2026 年 4 月 15 日

招银国际环球市场 | 睿智投资 | 行业研究

AI 主题研究

1Q26 大模型进展复盘：关注 Agent 化及 B 端突围

优于大市
（维持）

复盘 1Q26 AI 大模型行业最新进展，我们观察到：1）行业竞争加剧、模型迭代速度持续加速，行业竞争的重点从基础模型能力进一步向推理效率、智能体能力等方向衍生。过去一年，中美语言模型差距有所收窄，同时在视频等多模态模型中部分中国厂商表现亮眼。2）对比头部模型厂商，Anthropic 和 OpenAI 目前在收入体量方面仍维持全球领先。3）国内模型厂商目前估值驱动关键或已从单纯“模型能力排名”演进为关注“收入兑现速度与路径可持续性”。我们对国产 AI 原生板块维持积极判断，但同时建议投资者重点关注各家公司 2026-2030 年收入增长趋势，这将成为决定当前估值能否被基本面消化的关键因素。4）AI 商业化方面，我们持续看好 AI 编程、企业智能体与 AI 创意生成领域，三条赛道目前已探索出较为清晰的商业模式并形成规模化收入。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）。美国互联网软件板块推荐谷歌（GOOG US）、微软（MSFT US）和 Meta（META US）。

- **大模型发展：迭代竞争加速，中美差距有所收窄。**2026 年以来大模型行业竞争进一步加剧，中美顶尖语言模型智能水平差距由 25 年 3 月的约 20 分区间收窄至 26 年 4 月的个位数差距（据 Artificial Intelligence 模型智能水平评分）。整体来看，当前全球大语言模型行业演进主要呈现三大趋势：1）大模型迭代节奏进一步加快；2）推理效率成为更加关键的竞争维度；3）智能体能力成为大模型行业演进的核心主线之一，模型能力强调工具调用、编程与多模态能力，从而打造能解决真实世界问题的智能体。其他模态方面，图像模型竞争格局上仍由美国厂商主导，而视频模型竞争更趋多极化，中国厂商表现更优且已在部分前沿产品上领先。

- **海外 AI 模型平台：商业化竞争白热化。**1Q26 海外 AI 平台竞争格局出现显著分化：1）收入体量看，1Q26 OpenAI ARR 仍居首位，但 Anthropic 于 2026 年 4 月公布的 300 亿美金 ARR 体量已超过 OpenAI 在 26 年 2 月的 ARR 水平（据 Sacra 数据：250 亿美元），xAI 独立 AI 业务收入规模则仍偏小。2）收入增速看，Anthropic 维持快速增长，三个月内实现 ARR 翻倍，而 OpenAI 营收维持较快增长，增长驱动主要来自付费渗透率提升。3）盈利方面，Anthropic 预计最早实现盈利。展望 2H26，三家头部厂商核心关注指标主要包括：Anthropic 与 OpenAI 的 ARR 体量及增速，OpenAI 广告与 Codex 等新产品的商业化进展，及 xAI 在与 SpaceX 合并后的协同效应能否转化为可量化的收入增量。

- **国内 AI 模型平台：进入商业化验证期。**我们认为当前国内 AI 平台的定价变量已从更偏重追求“模型能力排名”转向“收入兑现速度与路径可持续性”。我们对国产 AI 原生板块维持积极判断，但建议投资者重点关注各家公司 2026-2030 年收入绝对值与增长趋势，这将是检验当前估值能否被基本面消化的核心窗口。建议持续关注：1）模型能力提升及新的商业化突破（即模型提价及新商业化模式探索）；2）各公司模型层面的差异化优势及维持情况；3）算力供给瓶颈及地缘因素对变现节奏带来的扰动。

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. 美国软件 NDR 要点总结：关注客户端增量价值创造 – 13 Mar 2026
2. 海外龙头 TMT 公司业绩启示 - 资本开支进一步上行，半导体表现占优趋势或延续 – 23 Mar 2026
3. Salesforce (CRM US) - Inline 4QFY26 results; strong AgentForce momentum to support 2HFY27 reacceleration – 27 Feb 2026
4. Palo Alto Networks (PANW US) - Results beat; AI-related security brings new waves of growth opportunity – 20 Feb 2026
5. Coinbase (COIN US) - 4Q25 results: navigate short-term headwinds amid soft crypto market sentiment – 16 Feb 2026
6. Datadog (DDOG US) - Robust usage growth to drive solid revenue growth outlook – 11 Feb 2026
7. Amazon (AMZN US) - Growth story unchanged with acceleration in AWS revenue growth better than expectation - 9 Feb 2026
8. Alphabet (GOOG US) - 4Q25 results: AI continues to drive strong search and cloud business performance - 5 Feb 2026
9. ServiceNow (NOW US) - 4Q25 results: solid AI business and margin expansion - 30 Jan 2026
10. Microsoft (MSFT US) - Results beat; long-term structural growth story remains unchanged - 30 Jan 2026
11. Meta (META US) - 4Q25 results beat; AI continues to drive ad business growth - 29 Jan 2026
12. 中国互联网 - AI 应用商业化快速落地 – 14 Jan 2026
13. 中国互联网 - 2026 展望：承前启后，关键之年 – 9 Dec 2025
14. 美国互联网与软件 - 2026 展望：应用持续起量，关注投资回报周期 – 9 Dec 2025

- **AI 应用商业进展：关注 AI 编程、智能体与创意生成赛道。**我们重点对比目前商业化相对领先的 AI 应用赛道：1）编程是当前 AI 应用中商业化确定性高、十亿美元级产品较多的赛道，包括 Claude Code， Cursor 等，开发者对 AI 编程工具的付费意愿和用量表现强劲。2）企业 AI 智能体商业化正从概念验证切入规模放量，Salesforce 4QFY26 Agentforce ARR 达 8 亿美元，ServiceNow 4Q25 Now Assist ACV 突破 6 亿美元。3）AI 创意生成赛道，头部产品早期商业化进展积极，但行业竞争逐渐加剧，头部厂商之间尚未拉开显著模型能力或用户差距，仍需持续投入研发和算力以维持领先。

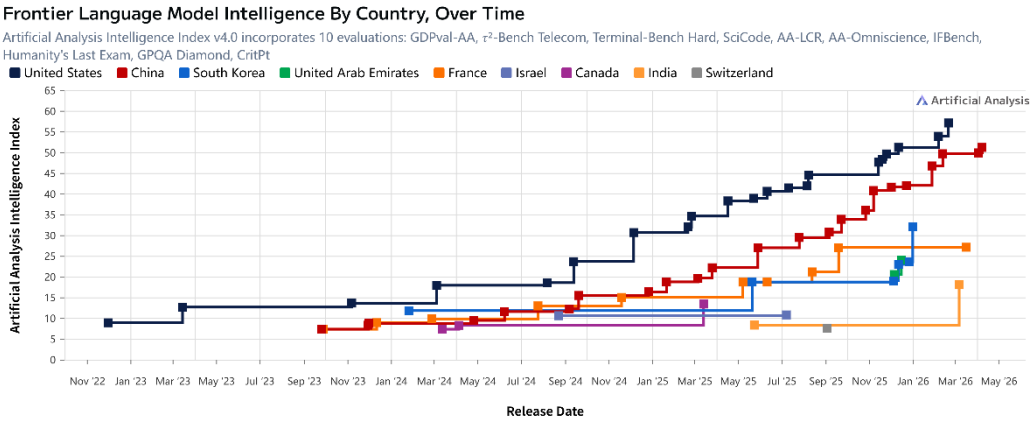
目录

大模型行业发展趋势：迭代持续加速，中美差距有所收窄

2026 年以来，大模型行业竞争进一步加剧，中美前沿模型能力差距持续缩窄。我们基于 Artificial Analysis 的数据测算，中国前沿语言模型智能水平已由 2025 年 3 月的约 20 分区间提升至 2026 年 4 月的约 50 分附近，而同期美国由约 35 分左右提升至接近 60 分，中美顶尖模型能力差距由此前约 10–15 分进一步收窄至 6 分（Google Gemini 3.1 Pro Preview 对比智谱 GLM-5.1）。整体来看，当前大语言模型行业演进主要呈现三大趋势：1）大模型迭代节奏进一步加快；2）推理效率正成为新的关键竞争维度；3）智能体成为大模型行业演进的核心主线之一，模型能力强调工具调用、编程与多模态能力，从而打造能解决真实世界问题的智能体。图像模型竞争格局上仍由美国厂商主导，视频模型竞争更趋多极化，中国厂商表现更强并已在部分前沿产品上领先。

从模型综合能力看，Google 在多个维度整体占优。分项来看，语言模型智能能力方面，国际厂商中以 Google Gemini 3.1 Pro Preview 领先，国内则以智谱 GLM-5.1 为代表；推理效率方面，国际 Google Gemini 3 flash，国内 DeepSeek V3.2、Minimax M2.7、Mimo-V2-Pro、Kimi K2.5 均进入最优象限；Agent 能力方面，国际第一梯队主要包括 OpenAI GPT 5.4 与 Anthropic Claude Opus 4.6，国内则智谱 GLM-5.1 和小米 MiMo-V2-Pro 领跑；图像生成方面，国际头部模型为 OpenAI、Google 和 Black Forest Labs，国内字节跳动表现较为突出；视频生成方面，国内厂商快手、生数科技和爱诗科技构成第一梯队领跑行业，海外厂商以 xAI 表现突出。

图 1: 各国前沿大模型的发布时间及能力对比



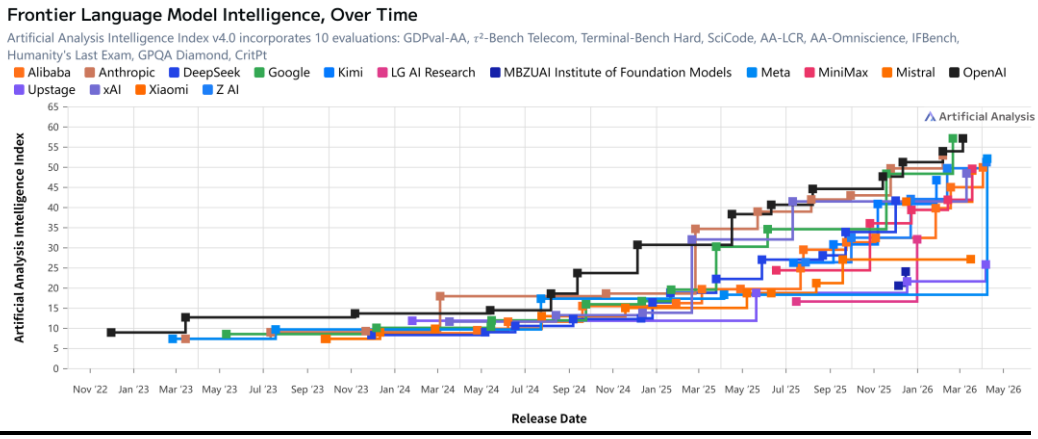
资料来源：Artificial Analysis, 招银国际环球市场

注：1）Artificial Analysis 智能能力指数涵盖 10 项评估维度，主要包括：GDPval-AA, τ2-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 4 月 8 日，除单独备注外，下同

大语言模型：模型能力高速迭代，Google 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 4 月 8 日，当前行业综合能力排名第一的模型为 Google Gemini 3.1 Pro Preview，但其领先幅度已较为有限，反映出前沿模型之间的能力差距正在持续收敛，行业竞争格局较为胶着。

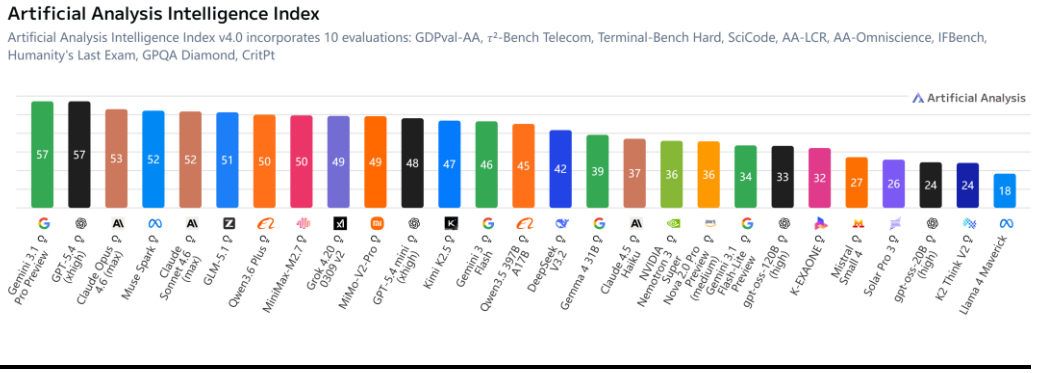
图 2: 主流大语言模型的模型能力对比



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 Google Gemini 3.1 Pro Preview、OpenAI GPT 5.4、Anthropic Claude Opus 4.6 等处于行业领先水平，而中国厂商中智谱的 Z.AI GLM-5.1、阿里巴巴的 Qwen3.6 Plus、MiniMax-M2.7、小米的 MiMo-V2-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

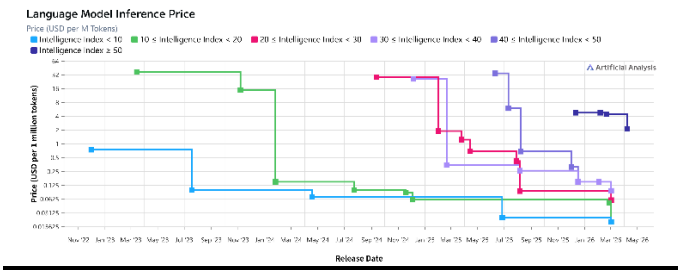


资料来源：Artificial Analysis, 招银国际环球市场

推理效率：行业竞争热度提升，小米与 Google 位于效率最优象限前列

根据 Artificial Analysis, 语言模型行业一方面推理成本持续降低，一方面输出速度持续提升。中高智能区间模型（智能能力指数 20-40 为例）的推理价格已由 2025 年 3 月的约 0.25–1 美元/百万 tokens 快速下探至 2026 年 3 月的约 0.0625–0.125 美元/百万 tokens，对应降幅约 75%–87.5%；与此同时，输出速度由约 150–200 tokens/秒 提升至 300 tokens/秒以上，对应增幅约 50%–100%以上。与此同时，我们观察到高智能能力的模型仍可持续维持较高定价甚至进一步提价。

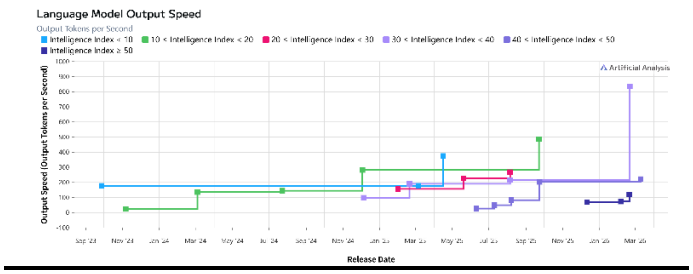
图 4: 语言模型推理成本



资料来源：Artificial Analysis, 招银国际环球市场

注：推理成本=输入/输出 token 价格按 3:1 权重加权后的每百万 tokens 成本

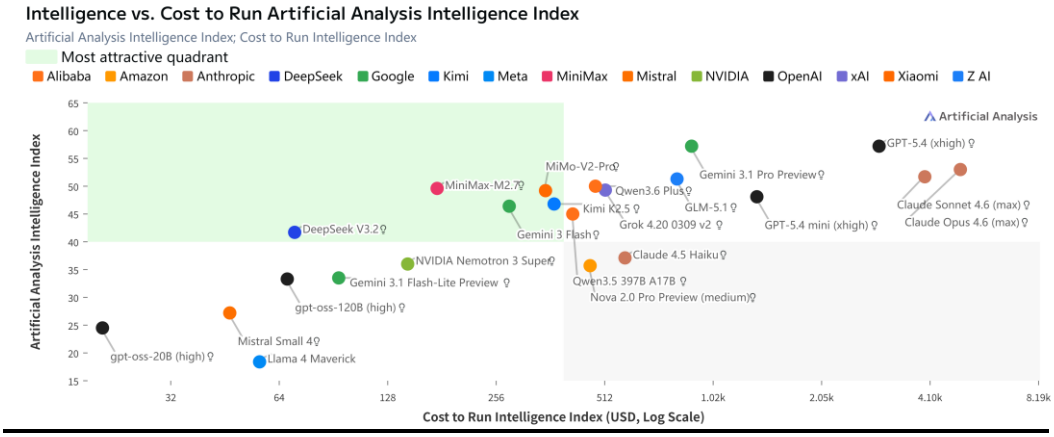
图 5: 语言模型输出速度



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，Flash 版本的 Google Gemini 3 Flash；非 Flash 版本的 DeepSeek V3.2、MiniMax M2.7、小米 MiMo-V2-Pro、Kimi K2.5 处于最具吸引力的象限。其中 MiniMax 2.7、MiMo-V2-Pro、Gemini 3 Flash 为智能能力范围接近情况下成本更优的模型，MiMo-V2-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 6: 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

行业最新进展：编程与多模态能力持续优化，强化工作流执行

截至 2026 年 3 月末，横向对比 12 家中美前沿大模型最新版本进展，我们观察到，以真实工作流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。具体来看，能力演进主要呈现三大趋势：1）持续强化工具调用能力，即围绕真实任务目标主动选择并调用外部工具，传递执行现实场景动作所需参数，例如 Google Gemini 3.1 Pro 已进一步优化面向定制化工具调用的接口能力；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 MiniMax M2.7 已可覆盖日志分析与缺陷定位、代码重构、代码安全、机器学习及安卓开发等真实研发任务；3）持续强化原生多模态理解与执行能力，即统一理解文本、图片、视频、文档等多种信息形态，并进行跨模态推理与任务执行，例如阿里巴巴千问 3.5 明确提出“原生多模态智能体”方向。

图 7：前沿大模型最新进展及参数

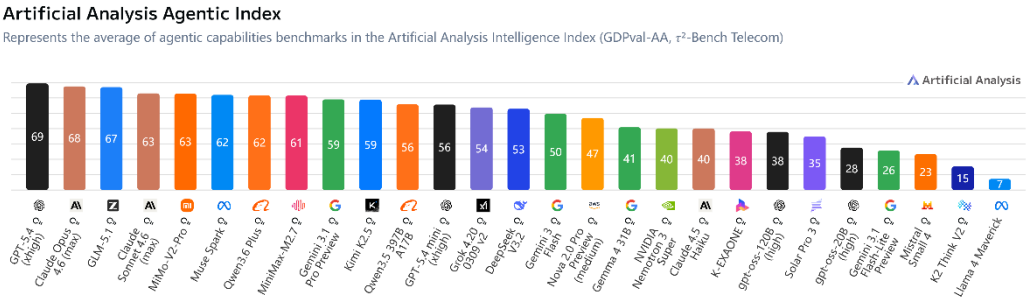
厂商	最新公开主模型	最新进展	较上一代版本的核心优化	更注重的功能方向	总参数/激活参数
Google	Gemini 3.1 Pro	2026 年 2 月 19 日发布，是 Gemini 3 Pro 的升级版，支持 1M token 输入、64k 输出。	相比 Gemini 3 Pro 系列，3.1 Pro 在 token 效率、事实一致性以及工具调用可靠性上进一步提升。	更偏向**复杂推理、智能体编程（agentic coding）、工具调用、长上下文多模态理解。**	未披露
OpenAI	GPT-5.4	2026 年 3 月 5 日发布，为 ChatGPT/API/Codex 的最新前沿通用模型；首次把 GPT-5.3-Codex 的前沿 coding 能力整合进主线模型，具备原生且顶尖计算机使用能力，支持高达 **1M token 上下文**。	对比 **GPT-5.2，**更强的专业知识工作、更强计算机使用能力、视觉理解与工具调用更高 token 效率；同时吸收了 5.3-Codex 的编程长板。	更偏向 **知识工作（文档/表格/PPT）、智能体、计算机使用、网页搜索、工具调用。**	未披露
Anthropic	Claude Opus 4.6	2026 年 2 月 5 日发布，是 Anthropic 最新 Opus 旗舰；首次给 Opus 级模型提供 **1M token context（beta）。**	相比 **Opus 4.5**，官方强调编程**更强、规划更谨慎、能维持更长智能体任务（agentic task）、在更大代码库中更可靠、代码审查与 debug 更强**。	更偏向智能体编程、**深度检索、专业知识工作、超长任务持续执行。**	未披露
xAI	Grok 4.20 Beta	2026 年 3 月 10 日发布，xAI 同步推出 **Grok 4.20 Beta** 与 **Grok 4.20 Multi-agent Beta**	**更强调智能体工具调用、严格遵循 prompt、降低幻觉，并推出多 Multi-agent Beta**	更偏向**工具调用、智能体执行、多工具协作。**	671B/37B
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。	相较此前以 Llama 系列为代表的公开主线，Muse Spark 被 Meta 定义为**新 Muse 路线下“从底层重建 AI stack”的首发模型。**	更偏向面向消费者的通用助手、多模态理解、健康信息处理、跨产品部署，以及服务大规模真实用户的产品化 AI 能力。	未披露
智谱 Z.AI	GLM-5.1	2026 年 3 月 27 日发布，聚焦通过强化后训练提升**编码与推理能力**，专为**更长时程**的智能体任务设计。	沿用 GLM-5 基础架构（744B 总参数/约 40B 激活参数，MoE，200K 上下文）。提升完全来自**后训练优化**（progressive alignment：多任务 SFT→多阶段 RL→跨阶段蒸馏），底座模型未变。	**编程 Agent 场景针对性强化。**	744B/40B
MiniMax	MiniMax M2.7	2026 年 3 月 18 日发布，MiniMax 将其定义为**首个“深度参与自身演进”的模型**，强调其可独立构建复杂智能体执行框架，并完成高复杂度生产力任务。	相比 M2.5，官方重点强化软件工程能力、办公生产力以及复杂环境交互能力，并提出模型参与自身学习流程与执行框架迭代的“自我演进”能力框架。	更偏向 **工作流执行 + 编程/工程型智能体**，尤其是软件工程、Office 多轮编辑、复杂工具环境交互。	官方暂未披露 M2.7 参数，已披露 M2.1 参数为 230B/10B
小米	MiMo-V2-Pro	2026 年 3 月 18 日正式发布，官方定位为最新旗舰；其早期匿名测试版“Hunter Alpha”在 OpenRouter 累计调用已超过 **1T tokens**。	相比前代 MiMo-V2-Flash，**模型参数规模扩展至 1T+/42B，约为前代的 3 倍**，并将**混合注意力机制的比例从 5:1 提升至 7:1**；同时，通过覆盖更广泛智能体任务的后训练，推动模型能力由对话式交互进一步走向智能体执行，更强调工具调用、多步推理与真实工作流执行。	侧重于**智能体场景下的任务处理、工具调用稳定性、多步推理以及长上下文真实任务执行。**	1T+/42B
阶跃星辰 StepFun	Step 3.5 Flash	2026 年 2 月 12 日发布，是其最新旗舰推理/Agent 开源基座模型；196B MoE、每 token 仅激活 **11B** 参数。	官方更强调“效率型跃迁”而不是对旧版逐项对标：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。	更偏向**高复杂度推理、**编程、智能体工具使用、搜索、办公任务。	196B/11B
月之暗面	Kimi K2.5	2026 年 1 月 27 日正式发布，为原生多模态、开源的 Kimi 新一代模型；从单 Agent 扩展到 **Agent Swarm。**	**K2.5** 引入更强的视觉编程与原生视觉理解能力，在约 15T 图文混合 token 上继续预训练，并面向并行化、生产级真实工作流执行。	更偏向**视觉编程、图/视频到代码、办公生产力、多智能体并行执行。**	1T/32B

幻方量化 DeepSeek	DeepSeek-V3.2	最新公开正式版为 2025 年 12 月 1 日发布的 **V3.2**；官方将其定义为“**面向智能体的推理优先模型**”，并同步推出更强推理版 V3.2-Speciale。	相比 **V3.1**，重点不只是答题能力，而是把思考**直接整合进工具使用。**	更偏向智能体、**思考与工具调用一体化、极限推理（Speciale）。**	671B/37B
字节跳动豆包	Seed2.0 Pro / Seed2.0 Code	字节在 2026 年 2 月正式发布 **Seed2.0** 系列，并明确写明 **Seed2.0 Pro 和 Code 已上线豆包 App 与 TRAE。**官方把这轮升级定义为“全面升级至多模态理解能力”。	相比 **Seed1.8**，官方明确强调在**视觉谜题、逻辑推理、非结构化文档理解、长上下文理解**上显著增强。	更偏向**多模态理解、视觉推理、文档理解、长上下文、科研级推理。**	Seed2.0 未披露参数，已披露 Seed1.6 参数为 230B/23B
阿里巴巴千问	Qwen3.5	阿里在 2026 年 2 月发布 **Qwen3.5，**官方定义为“转向原生多模态智能体”(Towards Native Multimodal Agents)。	官方将 Qwen3.5 的升级概括为四个方向：多模态学习、架构效率、强化学习规模与全球可及性，同时指出 Qwen3.5 在推理、编程、智能体及视觉理解等基准测试上表现优于 Qwen3-VL。在架构层面，Qwen3.5 采用了高效混合架构，并结合门控增量网络与稀疏混合专家模型，以提升推理吞吐、降低时延与成本开销。	更偏向**原生多模态、工具调用、深度研究、Web 开发、智能体工作流、长上下文知识任务。**	Qwen3.5-397B-A17B 参数为 397B/17B

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

从大模型智能体能力评分来看，基于 Artificial Analysis 数据，截至 2026 年 4 月 8 日，美国厂商中 Open AI GPT 5.4、Anthropic Claude Opus 4.6、Anthropic Claude Sonnet 4.6 处于行业领先水平，而中国厂商中智谱的 Z.AI GLM-5.1、小米的 MiMo-V2-Pro、阿里巴巴的 Qwen3.6 Plus 智能体能力处于行业领先水平。

图 8: 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，τ²-Bench Telecom）；2）排名并未穷尽所有大模型

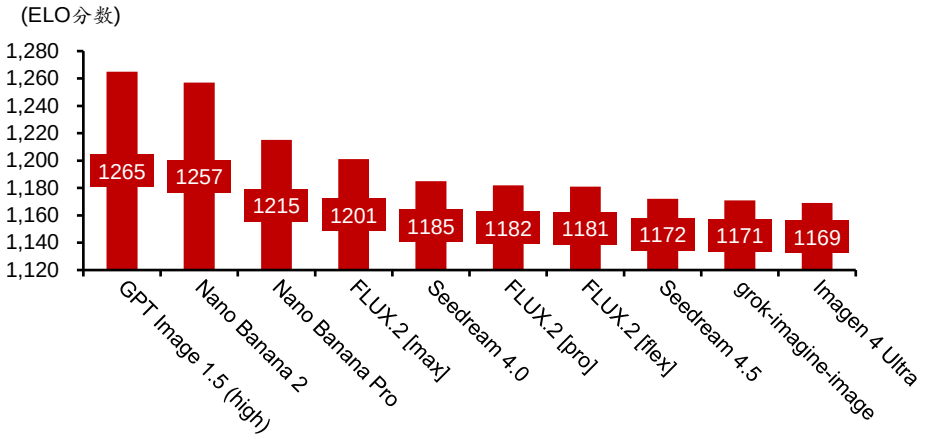
图像模型仍由美国领跑，但视频赛道中国厂商更具优势

我们观察到图像模型的功能演进也同样转向工作流执行，正由早期的“生成质量提升”转向：1）生成与编辑一体化，持续迭代的工作流工具，例如 OpenAI 的 GPT Image 1.5 强调精准编辑并保留重要细节，Google 的 Nano Banana 2 主打快速编辑与反复修改；2）强调文本渲染与排版精准度，例如 Google Nano Banana Pro 突出高精度文本渲染能力。

图像模型方面美国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 1.5 (high)、Google Nano Banana 2/Pro、Black Forest Labs FLUX.2 [max] 图像模型能力领先行业，中

国厂商则以字节跳动 Seedream 4.0 领先。此外，从图像模型的性价比来看（质量 vs. API 价格），字节跳动的 Seedream 4.0 与 Black Forest Labs 的 FLUX.2 [max] 最接近最优象限，处于相对高质量、低价格区间。

图 9: 主流图像模型文生图能力排名



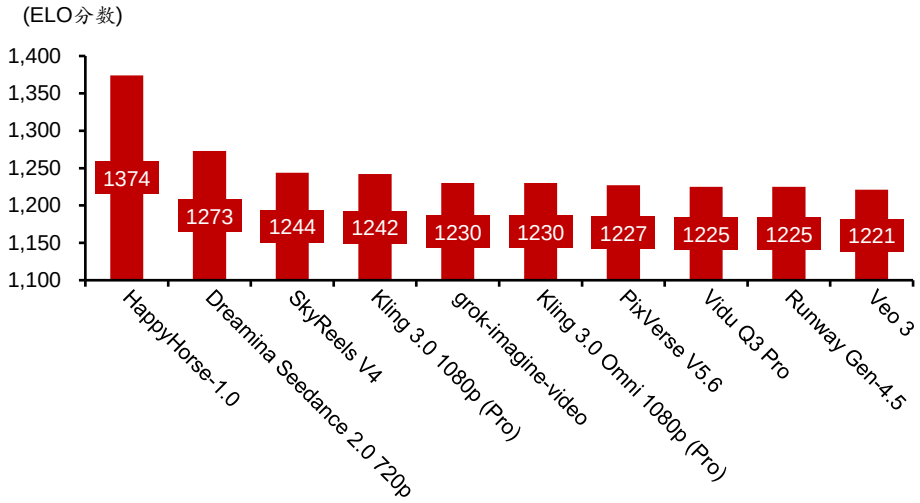
资料来源：Artificial Analysis, 招银国际环球市场

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）强调跨境头一致性，例如 Google Veo 3.1 强化了一致性与可控性，Runway Gen-4 强调不同视角下保持风格连贯；2）原生音视频一体化，快手可灵 3.0 及字节 Seedance 1.5 Pro 均强化原生视听内容生成；3）多镜头叙事，字节跳动 Seedance 2.0 进一步强化了导演级控制能力，可对表演、光线、阴影及镜头运动进行控制。

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 2025 年 9 月，根据 QuestMobile 数据，即梦 AI MAU 规模超 1000 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 no audio），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、昆仑万维 SkyReels V4、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、Google Veo 3 领先。从性价比来看（质量 vs. API 价格），xAI Grok-imagine-video 为最优象限中性价比最高的模型。

图 10: 主流视频模型文生视频能力排名

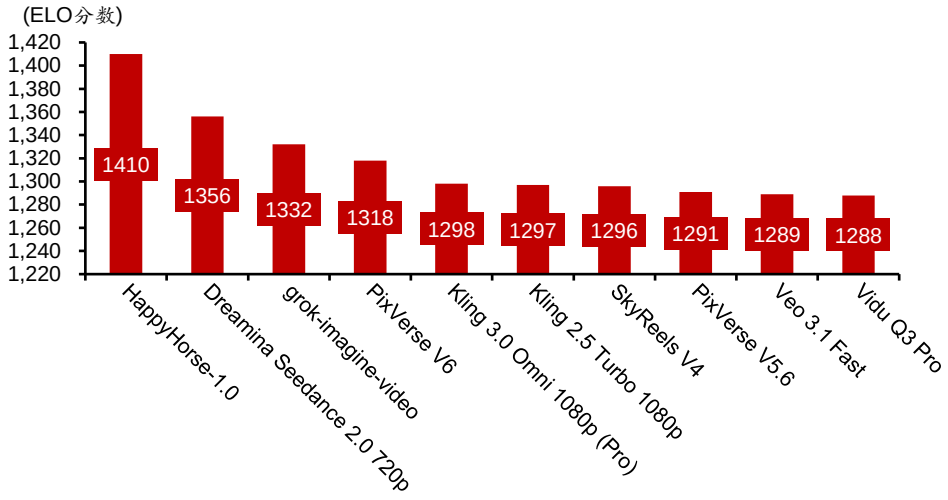


资料来源：Artificial Analysis, 招银国际环球市场

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，HappyHorse-1.0 与字节跳动即梦 Seedance 2.0 720p 位居行业前二，美国厂商以 xAI Gro-imagine-video 领先，其后多个中国厂商包含快手可灵、昆仑万维 SkyReels 及生数科技 Vidu 形成行业第一梯队。

图 11: 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：商业化快速推进

海外 AI 基础模型平台：ARR 竞赛白热化，Anthropic 增长迅速

1Q26 海外 AI 平台竞争格局出现显著分化，对比 OpenAI、Anthropic、xAI：1）**从收入体量看，Anthropic 和 OpenAI 维持行业领先水平：**Anthropic 2026 年 4 月公布其 ARR 已经超过 300 亿美元；据 Sacra 估算，OpenAI 2026 年 2 月 ARR 约 250 亿美元；Anthropic 紧随其后，据彭博社报道及 CEO Dario Amodei 确认，ARR 已于 2026 年 3 月初达到约 190 亿美元；xAI 独立 AI 业务收入规模则仍显著偏小，据 Sacra 估算，2025 年年化收入约 5 亿美元，2026 年预计增至 20 亿美元以上。**2）从增长斜率看，Anthropic 表现突出**：公司在年初至今 4 个月内 ARR 上涨超 200%，同期 Claude 付费订阅用户数亦实现翻倍（据 Forbes）。OpenAI 用户规模依旧领跑全球（据公司披露，2026 年 2 月 ChatGPT 周活用户达 9 亿），营收增长主要来自产品付费渗透的提升。据公司 4 月 8 日披露，当前企业收入占比超 40%，并有望在 2026 年末与 C 端收入占比齐平，B2B、平台业务及新产品（包括硬件）的营收贡献起量将是支撑中期营收增长的关键所在。**3）Anthropic 预计最早实现盈利：**据 The Information 报道，Anthropic 预计最早于 2027 年实现正现金流，毛利率亦有望从 2024 年的-94%改善至 2028 年的 77%——在三家公司中，目前其盈利兑现时间表最为清晰、经营杠杆释放最为可见。相比之下，据 The Information 报道，OpenAI 最新预测仅 2026 年一年现金消耗即可能达约 250 亿美元，2027 年进一步飙升至 570 亿美元，现金流预计在 2030 年前均难以转正，估值支撑来自 B 端及新业务营收起量预期及实现 AGI 的期待。xAI 方面，据彭博社报道，其截至 2025 年 6 月单月总成本已高达约 10 亿美元，管理层提出的 2027 年盈利目标兑现路径仍有较大不确定性。

我们认为，在三家公司均尚未盈利的背景下，2026 年下半年最关键的校验指标包括：ARR 绝对值体量及增速、OpenAI 广告与 Codex 等新产品的商业化进展、以及 xAI 与 SpaceX 合并后的协同效应能否转化为可量化的收入增量。

- **Anthropic：ARR 突破 300 亿美元，Claude Code + Cowork 双引擎驱动商业化效率领跑全球**

我们认为，Anthropic 正成为本轮基础模型竞争中增长最快、商业化效率最高、盈利兑现路径最清晰的公司之一。据公司披露，Anthropic ARR 已从 2025 年底的约 90 亿美元快速提升至 2026 年 4 月初的约 300 亿美元，不到四个月上涨超 200%；据 Forbes 2026 年 3 月报道，Claude 免费活跃用户数自 2026 年初以来增长了 60%，Claude Pro 和 Max 付费计划的订阅用户数翻了一番。

拆分来看，Anthropic 当前的增长主要由三大产品层共同驱动。1）**Claude Code 是现阶段用户采用度提升较快的产品线：**据 Anthropic 官方公告，Claude Code 自 2025 年 5 月公开发布后，仅 6 个月便于 2025 年 11 月达到 10 亿美元 ARR；据 Sacra 报道，至 2026 年 2 月 ARR 已进一步翻倍至 25 亿美元以上。据 Anthropic 披露，开发者平均日均支出约 6 美元，高频用户（使用量前 10%）日均超过 12 美元，显著高于传统 SaaS 开发工具 ARPU。据 Semi Analysis 报道，截至 2026 年 2 月，全球约 4%的公开 GitHub commits 已由 Claude Code 生成，预计年底有望超过 20%。据 Fortune 援引 Anthropic 发言人，在 Anthropic 内部，70–90%的代码由 Claude Code 产出；Claude Code 负责人 Boris Cherny 亦表示，其自身代码库的 90%由该工具完成，显示出较强的产品粘性与自我强化效应。2）**Cowork：面向通用办公场景的 AI Agent 产品。**Cowork 由 Anthropic 于 2026 年 1 月推出，定位为 "Claude Code for general computing"，聚焦电子表格、文件管理、报告起草及自动化工作流等场景。产品发布首月即推出 30 余项产品与功能，并配套 11 个开源插件，覆盖销售、法务、财务等多类岗位，目前已可直接运行于 Excel 和 PowerPoint 中，并支持按时间表自

动执行循环任务。3）**Opus 4.6 能力升级：夯实企业复杂任务与 Agent 协同的竞争优势。**Anthropic 于 2026 年 2 月更新 Opus 4.6，将上下文窗口由 20 万 token 大幅提升至 100 万 token，并内置 Agent Teams 功能，允许多个 Claude 实例协同执行复杂任务。API 定价为输入 5 美元/百万 token、输出 25 美元/百万 token，扩展上下文溢价至 10 美元/37.5 美元，为后续企业级调用和高复杂度任务变现打开更大空间。4）**Claude Mythos 首发：模型能力跃升，网络安全能力显著增强。**Anthropic 于 2026 年 4 月发布 Claude Mythos Preview，模型能力相较 Opus 4.6 实现明显提升，在编程、推理、人类最后考试、智能体任务等多个主流 AI 基准测试中的评分大幅高于 GPT-5.4、Gemini 3.1 Pro 等 SOTA 模型（SWE-bench Pro：Mythos Preview/Opus 4.6/GPT-5.4/Gemini 3.1 Pro 评分 77.8%/53.4%/57.7%/54.2%；GPQA Diamond：Mythos Preview/Opus 4.6/GPT-5.4/Gemini 3.1 Pro 评分 94.6%/91.3%/92.8%/94.3%）。此外，Mythos Preview 在网络安全能力方面显著增强，在 CyberGym 定向漏洞复现测试中，Mythos Preview 得分达 83.1%（vs Opus 4.6 得分 66.6%），并且在 OpenBSD（全球加固程度最高的操作系统之一）中发现了存在了 27 年的零日漏洞，基于其网络安全能力，Anthropic 提出 Project Glasswing，联合超 40 家关键软件基础设施组织，合作发现软件漏洞以保护全球关键软件基础设施。

从中长期**经营质量**看，Anthropic 的营收增长路径较为明确，变现主要来自付费意愿较高的 B 端客户对其价值认可而带来的增量付费渗透。据 The Information 报道，公司预计最早于 2027 年实现正现金流，2028 年实现约 170 亿美元现金流；毛利率也有望从 2024 年的-94%改善至 2025 年目标的 40%，并在 2028 年进一步提升至 77%，显示公司在收入快速放量的同时，经营杠杆和成本结构也在持续优化。**估值**层面，Anthropic 于 2026 年 2 月 12 日完成 300 亿美元 Series G 融资，投后估值达 3800 亿美元；若按融资时对应的约 140 亿美元 ARR 测算，估值约为 27x ARR，而若按公司披露的 2026 年 4 月 ARR 300 亿美元测算，则已降至约 12.7x ARR，收入快速增长助力估值消化。**上市进程**层面，据彭博报道，Anthropic 正考虑最早于 2026 年 10 月进行 IPO，融资规模或超 600 亿美元。

但与此同时，Anthropic 当前面临部分政策与环境的风险值得关注。在拒绝将 AI 用于大规模监控和自主武器后，Anthropic 被美国国防部认定为对美国供应链构成风险，特朗普政府亦于 2026 年 2 月 27 日下令联邦机构和军事承包商暂停与 Anthropic 往来。据 Anthropic 法庭文件披露，已有超过 100 家企业客户就此事与公司沟通，公司预计 2026 年相关政府不利行动可能带来数亿美元乃至数十亿美元的收入损失。

■ OpenAI：ChatGPT 周活用户达 9 亿，平台化优势较强

OpenAI 目前或仍是全球 AI 行业中用户规模最大、品牌心智最强、商业化天花板最高的平台型公司。据 Sacra，截至 2026 年 2 月，公司 ARR 已达到约 250 亿美元，较 2025 年底 CFO Sarah Friar 披露的超 200 亿美元进一步增长，过去三年收入实现约 10 倍扩张。用户侧，据 OpenAI 披露，截至 2026 年 2 月，ChatGPT 周活跃用户已达 9 亿，较 2025 年 10 月的 8 亿进一步增长；付费企业用户亦已超过 900 万，较 2025 年 8 月的 500 万明显跃升，反映其无论在消费者端还是企业端，仍具备最强的流量入口和分发能力。

拆分**收入结构**看，OpenAI 当前增长主要由 ChatGPT 订阅、Codex、广告与 API/企业服务共同驱动。1）**ChatGPT 订阅：核心收入来源，付费渗透率或仍有提升空间。**当前产品层级包括 Go（8 美元/月）、Plus（20 美元/月）及 Pro（200 美元/月）。据 The Information 披露，截至 2025 年 7 月，仅约 5%的用户选择付费订阅，未来付费渗透率或仍有进一步提升空间。2）**Codex：企业增长的新引擎。**据公司 2026 年 4 月 8 日披露，Codex 周活用户已突破 300 万，API 每分钟处理超过 150 亿 token；Fortune 报道 Cisco、Nvidia、Ramp、Rakuten 及 Harvey 等企业均已在开发团队中全面部署 Codex。据 CEO Sam Altman 于 2026 年 1 月披露，API 业务单月新增 ARR 已超过 10 亿美元，显示 OpenAI 正在基于其此

前积累的企业资源池进行较为迅速的交叉销售扩展。3）**广告业务：免费用户变现的第二增长曲线。**OpenAI 广告业务于 2026 年 2 月正式上线，面向美国 Free 和 Go 层级用户测试投放，初始 CPM 约 60 美元，最低投放门槛 20 万美元，首批广告主包括 Target、Ford、Adobe、Best Buy、AT&T 和 Expedia。

然而，目前 OpenAI 成本端亦面临较为迅速的增长趋势。据 The Information 报道，OpenAI 最新预测仅 2026 年一年现金消耗即可能达约 250 亿美元，2027 年进一步飙升至 570 亿美元。成本端压力主要来自推理费用持续上升：据 The Information，2025 年推理成本同比增长 4 倍至 84 亿美元，而 Sacra 预测 2026 年将进一步升至 141 亿美元，或部分意味着在当前阶段，收入增长带动成本节省的规模效应仍不够显著。

公司在基础设施端已开始为中长期成本优化提前布局。截至 2026 年初，OpenAI 已签署超过 5000 亿美元的云计算容量承诺，涵盖微软 Azure 约 2500 亿美元、AWS 扩展合作以及 Oracle 约 3000 亿美元/5 年的协议。我们认为，多云策略一方面有助于降低对单一供应商的依赖，另一方面也为后续推理成本谈判和资源调度优化提供了更大可能性。

估值层面，OpenAI 于 2026 年 3 月 31 日宣布完成 1220 亿美元融资；这将成为历史上最大规模私募融资之一，投资方包括 SoftBank（300 亿美元）、Nvidia（300 亿美元）、Amazon（500 亿美元）和 Andreessen Horowitz 等投资机构（120 亿美元），投后估值达 8520 亿美元。按 Sacra 估算的约 250 亿美元 ARR 计算，最新融资对应投后估值倍数约 34x，较 Anthropic 仍享有估值溢价。据路透社报道，公司最早可能于 2027 年以 1 万亿美元估值上市。整体而言，我们认为，OpenAI 在用户规模、品牌影响力和多元商业化能力上仍处于全球领先地位，但相较 Anthropic 更早可见的现金流拐点，目前 OpenAI 的现金流拐点能见度或相对而言更低，其估值支撑更多来自平台垄断式入口价值、远期利润兑现预期、及市场对 AGI 的期待。

■ **xAI（Grok）：市场对远期生态联动给出较高期待**

xAI 是当前龙头 AI 公司中市场估值相对较高的标的，除营收体量目前仍相对更小的原因外，市场对其的估值认可或也建立在马斯克生态整合、平台协同和远期业务发展的想象空间之上，而非当前独立 AI 业务的收入与盈利能力。xAI 成立于 2023 年，旗舰产品为 Grok 聊天机器人；此后资本运作节奏显著加快——先是于 2025 年 3 月通过全股票交易收购 X（原 Twitter），随后又于 2026 年 2 月被 SpaceX 以全股票方式收购，最终被纳入更大的马斯克资产版图之中。

从合并前的独立收入结构来看，xAI 的商业化主要来自消费订阅、企业 API 以及政府合同三条主线。1）**消费订阅：依托 X 流量体系的最直接变现路径。**当前产品矩阵已形成分层体系，包括免费版、X Premium（8 美元/月）、X Premium+（40 美元/月，含 Grok 4 访问及免广告）、SuperGrok（30 美元/月，独立订阅，含 DeepSearch 和高级推理）以及 SuperGrok Heavy（300 美元/月，提供 Grok 4 Heavy 预览及最高 8K token 记忆）。据 X 产品负责人 Nikita Bier 披露，X 平台订阅业务于 2026 年 2 月达到约 10 亿美元 ARR。用户侧，据 xAI 于 2025 年 9 月披露，Grok 全球月活用户约 6400 万，较 2025 年 4 月（据 Famewall 数据为 3510 万）增长约 80%，表明其在 X 流量体系加持下已具备一定的消费级分发能力。2）**企业 API：开始成为新增量来源。**企业订阅方面，xAI 于 2025 年 12 月推出 Grok Business（30 美元/用户/月）和 Grok Enterprise（定价未公开披露），分别面向中小团队和大型组织，后者侧重更严格的管控和安全审计功能。API 定价方面，Grok 4 价格为每百万输入 token 3 美元、每百万输出 token 15 美元，Grok 4.1 Fast 则低至 0.20 美元/0.50 美元，为主流前沿模型中最低，显示公司正尝试以低价策略切入企业 API 市场，但这一激进定价能否可持续仍有待观察。3）**政府合同：**xAI 已与美国总务管理局（GSA）签订协议，以每机构 42 美分的价格向联邦机构提供为期一年半的 Grok 服务。此外，美国国防部于 2025 年 7 月

向 OpenAI、Google、Anthropic 和 xAI 各授予最高约 2 亿美元的 AI 合同，旨在推动国防部采用先进 AI 能力。

盈利与估值层面，据彭博社报道，xAI 截至 2025 年 6 月的单月总成本已高达约 10 亿美元，或意味着现阶段收入规模仍不足以覆盖算力、训练和运营投入。管理层虽提出 2027 年实现盈利的目标，但兑现路径或存在较大不确定性。2026 年 2 月，SpaceX 以全股票方式收购 xAI，其中 xAI 估值为 2500 亿美元，SpaceX 估值为 1 万亿美元，合并后整体估值达 1.25 万亿美元。据彭博社报道，SpaceX 将于 2026 年 6 月上市，预计融资 750 亿美元，使得公司整体估值达到 1.75 万亿美元，届时资本市场买入的将不只是 SpaceX 本体，还将同时包含 X 与 xAI 两块业务。若将 X 与 xAI 2025 年约 38 亿美元收入（据 Sacra）合并计算，xAI 当前 2500 亿美元估值对应的合并收入倍数为约 66x。

图 12：AI 基础模型平台核心财务对照

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	**~300 亿美元** (2026.4)	**~250 亿美元** (2026.2)	独立 AI 业务**~5 亿美元** (2025 年化)；X 订阅约 **10 亿美元** ARR (2026.2)
增长轨迹	2025 年底~90 亿→2026 年 4 月初~300 亿，**不到四个月上涨超 200%**；Claude 免费活跃用户自年初**+60%** (2026.3)	过去三年收入**~10 倍**扩张；周活 **9 亿**（vs 2025.10 的 8 亿）；付费企业用户 **900 万+**（vs 2025.8 的 500 万）	Grok 月活 **6400 万** (2025.9)，较 2025.4 增长**~80%**；Sacra 预计 2026 年 AI 业务 **20 亿+**
ARR 目标	此前预计 2026 年 **200–260 亿美元**，2026 年 4 月已经达成	未公开披露具体目标	管理层目标 2027 年盈利
收入结构与核心产品			
引擎① 编程/开发	**Claude Code**：核心引擎。2025.11 达 **10 亿** ARR→2026.2 达 **25 亿+**；企业端订阅量自年初**增长 4 倍**，贡献 Claude Code 过半收入；开发者日均支出**~$6**，Top10%超 **$12**；全球约 **4%**公开 GitHub commits 由 Claude Code 生成 (2026.2)	**Codex**（GPT-5.3）：2026.2 推出，周活 **300 万**，API 每分钟处理超过 **150 亿**个 token (2026.4)；Cisco/Nvidia/Ramp 等已全面部署；API 业务单月新增 ARR 超 **10 亿美元** (2026.1)	Grok 4 API：**$3/$15**（输入/输出每百万 token）；Grok 4.1 Fast **$0.20/$0.50**——主流前沿模型最低价
引擎② 通用/办公	**Cowork**：2026.1 推出，"Claude Code for general computing"；首月 30+功能、11 个开源插件；运行于 Excel/PPT 内；发布后全球 SaaS 板块蒸发**~2 万亿美元**市值。	**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)，提升空间大	**消费订阅**（依托 X 流量）：免费/X Premium $8/Premium+ $40/SuperGrok $30/SuperGrok Heavy $300 每月
引擎③ 新曲线	**Opus 4.6**：上下文窗口 **100 万 token**；内置 Agent Teams 多实例协同；API 定价 $5/$25（输入/输出每百万 token），扩展上下文$10/$37.5	**广告业务**：2026.2 上线，面向 Free/Go 用户；初始 CPM**~$60**，最低投放**$20 万**；首批广告主含 Target/Ford/Adobe 等；广告业务 2026 年收入预计达**$25 亿美元**，并计划在 2030 年达**$1000 亿美元**	**政府合同**：GSA 协议每机构**$0.42**/一年半；国防部向四大 AI 公司各授最高**~2 亿美元**合同
盈利与成本			
盈利节奏	预计 **2027 年**现金流转正	2026 年现金消耗**~250 亿美元**；2027 年达 **570 亿美元**	月成本**~10 亿美元** (2025.6)；管理层目标 2027 年盈利
毛利率演进	2024 年**-94%** → 2025 年目标 **40%** → 2028 年目标 **77%**，经营杠杆持续优化	推理成本 2025 年同比**+4 倍**至 84 亿→2026 年预计 **141 亿**；收入与成本尚未有效脱钩	低价 API 策略+高基础设施投入，毛利率数据未披露
估值与融资			
最新估值	**3800 亿美元**（投后估值）(Series G, 2026.2.12)；融资 **300 亿美元**	**8520 亿美元**（投后估值）(2026.3)；融资轮次中 SoftBank 出资 300 亿/Nvidia 300 亿/Amazon 500 亿/Andreessen Horowitz 等投资机构 120 亿	SpaceX 全股票收购 xAI，xAI 估值 **2500 亿** (2026.2)；SpaceX 最新 IPO 合并估值 **1.75 万亿美元**
估值倍数	按 300 亿 ARR**~12.7x**；估值消化快于收入增长	按 250 亿 ARR 约**~34x**	含 X 合并收入倍数仍高达**~66x**
上市进度	最早于 2026 年 10 月上市，融资规模或超 600 亿美元	或最早于 **2027 年**上市	SpaceX 计划 **2026 年 6 月**上市，届时将同时包含 X 与 xAI

资料来源：彭博，Forbes，Sacra，The Information，Fortune，路透社，CNBC，Semi Analysis，Orbilon Tech，Famewall，Sacra，招银国际环球市场

注：数据截至 2026 年 3 月 31 日

国内 AI 基础模型平台：进入商业化验证期，营收维持高增状态

我们认为当前国内 AI 平台的定价变量已从更偏重追求“模型能力排名”转向“收入兑现速度与路径可持续性”：基于 2026 年彭博市场预期，智谱/MiniMax 截至 4 月 14 日收盘交易价格分别对应 126.2x/168.2x 2026E PS，彭博市场预期 2025-2027E 收入 CAGR 分别为 216.2/193.4%；月之暗面正寻求最高 10 亿美元融资、阶跃星辰潜在 IPO 计划或于 2026 年底启动。我们判断目前市场对四家公司的主要关注点为：1）未来几年的营收变现起量速度，

及营收增速是否可以有效为估值提供支撑；2）在行业整体竞争仍激烈的情况下如何有效维持模型层面的长期差异化优势；3）算力供给瓶颈和地缘因素如何影响模型研发进展及营收释放节奏。差异化方面，智谱的核心看点或在于 GLM-5 系列的定价权提升趋势；MiniMax 的关键变量在于从"AI+社交娱乐"向 Agent 平台的延伸带来的 ARPU 提升；月之暗面需证明 Kimi K2.5 的爆发式增长可以在更大基数上持续；阶跃星辰则需验证终端分发模式的收入转化效率。整体而言，我们对国产 AI 原生板块发展展望乐观，但也同时建议投资者重点关注各家公司 2026-2030 年收入绝对值与增长趋势，以及新业务的变现探索。

■ **智谱 AI（2513 HK，未评级）：模型提价助推商业化进一步提速**

智谱 AI 于 2026 年 1 月 8 日在港交所挂牌，成为全球首家以通用大模型为核心主业的上市公司，IPO 首日市值达 578.9 亿港元；截至 4 月 14 日，市值突破 4226 亿港元。

基本面方面，据智谱 2025 年业绩会披露，2025 年全年收入为人民币 7.243 亿元，同比增长 131.9%；2025 年付费开发者规模突破 24.2 万，MaaS 平台注册用户突破 400 万，业务覆盖全球 218 个国家和地区，与超 400 万中小企业及开发者共建生态，中国前 10 大互联网公司中已有 9 家深度集成 GLM，反映公司在政企、开发者及平台生态侧已形成初步规模化落地。盈利能力上，2025 年随着云端部署占比提升（从 15.5%升至 26.3%），综合毛利率从 56.3%降至 41.0%，但公司毛利润绝对额仍增长 68.7%至 2.967 亿元；公司在研发端持续投入，2025 年全年研发费用进一步增长至人民币 31.80 亿元（同比增长 44.9%），约为全年收入的 4 倍；2025 年全年经调整净亏损达人民币 31.82 亿元（同比扩大 29.1%）。

2026 年 3 月 27 日公司发布新一代旗舰模型 GLM-5.1，距 GLM-5 推出仅六周，进一步助推了公司的商业化。GLM-5.1 沿用 GLM-5 基础架构（744B 总参数/约 40B 激活参数，MoE，200K 上下文），主要针对编程 Agent 场景做了针对性强化。据公司披露，GLM-5.1 在 SWE-Bench Pro 上以 58.4 分超过 GPT-5.4（57.7）、Claude Opus 4.6（57.3）和 Gemini 3.1 Pro（54.2）；GLM-5.1 在综合能力与 Coding 能力上达到全球第一梯队，整体表现对齐 Claude Opus 4.6。更为重要的是，模型能力提升助推了商业化能力提升：公司于 2 月上调 GLM Coding Plan 套餐价格，其中国区涨价 30%，海外版涨价 30%–100%。2025 年报进一步确认，截至 2026 年 3 月，API 调用定价较 2025 年底提升了 83%。此外，2026 年 3 月推出的 Claw Plan 上线仅两天订阅用户即破 10 万，上线 20 天突破 40 万。

估值层面，按 4 月 14 日收盘 4226 亿港元市值和 2026/2027 年彭博市场预期 29.2/68.5 亿人民币收入测算，公司 PS 约为 126.2/53.8 倍。

■ **MiniMax（100 HK，未评级）：模型+应用产品矩阵清晰，关注营收成长空间**

MiniMax 于 2026 年 1 月 9 日在港交所上市，首日股价上涨近 110%；截至 4 月 14 日收盘，公司市值已升至 2983 亿港元。

基本面方面，据公司披露，MiniMax 服务已覆盖全球 200 多个国家和地区，超过 70%的收入来自海外。据公司财报，2025 年全年收入同比增长 159%，毛利率由 2024 年的 12.2%上涨至 2025 年的 25.4%。据公司披露，2025 年 AI 原生产品收入 5308 万美元，占比 67.2%；开放平台及企业服务 2596 万美元，占比 32.8%。

从产品结构看，MiniMax 已形成较为清晰的模型+应用产品矩阵。底层为 M 系列大语言模型（目前已演进至 M2.7）、视频生成模型 Hailuo 和语音生成模型 Speech 等，覆盖多模态生成能力；上层则是 Talkie（海外社交陪伴）与星野（国内版）等 C 端应用，形成从模型、工具到终端产品的较为完整的闭环。我们认为，这一结构的优势在于公司不仅具备底层模型能力，还能够通过上层产品直接触达用户、获取行为数据并反哺模型优化，从而形成更强的应用驱动型迭代路径。

在模型层，2026 年 3 月公司发布 MiniMax-M2.7，MiniMax 将其定义为首个“深度参与自身演进”的模型，强调其可独立构建复杂智能体执行框架，并完成高复杂度生产力任务。相比 M2.5，官方重点强化软件工程能力、办公生产力以及复杂环境交互能力。在生态层，OpenClaw 生态的爆发成为公司 2026 年初最重要的估值催化之一。MiniMax 被 OpenClaw 列为官方大模型提供商，并于 2 月 26 日推出基于 OpenClaw 的托管云端助手 MaxClaw。

在商业化层，上述多模态能力的成熟也在推动变现模式同步优化。2026 年 3 月 23 日，MiniMax 宣布将此前的 Coding Plan 全面升级为 Token Plan，升级后 Plus 及以上套餐用户在保留 M2.7 编程模型原有用量的基础上，额外获赠海螺视频、语音合成、音乐生成、图像生成等多模态模型调用额度，无需额外付费；同时面向专业开发者和企业用户推出语音和视频资源包，批量使用价格最高可优惠 20%。这一定价策略将多模态能力打包进统一订阅体系，以更低的边际成本撬动用户从单一文本场景向全模态使用迁移，从而进一步巩固产品矩阵的协同效应。

财务表现及市场估值方面，公司 2025 年收入约为 7,900 万美元，经调整净亏损 2.51 亿美元。若以 4 月 14 日收盘约 2983 亿港元市值测算，公司 2026E/2027E PS 为 168.2/56.2 倍。

■ **月之暗面（未上市）：Kimi K2.5 发布带动收入快速增长**

据彭博社报道，月之暗面于 2026 年 3 月已开始寻求最多 10 亿美元新一轮融资，目标估值约 180 亿美元。据创始人杨植麟披露，截至 2025 年末，月之暗面账面现金储备已突破 100 亿元人民币，融资金额已超过绝大部分 IPO 募资及上市公司定向增发规模。

2026 年，月之暗面发布 Kimi K2.5 模型。据澎湃新闻报道，模型发布近 20 天的累计收入已超过 2025 年全年总收入，增长主要受全球付费用户及 API 调用量大涨共同驱动。海外付费用户保持高速增长，Kimi 海外收入已超过国内。不过需要注意的是，月之暗面的付费订阅计划自 2025 年 9 月才正式开启，因此基数相对较小。据创始人杨植麟在内部信中披露，2025 年 9–11 月，海外和国内付费用户数月均环比增长超 170%，海外 API 收入同期增长 4 倍。据 Stripe 数据显示，Kimi 个人订阅用户支付订单量在 2026 年 1–2 月呈爆发式增长——1 月环比增长 8280%，2 月环比进一步增长 123.8%，进入 Stripe 全球榜单前十。短期公司收入高增或部分由于较低的基数，后续可持续关注新模型的发布节奏，及对 ARPU 和营收的拉动作用。

■ **阶跃星辰（未上市）：关注模型与硬件厂商联动的差异化商业化路线**

据《财经》报道，阶跃星辰 2025 年收入近 5 亿元人民币，2026 年预计收入约 12 亿元。根据 OpenRouter 统计，截至 3 月 12 日，阶跃星辰 Step 3.5 Flash 30 天 tokens 调用总量位居全球第一。落地层面，据公司披露，阶跃星辰模型已在 OPPO、荣耀等头部手机品牌装机超 4200 万台，日均服务近 2000 万人次。汽车端，据乘联会数据，2025 年 9–12 月与吉利合作的银河 M9 上市销量近 4 万辆，预计 2026 年全年"上车"将超过 100 万辆，显示其"AI+手机""AI+汽车"两大高频终端路径或已逐步开始进入放量阶段。2026 年 1 月，旷视科技联合创始人印奇正式加入阶跃星辰担任董事长；考虑到其同时担任吉利控股旗下千里科技董事长，这或意味着阶跃与吉利银河品牌年销超 120 万辆（2025 年口径）的终端分发体系有望形成更大合力，助力强化模型在车端的深化落地能力。

上市进程方面，据《财经》2 月 27 日报道，阶跃星辰正在推进 Pre-IPO 融资，第一拨投前估值约 40 亿美元，第二拨预计升至 50–60 亿美元，并计划于 2026 年 6 月 30 日前向港交所递表，预期年底完成上市。据公开信息，公司在 2026 年 1 月完成超 50 亿元 B+轮后，仅两个月内即开启 Pre-IPO 轮，累计融资已接近 80 亿元人民币。**估值**层面，若基于 2025 年约 5 亿元收入和第一拨投前估值 40 亿美元（约 280 亿元人民币）测算，对应的 2025 年 PS 倍数约 55x。

■ **DeepSeek（未上市）：持续推进模型迭代，关注国产算力适配**

DeepSeek 在持续推进模型迭代的同时，将国产算力适配作为阶段性重点。年初 DeepSeek 发布了 mHC 和 Engram 两篇技术论文：前者优化注意力机制以降低推理内存开销，后者引入跨会话长期记忆架构，为下一代模型奠定架构基础。上述技术成果于 2 月初开始应用：DeepSeek 将网页端和 APP 端的上下文窗口从 128K 大幅提升至 1M，同时将知识库截止时间从 2024 年 7 月更新至 2025 年 5 月，官方表示正在测试新的长文本模型结构。3 月 29 日，DeepSeek 经历了近 13 小时的服务中断，恢复后多数社区用户反馈输出质量提升，市场推测此次停机系底层模型迭代部署所致，外界推测此次迭代或与 DeepSeek 下一代旗舰模型 V4 相关。与此同时，据路透社 2 月底报道， DeepSeek 仅向国产芯片厂商开放 V4 的早期访问权限；The Information 4 月报道，过去数月 DeepSeek 团队将大量精力投入了华为与寒武纪芯片的适配与架构重写，未来公司与国产算力的适配进程或值得关注。

AI 应用商业化进展更新

当前，AI 应用商业化起量较为迅速的主要 C 端领域为 AI 编程、企业 AI 智能体、AI 创意生成和 AI 搜索四大领域。从 AI 应用层四大垂直赛道的横向对比中，我们观察到以下几条规律。

1）**编程是当前 AI 应用阶段中跑出十亿美元级产品较多的赛道**：据 Sacra 报道，Claude Code 截至 2026 年 2 月 ARR 已突破 25 亿美元；据彭博报道，Cursor 同期 ARR 突破 20 亿美元，从 1 亿增长至此仅用 14 个月——且两者均处于高速增长阶段，表明赛道仍以共同扩容为主而非存量竞争，开发者对 AI 编程工具的付费意愿和使用强度较强。

2）**企业 AI 智能体的商业化正从概念验证切入规模放量，但增长路径与编程工具有所差异**：其核心驱动力并非独立新产品爆发，而是 SaaS 巨头在存量客户基础上进行 AI 功能增购。Salesforce 4QFY26 Agentforce ARR 达 8 亿美元（同比+169%），超 60%预订来自老客扩展，ServiceNow 4Q25 Now Assist ACV 突破 6 亿美元。

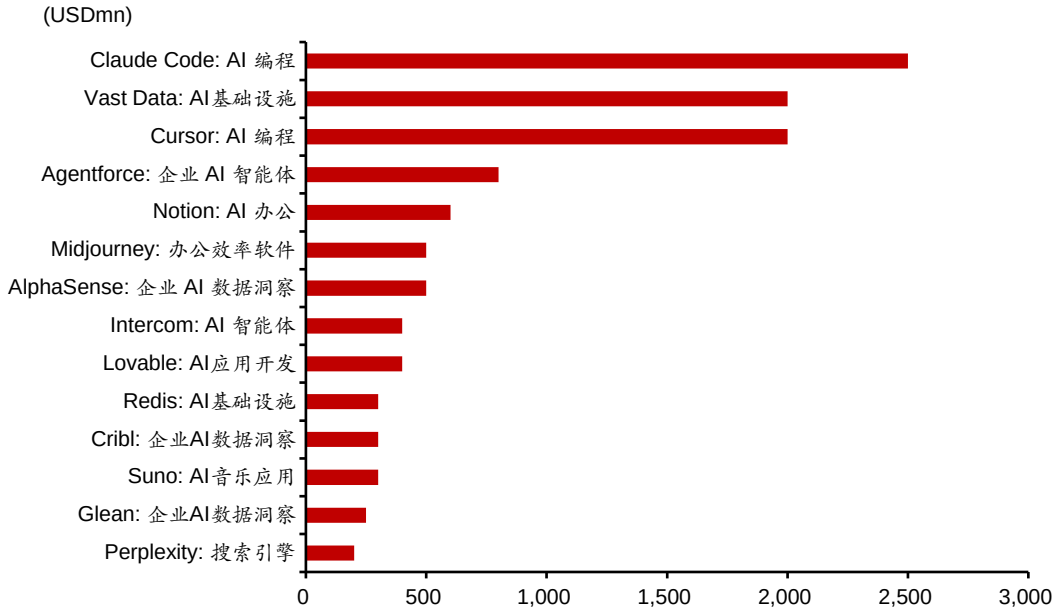
3）**定价模式逐步演进：从纯席位制向"席位+用量"混合制迁移。**编程工具按 token 及使用时长收费，据 Salesforce 官方公告，Agentforce 已从最初的按对话计费扩展为按对话、按动作和按席位三种模式并行，据 ServiceNow 官方文档，Now Assist 通过 assists 额度包实现按用量计费——我们认为，按产出及结果闭环的收费模式正在成为 AI 时代 SaaS 的趋势。

4）**AI 创意生成赛道，头部产品早期商业化进展积极，但行业竞争逐渐加剧：**得益于 AI 创意生成的丰富应用场景以及模型不断提升的多模态能力，头部 AI 创意生成应用早期商业化进展积极，2025 年末 MidJourney/Suno/Adobe Firefly/快手可灵 ARR 水平已达到 5 亿/3 亿/2.5 亿/2.4 亿美元。但同时我们看到行业竞争持续加剧，头部厂商之间尚未拉开显著的模型能力或用户差距，仍需持续投入研发和算力以维持领先，直到中长期市场份额向已形成正向商业循环的少数头部厂商集中。

5）**单一的 AI 搜索平台或面临较大的行业竞争：**据 Financial Times 披露，Perplexity 截至 2026 年 3 月 ARR 约 4.5 亿美元，且公司于 2026 年 2 月宣布放弃此前试行的 AI 整合广告战略，转向订阅优先模式。整体来看，目前 AI 搜索仍然在商业化探索的早期阶段，部分头部搜索平台基于其完善的广告基础设施及 AI 能力已实现 AI 搜索广告的初步商业化，但收入贡献仍较为有限。

6）纵观四条赛道，**我们认为 AI 产品的商业化进展与其对用户既有工作流的嵌入深度高度正相关。**编程工具直接替代开发者每天的核心动作，且价值创造较为明确和好衡量，因此变现最快；企业 Agent 嵌入已有 SaaS 业务流程，增速次之；创意生成满足明确的内容生产需求，稳步增长；而 AI 搜索仍在持续探索有效的商业化变现途径。

图 13: 主要 AI 应用最新*ARR



资料来源：ARR Club，TechChurch，公司披露，招银国际环球市场
注：数据基于各公司及新闻网站最新公开披露，摘录日期截至 2026 年 4 月 9 日，具体数据时点因公司而异。

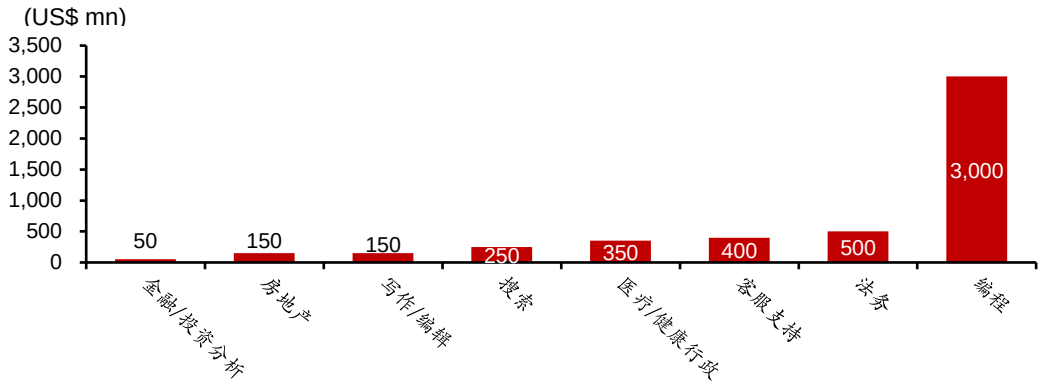
AI 应用细分赛道对比

■ **AI 编程：商业化领先的 AI 应用垂类赛道**

AI 编程是当前商业化较为领先的 AI 应用垂直赛道，已跑出多个十亿美元级 ARR 产品。根据 a16z 2026 年 4 月的测算，若仅考虑企业 AI 场景（剔除个人），AI 编程赛道当前年化收入规模约 30 亿美元，大幅领先其他企业 AI 应用场景。

整体而言，若合并考虑 C 端应用场景，AI 编程赛道已进入规模化变现阶段：1）**赛道增速远超其他 AI 垂直赛道**：据 Sacra 报道，Claude Code 截至 2026 年 2 月 ARR 已突破 25 亿美元；据彭博报道，Cursor 同期 ARR 突破 20 亿美元，从 1 亿美元增长至此仅用 14 个月。2）头部产品均实现 ARR 高速增长，表明**当前阶段仍以共同扩容为主而非存量竞争。**

图 14: 企业 AI 各应用赛道年化收入规模



资料来源：a16z，招银国际环球市场

Anthropic Claude Code：Claude Code 在发布后 6 个月即达到 10 亿美元 ARR，至 2026 年 2 月进一步翻倍至 25 亿美元以上，是当前增速最快的 AI 编程产品之一。从生态渗透度看，其在 GitHub 公开 commits 中的占比正快速攀升，据 Semi Analysis 报道，已从 2026 年 1 月的约 2%提升至 2 月的 4%，Claude Code 正在从"开发者工具"向"AI 原生工作流基础设施"演进。

Anysphere Cursor：据彭博报道，Cursor 在 2026 年 2 月 ARR 突破 20 亿美元，过去三个月内翻倍，从 1 亿美元 ARR 增长到 20 亿美元仅用了 14 个月，且截至 2026 年 2 月，企业客户目前贡献了约 60% ARR，由以 C 端为主的收入结构逐渐向 B 端为主转移。Cursor 在 2025 年 11 月完成 23 亿美元融资，估值达 293 亿美元。Cursor 的定价较高： 根据 Cursor 官网，当前定价为 Pro 20 美元/月，Pro+ 60 美元/月，Teams 40 美元/用户/月。我们认为，虽然 Cursor 定价相对较高，其 ARR 仍实现快速增长，反映出开发者对高端 AI 编程工具具有较强的付费意愿。

GitHub Copilot：据微软 2QFY26（财年结于 6 月）业绩会披露，GitHub Copilot 付费订阅用户数达 470 万，同比增长 75%，其中 Copilot Pro+个人订阅环比增长 77%。截至 2025 年 7 月，GitHub Copilot 累计用户数突破 2,000 万，较三个月前的 1,500 万（2025 年 4 月）显著增长。截至 2025 年 6 月，90%的 Fortune 100 企业已部署 GitHub Copilot。定价方面，据 GitHub Copilot 官网，当前订阅方案为：Pro 10 美元/月、Pro+ 39 美元/月、Business 19 美元/用户/月、Enterprise 39 美元/用户/月。虽然基于订阅用户数及订阅方案价格简单计算 GitHub Copilot 在微软总收入中占比尚不足 1%，但其业务营收起量较为迅速。

OpenAI Codex：Codex 自 2026 年初推出以来用户增长迅猛，周活跃用户与 token 使用量均实现数倍增长，已成为 OpenAI 在企业编程市场的核心增长引擎。产品层面，Codex 捆绑于 ChatGPT 订阅体系（Go/Plus/Pro），尚未独立拆分定价，收入贡献暂包含在 OpenAI 整体 ARR 中。

图 15：AI 编程赛道核心产品对照

海外主要 AI 编程工具对比（截至 2026 年 3 月）				
对比维度	Claude Code（Anthropic）	Cursor（Anysphere）	GitHub Copilot（Microsoft）	Codex（OpenAI）
ARR / 收入规模	Claude Code ARR 约 **25 亿美元**（2026.2），较年初翻倍以上；Anthropic 整体 ARR 达 **190 亿美元**，过去三年每年增长超 10 倍	ARR 约 **20 亿美元**（2026.2），过去 3 个月内翻倍；从 1 亿到 20 亿仅 14 个月	未单独披露	捆绑于 ChatGPT 订阅体系，未单独拆分；OpenAI 整体 ARR 约 **250 亿美元**（2026.2）
用户规模	GitHub 公开 commits 中 Claude Code 生成占比从 2026.1 的 **2%**升至 2026.2 的 **4%**；VS Code 扩展程序 30 天平均安装量均值达 **2,900 万**（2026.1），呈指数增长；业务订阅自年初以来**翻四倍**	未披露具体用户数	付费订阅用户 **470 万**，同比 +75%，Copilot Pro+个人订阅环比+77%（2025.12）；累计用户突破 **2,000 万**（2025.7），较 2025 年 4 月的 1,500 万显著增长	周活跃用户超 **160 万**（2026.2），年初以来增长超 3 倍；桌面端下载超 100 万次；周 token 用量增长约 5 倍
企业渗透	企业用户贡献超 **50%**的 Claude Code 收入（2026.2）；年消费超$10 万的客户数量同比增长 **7 倍**（2026.2）；Deloitte 部署覆盖约 47 万员工（迄今最大单笔部署）（2025.10）；超 30 万家企业客户（2025.9）	企业客户贡献约 **60%** ARR（2026.2），较早期以个人开发者为主的结构发生显著转变；企业付费增长有效抵消了部分个人用户向 Claude Code 的迁移	**90%**的 Fortune 100 企业已部署；超 5 万家组织使用（2025.6）	Cisco、Nvidia、Ramp、Rakuten、Harvey 等已在开发团队部署；**92%** 的 Fortune 500 使用 ChatGPT（2025.12）
定价体系	Free **$0**/月	Hobby **$0**/月	Free **$0**/月	
	Pro **$20**/月	Pro **$20**/月	Pro **$10**/月	Go **$8**/月
	Max 5x **$100**/月	Pro+ **$60**/月	Pro+ **$39**/月	Plus **$20**/月
	Max 20x **$200**/月	Ultra **$200**/月		Pro **$200**/月
	Teams **$20-$25**/用户/月	Teams **$40**/用户/月	Business **$19**/用户/月	Business / Enterprise 定制
	Enterprise 定制	Enterprise 定制	Enterprise **$39**/用户/月	
	（捆绑于 Claude 订阅体系）	自 2025 年 6 月起，Cursor 所有付费层级（Pro / Pro+ / Ultra）在耗尽每月包含的信用额度后，均可启用按 API 费率的超额按量计费，或切换至无限使用的 Auto 模式		（捆绑于 ChatGPT 订阅体系）

资料来源：公司官网，Bloomberg，CNBC， Sacra，Semi Analysis，Fortune，TechCrunch，招银国际环球市场

注：数据截至 2026 年 3 月

■ **企业 AI 智能体：海外落地迅速**

得益于海外企业客户良好的付费习惯以及成熟的 SaaS 订阅模式，海外 B 端 AI 智能体商业化进展迅速，头部产品 ARR/ACV 已接近十亿美元级别。企业 AI 智能体赛道的核心投资逻辑在于：1）**海外 SaaS 巨头的 AI 产品向上销售及增购正在从早期验证进入规模放量**，Salesforce Agentforce 4QFY26 ARR 达 8 亿美元（同比+169%），ServiceNow Now Assist ACV 突破 6 亿美元并朝 2026 年超 10 亿美元目标推进；xAI 虽以激进定价（Grok 4.1 Fast 仅 0.20 美元/百万 input token）切入企业市场，但客户基础和工作流渗透深度仍有明显差距。2）**收费模式从席位制向"席位+使用量"混合制转型是行业共性趋势**，Salesforce 已从最初的 2 美元/对话扩展为按对话、按行动（Flex Credits，0.10 美元/action）和按席位三种模式并行，ServiceNow 通过 assists 额度包实现按使用量计费。3）**现有客户扩容是核心增长引擎**，Salesforce 4QFY26 超 60%的 Agentforce 与 Data 360 订单来自老客增购；ServiceNow 4Q25 客服类 Now Assist 产品在续约时增售幅度超 70%，包含 5 款及以上 Now Assist 产品的订单数量同比增长超 10 倍。

Salesforce Agentforce: 据 Salesforce 4QFY26 业绩会（财年截至 2026 年 1 月 31 日），Agentforce ARR 达 8 亿美元，同比增长 169%，累计完成 29,000 笔交易，环比增长 50%。Agentforce 与 Data 360 的合并 ARR 超过 29 亿美元，同比增长超 200%。从客户结构来看，超过 60%的 Agentforce 与 Data 360 4QFY26 预订来自现有客户扩展，Agentforce 生产环境账户数环比增长近 50%，客户粘性较强。管理层表示 FY27 下半年有机收入增速有望重新加速，长期目标为 FY30 收入 630 亿美元。

ServiceNow Now Assist： 据 ServiceNow 4Q25 业绩会，Now Assist ACV 已突破 6 亿美元并朝着 2026 年超 10 亿美元的目标稳步推进，4Q25 Now Assist 净新增 ACV 同比翻倍以上，单季超 100 万美元的 Now Assist 交易达 35 笔。公司指引 2026 全年订阅收入为 155.3 亿至 155.7 亿美元，同比增长 19.5%-20%（固定汇率口径，含约 1 个百分点 Moveworks 并表贡献）。在生态合作方面，ServiceNow 持续扩大与 Microsoft、Anthropic、OpenAI 等的合作，推进 AI 集成和 LLM 模型选择多元化。在产品扩展方面，ServiceNow 于 2026 年初进一步推出 Autonomous Workforce 和 EmployeeWorks 等新产品，将自身定位为企业 AI 工作流的"控制塔"（AI Control Tower），而业绩会披露 AI Control Tower 在 2025 年的交易量已超出 2025 年目标的 4 倍以上。这一定位如果持续验证，ServiceNow 将从传统 ITSM 厂商升级为企业 AI agent 的调度中枢，打开更广阔的 TAM。

xAI Grok Business/Enterprise： xAI 于 2025 年 12 月 30 日正式推出 Grok Business 和 Grok Enterprise 两档企业订阅计划，分别面向中小团队和大型组织。从定价策略看，无论是订阅层（Business 30 美元/用户/月）还是 API 层（Grok 4.1 Fast 低至 0.20 美元/百万 input token），xAI 在主流前沿模型中均处于较低价格区间，与 Cursor、GitHub Copilot 等以产品力和生态粘性驱动定价的路径形成明显差异。但这一激进定价能否可持续，仍需观察其背后的资本补贴力度。

Google Gemini Enterprise（包含在 Google Cloud 中）： 据 Alphabet 4Q25 财报电话会，Gemini Enterprise 已累计售出超 800 万个付费席位，覆盖逾 2,800 家企业；使用 Gemini 的企业总数超 12 万家，涵盖 Airbus、Honeywell 等头部客户，95%的全球 Top 20 SaaS 公司已接入 Gemini。据 Google Cloud 官方定价，Enterprise 版起价 30 美元/用户/月，面向小型团队的 Business 版起价 21 美元/用户/月。据 CNBC 报道，该产品支持连接 Box、Microsoft、Salesforce 等 50 余种企业工具数据，并提供 Deep Research、NotebookLM 等预制 agent 及无代码 Agent Designer 工作台。我们认为，800 万付费席位相对 Google Workspace 超 30 亿用户安装基础而言渗透率仍较低，需关注后续席位增长和 ARPU 提升趋势。

图 16：企业 AI 智能体赛道核心产品对照

企业 AI 智能体：海外头部产品关键指标对比				
维度	Salesforce Agentforce	ServiceNow Now Assist	xAI Grok Biz/Enterprise	Google Gemini Enterprise
核心规模	ARR **8 亿美元**，同比 **+169%** (4QFY26, 截至 2026.1)	ACV 突破 **6 亿美元**，目标 2026 年超 **10 亿** (4QFY25, 截至 2025.12)	暂无公开数据 (2025.12.30 上线)	Gemini Enterprise 800 万+付费席位，12 万+企业使用 Gemini (4Q25)
交易量	累计 **29,000 笔**，环比 +50% (4QFY26)	Q4 百万美元大单 **35 笔** (4QFY25)	暂无	Gemini Enterprise 覆盖逾 2,800 家企业 (4Q25)
客户粘性	**60%+**预订来自老客扩展 (4QFY26)	续约率 **98%** (4QFY25)	客户基础与工作流渗透仍有差距	95%全球 Top 20 SaaS、80%+ Top 100 SaaS 已接入；头部客户含 Airbus、Honeywell (4QFY25)
收费模式	席位+使用量混合制：按对话、按行动（**$0.10**/action）、按席位三轨并行	席位+assists 额度包按量计费	Business **$30**/用户/月；API: Grok 4.1 Fast **$0.20**/百万 input token	Enterprise **$30**/用户/月；Business **$21**/用户/月 (据 Google Cloud 官方定价)

资料来源：公司资料，招银国际环球市场
注：数据截至 2026 年 3 月

■ **AI 创意生成：头部产品早期商业化进展积极，但行业竞争逐渐加剧**

AI 创意生成赛道涵盖图片、视频、音乐三个细分方向，早期商业化进展积极，但同时行业竞争亦在逐渐加剧。我们看好 AI 创意生成赛道的中长期市场空间，市场份额将逐渐向已形成正向商业循环的少数头部厂商集中：1）AI **在营销素材制作、视频/图片创作等场景已能大量替代人力**，使用场景明确、用户付费意愿强，Midjourney、Adobe Firefly、快手可灵等均已实现超 2 亿美元 ARR；2）**但行业竞争也在持续加剧：**从 Artificial Intelligence 相关榜单变化趋势看，全球头部厂商之间尚未在创意生成领域拉开显著的模型能力或用户差距，仍需持续投入研发和算力以维持领先；3）Midjourney 基于早期用户心智占领及良好的产品能力，已实现高人效且盈利的 AI 应用变现模式，但随着竞争加剧，我们认为公司或也将加大投入以维持竞争力；快手可灵在毛利率层面已经转正，但公司计划持续加大算力和研发投入，未来短期内仍将是持续亏损的态势。

- **Midjourney（AI 文生图）：**根据 Sacra 报道，公司 2025 年收入达到约 5 亿美元（2024 年 3 亿），DemandSage 预测公司 2026 年 ARR 将达 5 亿至 6 亿美元，增量由面向创意机构的企业级订阅驱动。Midjourney 于 2022 年 8 月（上线仅一个月后）即实现盈利，完全自筹资金运营，无外部风投融资。根据 Quantumrun 报导，公司目前员工 100+、零营销支出，是头部 AI 消费端产品中少有的已实现盈利的公司之一。定价体系：Basic 10 美元/月，Standard 30 美元/月，Pro 60 美元/月，Mega 120 美元/月（年付享约 20%折扣）。公司早期占领大量用户心智，2024 年 Midjourney 即以 26.8%的市场份额领先全球 AI 图像生成工具市场（据 Quantumrun）。但随着头部厂商持续加大多模态能力布局，行业竞争逐渐加剧，例如 Google 旗下 Nano Banana 的推出给 Midjourney 带来较大冲击：1）Google 以低价策略（2026 年 2 月推出的 Nano Banana 2 单张图片成本低至约 0.067 美元）侵蚀 Midjourney 在商业摄影和电商场景的付费用户；2）Midjourney V8 Alpha（2026 年 3 月发布）能否在写实和编辑能力上缩小差距，将是维持市场份额的关键。

- **快手可灵 AI（AI 视频生成）**：快手可灵 AI 商业进展积极，4Q25 可灵 AI 收入达 3.4 亿元，2026 年 1 月年化 ARR 突破 3 亿美元，管理层预计 FY26 可灵 AI 收入同比增长超一倍，达到 3 亿美元以上。在赛道竞争持续加剧的环境下，管理层预计将显著加大资本开支及研发人员投入，以支撑可灵模型与产品迭代。管理层预计 FY26 公司资本开支将达到约 260 亿元，同比增加 110 亿元，增量将主要投入到支撑可灵推理与训练的算力。

- **Suno（AI 音乐生成）：**据 Suno CEO Mikey Shulman 披露，截至 2026 年 2 月，Suno ARR 已达 3 亿美元，付费订阅用户突破 200 万，较 2025 年 11 月华尔街日报报道的 2 亿美元年收入实现快速增长。以 3 亿美元 ARR 和 200 万付费订阅者计算，平均 ARPU 约 150 美元/年。定价方面，Suno 目前提供免费版、Pro 版 10 美元/月（年付 8 美元/月）及 Premier 版 30 美元/月（年付 24 美元/月）三档方案。融资方面，Suno 于 2025 年 11 月完成 2.5 亿美元 Series C 融资，投后估值 24.5 亿美元。但 AI 音乐生成赛道需要关注的风险是版权合规。据公开信息，Warner Music 已于 2025 年 11 月与 Suno 达成和解，但 Suno 与 Universal Music 和 Sony Music 的诉讼仍在进行中。

- **Adobe Firefly（AI 创意工具）：**据 Adobe 1QFY26 财报披露，Firefly 相关产品（含 Firefly 应用、Credit pack 及 Firefly Enterprise）ARR 已突破 2.5 亿美元。其中，Firefly 订阅及 Credit pack ARR 环比增长 75%，Generative credit 消费量环比增长超 45%，视频生成动作同比增长 8 倍，音频生成动作同比翻倍，使用端增长强劲。Adobe 管理层提出 FY26 全年总 ARR 增速 10.2%的目标。目前，Firely 收入占公司总收入比例约 1%，绝对值仍偏低，但 Firefly 在订阅转化、credit 消费和多模态（视频/音频）场景的增速均呈加速态势，需持续观察 AI 产品能否从边际增量升级为企业增长引擎。

图 17：AI 创意生成赛道核心产品对照

AI 创意生成赛道：头部产品关键指标对比				
维度	Midjourney（文生图）	快手可灵 AI（视频生成）	Suno（音乐生成）	Adobe Firefly（创意工具）
核心规模	收入约 **5 亿美元** (2025 年, Sacra)；预计 2026 ARR **5–6 亿美元** (2026.1)	2025 全年收入 **10.4 亿人民币**，12 月单月超 **2000 万美元**，ARR 破 **2.4 亿美元** (2025.12)	ARR **3 亿美元**，付费用户突破 **200 万** (2026.2)	Firefly 相关产品 ARR 突破 **2.5 亿美元**；订阅及 credit pack ARR 环比 **+75%** (1QFY26,年结于 12 月 1 日)
增长轨迹	2022 年 5000 万→2023 年 2 亿→2024 年 3 亿→2025 年 5 亿	2025 年月活突破 **1200 万**；App 付费用户环比**+350%**	2025.11 ARR 2 亿→2026.2 ARR 3 亿，大幅跃升	Generative credit 消费环比 **+45%+**；视频生成同比**+8 倍**，音频生成同比**翻倍** (1QFY26)
盈利与融资	上线一个月即盈利 (2022.8)；**零外部融资**、零营销、100+人团队，人效比极高	P 端付费订阅贡献近 **70%**营收 (2025.3)；收入结构相对健康	Series C **2.5 亿美元**，估值 **24.5 亿美元** (2025.11)	Firefly 收入占公司总收入 **1%**；FY26 全年总 ARR 增速目标 **10.2%**
定价体系	Basic $10 / Standard $30 / Pro $60 / Mega $120 每月；年付约 8 折	—	Free / Pro $10/月（年付$8）/ Premier $30/月（年付 $24）；ARPU 约**$150**/年	Firefly 应用订阅 + credit pack + Enterprise 三条产品线
市场地位	全球 AI 图像生成市占率 **26.8%**（第一）(2024)；Discord 社区**~2100 万**成员 (2025.6)	40 余国艺术/设计类下载量**第一** (2026.1)；国内 AI 应用商业化最亮眼标的之一	AI 音乐生成赛道头部；**200 万**付费用户 (2026.2)	依托 Adobe 创意生态（Photoshop/Illustrator 等），企业级渗透优势明显

资料来源：公司资料，Sacra，DemandSage，Quantumrun，AIPRM，晚点 LatePost，Sensor Tower，WSJ，pxz.ai，midlibrary.io，招银国际环球市场；数据截至 2026 年 3 月

■ **AI 搜索：持续探索高效商业化路径**

AI 搜索赛道目前处于早期探索阶段。Perplexity 放弃广告转向纯订阅，压缩了短期变现空间，而纯订阅模式在 AI 搜索领域能否支撑足够大的收入规模仍未得到验证，商业化路径的确定性相对较弱。Google AI 搜索的商业化路径更为清晰（捆绑现有搜索广告生态），Google 方面，AI 搜索广告商业化效率已与传统搜索基本持平，AI 搜索量的增长带来增量广告商业化机会。

变现模式方面，Perplexity 于 2026 年 2 月放弃了此前试行的 AI 整合广告战略，转向订阅优先模式，管理层表示此举旨在维护用户对"答案引擎"的信任。当前定价为 Pro 订阅 20 美元/月、企业版 40 美元/月、Comet Plus 浏览器附加 5 美元/月。分发策略方面，2026 年初，Perplexity 与 Snapchat 达成合作，将为后者聊天界面提供 AI 搜索和答案功能——Snapchat 拥有近 10 亿月活用户，该合作有望显著扩大 Perplexity 的用户触达面。同月，Perplexity 与 Microsoft Azure 签署了 7.5 亿美元的三年承诺协议，获取多种 AI 模型的访问权限，进一步巩固基础设施能力。融资方面，据 PM Insights 数据，截至 2026 年初，Perplexity 在完成 Series E-6 融资轮后估值达到约 212.1 亿美元。

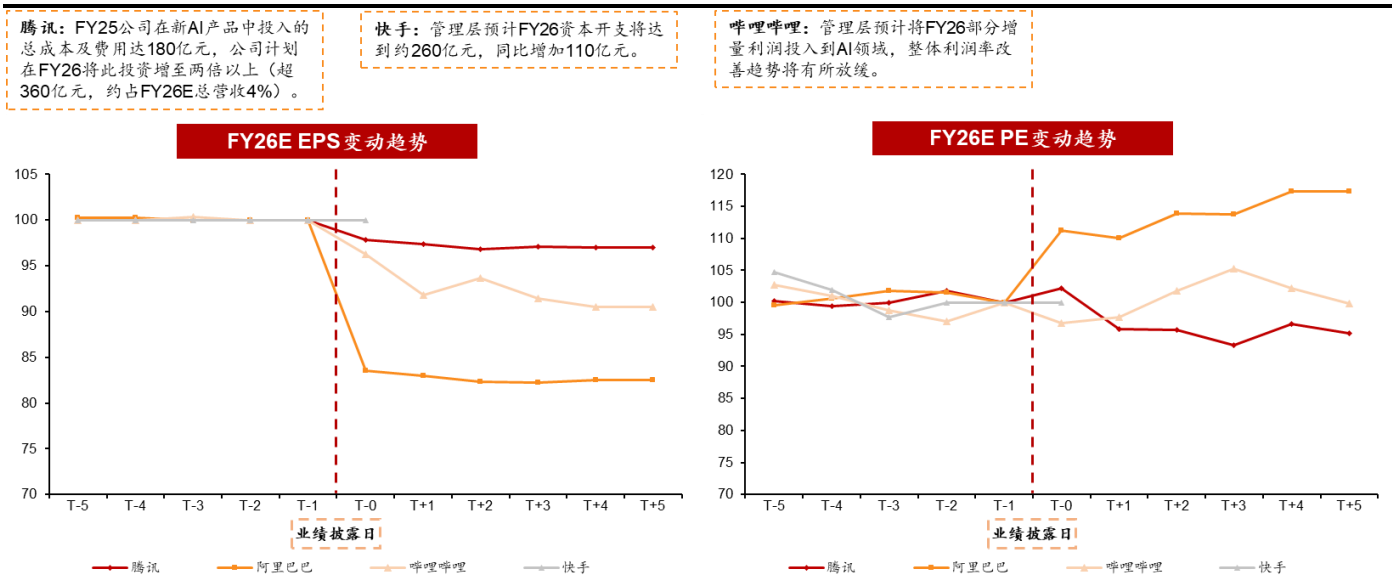
Google AI Overviews + AI Mode：据 Alphabet 4Q25 业绩电话会，Google Search 单季收入达 630.7 亿美元（+17% YoY），全年搜索收入增速逐季加速（1Q +10%→2Q +12%→3Q +15%→4Q +17%）；美国用户 AI Mode 日均查询量自推出以来已翻倍。据 Google 搜索业务 SVP Nick Fox 2025 年 12 月披露，AI Mode 日活用户已超 7,500 万。据 BrightEdge 统计，截至 2026 年 2 月，约 48%的查询现已触发 AI Overviews，同比增长 18 个百分点。广告变现方面，据 Business Insider 2026 年 1 月报道，Google 全球广告副总裁 Dan Taylor 确认 AI Overviews 广告与传统搜索广告的用户点击与互动率处于同一水平；AI 搜索量的增长带来增量广告商业化机会。

重点覆盖公司 1Q26 AI 业务进展更新

我们梳理了宣布将在 2026 年加大 AI 投入，导致 4Q25 业绩后股价波动较大的互联网平台的盈利预期与估值变化趋势，核心观点：1）业绩后股价下跌主要是反映短期盈利预期因 AI 投入而下调，多数公司（除腾讯外）业绩后估值基本维持稳定甚至有所上修，反映投资者并不否定 AI 投入的必要性，对于公司长期盈利与现金流的预期并未发生显著改变。2）腾讯业绩后估值有所下降，主要由于腾讯此前因盈利确定性享有一定估值溢价（风险溢价低），而短期 AI 投入和盈利波动一定程度上影响了该估值溢价（风险溢价有所提升），也并非由于投资者对于公司长期盈利预期有明显变化。

随着 AI 投入对利润的影响充分反映到 FY26 盈利预期当中，我们建议后续继续关注：1）有 AI 产品迭代催化以及商业化积极进展的公司；2）核心业务受 AI 影响有限或受益于 AI，盈利稳定的公司。我们推荐腾讯、阿里巴巴。

图 18：主要互联网平台：业绩后盈利与估值变动趋势



资料来源：彭博，招银国际环球市场
注：业绩披露日次日作为 T+0 日

腾讯：OpenClaw 相关产品开拓 AI 智能体业务机会

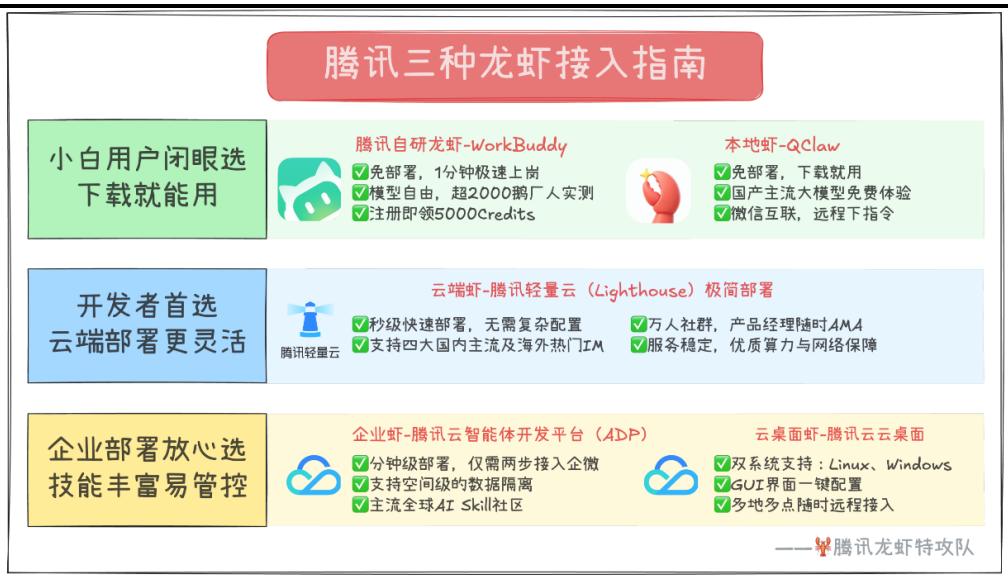
我们看好腾讯 2026 年在 AI 方面取得积极进展：1）微信生态+智能体+开源/三方大模型有望帮助腾讯抢占 AI 智能体赛道份额，进一步完善微信生态的同时带来潜在的智能体商业化机会，例如 3 月腾讯办公智能体 WorkBuddy 上线后用户量快速增长；2）2026 年公司加大人才投入，升级大模型研发组织架构，持续推进自有大模型研发迭代，公司计划 4 月推出混元 3.0 大模型，进一步构建 AI 领域的竞争壁垒。

2026 年腾讯加速在 AI 领域的投入和布局：春节期间元宝投入 10 亿现金红包加速 C 端应用渗透；3 月随着用户对于 OpenClaw 产品关注度提升，公司推出多款相关产品以加速切入 AI 智能体赛道；公司计划 2026 年在 AI 新产品的投入将至少翻倍（>360 亿元）。腾讯 OpenClaw 相关产品具体可细分为三类：1）面向消费者：AI 原生的 PC 智能体工作台 WorkBuddy、能够自动部署 OpenClaw 并连接微信的 QClaw；2）面向开发者：腾讯提供轻量云服务器租赁以帮助开发者在云端部署 OpenClaw；3）面向企业：腾讯提供针对企业

大规模部署协同，并满足企业数据安全的 OpenClaw 解决方案，包括腾讯云智能体开发平台（ADP）、腾讯云桌面等。

短期看，腾讯 OpenClaw 相关产品有望带来一定收入贡献，包括 WorkBuddy 订阅收入、ADP 按 Token 量收费以及服务器等相关租赁收入；长期看，微信+AI 智能体的产品组合有望发挥腾讯社交网络优势，进一步提升微信的用户粘性和生态壁垒，加强公司在国内 AI 应用领域的竞争力。

图 19: 腾讯：OpenClaw 相关产品体系

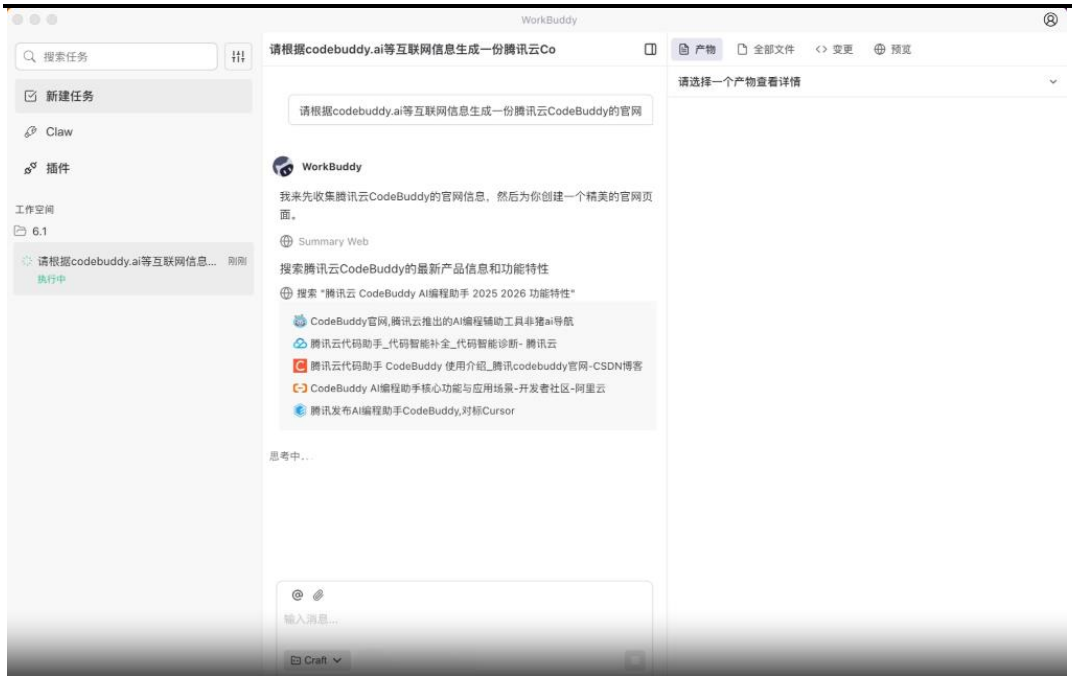


资料来源：公司资料，招银国际环球市场

基于腾讯在 C 端的流量优势以及社交网络，我们看好腾讯面向消费者端的 WorkBuddy 和 QClaw 等 AI 智能体产品。腾讯 WorkBuddy 是 AI 原生的电脑桌面智能体工作台，用户可以通过自然语言驱动办公自动化，一句指令即可让智能体完成数据处理、内容创作与深度分析等工作。WorkBuddy 完全兼容 OpenClaw 技能的同时免去了 OpenClaw 繁琐的部署环节，并且能够快速接入微信、QQ、企业微信、飞书、钉钉等工具，产品内置超 20 种 Skills 技能包与 MCP 协议，支持混元、DeepSeek、GLM 等多家模型，并可以实现企业级的安全与管理。WorkBuddy 目前主要采用订阅制收费，个人专业/企业旗舰/企业专享版对应价格分别为每人每月 58/78/158 元，中长期 WorkBuddy 有望帮助腾讯切入 AI 办公应用赛道，创造增量收入机会。腾讯还推出了 QClaw，基于 OpenClaw 打造可在电脑本地部署的 AI 助手，覆盖 5000+skills，产品优势主要在于可以接入微信，用户通过与微信对话，QClaw 即可完成一系列动作，产品有望从办公智能体领域进一步向个人生活智能体衍生，提升 AI 智能体的用户渗透率。

此外微信在 3 月正式推出 ClawBot 插件，用户扫码或复制命令即可将 OpenClaw 接入微信，接入后用户能通过微信聊天的方式调用自己的 OpenClaw 高效互动。微信 ClawBot 插件还支持接入腾讯自研产品 Lighthouse、WorkBuddy 与 QClaw 等。我们认为微信生态+智能体+开源/三方大模型有望帮助腾讯抢占 AI 智能体赛道份额，同时加强微信的竞争壁垒，并为微信生态内各产品带来潜在商业化机会。

图 20: 腾讯：WorkBuddy



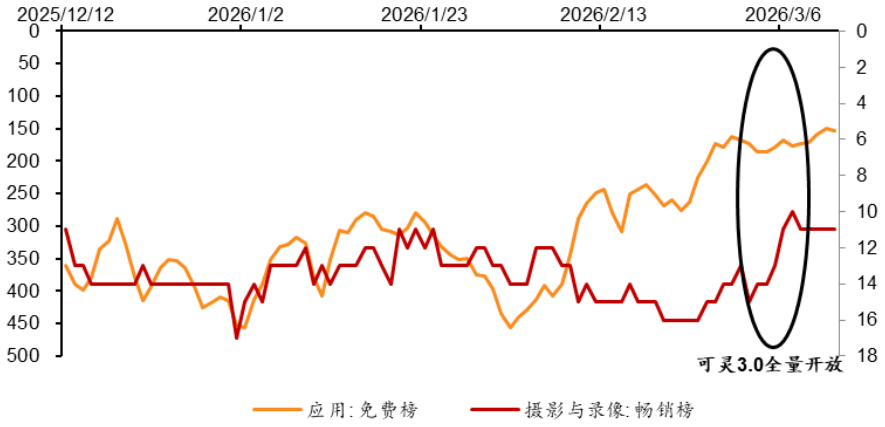
资料来源：公司资料，招银国际环球市场

快手：维持视频模型赛道领先梯队

快手 2026 年 3 月正式面向公众开放可灵 3.0 模型，发布后模型登顶 Artificial Analysis 文生视频模型榜单，后续也维持在榜单的领先梯队。相比可灵 2.5，可灵 3.0 模型实现多项能力的升级：1）智能分镜：让 AI 深度理解剧本后自动调度景别与机位，一键生成电影级影像叙事；2）动作控制 3.0：用户可上传动作参考视频、首帧图、主体视频、提示词作为统一的控制信号，显著提升视频动作一致性与主体稳定性；3）全能音画：支持中、英、日、韩、西多语种生成；4）15 秒超长生成：支持用户生成 3-15 秒灵活时长。

快手在 AI 商业化方面亦取得扎实进展：1）4Q25 可灵 AI 收入达 3.4 亿元，2026 年 1 月年化 ARR 突破 3 亿美元。管理层预计 FY26 可灵 AI 收入同比增长超一倍（至超 3 亿美元）；2）在线营销领域：快手运用生成式推荐模型与智能出价模型，带动 4Q25 国内在线营销服务收入增速提升约 5%；3）电商领域：快手迭代生成式检索架构 OneSearch，4Q25 带动商城搜索订单量增长 3%。

图 21: 快手可灵：中国 iOS 榜单趋势



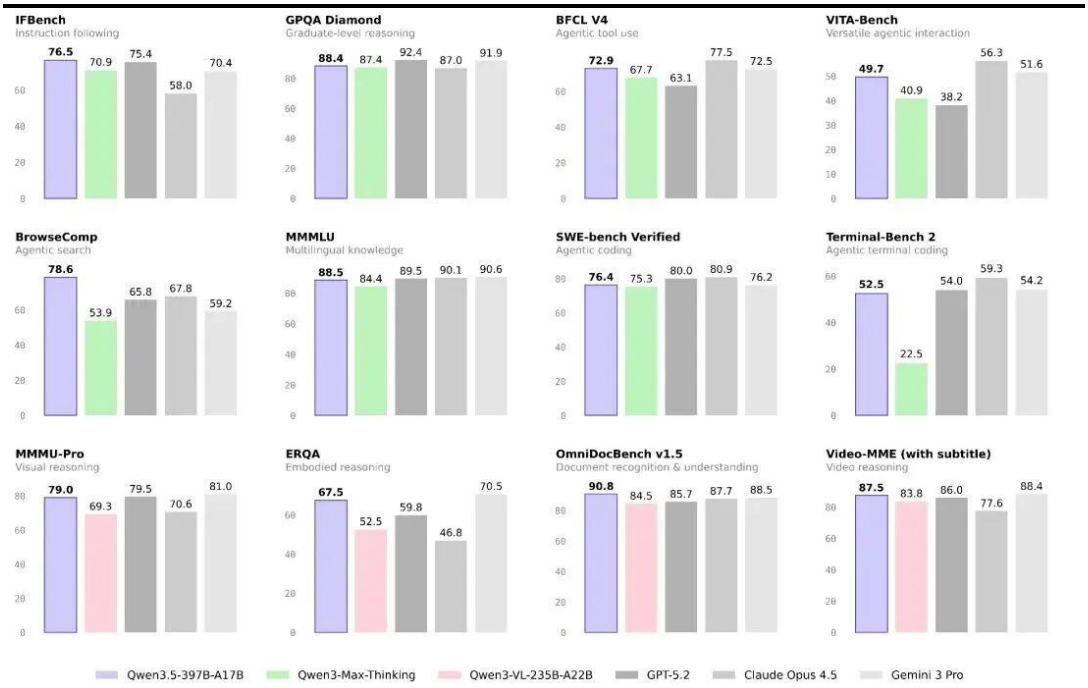
资料来源：公司资料，招银国际环球市场

阿里巴巴：千问系列模型持续迭代，全栈 AI 商业化提速

模型层面，2026 年 1 月 26 日，阿里正式发布千问旗舰推理模型 Qwen3-Max-Thinking。该模型总参数量超万亿（1T），预训练数据量达 36T Tokens。技术层面，Qwen3-Max-Thinking 引入两项核心创新：1）自适应工具调用能力，可自主调用内置工具缓解模型幻觉、提升回复质量；2）测试时扩展技术，能在控制算力的同时大幅提升推理性能。据公司披露，在涵盖事实知识、复杂推理、指令遵循、人类偏好对齐、Agent 能力等 19 个公认的大模型基准测试中，千问旗舰推理模型刷新了数项最佳表现（SOTA）纪录，整体性能可对标国际顶尖 AI 大模型。1 月 29 日，平头哥官网上线"真武 810E"高端 AI 芯片，据澎湃新闻 1 月报道，业内人士透露，对比关键参数，"真武"PPU 的整体性能超过了英伟达 A800 和主流国产 GPU，与英伟达 H20 相当。

2 月 16 日除夕，阿里开源全新一代模型 Qwen3.5-Plus，总参数为 3970 亿，推理时仅激活 170 亿，性能超过万亿参数的 Qwen3-Max 模型，部署显存占用降低 60%，最大推理吞吐量可提升至 19 倍。API 价格每百万 Tokens 输入低至 0.8 元，相当于同级别 Gemini 3 Pro 的约 1/18。3 月 2 日，阿里继续开源 Qwen3.5-0.8B/2B/4B/9B 四款小尺寸模型，完成从大到小的全谱系覆盖。据 Hugging Face 数据，截至 2026 年 2 月 24 日，阿里千问开源大模型的衍生模型数量已突破 10 万，稳居全球最大开源模型榜首。

图 22: 阿里云：千问系列模型与 GPT、Claude Opus 4.5 及 Gemini 3 Pro 能力对比



资料来源：公司资料，招银国际环球市场

组织与应用层面，3 月 16 日，阿里巴巴 CEO 吴泳铭发出全员信，宣布正式成立 Alibaba Token Hub（ATH）事业群。 ATH 事业群整合通义实验室、MaaS 业务线、千问事业部、悟空事业部及 AI 创新事业部。据量子位，悟空已在阿里云实现多个万卡集群部署，服务了国家电网、中科院、小鹏汽车、新浪微博等 400 多家客户。悟空同步发布 OPT（One Person Team，一人团队）十大行业解决方案，首批覆盖电商、跨境电商、知识类博主、开发、门店、设计、制造、法律、财税、猎头等场景。C 端方面，据阿里 3QFY26 财报，截至 2026 年 2 月，千问 App 月活跃用户数已突破 3 亿；春节活动期间超 1.4 亿用户通过千问 App 的智能体功能完成首次 AI 购物。

3 月 30 日，阿里发布千问新一代全模态大模型 Qwen3.5-Omni，将千问系列能力边界从文本推理进一步拓展至原生全模态交互。该模型采用混合注意力 MoE 架构，基于海量文本、视觉及超 1 亿小时音视频数据完成原生多模态预训练，支持文本、图片、音频、视频的全模态输入与输出，最长可处理 256k 上下文及超 10 小时纯音频输入。性能层面，据千问官方披露，Qwen3.5-Omni-Plus 在音视频理解、语音识别、跨模态推理等 215 项第三方基准测试中取得 SOTA，整体表现超越 Gemini-3.1 Pro；在多语言语音稳定性测试中击败 ElevenLabs、GPT-Audio 及 Minimax。多语言方面，语音识别覆盖 113 种语种与方言，语音生成支持 36 种；新引入的 ARIA 技术从根源上解决了漏读、误读等语音不稳定问题。值得关注的是，新模型涌现出音视频 Vibe Coding 能力——用户仅需对着镜头阐述需求，模型即可自主生成 APP、网页等复杂产品代码。商业化层面，官网显示当前该模型仅支持 API 调用，暂时未表明开源计划。

投入与财务层面，2025 年 2 月，阿里 CEO 吴泳铭宣布未来三年将投入超过 3800 亿元用于云和 AI 硬件基础设施建设。据阿里巴巴财报电话会，平头哥自研 GPU 芯片截至 2026 年 2 月已累计规模化交付 47 万片，60%以上服务于外部商业化客户，支持了 400 多家企业客户的 AI 任务。阿里云 3QFY26 营收达 432.8 亿元，同比增长 36%（含 AI 相关产品收入连续第十个季度实现三位数增长）；财报电话会上，CEO 预计商业化 MaaS 收入将成为阿里云最大的收入产品，未来五年，目标 MaaS 在内的云和 AI 商业化年收入突破 1000 亿美元。

谷歌：打造全球领先的全栈 AI 产品体系

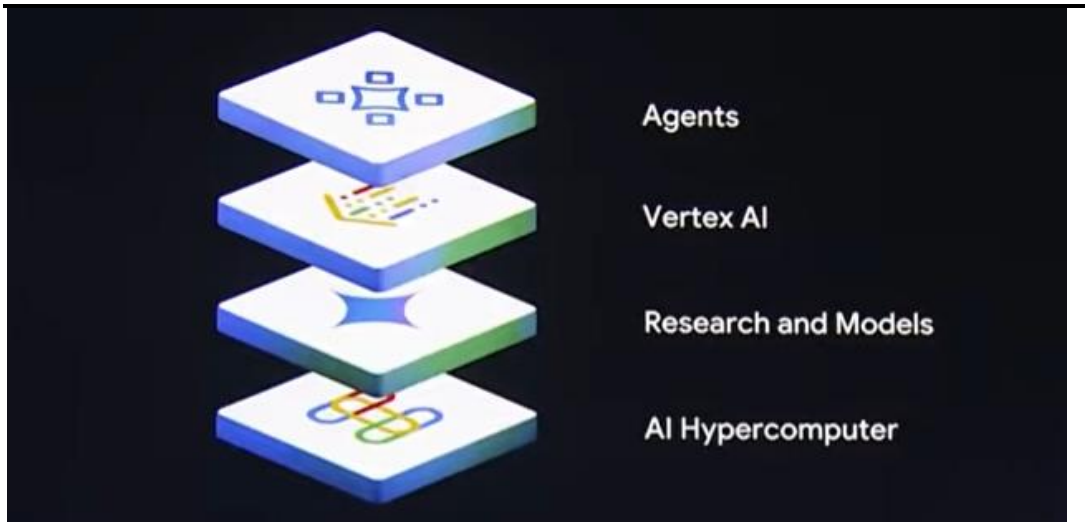
谷歌目前已打造全球领先的全栈 AI 产品体系，覆盖 AI 计算基础设施+大模型+AI 开发平台+AI 智能体与应用生态。基于公司全栈 AI 产品能力，我们看好谷歌多方面持续受益于 AI 发展：1）AI 赋能搜索与广告业务：4Q25 谷歌搜索业务营收同比增加速至 17%，AI 已成为公司提升搜索体验与广告效果的核心驱动力，AI Mode 自上线以来用户日均问询量实现翻倍；2）AI 显著加速谷歌云业务增长：4Q25 谷歌云营收同比增长 48%，在手订单金额环比增长 55%、同比增长超 100%至 2400 亿美元。目前客户通过直连 API 调用 Gemini 的消耗量已达每分钟 100 亿 tokens。

基础设施层面，谷歌 2025 年推出的 Ironwood TPU 在推理任务上实现性能的显著提升，单芯片峰值算力达到 4614 TFLOPS（FP8 精度），最大集群 9216 颗芯片，总算力可达 42.5 EFLOPS。据博通披露，头部模型厂商 Anthropic 已下单了 210 亿美元 TPU 订单，Meta 亦与谷歌协商数十亿美元的 TPU 订单。

模型层面，谷歌 2026 年 2 月发布新一代旗舰模型 Gemini 3.1 Pro，在 12 项测试中超过 Claude、GPT 等旗舰模型，全球排名第一。Gemini 3.1 Pro 的推理能力显著提升，在 ARC-AGI-2 通用智能基准测试中，Gemini 3.1 Pro 评分达到 77.1%，成绩相较 Gemini 3 Pro 实现翻倍提升。

应用层面，Gemini Enterprise 服务自推出 4 个月后已售出 800 万付费席位（标准版 30 美元/席位/月），谷歌云成为企业/开发者打造智能体的首选平台之一。C 端应用方面，4Q25 Gemini App 月活跃用户已超 7.5 亿，AI 支持的搜索功能 AI Mode 自上线以来用户日均问询量实现翻倍。

图 23：谷歌云：全栈 AI 产品体系



资料来源：公司资料，招银国际环球市场

Meta：Muse Spark 聚焦 C 端 AI 应用场景

Meta 2026 年 4 月正式发布大模型 Muse Spark，该模型是 Meta Superintelligence Labs（MSL）成立后发布的首个大模型，具备原生多模态推理以及视觉思维链能力、能够使用工具、并且支持多智能体协作。从评测标准上看，Muse Spark 主要在多模态、健康等领域展示出优于 Opus 4.6 和 Gemini 3.1 Pro 等 SOTA 模型的能力，但在复杂智能体任务和编程工作流等方面的表现仍然弱于 SOTA 模型。

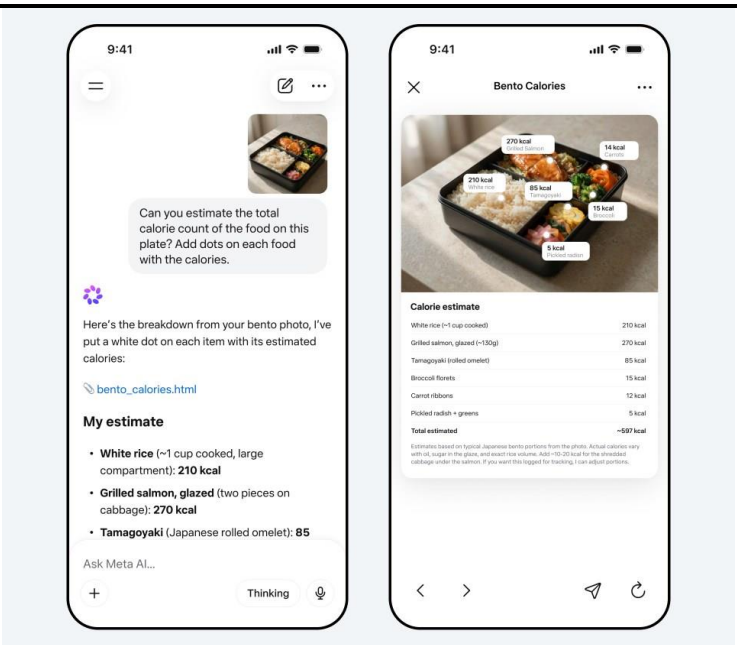
图 24：Meta Muse Spark：主要测评表现

	评测标准	Muse Spark (Thinking)	Opus 4.6 (Max)	Gemini 3.1 Pro (High)	GPT 5.4 (Xhigh)	Grok 4.2 (Reasoning)
多模态	CharXiv Reasoning Figure Understanding	86.4	65.3	80.2	82.8	60.9
	MMMU Pro Multimodal Understanding	80.4	77.4	83.9	81.2	75.2
	ERQA Embodied Reasoning	64.7	51.6	69.4	65.4	54.1
	SimpleVQA Visual Factuality	71.3	62.2	72.4	61.1	57.4
文本/推理	Humanity's Last Exam Multidisciplinary Reasoning (No Tools)	42.8	40.0	45.4	43.9	31.6
	Humanity's Last Exam Multidisciplinary Reasoning (With Tools)	50.4	53.1	51.4	52.1	—
	ARC AGI 2 Abstract Reasoning Puzzles (Public)	42.5	63.3	76.5	76.1	53.3
健康	HealthBench Hard Open-Ended Health Queries	42.8	14.8	20.6	40.1	20.3
	MedXpertQA (Text) Medical Multiple Choice	52.6	52.1	71.5	59.6	50.2
	MedXpertQA (MM) Medical Multiple Choice	78.4	64.8	81.3	77.1	65.8
	DeepSearchQA Agentic Search	74.8	73.7	69.7	73.6	62.8
智能体	SWE-Bench Verified Agentic Coding	77.4	80.8	80.6	—	76.7*
	SWE-Bench Pro Diverse Agentic Coding	52.4	53.4	54.2	57.7	51.8*
	Terminal-Bench 2.0 Agentic Terminal Coding	59.0	65.4	68.5	75.1	47.1*
	GDPval-AA Elo Office Tasks (Artificial Analysis)	1444	1606	1320	1672	1055

资料来源：公司资料，招银国际环球市场

Muse Spark 主要面向 Meta 的 C 端产品体系打造，包括 Meta AI、Facebook 和 Instagram 等应用，模型目前主要强调两大核心应用场景：**1）多模态：**Muse Spark 从模型训练的底层架构开始便实现跨领域、跨工具的视觉信息深度融合。它在视觉类 STEM 问题、物体识别与定位任务中均取得出色表现。这些能力共同支撑起丰富的交互式体验，例如创作趣味小游戏、或通过动态标注为家用设备排查故障等。**2）健康：**Meta 认为 C 端 AI 应用的核心场景之一便是帮助用户了解并改善自身健康状况。为提升 Muse Spark 的健康推理能力，公司与超过 1000 名医师合作，共同打造专属训练数据集，使模型能够输出更具事实依据、更全面的健康解答。Muse Spark 可生成交互式可视化内容，解读各类健康信息，例如不同食物的营养成分，或是运动过程中激活的肌肉群等。

图 25: Meta AI：基于 Muse Spark 多模态能力的健康应用场景

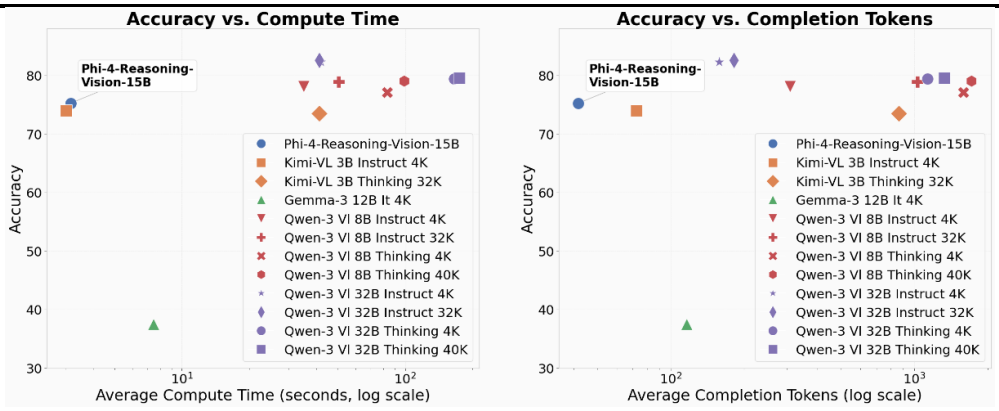


资料来源：公司资料，招银国际环球市场

微软：Phi-4 系列聚焦“小而精”模型定位

微软自研模型延续"小而精"定位，Phi-4 系列聚焦端侧部署与低延迟场景。2026 年 3 月，微软发布多模态推理模型 Phi-4-reasoning-vision（15B），逐步补齐从文本推理到多模态推理、从标准模型到 mini/flash 轻量化的能力拼图。Phi-4 系列的核心竞争力在于极高的训练与推理效率；据微软技术报告披露，Phi-4-reasoning-vision 仅使用约 2000 亿多模态 token 完成训练，数据用量约为同类竞品的五分之一，这意味着更低的训练成本和更快的迭代周期。同时，15B 的参数规模使其天然适配端侧及边缘设备部署，与微软正在推进的 Copilot+PC 硬件生态形成协同。

图 26: 微软：Phi-4-reasoning-vision-15B 以较低训练消耗达到同等数学和科学推理能力



资料来源：公司资料，招银国际环球市场

短期看，Phi-4 有望通过 Azure AI Foundry 平台吸引开发者，带动 Azure 推理算力消耗增长，同时提升微软在开源社区的品牌影响力。长期看，我们认为 Phi-4 系列的战略价值在于：一方面为 Copilot+PC 提供本地化 AI 推理能力，改善用户体验和响应速度；另一方面，小模型的高效训练范式为微软在垂直场景（如科学推理、GUI 代理等）快速定制专用模型提供了可复用的方法论。

股票推荐

1） **腾讯（700 HK；买入；目标价：750.0 港元）：**2026 年将是腾讯 AI 的关键投资年，管理层计划将 FY26 新 AI 产品的投入增至 FY25 的两倍以上（超 360 亿元）。尽管 AI 投入可能拖累短期盈利增长，但我们认为此举将强化腾讯核心业务，把握 AI 智能体等新机会，并缓解投资者对腾讯因投入不足可能在 AI 竞争中落后的担忧。AI 时代下，公司在社交/游戏/广告等赛道的竞争壁垒仍然稳固，我们仍看好公司 FY26-28 延续稳健收入增长（同比增长 10%/7%/5%），而非 IFRS 运营利润增速将在 FY26E 放缓至+6%，但在 FY27/28E 将重回加速轨道（分别+10%/8%）。FY26 我们预计公司在 AI 领域将迎来多项催化：1）大模型：腾讯 AI 团队重建后的首款旗舰大模型混元 3.0 已进入内测阶段，预计 4 月对外正式发布；2）微信 AI：公司正在微信内构建下一代智能体服务，有望进一步提升生态活跃度并创造增量营收；3）生产力 AI：公司正为其社交平台引入 WorkBuddy、Qclaw 等多款自主 AI 智能体，探索新的商业化机会。我们基于 SOTP 的目标价为 750.0 港元。“买入”评级。

2） **阿里巴巴（BABA US；买入；目标价：206.1 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。凭借强大的技术能力和丰富的人工智能应用场景，阿里巴巴有望通过“消费”与“AI +云”这两大战略业务支柱推动其长期收入和利润增长。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 206.1 美元。“买入”评级。

3） **谷歌（GOOG US；买入；目标价：396.0 美元）：**AI 驱动谷歌 2025 年搜索及云业务收入增长呈现逐季度加速趋势，4Q25 谷歌云在手订单金额同比增长超 100%至 2,400 亿美元。尽管部分投资者对公司 FY26 较快的资本开支（资本开支 1,750-1,850 亿美元，同比增长 91%-102%）及费用增长存在担忧，但我们认为在 AI 的驱动下，FY26 谷歌搜索和云业务的强劲表现将驱动公司营业利润的稳健增长，我们预计谷歌 FY26 运营利润同比增长 20%。谷歌目前已打造全球领先的全栈 AI 产品体系，覆盖 AI 基础设施+大模型+AI 开发平台+AI 智能体与应用生态。基于其全栈 AI 产品能力，我们看好谷歌多方面持续受益于 AI 发展。我们的目标价为 396.0 美元，基于 35 倍 FY26E PE。“买入”评级。

4） **Meta（META US；买入；目标价：880.0 美元）：**AI 持续赋能 Meta 广告业务增长，支撑公司利润端表现好于市场的悲观预期。管理层指引公司 FY26 总费用 1620 亿-1690 亿美元（同比增长 38%-44%），但同时预计 FY26 营业利润仍将维持正增长，隐含 FY26 总营收同比增速将超 22%，AI 对于广告业务的赋能将在 FY26 延续，缓解部分投资者对于 FY26 盈利增长的担忧。公司当前现价仅对应约 19x FY26E PE，相较谷歌 26x 仍有明显折让，收入增长加速、盈利确定性提升、叠加新模型 Muse Spark 带来增量催化，有望支撑公司估值修复。

5） **微软（MSFT US；买入；目标价：614.6 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（同比 + 110%；若剔除 OpenAI 贡献则同比 + 28%）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。我们基于现金流折现模型的目标价为 614.6 美元，对应 35x/31x FY26/FY27 财年市盈率。“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。

如需索取更多有关证券的信息，请与我们联络。

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。