10 亿元人民币，已有 100 家安全企业将 GLM 接入产品及业务流程。安全企业将 GLM 接入产品及日常业务流程，有望带来持续的模型调用需求。

约 50 亿美元融资完成，六成资金投向模型研发及算力基础设施。智谱于 9 月 16 日以每股 714 港元完成配售 2,196.5 万股 H 股，9 月 18 日完成发行本金 201.4 亿元人民币的美元结算零息可转债，两项净所得合计 392.74 亿港元（约 50.1 亿美元）。公司计划将约 60%资金用于下一代 GLM、完全自训练体系及训练推理基础设施。配售新股占扩大后股本 4.50%；可转债初始转股价为 892.50 港元，于 2027 年到期。

■ **MiniMax（100 HK）：开放平台为主要增长驱动，多模态模型与推理降本持续推进**

MiniMax 商业化重心向企业及开发者转移，开放平台成为主要收入来源。M3、Token Plan 及开放平台带动 API 调用，H3 与 Music 3.0 继续拓展多模态内容生产。1H26 开放平台收入占比及整体毛利率均提升，研发开支随模型投入增加。预计长上下文及多模态能力升级、推理成本下降，将继续推动企业用户和调用量增长。

M3 以长上下文与原生多模态支撑编程 Agent，模型效率构成 API 商业化基础。据 MiniMax 2026 年 6 月 1 日发布及官方模型卡，M3 总参数约 428B、激活参数 23B，支持 100 万 Token 上下文、原生图像及视频理解，并可执行桌面操作；采用 MSA 稀疏注意力后，在 100 万 Token 上下文条件下，每 Token 计算量约为上一代模型的 1/20，预填充及解码速度分别提升超过 9 倍及 15 倍。公司评测的 SWE-Bench Pro 及 Terminal-Bench 2.1 得分分别为 59.0%及 66.0%，模型通过 MiniMax Code、Token Plan 及 API 提供服务，为三季度企业及开发者业务增长提供能力与产品基础。9 月下旬，公司进一步上线 M3.1-Flash-Preview，延续百万 Token 上下文及多模态输入，新增五档推理深度设置，可根据任务复杂度调整推理投入，覆盖代码修复、调试及完整功能开发等场景。

H3 强化商业视频生成，Music 3.0 将完整歌曲生成扩展至 5 分钟。H3 于 7 月 31 日正式发布，可以统一处理文本、图像、视频及音频输入，支持最长 15 秒、2K 分辨率及原生立体声视频，强化广告、电商、品牌展示等场景中的文字呈现、运动迁移及视听一致性。公司称其 2K 视频每秒价格低于所比较主流模型的三分之一，768p 价格低于主流 720p 模型的一半；H3 已开放权重。公司于 2026 年 8 月 13 日发布 Music 3.0，根据创作概念及可选歌词，一次完成作曲、编曲、演唱及制作，支持最长 5 分钟完整歌曲，并可按段落控制情绪、配器及人声表现，扩展商业内容生产场景。

8 月 ARR 超过 8 亿美元，企业及开发者业务占比提升。公司此前披露 2 月 ARR 超过 1.5 亿美元；管理层于 1H26 业绩会披露，8 月 ARR 超过 8 亿美元，其中 To B 约占 80%、To C 约占 20%，去年同期约为 30%和 70%。公司 5 月底企业及开发者客户超过 100 万，较约半年前的 20 万增至五倍，为调用增长提供客户基础。

开放平台收入快速增长，毛利率改善。MiniMax 1H26 收入 1.166 亿美元，同比+283.1%；开放平台及其他 AI 企业服务收入 7,393 万美元，同比+703.1%，占比由 1H25 的 30.3%升至 63.4%，受付费客户、API 调用及 Token Plan 增长推动。AI 原生应用收入 4,264 万美元，同比+100.9%；中国内地以外收入 7,083 万美元，占 60.8%。整体毛利率由 12.1%升至 17.9%，研发开支增至 2.969 亿美元、同比+138.8%，经调整净亏损为 2.930 亿美元、同比+111.2%。

推理降本继续聚焦 M3.1 与国产算力。管理层表示，文本模型训练资源投入约为视频模型的四倍，M3 及 H3 正在推进国产芯片适配，并计划上线大规模国产算力集群。M3.1 推理成本目标为 M3 刚上线时的约三分之一，公司计划通过降价扩大用户及 Token 规模，沿模型效率与服务系统两端提升算力利用率。

7 月完成约 160.4 亿港元融资，科创板上市筹备继续推进。 据 MiniMax 7 月 16 日公告，公司分别于 7 月 14 日及 16 日完成配售和可转债发行，配售 3,560 万股、每股 268 港元，融资总额 95.408 亿港元，可转债本金 65 亿港元，合计约 160.4 亿港元。A 股方面，公司 2026 年 5 月 31 日公告探索科创板上市，融资及上市进程有望支持后续模型训练、推理基础设施及全球业务投入。

■ **DeepSeek（未上市）：架构优化压缩长程任务成本，API 商业化与融资同步推进**

DeepSeek 以架构优化提高算力利用效率，API 商业化规模继续扩大。V4.1-Flash 在强化编程和原生视觉的同时，降低长上下文缓存和计算开销，峰谷定价进一步推动持续调用。媒体披露的年化收入运行率已达 10 亿美元，公司同步推进外部融资及科创板 IPO 准备，为后续模型训练和基础设施投入提供资金支持。

V4.1-Flash 强化编程与自动化任务，长上下文效率继续提升。 据 DeepSeek 2026 年 9 月 10 日发布及官方模型卡，V4.1-Flash 采用非对称编码器—解码器架构，主干参数 552B，另含 196B Engram 条件记忆模块，预填充及解码阶段分别激活 8B 及 16B 参数，加入原生视觉能力。模型支持 100 万 Token 上下文及最高 384K 输出；在公司 max 推理档评测中，Terminal-Bench 2.1 由 V4-Flash 的 82.7%提升至 90.6%，DeepSWE 1.1 由 54.4%提升至 74.2%。KV 缓存压缩至每 Token 约 890 字节，高带宽内存及 SSD 缓存空间分别降至前代的 1/4 及 1/8；上下文从 4K 扩大至 1M 时，解码计算量仅增加约 25%。

峰谷定价叠加缓存折扣，降低持续调用成本。新价格于北京时间 9 月 10 日 12 时生效：V4.1-Flash 峰时每百万 Token 缓存命中输入/未命中输入/输出价格分别为 0.006/0.30/1.20 美元，谷时分别为 0.003/0.15/0.60 美元。公司同步宣布 WorkBuddy（含 CodeBuddy）及 OpenCode 完成集成，模型能力通过编程及办公 Agent 直接触达用户。

国产算力部署继续拓展，长程任务具备更低资源门槛。V4.1-Flash 已可通过 vLLM-Ascend 在昇腾 910B 集群部署，采用 W8A8 量化及双节点方案。结合模型侧的缓存压缩，国产算力适配进一步拓展了企业私有化部署的选择。

API 商业化据报快速增长，年化收入运行率达到 10 亿美元。Reuters 于 2026 年 9 月 24 日转述 The Information 援引知情人士报道，DeepSeek 年化收入运行率已达 10 亿美元，较数月前翻倍以上。据新浪财经 8 月 26 日援引知情人士报道，公司 1-7 月累计收入约 4.75 亿元人民币，约为 2025 年全年的十倍；同期整体毛利率 44.6%、API 业务毛利率 82.9%，净亏损约 7.15 亿元人民币。架构与定价优势正在带动 API 调用及收入规模扩大。

外部融资与科创板 IPO 准备并行。据 Reuters 2026 年 9 月 9 日报道，DeepSeek 于 6 月完成约 74 亿美元首轮外部融资，投后估值超过 500 亿美元，并准备潜在科创板 IPO。Reuters 9 月 24 日转述 The Information 称，公司下一轮融资目标金额约 500 亿元人民币、讨论估值约 5,000 亿元人民币，计划于 10 月底前完成。持续融资支持训练及基础设施扩张，也反映公司在 API 商业化放量阶段增加模型与算力投入。

新一轮融资规模有望接近千亿元。 据彭博于 2026 年 10 月 6 日报道，DeepSeek 新一轮融资规模预计超过 800 亿元人民币，按已签署的投资条款书，最终金额可能接近 1,000 亿元，预计于 10 月内完成。腾讯及宁德时代为本轮主要投资方，融资目标估值约 5,000 亿元人民币。公司同时推进 2027 年初 IPO 准备。融资规模扩大将增强后续模型训练及基础设施建设的资金储备，为持续迭代和 API 业务扩张提供支持。

■ **月之暗面（未上市）：K3 拓展长程工作能力上界，Coding 及企业渠道承接商业化**

月之暗面三季度以 K3 的长程任务能力为基础，持续拓展 Kimi Code 及企业产品。9 月 Kimi K3 接入 AWS 并落地按调用量分成，海外商业化由自营 API 进一步延伸至云平台渠道。公司 ARR 据报已达约 10 亿美元，最新一轮私募估值升至约 500 亿美元，资本与渠道扩张共同支持后续模型迭代和企业业务发展。

Kimi K3 升级至 2.8T 参数，长程编程与电脑操作能力同步提升。K3 于 2026 年 7 月 16 日发布、7 月 27 日开放完整权重。模型总参数 2.8T、激活参数 104B，每次从 896 个专家中选择 16 个，支持原生视觉及 100 万 Token 上下文。在公司评测中，K3 的 SWE-Marathon 得分 42.0%，高于 GPT-5.6 Sol 的 39.0%及 Claude Opus 4.8 的 40.0%；FrontierSWE 为 81.2%，OSWorld-Verified 为 84.8%。

长程能力向知识工作产品落地，缓存命中降低重复上下文成本。K3 通过 Kimi、Kimi Work、Kimi Code 及 API 提供服务；Kimi Work 增加 Widgets 交互组件及 Dashboard，支持连接本地和外部插件、生成可持续使用的工作面板。公司展示 K3 连续自主运行 48 小时完成芯片设计及仿真验证的案例。API 每百万 Token 缓存命中输入/未命中输入/输出价格分别为 0.30/3/15 美元，公司称官方编程 API 的缓存命中率超过 90%。国产算力方面，其已完成 K3 在昇腾 910C 上的缓存卸载适配及部署验证。

Kimi Code 补齐 Windows 电脑操作与桌面入口，持续扩大长程编程产品覆盖。2026 年 8 月 6 日开放 Windows x64 电脑操作能力；9 月 9 日 Remote Control 正式上线，支持通过手机或其他电脑继续控制本地会话；9 月 11 日向用户开放 K2.8 Preview 及 100 万 Token 上下文，提供 low/high/max 推理档位。9 月 17 日发布的 Kimi Code Desktop 覆盖 macOS 及 Windows，支持任务目标与计划审阅、长任务后台运行、内嵌浏览器及可视化标注，降低复杂编程任务的操作门槛。

ARR 由 3 月超过 1 亿美元升至 8 月超过 10 亿美元，API 贡献主要收入。据 Bloomberg 9 月 11 日援引知情人士报道，月之暗面 ARR 由 2026 年 3 月初超过 1 亿美元，增至 6 月 3 亿美元、8 月超过 10 亿美元，公司年底目标为 20 亿美元。据观察者网 7 月 22 日报道，彼时 API 收入占比超过 70%，海外开发者及企业客户为主要收入来源。

海外云分成已落地，AWS 及阿里云扩大企业调用覆盖。9 月 18 日，Kimi K3 正式上线 Amazon Bedrock；《每日经济新闻》9 月 21 日向公司确认，AWS 与月之暗面的收入分成合作已落地，按模型调用量结算。同期 Kimi K3 也已上架阿里云百炼，并达成类似合作。云平台提供推理资源、企业客户及结算渠道，使月之暗面能够在自营 API 之外分享模型使用收入。国内交付方面，公司 9 月 10 日推出企业合作伙伴“登月计划”，以 FDE 驻场工程团队协同系统集成商交付企业 Agent。

据报完成上市前融资，估值升至约 500 亿美元。Bloomberg 于 10 月 6 日报道，月之暗面已完成上市前最后一轮私募，估值约 500 亿美元，较今年夏季一轮融资的 315 亿美元提高约 59%；公司正以 2027 年一季度香港上市为目标，并考虑 IPO 募资最多 50 亿美元。K3 及 Kimi Code 带动自营 API 与订阅增长，海外云分成增加新的变现渠道，融资补充模型迭代和商业化所需的投入，支持公司在上市前继续扩大收入规模。

■ **阶跃星辰（未上市）：全双工语音强化任务执行，自有终端拓展分发入口**

阶跃星辰以专业模型与终端合作拓宽 Agent 应用。Step 5 Preview 强化长程编程和金融工作，StepAudio 3 将实时语音延伸至工具执行，STEPX、Step AOS 及 Amoo 提供个人终端入口。手机和汽车合作扩大模型触达，专业任务及语音交互则增加用户使用场景，形成模型能力与渠道协同推进的路径。

Step 5 Preview：旗舰模型向复杂编程及金融工作延伸，长程执行与低价 API 改善使用效率。 阶跃星辰于 2026 年 9 月 20 日发布新一代旗舰基座模型 Step 5 Preview，采用 600B 总参数、每 Token 激活 27B 参数的 MoE 架构，支持 100 万 Token 上下文及视觉输入，重点强化软件工程及专业知识工作。在公司评测中，High 档 StepCodeBench 成绩为 49.0%，高于 Kimi K3 Max 的 43.9%及 GLM-5.3 Max 的 40.2%；FrontierFinance 得分 66.4%，高于 GLM-5.3 Max 的 64.1%及 GPT-6 Astra Max 的 55.0%，低于 Claude Opus 5 Max 的 69.7%。长程任务方面，在 H100 上优化 MLA GPU 算子的 24 小时测试中，各模型分别进行四次独立尝试并取最佳结果，Step 5 Preview 持续迭代约 22 小时后达到 508 TFLOPS，高于 Claude Opus 5 的 493 TFLOPS，体现通过执行、测试及反馈持续优化计算效率的能力。定价方面，每百万 Token 输入/输出价格为 1/2.70 美元，缓存读取价为 0.05 美元，有助于降低长上下文及重复调用成本。公司表示，模型已通过产品及 API 提供服务，计划于 2026 年 10 月 15 日开放权重。

StepAudio 3 系列发布，全双工语音向任务执行延伸。公司于 9 月 15 日发布 StepAudio 3 系列，覆盖 Realtime、ASR、TTS、Gen 及 Music。Realtime 支持边说边思考、实时打断及异步工具执行，在公司评测中 AA Full-Duplex Bench 得分 98.9 分；τ-Voice 语音 Agent 任务成功率 56.0%，接近 Grok Voice Think Fast 2.0 High 的 56.5%，高于 GPT Realtime 2.1 High 的 45.7%。

STEPX 与 Step AOS 亮相，个人智能体向系统级执行拓展。阶跃于 7 月 13 日发布 STEPX 品牌、Step AOS 操作系统及 Amoo 个人智能体，并展出 STEPX Neo 手机。Step AOS 将设备能力与数据接口开放给模型，支持智能体理解需求后调用系统功能；端侧模型 StepEdge 承担轻量任务，复杂任务调度至云端，形成端云协同的执行方式。

手机和车载渠道扩大模型覆盖。据《科创板日报》2026 年 1 月 26 日报道，截至 2025 年底，阶跃合作覆盖 OPPO、荣耀、中兴等品牌，模型装机量超过 4,200 万台，日均服务近 2,000 万人次；公司与千里科技、吉利联合推出 AgentOS 智能座舱，彼时预计 2026 年模型上车规模超过 100 万辆。终端入口为实时语音、个性化助手及跨应用执行提供分发基础。

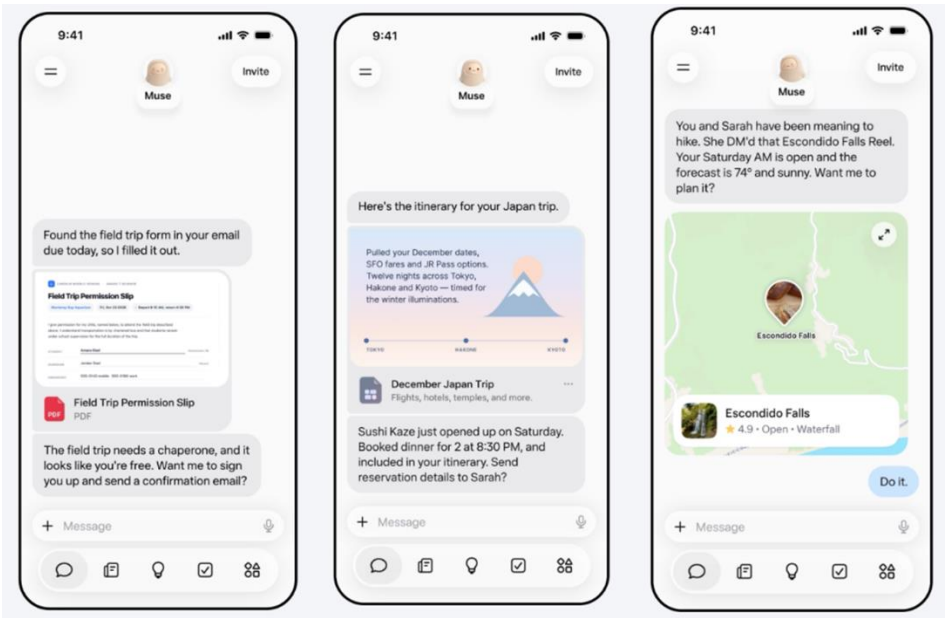
收入增长与 Pre-IPO 融资提供投入支持。 据《财经》2026 年 2 月 27 日报道，阶跃 2025 年收入近 5 亿元人民币，彼时预计 2026 年收入约 12 亿元人民币。据道琼斯 2026 年 6 月 8 日报道，公司于 5 月完成超过 20 亿美元 Pre-IPO 融资，主要投资者提出的上市估值最高达 120 亿美元；公司当时计划递交香港上市申请。

重点覆盖公司 3Q26 AI 业务进展更新

Meta：C 端 Agent 加速落地

Meta 在 9 月推出个人 Agent 应用 Muse，让每个用户的个人 Agent 能够运行在安全独立的云端虚拟机内，帮助用户完成日常工作生活中的各类任务，包括购物、发送邮件、制定计划等。

图 22：Meta Muse



资料来源：公司资料，招银国际环球市场

根据 FundAI 数据，Meta Muse 上线后首月内 DAU 即突破 200 万，但由于产品目前仅上线北美地区，DAU 在达到 200 万后整体增速放缓；14 日留存率在上线初期较低仅 9%，但后续逐渐改善至 16%。

图 23：Meta Muse：DAU

资料来源：FundAI，招银国际环球市场

从商业化维度看，短期 Muse 主要通过会员订阅变现，Free/Power/Maximum 用户每月会员价格对应 0/20/100 美元，每周 Token 额度对应 1 亿/5 亿/30 亿；中长期 Muse 或考虑对交易抽取佣金的方式进行变现。我们按照 Muse 长期 DAU 渗透率能达到 Meta 全球 DAU 的

10%测算 Muse 的长期收入空间，Muse 收入空间有望达到 76 亿美元（订阅/佣金收入分别 43/33 亿美元），约为公司 FY26E 总收入的 3%。

图 24：Meta Muse：年收入空间测算

主要假设	收入空间测算
Muse DAU (百万) - Meta 全球 DAU*10%	180
1) Power 付费用户 - 5%付费渗透率（百万）	9
年 ARPPU (美元)	240
2) Maximum 付费用户 - 1%付费渗透率（百万）	1.8
ARPPU (美元)	1,200
Muse 订阅收入 (亿美元)	**43**
Muse DAU (百万) - Meta 全球 DAU*10%	180
1) 每 DAU 日均商业化问询量（次）	0.5
2) 商业化问询变现费率（美元）	0.1
Muse 交易佣金收入（亿美元）	**33**
Muse 潜在年收入空间（亿美元）	**76**

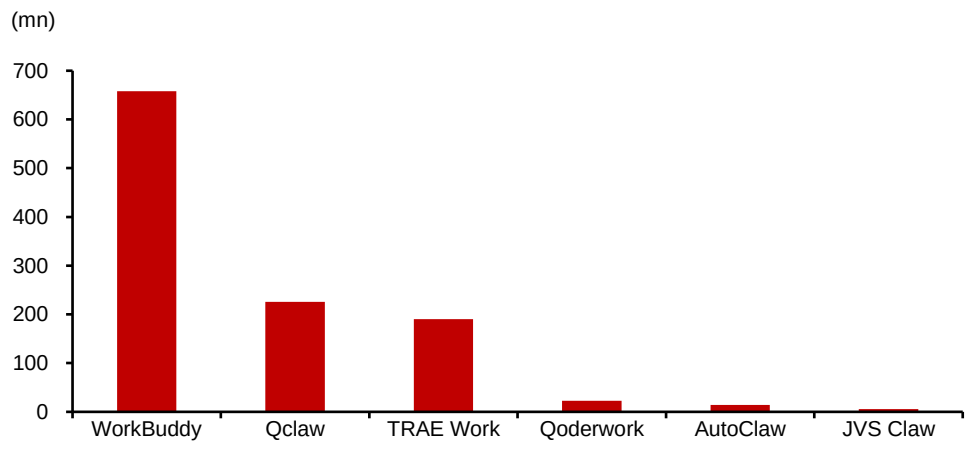
资料来源：公司资料，招银国际环球市场预测

Muse Wonder 系列发布，强化个人 Agent 及长程软件工程能力。Meta 于 10 月 12 日推出 Muse Wonder 及 Muse Wonder Full，支持 100 万 Token 上下文，通过 Active Memory 保留并调用跨会话信息，通过 Active Focus 在长程任务中持续推进、调整方案及修复错误。公司评测中，Wonder Full 的 DeepSWE 1.1 及 SWE-Marathon 成绩分别为 79.0%及 57.5%（后者为 xhigh 档）。Wonder 已通过 Meta Model API 及 Muse Code 提供服务，输入/输出价格为 4/20 美元每百万 Token；Full 面向 Muse Code 订阅用户开放。Wonder 后续将接入 Muse 并优先覆盖订阅用户，提升个人 Agent 承接复杂工作及持续执行任务的能力。

腾讯：WorkBuddy 持续领跑国内办公 Agent 市场

腾讯 WorkBuddy 用户量持续领跑中国办公类 Agent 市场。根据 QuestMobile 数据，2026 年 7 月腾讯 WorkBuddy 的 MAU 达到 658 万，相较 4 月增长 269%，显著领先于其他办公类 Agent 应用。

图 25：中国典型办公类 Agent：MAU 规模（2026 年 7 月）



资料来源：QuestMobile，招银国际环球市场

WorkBuddy 持续加强生态建设，将 WorkBuddy 打造成面向 Agent 时代的操作系统和平台。公司在 2026 年 9 月正式上线 WorkBuddy 开放平台，首批引入超百家生态伙伴，面向智能硬件、行业应用与开发者等全面开放底层 Agent 能力：1）围绕“听、看、记、聊、协”五类触点，WorkBuddy 已接入智能眼镜、录音卡片、耳机、键鼠、PC 与平板、智能穿戴、手机等超十个品类；2）围绕行业伙伴，WorkBuddy 上线“Buddy 应用”，行业伙伴可以基于 WorkBuddy Agent 底座能力，结合自身行业知识和经验积累，构建自己的品牌化 AI 工作台；3）围绕开发者，WorkBuddy 开放平台同步开放 Skill、Expert、Connector 三大能力，调用标准 API 即可获得完整 Agent 能力。

图 26：腾讯 WorkBuddy：Agent 开放平台

资料来源：公司资料，招银国际环球市场

以中国白领用户总规模作为 WorkBuddy 可触达市场空间，假设 WorkBuddy 在白领人群的用户渗透率达到 10-20%，我们测算 WorkBuddy 长期收入空间为 264-720 亿元，对应约占腾讯 FY26 总收入 3-9%。

图 27：腾讯 WorkBuddy：收入空间测算

WorkBuddy收入空间测算（亿元）						
	付费用户规模(百万) - 10-20%中国白领付费渗透率					
		20	25	30	35	40
年ARPPU（元）-对应不同档位会员比例	1320	264	330	396	462	528
	1440	288	360	432	504	576
	1560	312	390	468	546	624
	1680	336	420	504	588	672
	1800	360	450	540	630	720

资料来源：公司资料，国家统计局，招银国际环球市场预测

阿里巴巴：深化全栈布局、拓展全球数据中心规模

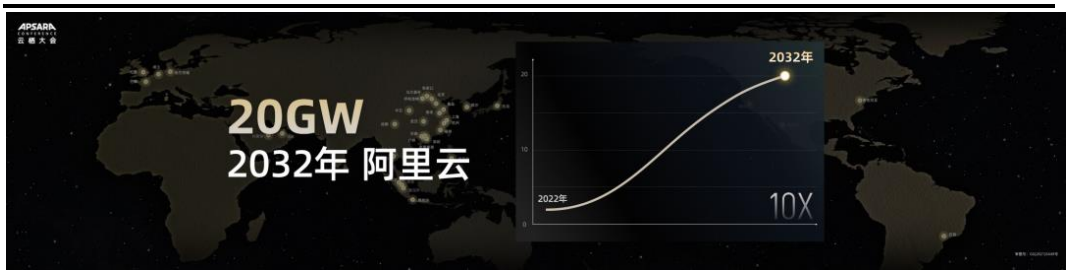
2026 年 9 月云栖大会，阿里巴巴 CEO 吴泳铭提出，到 2032 年阿里云运营的全球数据中心规模要超过 20GW。模型方面，公司更新 Qwen4 正在训练，Qwen4.5 和 Qwen5 规划扩展到 5–10T 参数。芯片方面，平头哥发布真武 V900，性能是 M890 的 3 倍，单集群可扩到 50 万卡，预计于 1Q27 量产。倚天 720 和 730 CPU 将在 2027 年推出。

图 28：阿里巴巴：Qwen 基座模型迭代路线及计划

资料来源：公司资料，招银国际环球市场

同时，阿里云在云栖大会上宣布新一轮全球基础设施扩建计划：未来 12 个月，将于土耳其、芬兰和荷兰首次设立云计算地域节点，并扩建马来西亚、德国、阿联酋、法国和中国香港的数据中心，进一步提升全球算力供给和本地化云与 AI 服务能力。同时，阿里云还将升级 SmartFusion、SmartVideo 等面向国际市场的 AI 产品。

图 29：阿里巴巴：规划全球数据中心规模预计将在 2032 年超过 20GW



资料来源：公司资料，招银国际环球市场

股票推荐

- **微软（MSFT US，买入，目标价：634.2 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。我们认为，微软的资本开支投资所带来的可量化回报正在云业务和自有应用中变得愈发显著，同时管理层进一步强调资本开支效率而获得的市场正反馈有望维持，管理层指引 FY27 自由现金流将为正。我们基于 DCF 的目标价 634.2 美元对应 32 倍 FY27E 市盈率。"买入"评级。
- **亚马逊（AMZN US，买入，目标价：301.0 美元）：**AWS 收入增长稳健， 利润率持续提升。我们仍看好 AWS 业务营收同比增速环比进一步加速的趋势，同时认为经营利润的稳健增长（2026E 招银国际预期：35% YoY）有望支撑股价增长。我们的目标价为 301 美元，基于 17.2 倍 EV/EBITDA。"买入"评级。
- **谷歌（GOOG US，买入，目标价：428.0 美元）：** 我们看好谷歌全栈 AI 解决方案将支撑公司核心搜索和云业务的快速增长，让公司在显著加大 AI 投入的同时维持稳健盈利增长与现金流，持续受益于强劲的 AI 需求。中短期维度，我们认为两项核心催化有望支撑公司基本面和估值，包括 TPU 订单加速落地推动云业务增长、Gemini 4 模型系列的发布进一步强化公司全栈 AI 战略优势。我们的目标价为 428.0 美元，基于 35 倍 FY26 市盈率。"买入"评级。
- **阿里巴巴（BABA US，买入，目标价：220.1 美元）：**公司云计算收入增长前景强劲，同时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。阿里云外部收入 1QFY27 同比增速进一步加快至 45%（4QFY26：40%），在强劲需求及持续扩充产能的支持下，未来几个季度增速加快的势头有望延续。此外，AI 相关资本开支可实现约 3 年的回收期，随着 AI 相关产品毛利率提升，叠加规模经济改善、自研芯片部署比例提升以及资本效率改善，回收期有望进一步缩短。我们继续看好阿里巴巴作为我们覆盖范围内 AI 主题的核心受益者之一。我们认为，传统电商业务产生的稳定现金流以及阿里巴巴强劲的资产负债表，将支持公司持续投入，以更好把握 AI 时代的发展机遇；与此同时，清晰且具吸引力的 ROIC 实现路径有望为长期股东回报提供有力支撑。我们基于 SOTP 的目标价为 220.1 美元。"买入"评级。
- **腾讯（700 HK，买入，目标价：705.0 港元）：**腾讯主业良好的盈利与现金造血能力，叠加公司庞大且可变现的投资组合，有望支撑公司持续的 AI 投入。我们预计公司 FY26 全年 Non-IFRS 运营利润同比增长 5%。我们认为公司后续需凭借 AI 方面的积极进展以证明 AI 投入的 ROI，短期推动公司估值回升，中长期推动盈利增长重新加速：1）WorkBuddy 用户数据/调用量/商业化进展更新；2）HY4 正式版接近国内模型性能-成本前沿；3）微信小微全量推出并探索商业化。公司当前估值仅对应 12x FY26 Non-IFRS PE，仍有较大修复空间。我们基于 SOTP 的目标价为 705.0 港元，对应 21x FY26E Non-IFRS PE。"买入"评级。
- **智谱（2513 HK，买入，目标价：1985.0 港元）：**我们认为，智谱在模型迭代速度、智能水平与单任务成本之间实现了有效平衡，展现出持续处于大语言模型行业帕累托前沿的强大能力，有望为长期商业化打开可观空间。我们目标价为 1,985.0 港元，基于 28 倍 2028E 市销率（P/S）折现至 2026E。"买入"评级。
- **MiniMax（100 HK，买入，目标价：580.0 港元）：**我们看好公司维持强劲 ARR 增长，得益于模型的持续迭代及多模态战略带来增量变现机会。展望 4Q26，我们建议关注三大催化，有望推动公司估值修复：1）9 月 MiniMax Code 上线 M3.1-Flash-Preview 模

型，基于 4300 亿参数的 M3 模型底座，通过后训练在工程开发、金融分析、视觉创作等场景下实现模型性能的提升，后续正式版模型上线有望推动 ARR 增长加速；2）M3 Pro 模型计划 4Q26 上线，模型参数规模扩大至 3 万亿，同时借助架构创新以保持成本与速度优势；3）H3.1 视频模型基于用户社区反馈进一步强化视频生成能力。我们的目标价为 580.0 港元，基于 50 倍 FY26E PS。“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之分析员，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于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 个月内与本报告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定。

本报告包含的任何信息由招银国际环球市场按照香港证券及期货事务监察委员会对其获发牌进行受规管活动的监管框架，编写和在香港发布报告。本报告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定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着区别，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，任何机构或个人未经本公司书面授权的情况下，不得以任何形式、以任何目的，直接或间接向特定读者以外的人士翻版、复制、转售、转发及或发布本报告的全部或部分内容，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，请与我们联络。

若本重要披露的中英文版本存在歧义，概以英文版本为准。

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之分析员，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个人。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篇制的报告。 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。