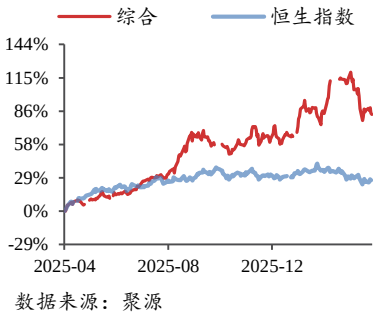


综合

2026 年 04 月 06 日

投资评级：看好（维持）

行业走势图



相关研究报告

《OpenClaw 生态持续深化发展，国产大模型全球市占率再攀新高——行业周报》-2026.3.29

《AI Token 消耗指数级增长，云计算行业提价——行业周报》-2026.3.22

《国内大厂抢占市场，“龙虾热”引爆国产模型增量红利——行业周报》-2026.3.15

Agent 驱动模型调用进入加速拐点，中国 AI 军团定义智能时代新范式

——港股行业深度报告

初敏（分析师）

chumin@kysec.cn

证书编号：S0790522080008

杨哲（分析师）

yangzhe@kysec.cn

证书编号：S0790524100001

● AI 市场：Agent 有望驱动新 token 一轮，国产模型开始 S 型跃迁

市场规模：根据 Gartner 数据，2025 年全球 AI 产业规模 1.8 万亿美元，其中 AI 模型占比 1%，未来 2 年 CAGR 为 74%；根据 IDC 数据，活跃 Agent 数量将从 2025 年的约 2860 万快速增长至 2030 年的 22.16 亿，潜在空间较大。**市场格局：**top3 份额超 50%，国产大模型成为新的搅局者，2 月中旬 MiniMax M2.5 周 tokens 消耗量位列 OpenRouter 榜单第一。在实际场景方案解决、长上下文学习上主流大模型仍有较大提升空间，AI 的下半场评估或大于训练，行业正从“大而全”到“精而准”的范式转移。

● OpenAI vs Anthropic：B 端+Coding 模式的增长动能、算力利用效率更优
增长：2023-2025 年 AnthropicARR 基本保持每年 10 倍速度增长，OpenAI 在 3 倍左右，2026 年 2 月 OpenAI 实现 ARR 250 亿美元，3 月初 Anthropic ARR 达到 200 亿美元，Anthropic 在持续加速追赶 OpenAI；**盈利能力：**二者毛利率中长期展望 70+%，推理算力超预期导致毛利率展望下调，2024 年 OpenAI 训练（研发费用）+运营（推理成本/营运成本）带来的算力成本占到收入的 125%，且 OpenAI 算力投入明显高于 Anthropic，到 2028 年 Anthropic 的收入体量有望接近 OpenAI，而算力投入只需有 OpenAI 的 1/3 左右。

● B/C 端：C 端春节红利下大厂加速渗透，B 端国产模型跻身第一梯队

C 端：ChatGPT 的月访问量一直稳居 55 亿次以上，明显领先其他竞品，Gemini 依托多模态能力持续追赶。中国 AI 大模型市场快速增长，企业级市场千问杀出重围登顶第一，2025H2 市场份额 32%。春节红包红利与 Agent 深度集成驱动月活抬升、阿里、字节、腾讯依托各自资源禀赋打法各异；**B 端：**API 调用量市场集中度高，Anthropic 占据半壁江山，国产大模型企业逐步跻身第一梯队，当前格局有望松动，依托“开源+成本”优势，国产模型有望维持长期竞争力；**估值：**Anthropic 每次估值更新基本是前次的 2-3 倍，但更为快速的 ARR 增长成为消化估值的基础，在早期（2023 年 5 月）业务基本处于 0-1 的过程，当年估值达到 46 倍，2025 年以来 Anthropic 的当年估值最高位 20+倍，且通常次年估值会降到 10 倍以内。对应到国内大模型，以智谱为例，若以 Anthropic 的估值为锚，则智谱 2025-2027 年收入 CAGR 需达到 176%。考虑到 agent 带动的模型需求并非简单的线性增长，国产大模型的 ARR 增长有望充分消化当前的估值水平。

● 投资建议：看好 Agent 指数跃迁背景下模型、AI Infra、算力端的投资机会
OpenClaw 加速了 Agent 的渗透能力，国产大模型能力已逐步达到全球第一梯队水平，受益标的有阿里巴巴-W、百度集团-SW、MiniMax-W、智谱，AI Infra 层的受益标的第四范式、百融云-W，算力端的受益标的金山云、万国数据等。

● 风险提示：模型迭代不及预期、行业竞争加剧、agent 落地不及预期。

目 录

1、 AI 市场：从“线性对话”到“指数执行”，Agent 驱动模型调用进入加速奇点	4
1.1、 市场规模：AI 产业快速增长，Agent 时代有望带动 tokens 消耗数量级跃迁	4
1.2、 市场格局：海外龙头领跑市场，国产模型成为新的搅局者	5
1.3、 国内：企业级模型市场高速增长，千问市占率节节攀升	9
2、 OpenAI vs Anthropic：B 端+Coding 模式的增长动能更为强劲，算力利用效率亦是更优	11
3、 C 端（Chatbot 为主）：ChatGPT 月活明显领先，春节红利国内大厂依托资源禀赋打法各异	16
4、 B 端（AI Coding/Agent）：行业或将进入加速拐点，国产模型跻身第一梯队	17
4.1、 市场分析：Anthropic 主导市场，国产大模型逆袭而上	17
4.2、 Anthropic：ARR 增速远高于估值增速，当年估值通常在 20 倍左右	22
5、 国产初创大模型：新厂商性能达到 opus 4.5 水平，点燃 Agent 需求想象空间	24
6、 投资建议：	30
7、 风险提示：	31

图表目录

图 1： 根据 Gartner 报告，2025 年全球 AI 总支出达 1.8 万亿美元	4
图 2： 2025 年在于大模型厂直接相关的市场中，纯 AI 模型市场占比仍然较小	5
图 3： AI 模型收入预计将保持快速增长（单位：百万美元）	5
图 4： 活跃 Agent 将呈现指数级上升	5
图 5： 预计 2029 年大模型市场规模 2065 亿美元（单位：十亿美元）	6
图 6： 2024 年各模型公司市占率	6
图 7： MiniMax 和 Kimi 领跑全模型，带动 AI 大模型市场飞速增长	6
图 8： OpenClaw 平台 token 日消耗量翻倍，国产模型占领前三名	7
图 9： 全球 AI 大模型周 token 消耗量超 7 万亿	8
图 10： 全球开源模型中 DeepSeek 市占率逐步下降	8
图 11： 2025H2 中国开源模型占比过半	9
图 12： 2025H2 中国开源模型占全球开源模型超 90%	9
图 13： 中国 AI 大模型市场快速增长	9
图 14： 2025H1 千问成为中国企业级大模型市场龙头	10
图 15： 2025H2 中国企业级大模型市场千问增长最快	10
图 16： 2025H2 中国企业级大模型市场较 2025H1 增长近 3 倍（单位：万亿 token）	10
图 17： AI 应用全球月活榜上，ChatGPT 明显领先（2026 年 2 月）	11
图 18： 国内榜单上，豆包领先市场，千问快速增长（2026 年 2 月）	11
图 19： 2026Q1 OpenAI 对于 2029-2030 的收入展望大幅上调（亿美元）	11
图 20： 2026Q1 OpenAI 最新收入展望继续与 Anthropic 拉开差距（亿美元）	11
图 21： Anthropic ARR 增速明显高于 OpenAI	12
图 22： Anthropic 的 ARR 几乎维持每年 10 倍的增长（图中数据为当年 ARR/上一年度 ARR 的倍数）	12
图 23： 2024 年秋季 OpenAI 展望未来营收加速增长（单位：亿美元）	13
图 24： 2026 年 OpenAI 展望未来 C 端贡献占据营收主体（单位：亿美元）	13
图 25： OpenAI 月度毛利率企稳约 70%	14
图 26： OpenAI 未来毛利率展望有所下调	14
图 27： OpenAI 和 Anthropic 预计未来毛利率逐步上升	14
图 28： 从 2024 年 OpenAI 的盈利结构俩看，算力为最主要的亏损来源	15

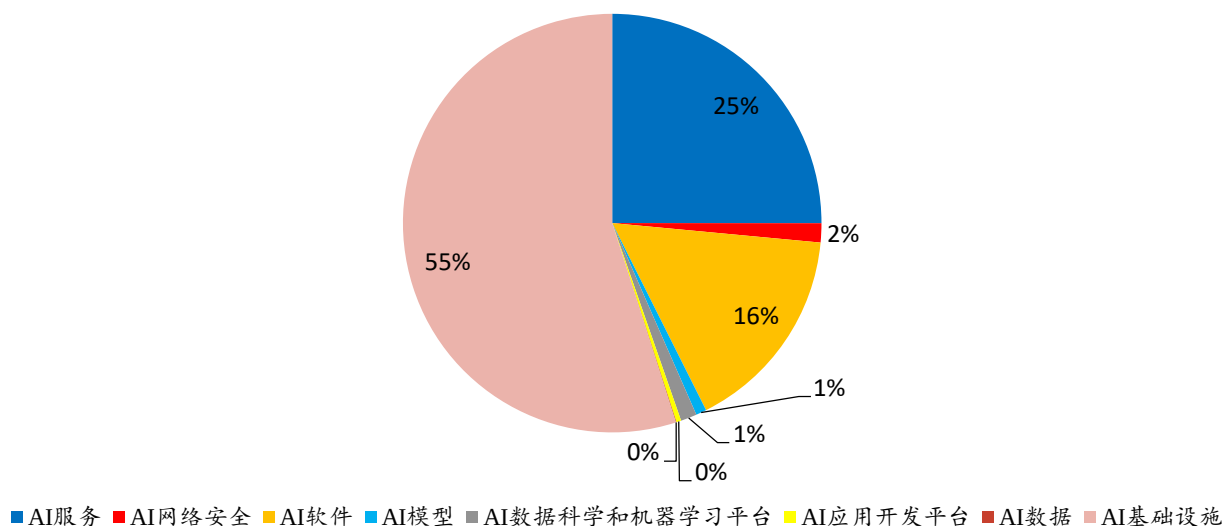
图 29: OpenAI 服务器成本预计到 2028 年查过 1000 亿美元 (单位: 亿美元)	15
图 30: Anthropic 服务器成本预计到 2028 年为 270 亿美元 (单位: 亿美元)	15
图 31: Anthropic 的运用效率优于 OpenAI (数据: 服务器投入/收入)	15
图 32: OpenAI 现金流转正时间推到 2030 年 (单位: 亿美元)	16
图 33: Anthropic 现金流转正时间快于 OpenAI (单位: 亿美元)	16
图 34: Chatbot 访问量 ChatGPT 领跑全模型, Gemini 紧随其后 (单位: 十亿次)	16
图 35: 企业大模型支出快速增长	18
图 36: Anthropic 的 API 调用量份额快速提升	18
图 37: 2025 年中期 Anthropic Coding 市场份额为 42%	18
图 38: 2025 年底 Anthropic 的 Coding 市场份额为 54%	18
图 39: 企业持续追求最前沿的模型	19
图 40: 2026 年在软件工程任务上, 已有模型能在半小时内达到专业程序员 80% 的水平	19
图 41: OpenRouter (2026.02.12-2026.02.22) 大模型排名上, 前三为国产大模型	20
图 42: 2025 年底到 2026 年初, 开源模型相较闭源模型的差距进一步缩小	21
图 43: 国产大模型较海外模型在成本上具有显著竞争力 (单位: 美元/百万 token)	21
图 44: OpenAI 收入构成以 chatbot (C 端) 为主	22
图 45: AnthropicB 端收入占比 80% 以上	22
图 46: Anthropic 估值快速提升	23
图 47: AnthropicARR 持续提升	23
图 48: Anthropic 收入预期持续上调 (单位: 亿美元)	23
图 49: 国产大语言模型 SWE-bench Verified 测试 MiniMax M2.5 表现最好	26
图 50: 国产大语言模型 HLE 测试 Seed 2.0 Pro 表现最好	26
图 51: 国产大语言模型 BrowseComp 测试 Qwen 3.5 表现最好	26
图 52: 国产大语言模型 Step 3.5 Flash 输出成本最低 (单位: 人民币元)	27
图 53: 国产大语言模型 Step 3.5 Pro 输出速率最快	27
表 1: 腾讯、字节、阿里在 C 端战略打法有所区别	17
表 2: 2025 年国内外大语言模型企业人均收入对比 (OpenAI、智谱、Moonshot 为预测数据)	22
表 3: P/ARR 口径下 Anthropic 估值逐步消化	23
表 4: PS 口径下 Anthropic 估值逐步消化	24
表 5: 国产大模型近期更新	24
表 6: 国产大语言模型参数对比	25
表 7: 从行业主要的测评上看, Vibe Coding 逐渐转向 Agentic Engineering	27
表 8: ApexAgent 考察了投行、法务、咨询管理的实际任务处理能力	29
表 9: 主要模型的通过率仍然较低	29
表 10: 根据 CL-bench 排名, 当前大模型的上下文学习能力	30
表 11: 主要标的盈利预测及估值	31

1、AI 市场：从“线性对话”到“指数执行”，Agent 驱动模型调用进入加速奇点

1.1、市场规模：AI 产业快速增长，Agent 时代有望带动 tokens 消耗数量级跃迁

2025 年全球 AI 产业规模 1.8 万亿美元，其中 AI 模型规模 144 亿美元，占比 1%：
根据 Gartner 在 2026 年 1 月的数据显示，2025 年全球 AI 相关产业市场规模在 1.8 万亿美元，其中最大的项目为 AI 基础设施（占比 55%），AI 大模型相关的市场包括 AI 模型 144 亿美元（占比 1%）、AI 服务 4394 亿美元（占比 25%）、AI 软件 2831 亿美元（占比 16%）。

图1：根据 Gartner 报告，2025 年全球 AI 总支出达 1.8 万亿美元

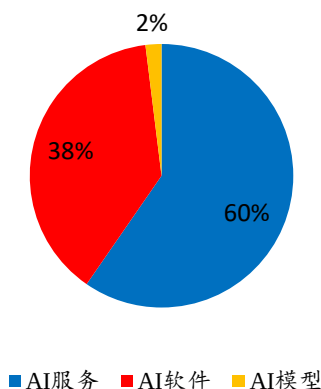


数据来源：Gartner、开源证券研究所

若仅仅聚焦于大模型厂商密切度更高的 AI 模型（订阅费、API 调用）、AI 服务（落地实施服务等）、AI 软件（如 Agent、AI 开发平台、套件等）：

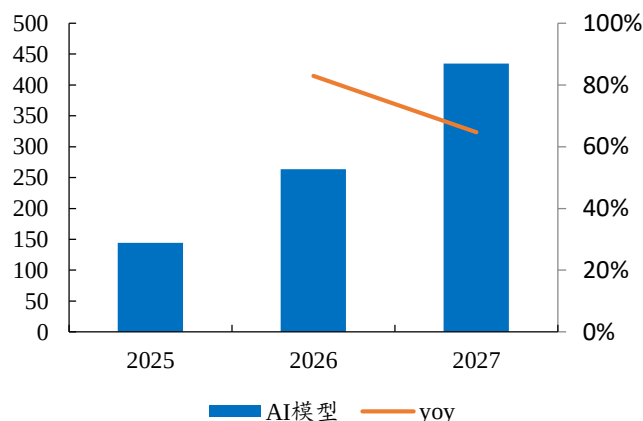
- 收入增长上：**各业务仍然保持快速增长，其中 AI 模型收入有望从 2025 年的 144 亿美元增长至 2027 年的 434 亿美元，CAGR 为 74%；
- 各企业业务构成上：**我们认为不同企业的业务构成差距较大，例如智谱可能具有较多比例的 AI 服务收入，MiniMax 的海螺、星野等带来较多 AI 软件业务，而头部的 OpenAI、Anthropic 当前收入构成中主要以订阅费、API 调用为主；
- 模型即应用：**自 Claude Cowork 发布以来，模型即应用的演绎趋势开始逐步加强，尽管从绝对值上看 AI 模型相对整个 AI 市场较低，但是模型厂商向 AI 软件渗透所带来的空间十分可观。

图2：2025 年在大模型厂直接相关的市场中，纯 AI 模型市场占比仍然较小



数据来源：Gartner、开源证券研究所

图3：AI 模型收入预计将保持快速增长（单位：百万美元）



数据来源：Gartner、开源证券研究所

未来 5 年全球活跃 Agent 将呈现指数级增长，token 消耗量呈现数量级跃迁。根据 IDC 预测，活跃 Agent 数量将从 2025 年的约 2860 万快速增长至 2030 年的 22.16 亿，Agents 年度执行量将从 2025 年的 440 亿次攀升至 2030 年的 415 万亿次，年度 token 消耗量将从 2025 年的 0.0005PetaTokens 暴增至 2030 年的 152,667 PetaTokens。相较于个人的 AI chatbot 使用，Agent 所带来的 tokens 消耗量将高出多个数量级，有望成为下一阶段 AI 需求的重要力量。

图4：活跃 Agent 将呈现指数级上升



数据来源：IDC

1.2、市场格局：海外龙头领跑市场，国产模型成为新的搅局者

根据 MiniMax 招股说明书，2024 年大模型市场规模达到 10.7 亿美元，其中应用规模为 7.1 亿美元，占比 66.36%，我们估计 2025 年又有明显增长。应用类主要包括如 ChatGPT、DeepSeek、豆包、Sora 等等应用，用户可以直接利用应用解决特定任

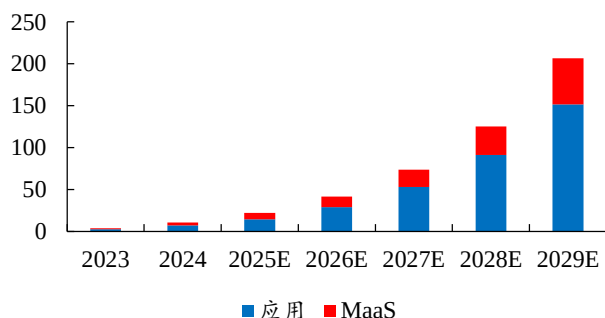
请务必参阅正文后面的信息披露和法律声明

5 / 33

务，该类型的收费模式以订阅制为主。MaaS 类是通过 API 接口或云端工具将大模型集成到用户需要的软件中，大幅降低用户的使用成本。根据 MiniMax 招股说明书，2029 年大模型市场规模有望达到 2065 亿美元，其中应用类市场规模为 1515 亿美元，占比 73.37%。

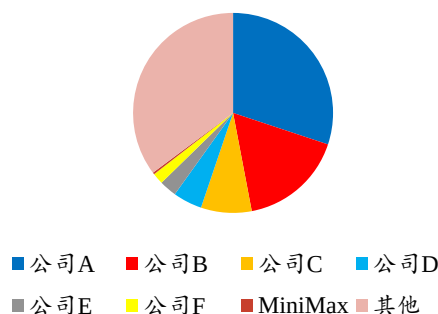
大模型市场分化严重，规模前三企业占据超 50% 市场份额。根据 MiniMax 招股说明书，截至 2024 年末，公司 A、公司 B、公司 C 市占率分别为 30.1%、16.9%和 8.2%，共计占比 55.2%。公司 A 断层领先其他模型，除前三家企业外其余公司模型市占率均不足 5%。

图5：预计 2029 年大模型市场规模 2065 亿美元（单位：十亿美元）



数据来源：MiniMax 招股说明书、开源证券研究所

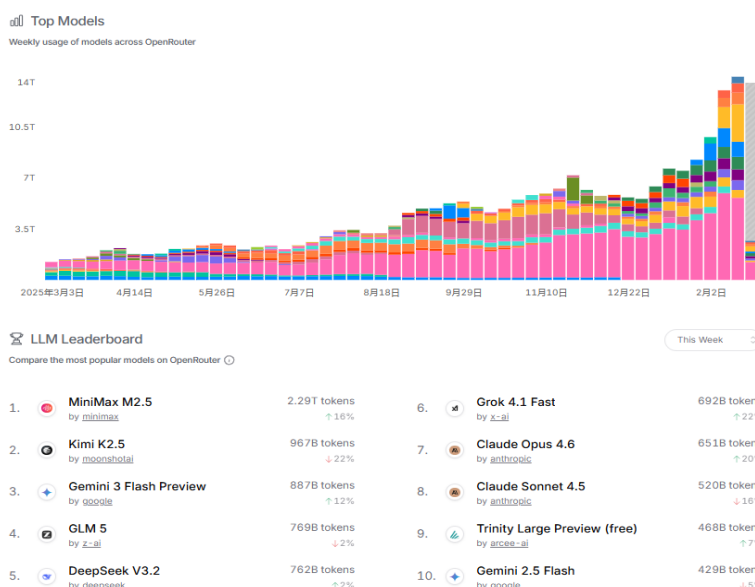
图6：2024 年各模型公司市占率



数据来源：MiniMax 招股说明书、开源证券研究所

AI 大模型市场跨入飞速增长期，全球各大模型周 token 消耗量总额已过十万亿。得益于 Gemini3.1 Pro、GLM-5、MiniMax2.5、Claude Sonnet 4.6 相继推出，AI 大模型市场迎来增量热潮。截至 2026 年 2 月 16 日，所有模型周总 token 消耗量达到 14 万亿，创下历史新高，AI 大模型市场正向“千万亿级”时代迈进。其中，MiniMax 周 token 消耗量达到 2.29 万亿，位列榜单第一。

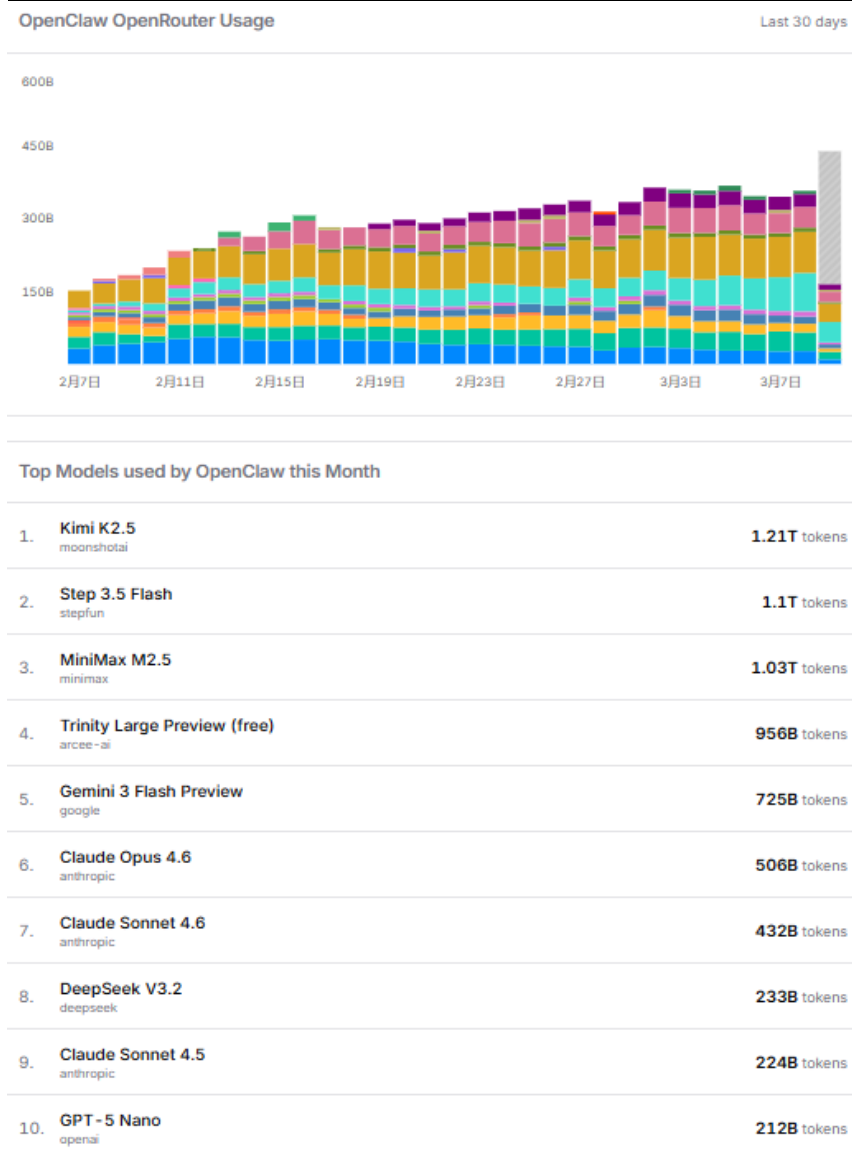
图7：MiniMax 和 Kimi 领跑全模型，带动 AI 大模型市场飞速增长



数据来源：OpenRouter

各大模型纷纷对接 OpenClaw，日 token 消耗量翻倍。春节前后，中国大语言模型纷纷对接 OpenClaw，推出自有的 AI 集成平台。Kimi K2.5 最先推出 Kimi Claw，并迅速占领市场。在 OpenClaw 平台 token 消耗量中，Kimi 日 token 消耗量约占当日总额的 20%，最高时可达 25%。随后，MiniMax M2.5 和 Claude 相继接入 OpenClaw，不到半月就将 OpenClaw OpenRouter 平台的总 token 消耗量从 150B/天增长至 300B/天。其中，MiniMax 与 OpenClaw 结合的 MaxClaw 在中国取得良好成果，日 token 消耗量稳定在总消耗量的 15% 左右。后起之秀阶跃星辰的 Step 模型则在 token 消耗量对比中也占据较高份额。

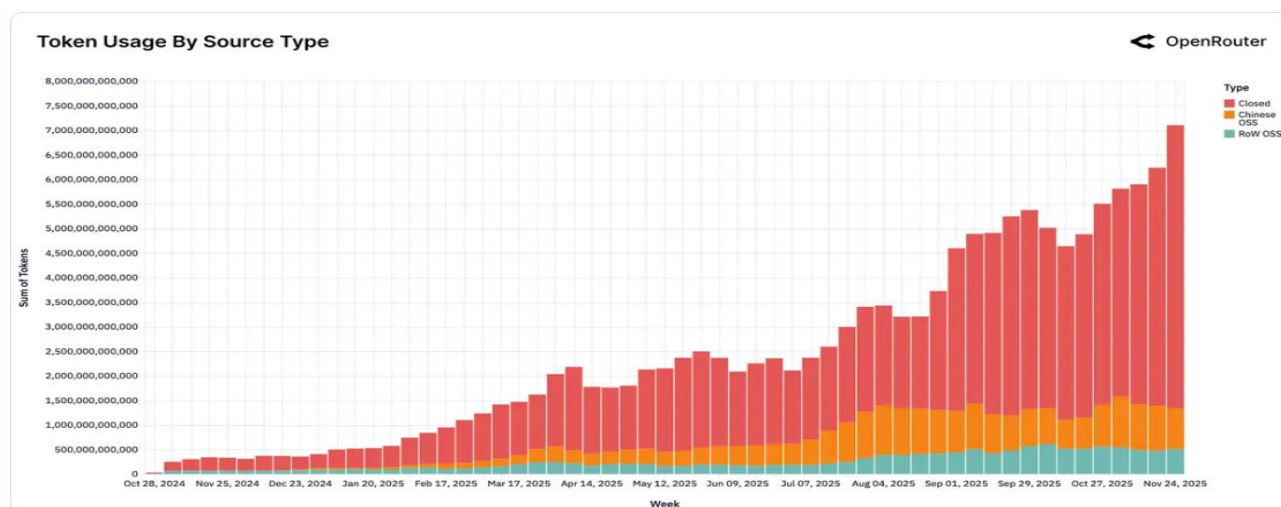
图8：OpenClaw 平台 token 日消耗量翻倍，国产模型占领前三名



数据来源：OpenRouter

AI 大模型 token 使用量快速增长，中国开源模型贡献较大。根据 OpenRouter，截至 2025 年 11 月 24 日，全球 AI 大模型市场周 token 消耗量已超 7 万亿。其中，闭源模型平均占比约 70%，开源模型仅为 30%。中国开源大模型从 2024 年底的 1.2% 上升至约 13%，某些周内甚至达到所有模型总使用率的近 30%。

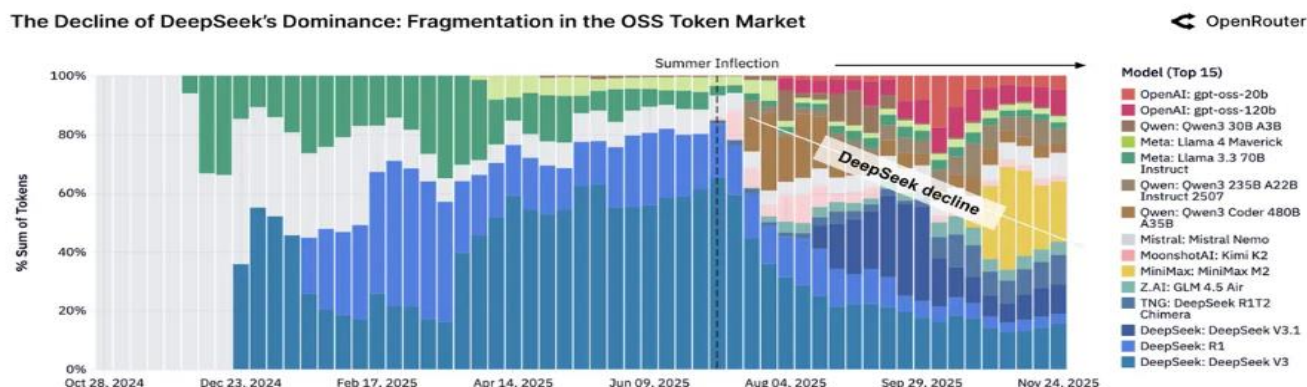
图9：全球 AI 大模型周 token 消耗量超 7 万亿



数据来源：OpenRouter

DeepSeek 仍为开源模型中 token 消耗量贡献最大的模型，但其主导地位已被削弱。自 2025 年 7 月以来，AI 大模型市场涌入大量新兴竞争者如千问、Moonshot、OpenAI、MiniMax、智谱等模型，DeepSeek 的市场份额随之下降，占比约 40%。对于 DeepSeek 而言，在开源领域重要竞争者包括 OpenAI、MiniMax、千问等。

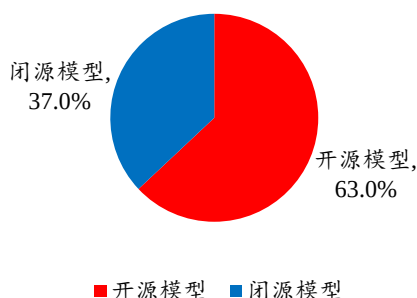
图10：全球开源模型中 DeepSeek 市占率逐步下降



数据来源：OpenRouter

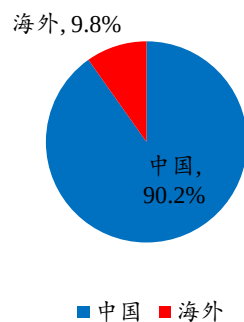
中国开源模型紧随海外 AI 大模型步伐，进一步强化开源模型市场主导地位。截至 2025H2，中国开源模型市占率已达全球开源模型市场的 90.2%，显著领先海外模型，占据全球开源模型市场主导地位。而在中国国内，开源模型发展程度也超过闭源模型，市占率达 63%。

图11：2025H2 中国开源模型占比过半



数据来源：沙利文、开源证券研究所

图12：2025H2 中国开源模型占全球开源模型超 90%

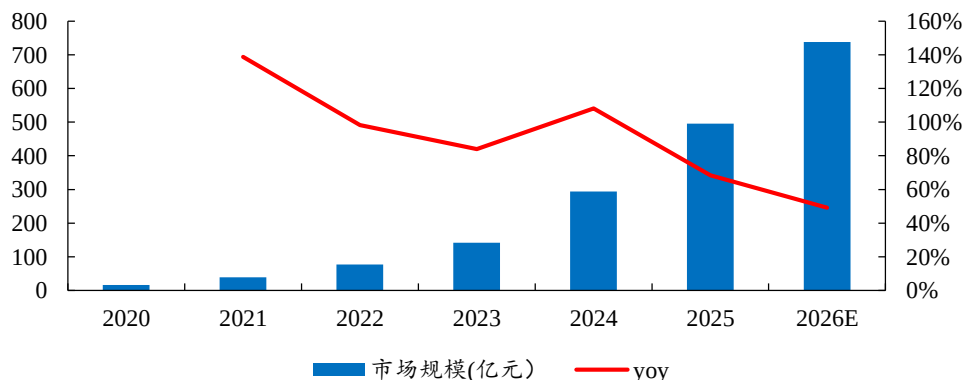


数据来源：沙利文、开源证券研究所

1.3、国内：企业级模型市场高速增长，千问市占率节节攀升

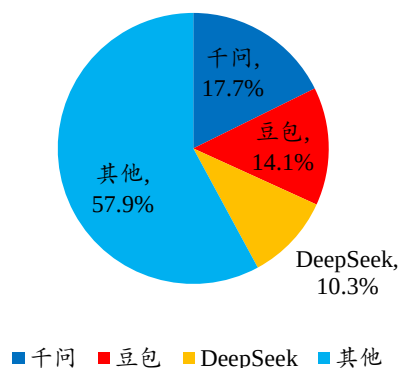
中国 AI 大模型市场快速增长，企业级市场千问杀出重围登顶第一。2025 年中国 AI 大模型市场规模接近 500 亿元,2022-2025 年的 CAGR 为 86%,保持快速增长。2025 年上半年 API 成本大幅下降，企业开始初步接入模型。2025H2 中国企业级大模型日均调用量 37.0 万亿 tokens，较 2025H1 增长 263%。随着下半年 AI 模型频繁更新功能，各类自动化智能体与流程机器人逐步推广，模型调用进入高速增长期，于年底强势拉升，12 月日均 token 调用量达 49.0 万亿。其中，千问大模型调用量占比 32.1%，较上半年与豆包的差距逐步拉大。DeepSeek 先发优势逐渐减少，日均调用量占比 18.4%，排名第三。

图13：中国 AI 大模型市场快速增长



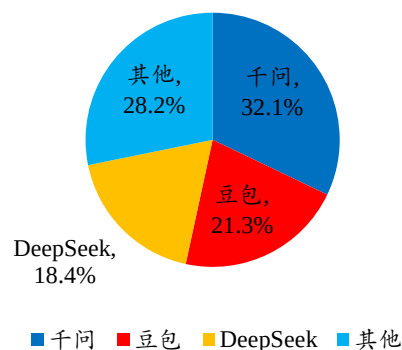
数据来源：艾瑞咨询、开源证券研究所

图14: 2025H1 千问成为中国企业级大模型市场龙头



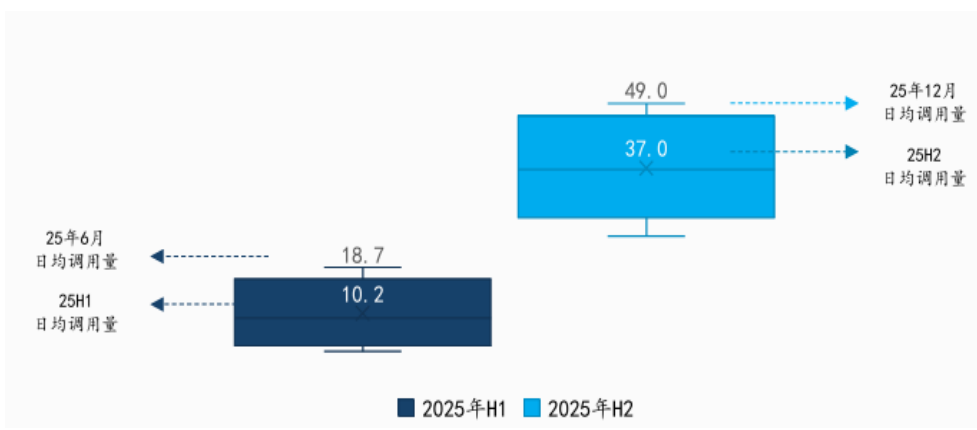
数据来源: 艾瑞咨询、开源证券研究所

图15: 2025H2 中国企业级大模型市场千问增长最快



数据来源: 沙利文、开源证券研究所

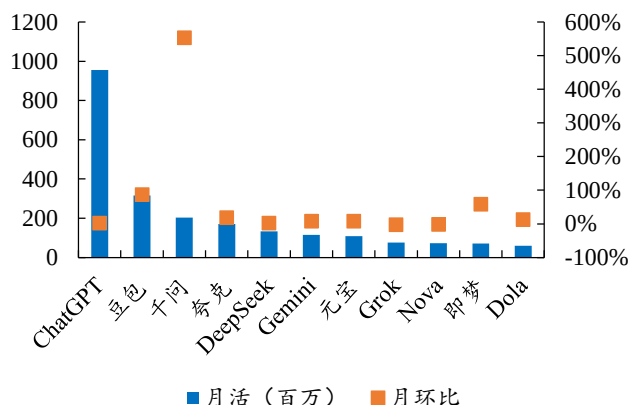
图16: 2025H2 中国企业级大模型市场较 2025H1 增长近 3 倍 (单位: 万亿 token)



数据来源: 沙利文

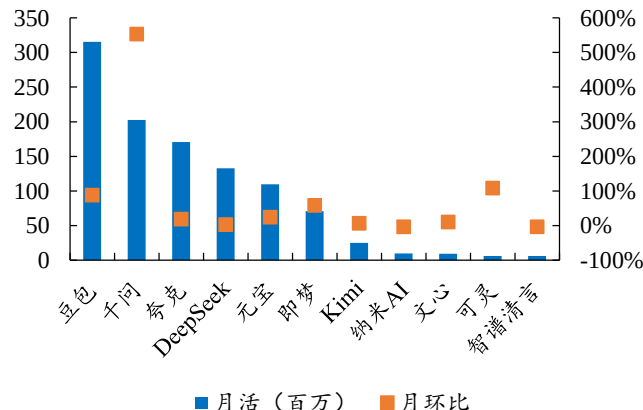
中国 AI 大模型市场在 2026 年初迎来新一轮催化，春节红包红利与 Agent 深度集成驱动月活抬升。从全球月活榜单上看，截至 2026 年 2 月，ChatGPT 拥有 9.5 亿月活用户，明显领先其他 AI 应用，但诸如海外的 Gemini、Grok，以及国内的豆包、千问等大厂产品仍有积极的追赶之势。尤其在 2026 年春节期间，各大厂投入超 45 亿元开启“红包大战”，豆包凭借春晚独家互动优势，除夕当天 AI 互动量冲至 19 亿次；千问大模型因深度接入本地生活闭环并实现近 2 亿次“一句话下单”，全国平均每 10 人就有 1 人在千问下单。根据市场研究机构 QuestMobile 的数据，春节红包活动开启后，千问的 DAU 日活在 2 月 6 日出现明显增长，单日上涨 727.7%，达到 5848 万；腾讯元宝宣布日活跃用户突破 5000 万、月活超过 1.14 亿。

图17: AI 应用全球月活榜上, ChatGPT 明显领先 (2026 年 2 月)



数据来源: aipcb.com、开源证券研究所

图18: 国内榜单上, 豆包领先市场, 千问快速增长 (2026 年 2 月)

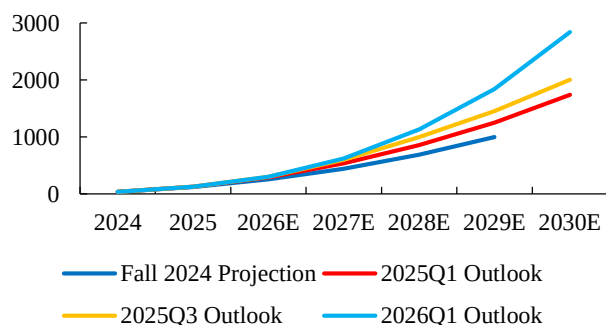


数据来源: aipcb.com、开源证券研究所

2、OpenAI vs Anthropic: B 端+Coding 模式的增长动能更为强劲, 算力利用效率亦是更优

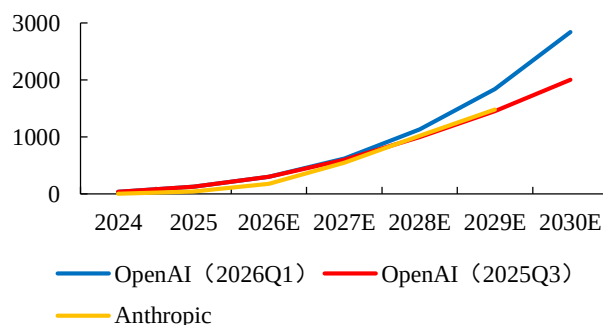
中期维度 Anthropic 与 OpenAI 的收入差距持续收窄, OpenAI 上调远期收入展望: 根据 The Information, 当前阶段 OpenAI 远期收入展望 2029-2030 高于 Anthropic, 而中期 (2026-2028 年) Anthropic 与 OpenAI 收入差距持续收窄, 而在 2025 年下半年的展望中, Anthropic 收入有望在 2028 年实现 OpenAI 的反超。在 2026 年 2 月 OpenAI 最新的展望中, 其 2029-2030 年收入较前次大幅上调, 当前 2030 年最新收入展望为 2840 亿美元 (较前次+42%), 进一步拉开与 Anthropic 的差距。

图19: 2026Q1 OpenAI 对于 2029-2030 的收入展望大幅上调 (亿美元)



数据来源: The Information、开源证券研究所

图20: 2026Q1 OpenAI 最新收入展望继续与 Anthropic 拉开差距 (亿美元)

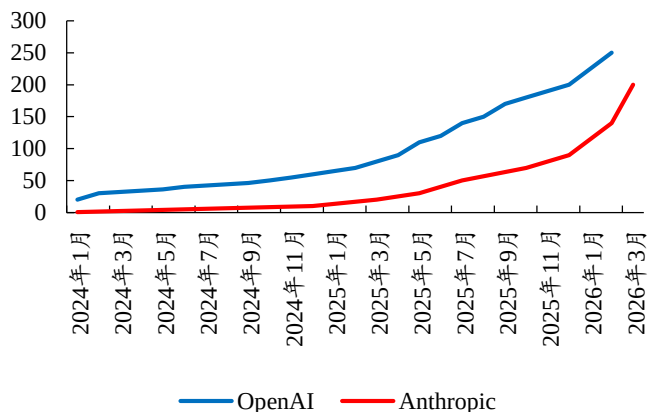


数据来源: The Information、开源证券研究所

从 ARR 角度看 Anthropic 对 OpenAI 的追赶或更为直观: 2023-2025 年 Anthropic 的 ARR 基本保持每年 10 倍的速度增长, 而 OpenAI 在 3 倍左右, 截至 2025 年 12 月 OpenAI/Anthropic 的 ARR 为 200/90 亿美元, 而得益于 Claude Code 及其 Cowork 插件驱动, 2026 年 2 月 Anthropic 的 ARR 进一步抬升至 140 亿美元 (其中 Claude Code 为 25 亿美元, 占比 18%), 较 2025 年底进一步提升 56%, 而 3 月初, 或因 Anthropic

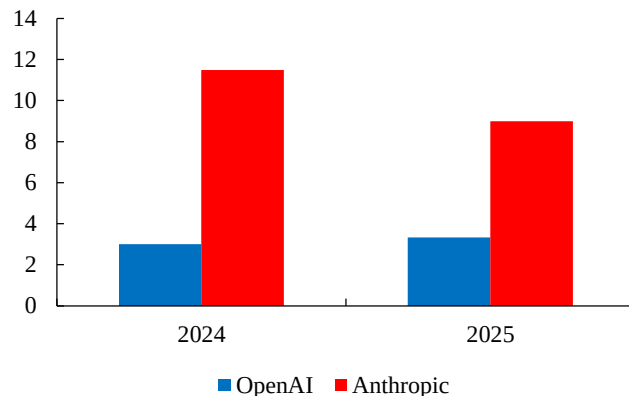
禁止五角大楼将其 AI 用来规划动能打击而被美国政府贴上“供应链风险”的标签，受到用户支持，Anthropic ARR 进一步提升至 200 亿美元，增长仍未见降速。

图21: Anthropic ARR 增速明显高于 OpenAI



数据来源: The Information、开源证券研究所

图22: Anthropic 的 ARR 几乎维持每年 10 倍的增长(图中数据为当年 ARR/上一年度 ARR 的倍数)



数据来源: The Information、开源证券研究所

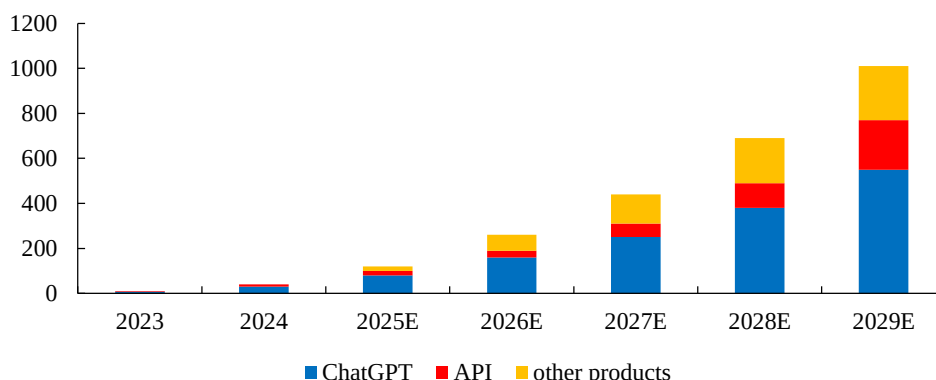
由于 Anthropic 80+% 的收入为 B 端的 API 收入，我们重点分析 OpenAI 的收入构成，通过对比 2024 与 2026 年 OpenAI 对收入结构的展望，得到如下结论：

C 端：C 端仍然是最重要的收入来源，2025 年占比 69%。但除了订阅费，OpenAI 在 2025 年推出了广告业务，预计到 2030 年将成为收入的重要构成部分。而 Anthropic 则明确提出 Claude 将保持无广告状态，以避免广告对用户的参与度影响，但这本质是 C 端与 C 端用户画像区别所导致的商业模式区别，我们认为并无高下之分；

B 端：相较 2024 年的收入拆分版本，OpenAI 单独强调了 agent 带来的 B2B 业务，2025 年 B2B/平台收入占比分别为 15%/15%，整体 B 端合计占比 30%，对应收入绝对值为 40 亿美元左右，Anthropic B 端 API 调用体量相当；

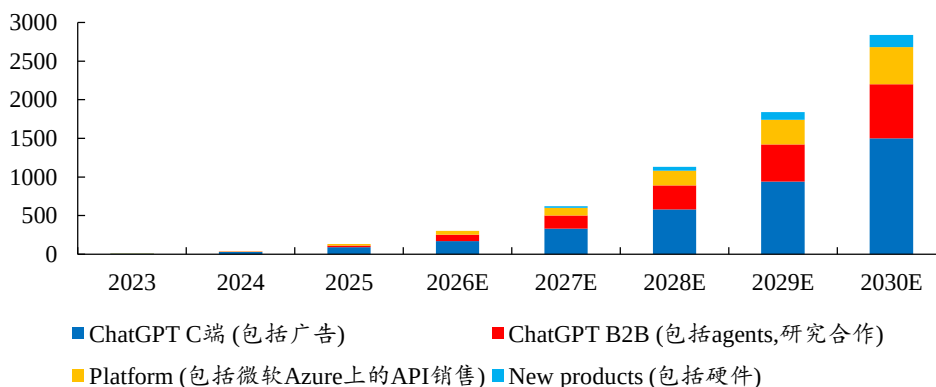
新产品：包括硬件产品，在收入展望中，新产品或在 2027 年才贡献收入。

图23：2024 年秋季 OpenAI 展望未来营收加速增长（单位：亿美元）



数据来源：The Information、开源证券研究所

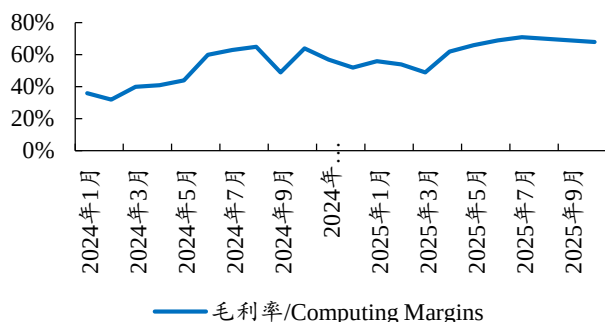
图24：2026 年 OpenAI 展望未来 C 端贡献占据营收主体（单位：亿美元）



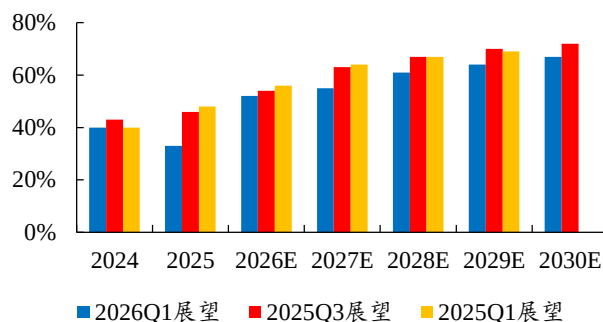
数据来源：The Information、开源证券研究所

毛利率中长期展望 70+%，推理算力成本超预期导致毛利率展望下调：OpenAI 与 Anthropic 的毛利率均在逐年提升，预计在 2026 年分别达到 55%/70%，中长期均展望 70+%水平，但另一方面，二者盈利能力改善程度亦在逐步推后，以 OpenAI 为例，从 2025 年至 2026 年，其对未来的毛利率展望基本在持续下调，且 2025 年因推理算力成本超预期，毛利率从原来的 46%下调至 33%，Anthropic 也从原本的 50%下调至 40%

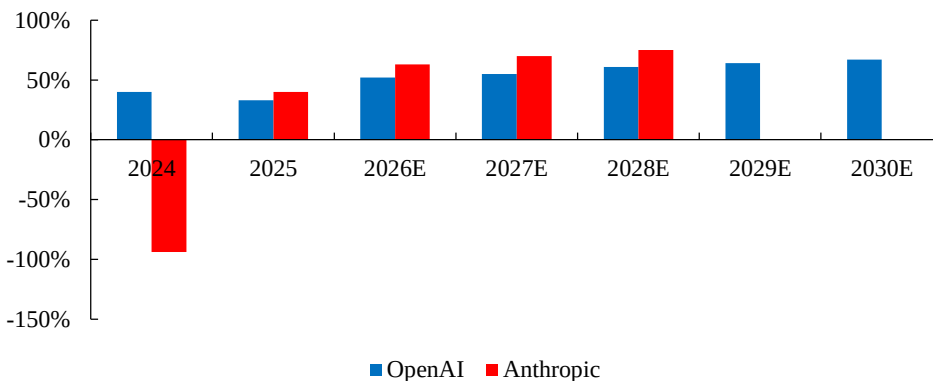
注：OpenAI 与 Anthropic 毛利率较大的差距系口径关系，Anthropic 仅考虑付费用户的毛利，而 OpenAI 的毛利计算还包括了免费用户。

图25: OpenAI 月度毛利率企稳约 70%


数据来源: The Information、开源证券研究所

图26: OpenAI 未来毛利率展望有所下调


数据来源: The Information、开源证券研究所

图27: OpenAI 和 Anthropic 预计未来毛利率逐步上升


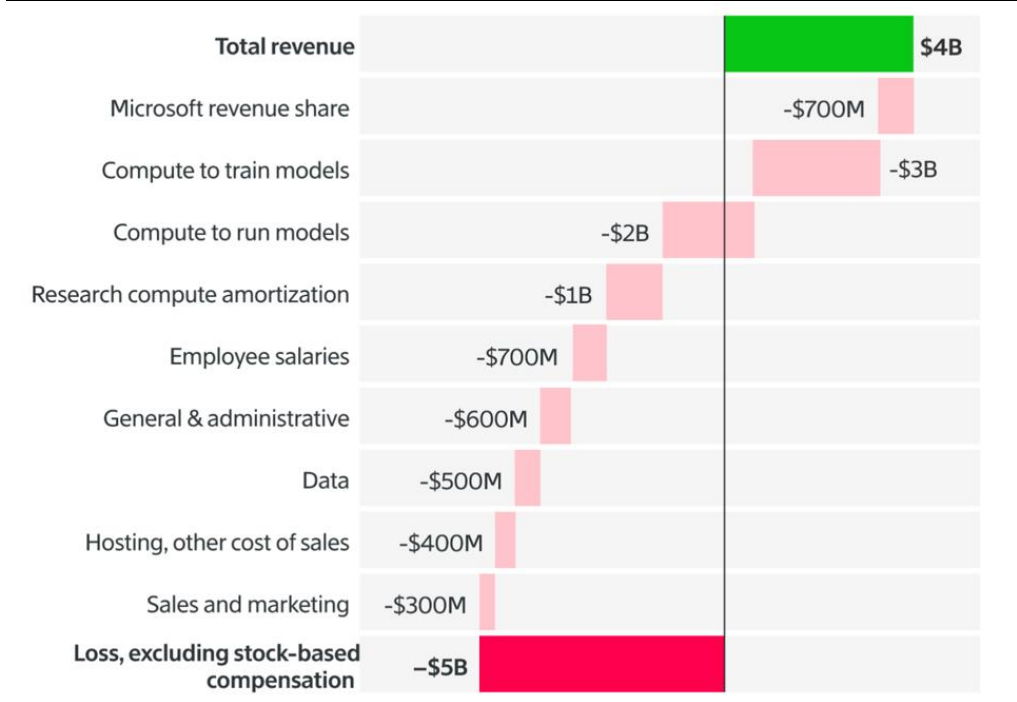
数据来源: The Information、开源证券研究所

算力成本是最主要的损耗来源, Anthropic 的算力利用效率优于 OpenAI: 以 OpenAI 2024 年的盈利拆解为例, 算力成本为最主要的费用来源, 在经营过程中, 训练 (研发费用) + 运营 (推理成本/营运成本) 带来的算力成本就占到收入的 125%, 此外还有 25% 的算力固定成本摊销。而 OpenAI 的算力成本投入明显高于 Anthropic, 我们以服务器投入/收入来衡量, 根据各个公司的展望数据, 2026-2028 年, OpenAI 该值将从 135% 降到 112%, 而 Anthropic 将从 88% 将至 40%, 即到 2028 年, Anthropic 的收入体量有望接近 OpenAI, 而其算力投入只需有 OpenAI 的 1/3 左右。

造成这样的结果可能有几方面原因:

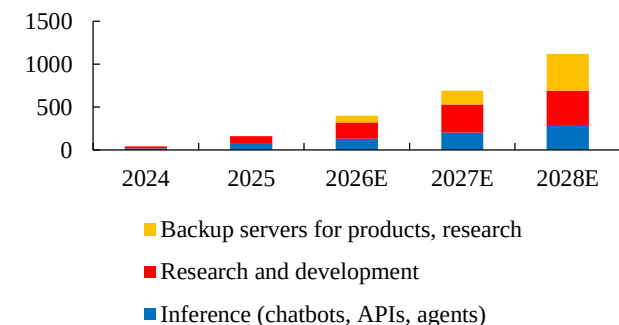
- 1、OpenAI 要展开更广泛的研究, 包括多模态、硬件、搜索等;
- 2、OpenAI 计划在备用服务器上投入巨资, 以确保在出现意外突破或爆款产品时不会出现计算能力不足的情况;
- 3、用户结构不同, Anthropic 多为 B 端企业, 免费用户的维持费用相对较小, 而 OpenAI 有着较大体量的免费用户 (付费率仅 5%)。

图28：从 2024 年 OpenAI 的盈利结构俩看，算力为最主要的亏损来源



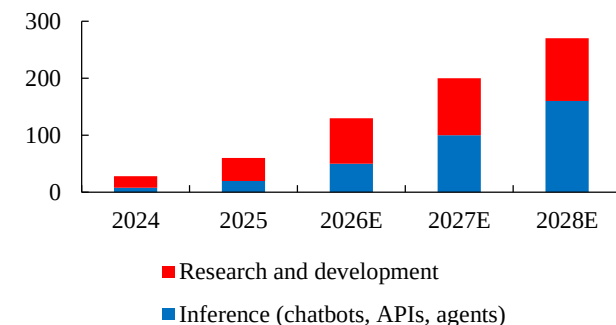
数据来源：The Information

图29：OpenAI 服务器成本预计到 2028 年查过 1000 亿美元（单位：亿美元）



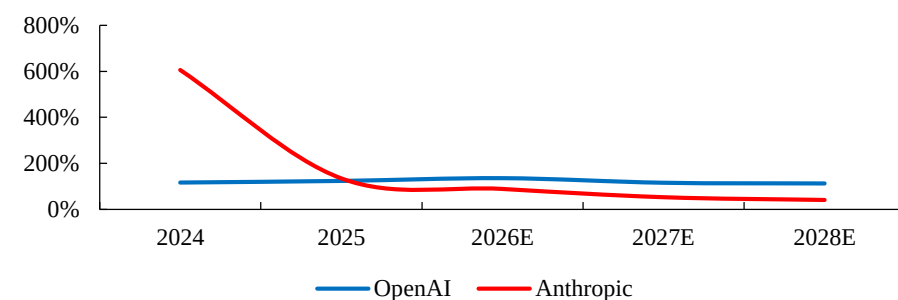
数据来源：The Information、开源证券研究所

图30：Anthropic 服务器成本预计到 2028 年为 270 亿美元（单位：亿美元）



数据来源：The Information、开源证券研究所

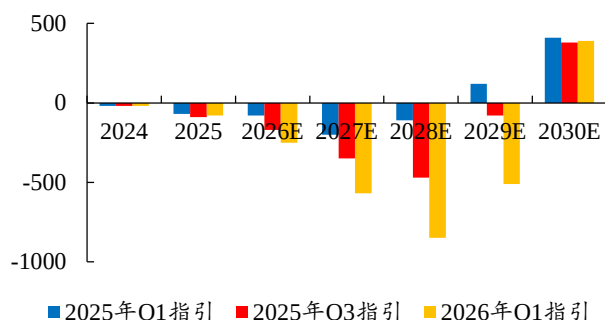
图31：Anthropic 的运用效率优于 OpenAI（数据：服务器投入/收入）



数据来源：The Information、开源证券研究所

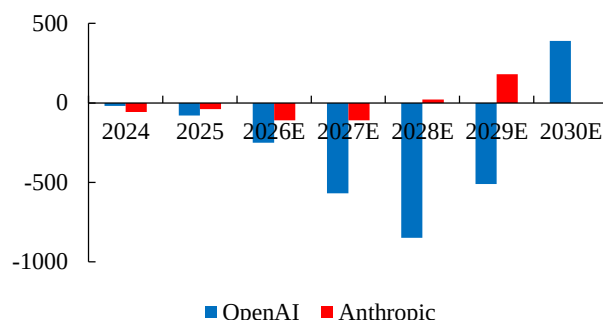
由此，在最新的 2026 年 Q1 的展望中，OpenAI 明显上调了 2028-2029 年的现金消耗量，Anthropic 也同样推后了其现金流转正时间（从 2027 推后至 2028 年），但从现金消耗的绝对值上看，Anthropic 明显优于 OpenAI，尤其在 2027 年之后。

图32：OpenAI 现金流转正时间推到 2030 年（单位：亿美元）



数据来源：The Information、开源证券研究所

图33：Anthropic 现金流转正时间快于 OpenAI（单位：亿美元）

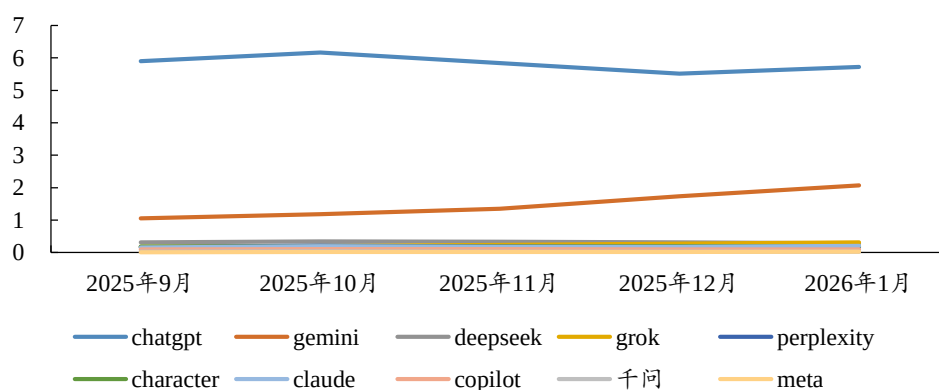


数据来源：The Information、开源证券研究所

3、C 端（Chatbot 为主）：ChatGPT 月活明显领先，春节红利国内大厂依托资源禀赋打法各异

ChatGPT 和 Gemini 稳居大模型前二宝座，相较其他模型具备显著竞争力。用户活跃度相对稳定。Gemini 的月访问量自 2025 年 9 月的 10.6 亿次上升至 2026 年 1 月的 20.7 亿次，呈现出良好的上涨态势，市场竞争力不断增强。其余模型较为稳定并保持较低月访问量，与头部模型存在明显差距，未来发展情况有待观察。

图34：Chatbot 访问量 ChatGPT 领跑全模型，Gemini 紧随其后（单位：十亿次）



数据来源：Tensor Tower、开源证券研究所

相比海外，国内大模型 C 端较难收取订阅费，因而主要模型厂商在 C 端的布局

应更多从生态角度去考虑，而腾讯、字节、阿里在 C 端战略打法有所区别：

腾讯——依托微信社交关系链，主打“嵌入型 AI”战略：其核心是让 AI 以群成员的身份自然融入真实社交场景，成为强化人际连接的“润滑剂”而非替代者；

字节——凭借深厚的技术底座与国民级流量，走“广撒网”的多模态 AI 平台路线：它借助大模型技术突破与春晚超级流量的双重驱动，在软硬件多端广泛布局试探，让模型成为全生态的底层支撑；

阿里——立足于自身强大的商业与本地生活生态资源，致力于打造重构商业终极入口的“办事型 AI”：它通过主客体倒置让全域商业资源为 AI 服务，从而实现了用户“一句话办事”的去中介化交易闭环。

表1：腾讯、字节、阿里在 C 端战略打法有所区别

核心禀赋/优势	战略定位	核心 AI 产品/模型	差异化打法与具体策略
月活超 14 亿的微信社交关系链	“嵌入型 AI”	元宝、元宝派	围绕“关系”而非“任务”：让 AI 作为人际连接的强化者和社交的“润滑剂”，而非替代者 场景融入：推出“元宝派”，让 AI 以“群成员”的平等身份融入家族群、读书群、观赛群等真实的微信社交场景 技术攻坚：通过联合打造 CL-bench 基准，解决上下文理解难题，让 AI 能读懂碎片化聊天和跨窗口需求
技术底座+国民级流量的规模化触达能力	多模态 AI 平台	豆包、Seedance2.0	“广撒网”路线与生态扩张：豆包作为 C 端主力，同时布局 AI 手机、耳机、眼镜等硬件及多模态产品线，让模型作为底层底座，等待市场筛选最优解 技术+春晚双驱动：以 Seedance2.0 模型占据技术舆论高地，并将春晚变成豆包的“产品发布会”，结合发红包、送科技好礼实现最大规模国民触达
能够降低交易成本的本地生活与商业生态资源	“办事型 AI”	千问	重构商业入口：从“连接人货”升级为“连接人与即时服务”，试图将千问打造成取代搜索框的 AI 时代“终极入口” 打造交易闭环：凭借“一句话办事”的 Agent 能力，直接调用淘宝、飞猪、盒马等阿里生态资源，实现从需求到交易的“去中介化”闭环体验 主客体倒置：让阿里旗下所有生态业务团队（淘宝、高德等）为千问 APP 服务，成为其能力支撑

数据来源：开源证券研究所

4、B 端（AI Coding/Agent）：行业或将进入加速拐点，国产模型跻身第一梯队

4.1、市场分析：Anthropic 主导市场，国产大模型逆袭而上

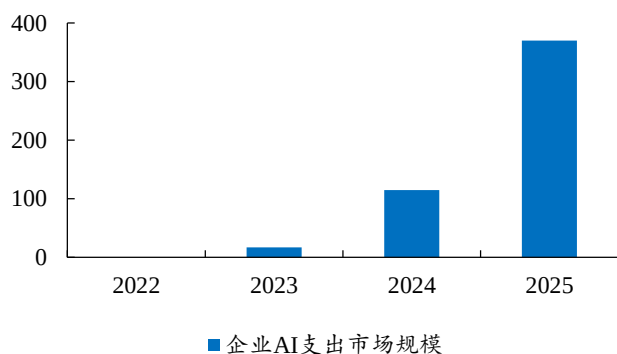
企业级 API 调用量市场集中度高，Anthropic 占据半壁江山。根据 Menlo Ventures 数据，2025 年全球企业 AI 支出达到 370 亿美元，为 2024 年的 3.2 倍，从市场份额上看，Anthropic 市场份额持续提升，2025 年达到 54%，地位持续稳固，从格局演变上看我们认为有两个启示：

1、API 调用/Coding 市场集中度高，腰部及尾部模型空间有限：从 API 调用数据上看，前四大模型已经占据 96% 的市场份额，留给腰部及尾部的模型市场

份额较小，因此持续追求 SOTA 是当前大模型的主要方向，从企业对 Claude 产品的采用率上看，企业持续在追求最前沿的模型使用，而非更便宜、更快的替代品；

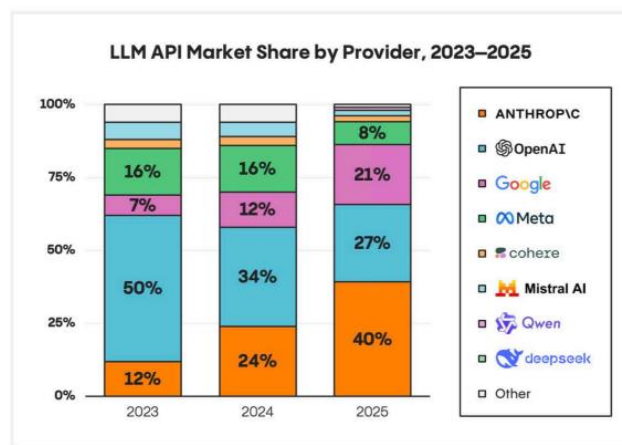
2、API 调用/Coding 市场变化轮动较快：尽管 OpenAI 的 coding 产品仍然处于行业第一梯队，但市场竞争激烈，以程序员为主的群体会积极尝试并使用 SOTA 模型，模型份额变化较快，2023-2025 年 OpenAI 市场份额从 50%下滑至 27%，因此随着国产大模型企业如智谱、MiniMax、月之暗面等逐步跻身第一梯队，当前格局有望松动。

图35：企业大模型支出快速增长



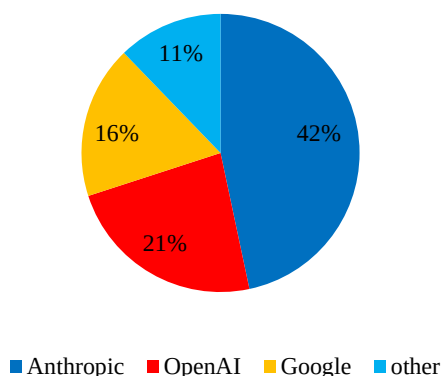
数据来源：Menlo Ventures、开源证券研究所

图36：Anthropic 的 API 调用量份额快速提升



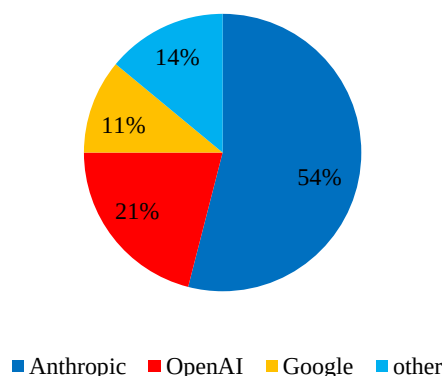
数据来源：Menlo Ventures

图37：2025 年中期 Anthropic Coding 市场份额为 42%



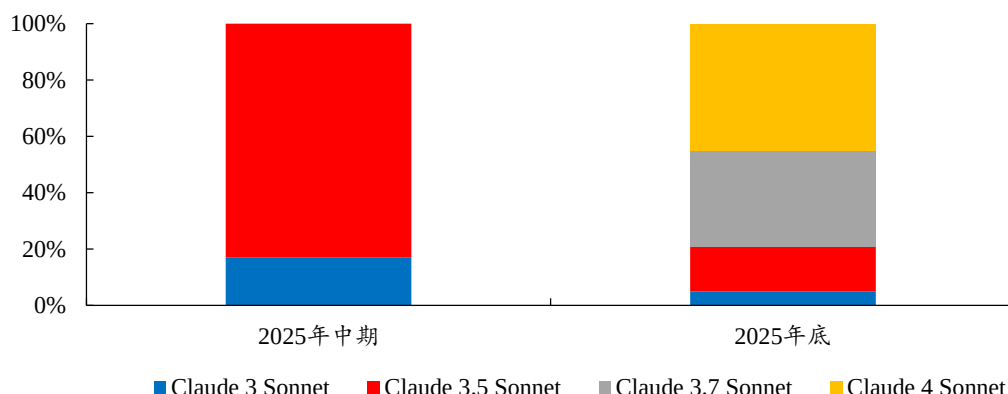
数据来源：Menlo Ventures、开源证券研究所

图38：2025 年底 Anthropic 的 Coding 市场份额为 54%



数据来源：Menlo Ventures、开源证券研究所

图39：企业持续追求最前沿的模型

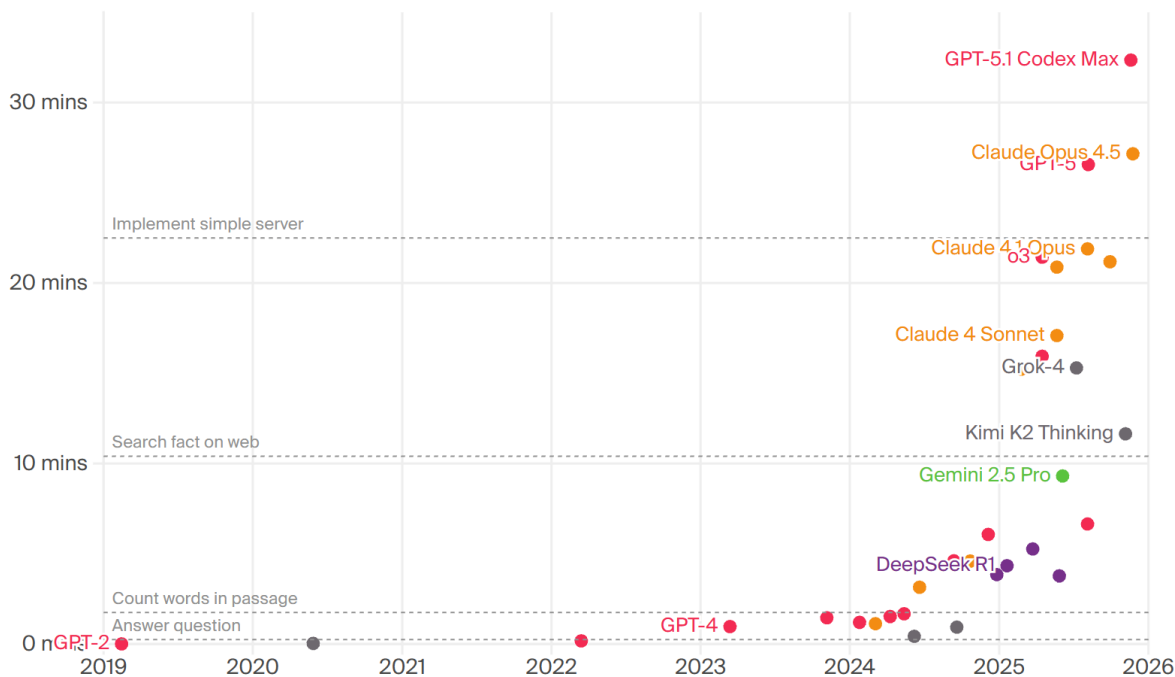


数据来源：Menlo Ventures、开源证券研究所

前沿大模型已突破半小时专业业务高完成度，有望带动需求增长及定价模式转换。根据 The Information 报道，2026 年初，GPT-5.1 Codex Max 与 Claude Opus 4.5 已能以 80% 的高成功率接管原本需要专业人士处理 30 分钟的复杂工程任务。这意味着 AI 终于突破了长链条逻辑的“断点”瓶颈，实现了从“对话框”向“Agent”的跨越。后续市场或将不再单纯为 Token 的生成速度买单，而是为“人类劳动时间的替代率”定价。

图40：2026 年在软件工程任务上，已有模型能在半小时内达到专业程序员 80% 的水平

Can complete software eng. task taking up to this long 80% of the time



数据来源 The Information 官网

国产大模型跻身前列，后续 tokens 调用量带来的想象空间较大。2 月 20 日，智

谱/MiniMax 股价逆势大涨 42.7%/14.5%，足以体现市场对二者在 2 月发布的新版大模型的认可，从近期 OpenRouter 大模型排名上看，国产大模型包揽前三名，而从实际的一些体验上看，智谱 GLM 5、MiniMax M2.5 差不多有 Claude Opus 4.5 的水平，在 Agent 方面，国产 AI、开源 AI 在算力和数据受限的前提下依然跻身全球第一梯队水平，头部玩家市场格局或将有所变化，后续 tokens 调用量带来的想象空间较大。

图41：OpenRouter（2026.02.12-2026.02.22）大模型排名上，前三为国产大模型

LLM Leaderboard

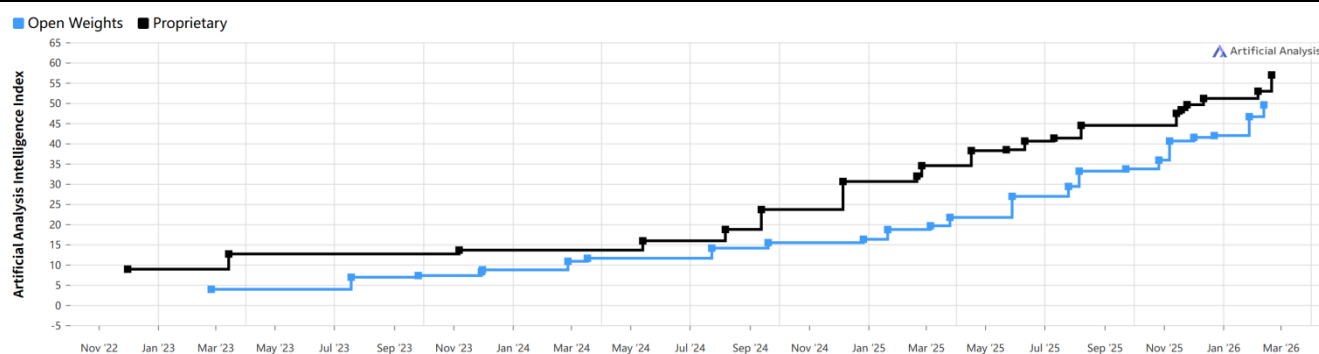
Compare the most popular models on OpenRouter

1.	 MiniMax M2.5 by minimax	3.07T tokens ↑524%
2.	 Kimi K2.5 by moonshotai	1.16T tokens ↓21%
3.	 GLM 5 by zai	1.03T tokens ↑462%
4.	 Gemini 3 Flash Preview by google	821B tokens ↑8%
5.	 DeepSeek V3.2 by deepseek	739B tokens ↓0%
6.	 Claude Opus 4.6 by anthropic	617B tokens ↑26%
7.	 Grok 4.1 Fast by xai	596B tokens ↑10%
8.	 Claude Sonnet 4.5 by anthropic	559B tokens ↓14%
9.	 Gemini 2.5 Flash by google	443B tokens ↓1%

数据来源：OpenRouter

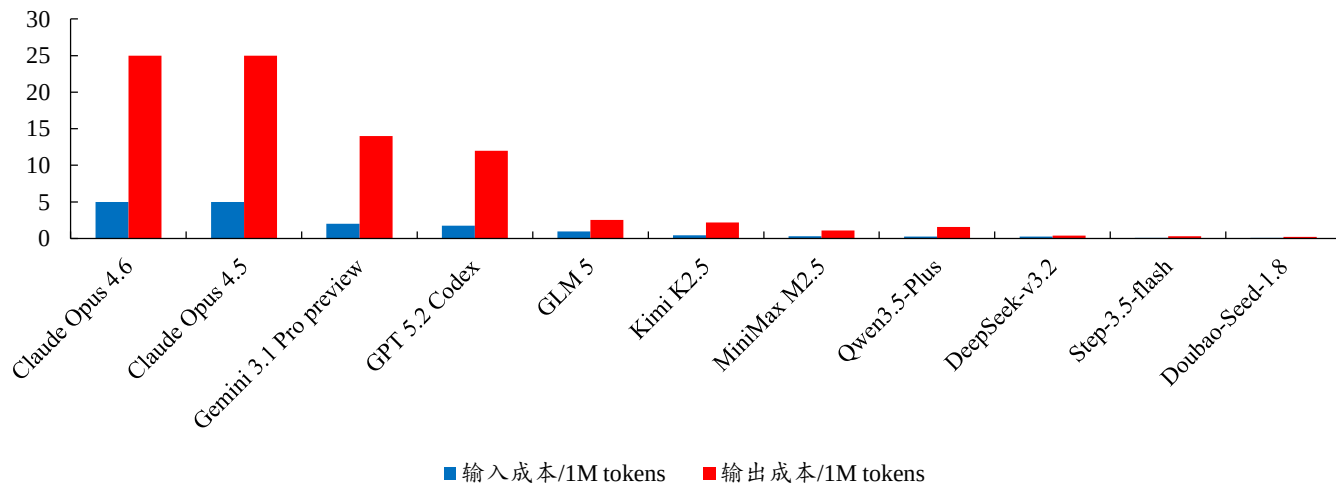
相比于 Claude 及 ChatGPT，尽管国产大模型技术力上仍有追赶空间，但有两个明显的优势：开源+成本：

开源：目前国产大模型多采用开源模式，由此丰富软件生态及训练反馈，从而缩小与闭源头部模型的差距，对企业而言，开源模型可定制性更强、有潜在的成本效益，根据 Artificial Analysis，得益于 GLM-5、MiniMax M2.5、Kimi K2.5 等模型发布，开源 2025 年底到 2026 年初，开源模型相较闭源模型的差距进一步缩小。

图42：2025 年底到 2026 年初，开源模型相较闭源模型的差距进一步缩小


数据来源：Artificial Analysis 官网

成本：目前国产前沿大模型已达到 Claude Opus 4.5 的水平，以 Claude Opus 4.5 为例，其输出成本高达\$25/1Mtokens，而国产 MiniMax M2.5 的输出成本仅约为其 1/25，成本优势明显。国内的低成本并非赔本赚吆喝，主要得益于算力限制下对模型架构的底层创新、对算力性能工程化的“压榨”与优化。此外，2 月 12 日，智谱率先将 GLM Coding Plan 套餐价格结构性上调，整体涨幅自 30%起，表明行业商业模式或逐步从流量获取转向实际工作流的完成度，本次涨价也是大模型定价逐步健康化、可持续化的体现。

图43：国产大模型较海外模型在成本上具有显著竞争力（单位：美元/百万 token）


数据来源：pricepertoken、开源证券研究所

OpenAI 稳居龙头，国产模型收入承压。根据五家企业的数据，OpenAI 凭借 4000 人规模、37 亿美元营收和 328 百万美元的人均营收稳居龙头。而 Anthropic 以 1097 名员工实现人均创收 200 万美元，暂居第二。中国公司智谱、MiniMax、Moonshot 营收规模均在 1 亿美元上下，人均创收普遍低于 20 万美元，与头部企业存在量级差距。这可能受制于国内市场定价环境，也表明当前国产模型仍处于投入期、收入尚未规模增长。

表2：2025 年国内外大语言模型企业人均收入对比（OpenAI、智谱、Moonshot 为预测数据）

	成立时间	员工数	营业收入（百万美元）	人均创收（千美元）
OpenAI	2015 年	4000	13100	3275
Anthropic	2021 年	1097	2200	2005
智谱	2019 年	883	109	123
MiniMax	2021 年	428	79	185
Moonshot	2023 年	300	17	57

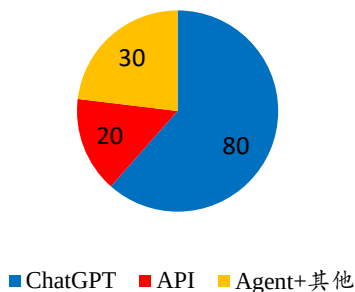
数据来源：各公司官网、开源证券研究所

注：智谱 2025 年收入为 wind 一致预期，汇率为 1USD=6.8902CNY

4.2、Anthropic: ARR 增速远高于估值增速，当年估值通常在 20 倍左右

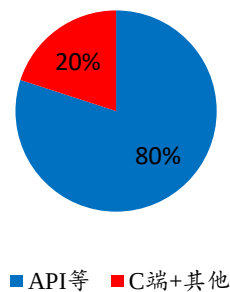
Anthropic 收入 B 端占比超八成，对国产大模型标的参考性更强。Anthropic 约 80%以上的收入来自于企业端，且主要为依托 API 及相关应用程序向企业提供算力服务，C 端占比较低，OpenAI 当前 C 端为主的 chatbot 收入占到收入的 60+%，且周活 9 亿用户的付费率或仅有 5%。从创收的角度上讲，B 端的前景较 C 端更为明确，付费粘性及付费金额也更为占优，也对智谱、MiniMax 等后续依托 agent 业务的标的更有参考价值。

图44：OpenAI 收入构成以 chatbot（C 端）为主



数据来源：The Information、路透社等、开源证券研究所

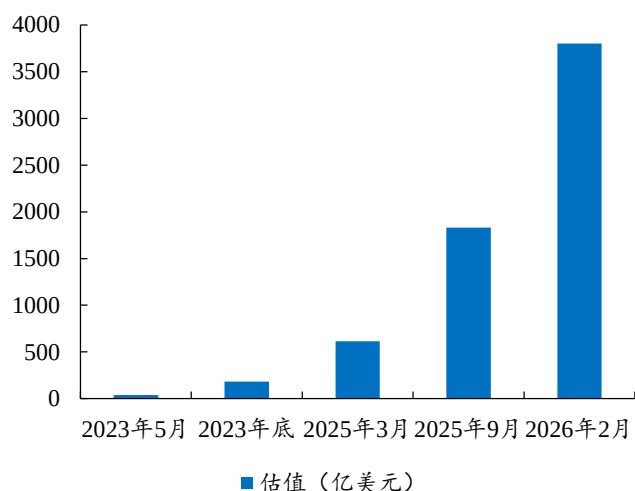
图45：Anthropic B 端收入占比 80%以上



数据来源：The Information、开源证券研究所

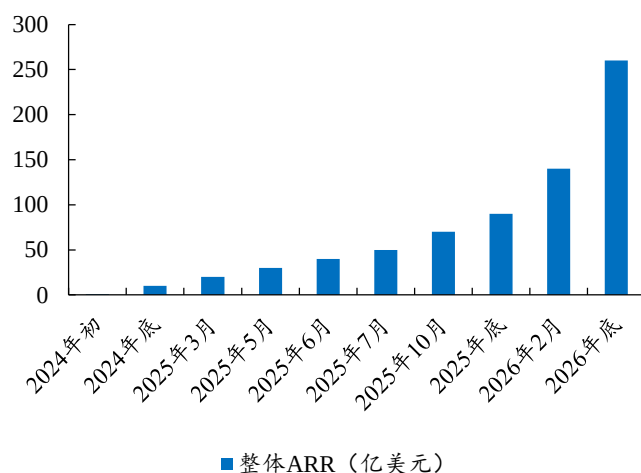
早期 ARR 增长远高于估值提升速度，充分消化估值。我们梳理了 2023 年以来 Anthropic 的估值变化数据，尽管 Anthropic 估值拉升较快，每次估值更新基本是前次的 2-3 倍，但更为快速的 ARR 及收入增长成为消化估值的基础，例如 2025 年 3 月 Anthropic 估值提升至 615 亿美元，为 2023 年底的 3.3 倍，同期 ARR 为 23 倍，最新数据为 2026 年 2 月，Anthropic 估值 3800 亿美元，为 2025 年 9 月的 2.1 倍，同期 ARR 为 2-3 倍。

图46: Anthropic 估值快速提升



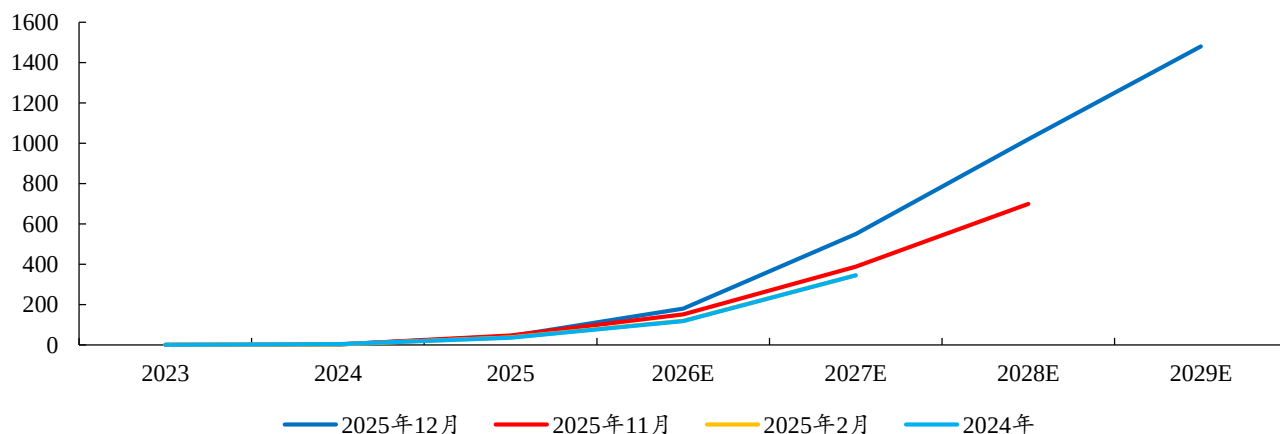
数据来源: The Information、路透社等、开源证券研究所

图47: AnthropicARR 持续提升



数据来源: The Information、开源证券研究所

图48: Anthropic 收入预期持续上调 (单位: 亿美元)



数据来源: The Information、开源证券研究所

我们采用市值/ARR 的性质梳理各个阶段 Anthropic 的估值变化过程, 有如下结论:

- 1、在早期 (2023 年 5 月): 业务基本处于 0-1 的过程, 当年估值达到 46 倍, 次年估值 4 倍;
- 2、2025 年以来: Anthropic 的当年估值最高位 20+倍, 且通常次年估值会降到 10 倍以内;

表3: P/ARR 口径下 Anthropic 估值逐步消化

P/ARR 估值	2023	2024	2025	2026	2027	2028
2023 年 5 月	46.0	4.0				
2023 年 12 月	211.5	21.6	2.0			
2025 年 3 月			6.8	2.4		

请务必参阅正文后面的信息披露和法律声明

23 / 33

P/ARR 估值	2023	2024	2025	2026	2027	2028
2025 年 9 月			20.3	7.0		
2026 年 2 月				14.6	4.9	1.9

数据来源：The Information、开源证券研究所

而在 PS 口径下, 当年估值通常为 20-30 倍, 次年估值在 5-10 倍, the information 对 Anthropic 最新收入预测已展望至 2029 年, 当前对应 1480 亿美元, 对应 PS 估值 2.6 倍。

另一方面, 对于处于发展早期的标的估值弹性可以相对较大, 例如 2025 年 9 月 perplexity 估值 200 亿美元, 而 ARR 接近 2 亿美元, 假设月均增加 2000 万美元, 则 P/ARR 仍有 70+倍。

表4: PS 口径下 Anthropic 估值逐步消化

估值更新时间	收入预测更新时间	2025	2026	2027	2028	2029
2025 年 3 月	2025 年 2 月	16.5	5.1	1.8		
2025 年 9 月	2025 年 11 月	38.9	12.0	4.7	2.6	
2026 年 2 月	2025 年 12 月		21.1	6.9	3.7	2.6

数据来源：The Information、开源证券研究所

5、国产初创大模型：新厂商性能达到 opus 4.5 水平，点燃

Agent 需求想象空间

国产大模型频频更新，多家模型接入 OpenClaw。在 GLM-5、MiniMax M2.5 和 Kimi K2.5 相继推出后，这三家企业也同步更新了大量辅助性功能。以 OpenClaw 为例，国产大模型均走向集成式云端部署平台，利用 OpenClaw 的第三方平台，无需用户本地部署就可以发挥模型更加全面的能力。MiniMax M2.5 在多模态领域推出 Music 2.5，进一步完善其多模态生态。Kimi K2.5 专注 Agents，在模型中更新可让 100 个子代理协助工作的功能，强化其在海量资料处理领域的优势。豆包、千问、阶跃星辰也分别推出全新模型，较先前做出提升。其中，Seedance2.0 在 AI 视频生成领域支持“复杂人机交互”运动生成，做出重大革新。

表5: 国产大模型近期更新

产品	模型	功能
GLM Coding Plan		适配 OpenClaw 和 Excel 环境下的 AI 插件
ZCode	GLM-5	能够理解整个仓库的依赖关系，自主完成模块解耦、跨文件重构以及复杂的系统集成任务
DSA 稀疏注意力机制		能够根据 Token 的重要性动态分配注意力资源，大幅降低算力成本
MaxClaw	MiniMax	无需部署就可以在网页浏览器运行
Music2.5	M2.5	支持多达 14 种音乐结构变体，具备物理级高保真音质
Kimi Claw	Kimi K2.5	无需部署就可以在网页浏览器运行
Agent Swarm		最多可部署 100 个子代理进行工作

请务必参阅正文后面的信息披露和法律声明

24 / 33

Qwen3.5-397B-A17B	Qwen 3.5	实现从“纯文本”向“原生多模态智能体”的跨越
Seedance 2.0	Seed 2.0 Pro	支持高可用率的“复杂人机交互”运动生成
上下文上限	DeepSeek v3.2	DeepSeek-V3 的网页端与 App 端正式支持 1M 上下文
NextStep-1.1	Step 3.5 Flash	创新的“自回归图像生成”模型，支持文本与图像 Token 的联合预测

数据来源：各公司官网、开源证券研究所

各家模型均在 AI Coding 及 Agentic Engineering 领域有所发力，但也各有特点：

从参数角度来看：GLM 和 Kimi 都具备处理海量资料的能力，总参数量分别为 744B 和 1T。MiniMax 总参数量为 229B，但凭借其 10B 的激活参数量，模型处理速度最快可达 100TPS，适用于高频 Agent 操作。GLM-5 预训练数据为 28.5T，是 Kimi K2.5 的近两倍，知识储备量大，相较而言更为适合大型企业或全栈工程师。

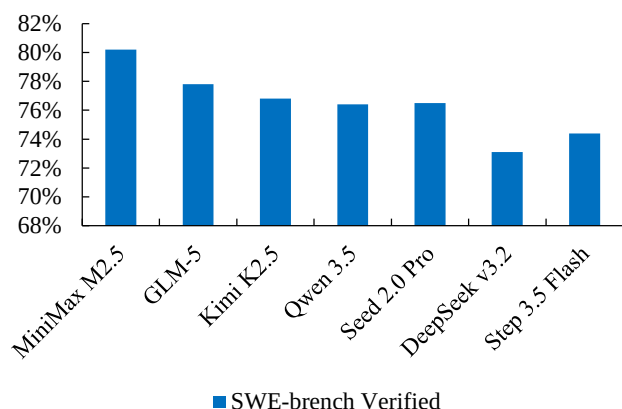
表6：国产大语言模型参数对比

	GLM-5	MiniMax M2.5	Kimi K2.5	Qwen 3.5	Seed 2.0 Pro	DeepSeek v3.2	Step 3.5 Flash
总参数量	744B	229B	1T	397B	400B	671B	197B
激活参数量	40B	10B	32B	17B	20B	37B	11B
上下文窗口	200K	200K	256K-1M	1M	256K	128K	256K
预训练数据	28.5T		15T			14.8T	

数据来源：各公司官网、开源证券研究所

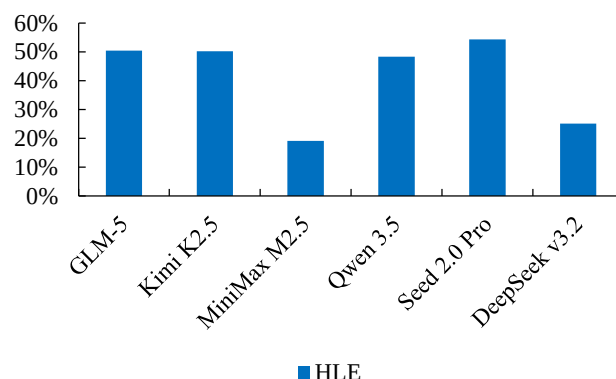
从各项测试结果来看：MiniMax M2.5 在 SWE-brench Verified 测试中占优，HLE 测试逊于其他模型。Qwen 3.5 和 Seed 2.0 Pro 在 BrpwsComp 测试中取得前两名。SWE-brench Verified 测试、BrowseComp 测试和 HLE 测试分别为针对模型代码编写能力、网络检索能力和专家级多学科难题的测试。根据测试结果，MiniMax M2.5 在代码处理方面具有显著优势，但由于较低的总参数量和预训练数据，模型的知识储备可能相对不足，因而未能在 HLE 测试中取得头筹，这表明在专业领域，GLM-5 与 Kimi K2.5 等模型仍具备足够的竞争力。HLE 测试中 GLM、Kimi、Qwen 和 Seed 表现差距不大，由于千问和豆包依托自身集团的海量数据，尽管未公布预训练数据，但预训练 token 消耗量可能不低于其他模型。Kimi K2.5 的预训练数据仅为 15T，这说明 Kimi 的预训练数据含金量较高，可以有效降低预训练成本。Qwen 3.5 和 Seed 2.0 Pro 更专注于搜索，因而在 BrowseComp 测试中领先其余国产模型。

图49：国产大语言模型 SWE-brench Verified 测试 MiniMax M2.5 表现最好



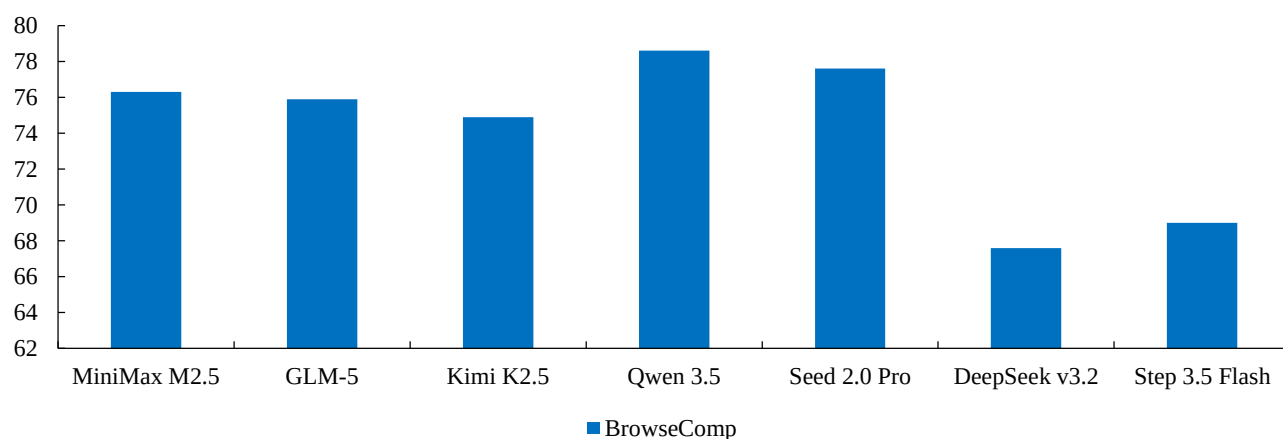
数据来源：各公司官网、开源证券研究所

图50：国产大语言模型 HLE 测试 Seed 2.0 Pro 表现最好



数据来源：各公司官网、开源证券研究所

图51：国产大语言模型 BrowseComp 测试 Qwen 3.5 表现最好

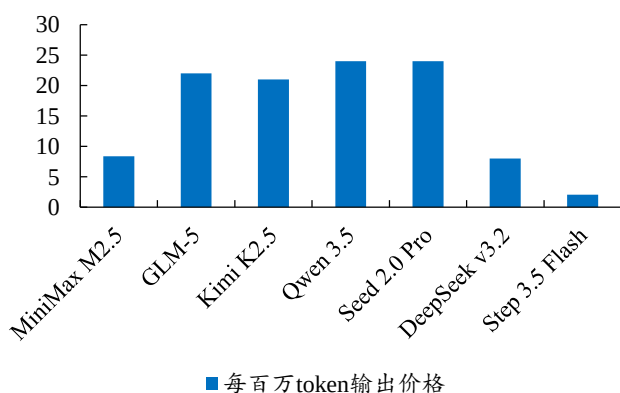


数据来源：各公司官网、开源证券研究所

从使用成本来看：Step 3.5 Flash 每百万 token 输出价格为 2 元，仅为 GLM-5 和 MiniMax M2.5 和 DeepSeek v3.2 的约 1/4，是其他大模型的 1/10 以下。考虑到 Step 3.5 Flash 在各项测试中的良好表现，该模型是国产模型中性价比最高的选择。但 Step 系列目前仍处于拓展市场阶段，不排除未来 API 涨价风险。而综合各能力测试，MiniMax M2.5 是绝大多数情况下的最优选。

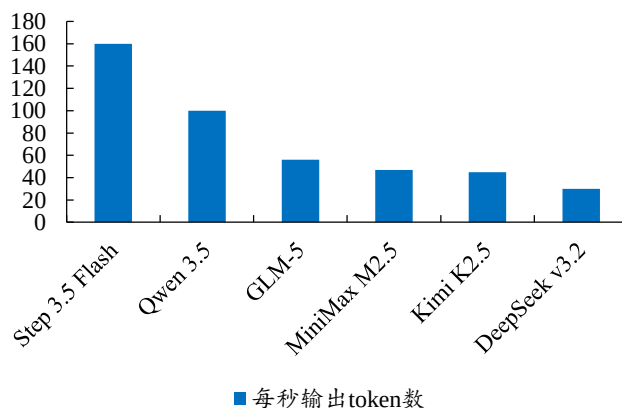
从运行速度来看：Step 3.5 Flash 凭借每秒输出 160 token 的速度领先其他国产模型，在处理海量数据工作中具有显著优势。

图52: 国产大语言模型 Step 3.5 Flash 输出成本最低(单位: 人民币元)



数据来源: 各公司官网、开源证券研究所

图53: 国产大语言模型 Step 3.5 Pro 输出速率最快



数据来源: 各公司官网、开源证券研究所

总结: MiniMax M2.5 在各方面表现良好, 但受制于其总参数量和知识储备, 不利于用于海量数据处理的工作, 但凭借其多模态优势, 该模型在视频生成、图片生成等领域仍存在广泛应用空间, 更适合初创公司和成本敏感性公司或用户。GLM-5 系统稳健、总参数量大、处理速度快, 并具备 ZCode 支持, 适用于大型企业和全栈工程师。Kimi K2.5 总体表现均衡, 但拥有集群模式的 Agents 辅助, 适合处理海量资料, 适合前端工程师和设计师、研究员等。

AI 上半场是训练大于评估, 而下半场是评估大于训练——这是 2025 年 8 月姚顺雨在 OpenAI 期间发表的博文《下半场》曾提出的一个被反复引用的判断。从投资角度而言, 与其去研判未来格局演绎后的胜败, 不如聚焦于当前状态主要模型的位置, 以及下一阶段的运动方向, 而评估或是一个有效的抓手。在这其中, AI coding、Agent 或将是率先取得突破的重要方向。

我们整理目前主要模型在 coding、agent 相关的评估指标, 且大致将指标分为三类:

工程交付类指标: 用于解决实际工程问题, 其中关注度较高的如 SWE-Bench Verified, 用以修复真实的 GitHub 项目 Bug;

认知进化类指标: 测试模型复杂思考的能力, 用以衡量模型的 AGI 的程度;

基础门槛类指标: 主要模型评分基本做到 90%以上, 已能较好实现商业化, 无需做过多讨论。

表7: 从行业主要的测评上看, Vibe Coding 逐渐转向 Agentic Engineering

能力评估	指标	Opus4.6	Opus4.5	Sonnet 4.5	Gemini3 Pro	GPT-5.2	GPT-5.3-Co dex(xhigh)	GLM4.7	GLM 4.6V	DeepSeek V3.2
工程交付指标	SWE-Bench Verified	80.8%	80.9%	77.2%	76.2%	80.0%		73.8%		73.1%
	Agentic coding									
	SWE-Bench Pro (Public)						56.8%			

		LiveCode			64.0%	91.5%		84.9%	83.3%
		Bench v6							
Agentic terminal coding	Terminal-Bench	h2.0	65.4%	59.8%	51.0%	56.2%	64.7%	77.3%	41.0%
									46.4%
Agentic computer use	OSWorld		72.7%	66.3%	61.4%	--	--		37.2%
	OSWorld Verified							64.7%	
Agentic search	BrowseComp		84.0%	67.8%	43.9%	59.2%	77.9%		67.5%
Scaled tool use	MCP Atlas		59.5%	62.3%	43.8%	54.1%	60.6%		
Agentic financial analysis	Finance Agent		60.7%	55.9%	54.2%	44.1%	56.6%		
Office tasks	GDPval-AAEL	o	1606	1416	1277	1195	1462	1193	1187
认知进化指标	Novel problem-solving	ARCAGI2	68.8%	37.6%	13.6%	45.1%(Thinking)	54.2%(Pro)		
	Multidisciplinary reasoning	Humanity's Last Exam	40.0%/53.1%	30.8%/4.4%	17.7%/3.6%	37.5%/4.8%	36.6%/50.0%	24.8%/4.8%	25.1%/40.8%
基础门槛指标	Visual reasoning	MMMU Pro	73.9%/77.3%**	70.6%/7.3%	63.4%/6.9%	81.0%/--	79.5%/80.4%		66.0%
	Multilingual Q&A	MMMLU	91.1%	90.8%	89.5%	91.8%	89.6%		
基础门槛指标	Agentic tooluse	t2-bench (Retail/Telecom)	91.9%/99.3%	88.9%/9.2%	86.2%/9.0%	85.3%/98.0%	82.0%/98.7%	87.4%	85.3%
	Graduate-level reasoning	GPQA Diamond	91.3%	87.0%	83.4%	91.9%	93.2%	85.7%	

数据来源：开源证券研究所

根据该榜单，我们可以得到以下结论：

Anthropic 与 OpenAI 代码工程能力处于第一梯队，国内主要模型已进入“生产级”阶段。领先 SWE-bench 作为业内评价软件工程的黄金标准，目前 Claude Opus 4.6、GPT-5.2 均已迈入 80% 分值门槛，处于行业第一档位。国产模型中智谱、DeepSeek、千问、Kimi 等均突破 70% 分段，意味着已经从“玩具级”向“生产级”过渡。此外 OpenAI 提出新的评判标准 SWE-Bench Pro，当前 Codex 5.3 得分仅 56.8%，或成为后续代码能力评测的重要标准。

Vibe Coding 逐渐转向 Agentic Engineering，随着 agent 自主执行的能力达到 PMF，2026 年或有望看到诸多垂类领域呈现具备商用价值的 Agent。Claude Code、Cowork、Clawdbot 的出现受行业关注较大，核心在于大模型从被动问答向主动执行、自主决策的转变，并实现向实际办公场景延伸，AI Coding 的重点也逐渐从 Vibe Coding 进化到 Agentic Engineering，而从 Terminal-Bench2.0（终端）、OSWorld（桌面）、BrowseComp（网页）这类指标上看，Claude 与 ChatGPT 已初步达到 PMF，此外本次的 Opus4.6 更新特意披露了一项 Finance Agent（评分 60.7%），与 cowork 的财务插件功能相呼应。

然而从 Agent 在金融及法务等实际复杂环境的运用，我们认为仍处于早期。我

们认为金融、法务、咨询类行业受益于数据标准化、信息化程度较高的行业，将首先受到 Agent 赋能，亦可作为我们衡量 Agent 在现实商业环境中渗透进度的观测指标。基于此，Mercor 开发了一项基准测试——APEX-Agents，旨在评估 Agent 在长周期、跨应用的专业任务重的效果，这些任务由投行分析师、律师、咨询顾问创建，需要 Agent 在真实环境中操作如文档、邮箱、PDF、聊天记录、日历等工具完成任务。

从最终的测试结果上讲仍不容乐观，在主要模型的一次通过率最高仅为 24%，即每 4 个任务仅成功 1 个，距离可商业化的 60%-70% 以上的水平仍相去甚远。

表8: ApexAgent 考察了投行、法务、咨询管理的实际任务处理能力

职位	项目数	文件数/项目	任务数	平均标准/任务	平均小时数	带有文件输出的任务
投资银行	10	172	160	2.93	1.36	27(16.9%)
法务	12	161	160	4.57	2.4	20(12.5%)
咨询管理	11	165	160	4.68	1.69	11(6.9%)
合计	33	166	480	4.06	1.82	58(12.1%)

数据来源: Huggingface 官网、开源证券研究所

表9: 主要模型的通过率仍然较低

Model	1 次通过率	8 次通过率	8 次都通过的 概率	平均得分	投行: 一次通 过率	咨询: 一次通 过率	法务: 一次通 过率
ClaudeOpus4.5	18.4%	34.0%	8.8%	34.8%	21.6%	13.2%	20.2%
Gemini3Flash	24.0%	36.7%	13.4%	39.5%	26.7%	19.3%	25.9%
Gemini3Pro	18.4%	37.3%	6.5%	34.1%	18.8%	12.4%	23.9%
GPT-5	18.3%	31.0%	7.7%	32.9%	27.3%	12.3%	15.3%
GPT-5.2	23.0%	40.0%	11.0%	38.7%	27.3%	22.7%	18.9%
GPT-OSS-120B	4.7%	11.5%	1.2%	14.5%	2.7%	3.5%	7.8%
Grok4	15.2%	32.9%	4.7%	30.3%	17.0%	12.0%	16.5%
KimiK2Thinking	4.0%	14.4%	0.3%	11.5%	1.2%	2.9%	8.0%

数据来源: Huggingface 官网、开源证券研究所

这背后目前存在的技术瓶颈可归纳为 2 个原因:

- (1) **信息碎片化:** 真实工作场景中信息碎片化高，一方面需要 AI 有较高的完整认知能力，另一方面需要较高的长上下文处理能力，这涉及到 agent 领域的长期目标偏移问题;
- (2) **对复杂问题的规划能力:** 当前生成式 AI 的本质是“预测下一个 token”，而规划能力需要将复杂问题拆成多步，并根据中间出现的错误进行调整，这涉及到 agent 领域长期存在的幻觉叠加的问题。

此外，通过上下文学习的衡量，亦可观测当前大模型、agent 的发展进度，根据腾讯姚顺雨提出的 CL-bench 排名，当前主流头部大模型在上下文学习能力上仍还有充足的提升空间。

表10：根据 CL-bench 排名，当前大模型的上下文学习能力

排名	模型名称	所属机构	总解决率
1	GPT 5.1(High)	OpenAI	23.7%±0.5%
2	GPT 5.1	OpenAI	21.1%±0.2%
3	Claude Opus 4.5 Thinking	Anthropic	21.1%±1.4%
4	Kimi K2.5	Moonshot	19.4%±0.6%
5	Kimi K2.5 Thinking	Moonshot	19.2%±0.6%
6	Claude Opus 4.5	Anthropic	19.1%±0.4%
7	GPT 5.2	OpenAI	18.2%±0.5%
8	GPT 5.2(High)	OpenAI	18.1%±0.8%
9	o3(High)	OpenAI	17.8%±0.2%
10	Kimi K2Thinking	Moonshot	17.6%±0.6%
11	HY 2.0 Thinking	Tencent	17.2%±0.6%
12	Gemini 3 Pro(High)	Google	15.8%±0.3%
13	HY 2.0	Tencent	15.0%±0.8%
14	Gemini 3 Pro	Google	14.8%±0.4%
15	Qwen 3 Max	Alibaba	14.5%±0.6%
16	Doubao 1.8(High)	ByteDance	14.2%±0.4%
17	Qwen3 Max Thinking	Alibaba	14.1%±0.1%
18	Doubao 1.8 (Minimal)	ByteDance	13.5%±0.5%
19	Doubao 1.6 Thinking	ByteDance	13.4%±0.1%
20	DeepSeek V3.2 Thinking	DeepSeek	13.2%±0.4%
21	Doubao1.6	ByteDance	13.1%±0.1%
22	DeepSeek V3.2	DeepSeek	12.4%±0.5%
23	Kimi K2	Moonshot	11.9%±0.5%

数据来源：clbench.com、开源证券研究所

总结：尽管当前在实际场景方案解决、长上下文学习上主流大模型仍有提升空间，但还是回到那句话——AI 的下半场是评估大于训练。明确了评估的方向，意味着明确了后续技术投入的发展方向，这种从“大而全”到“精而准”的范式转移，不仅将重塑资本配置的效率，更决定了谁能率先穿越技术泡沫。只有建立起穿透行业深度的度量衡，大模型才能真正从实验室里的“高分标本”转化为深耕产业的“生产工具”，从而在估值逻辑上实现从算力堆砌向真实商业红利的质变。

6、投资建议：

模型端：OpenClaw 加速了 Agent 的渗透能力，国产大模型能力已逐步达到全球第一梯队水平，商业模式有望逐步切换至订阅为主的持续性收入，增长空间可观，受益标的有阿里巴巴-W、百度集团-SW、MiniMax-W、智谱；

AI Infra：AI Infra 与 AI Agent 有较好的协同共振效应，且随着 AI 发展对安全、

高效工作流的诉求也同样提升，受益标的为与生成式 AI 互补的决策式 AI 龙头范式智能、推出硅基员工开发平台的百融云；

算力端：Agent 行情下 tokens 消耗有望有明显提升，智谱提价进一步维护了产业链盈利健康，受益标的金山云、万国数据等。

表11：主要标的盈利预测及估值

证券	代码	市值(亿港元)	股票基金货币	收入 (亿元)				调整后可比净利润 (亿元)				P/S			P/E			评级
				2025A	2026E	2027E	2028E	2025A	2026E	2027E	2028E	2026E	2027E	2028E	2026E	2027E	2028E	
百融云-W	6608 HK	37	CNY	29.2	35.8	40.4	38.7	0.7	2.9	3.6	3.8	0.9	0.8	0.8	11.3	9.2	8.6	未评级
Minimax-W	100 HK	3,180	USD	0.8	2.3	6.7	16.0	-2.5	-4.4	-5.1	-2.8	168.2	56.6	23.7	/	/	/	未评级
智谱	2513 HK	3,270	CNY	7.2	30.1	72.1	173.9	-31.8	-40.9	-41.2	-21.5	101.8	42.5	17.6	/	/	/	未评级
范式智能	6682 HK	189	CNY	71.4	89.1	116.4	161.6	0.1	2.9	5.6	3.5	1.9	1.5	1.1	58.9	30.3	49.4	未评级
金山云	3896 HK	324	CNY	95.6	120.5	144.6	169.8	-5.6	-5.7	-3.5	-1.4	2.3	1.9	1.7	/	/	/	未评级
万国数据	9698 HK	668	CNY	114.3	126.6	145.0	172.4	9.0	6.6	11.1	10.5	4.5	4.0	3.3	87.6	51.9	54.8	未评级
阿里巴巴-W	9988 HK	22,633	CNY	9,963.5	10,292.9	11,452.4	12,778.6	1,579.4	845.0	1,243.6	1,614.6	1.9	1.7	1.6	23.6	16.0	12.3	买入
百度集团-SW	9888 HK	2,962	CNY	1,290.8	1,326.1	1,415.5	1,534.0	189.4	192.9	220.5	274.2	2.0	1.8	1.7	13.5	11.8	9.5	买入

资料来源：Bloomberg、开源证券研究所。汇率 CNYHKD=1.135, USDBHKD=7.8381

7、风险提示：

模型迭代不及预期：若模型微调、轻量化部署、国产化适配优化效果无法达到商用落地标准，头部企业技术领先优势收窄，相关企业技术壁垒与产品竞争力弱化，进而引发 AI 相关业务营收放量、盈利改善进度不及市场预期，板块及个股业绩存在承压风险。

行业竞争加剧：若赛道内出现低价竞标、API 调用降价、私有化解决方案让利接单等内卷行为，同时行业人才争夺白热化推高研发与人力成本，中小厂商盈利空间持续收窄，头部企业存量市场份额争夺难度加大，行业整体盈利修复节奏放缓，企业盈利能力与估值中枢存在下行压力。

agent 落地不及预期：若下游客户对 Agent 降本增效的实际价值认可度偏低，付费意愿与落地验收标准严苛，试点项目商业化转化效率不足，行业插件生态、合规监管体系、运维配套体系尚未健全，易出现“试点落地多、大规模放量少”的情况，导致相关企业 AI Agent 业务订单增速、收入兑现不及预期，商业化进程放缓引发估值回调风险。

特别声明

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

分析师承诺

负责准备本报告以及撰写本报告的所有研究分析师或工作人员在此保证，本研究报告中关于任何发行商或证券所发表的观点均如实反映分析人员的个人观点。负责准备本报告的分析师获取报酬的评判因素包括研究的质量和准确性、客户的反馈、竞争性因素以及开源证券股份有限公司的整体收益。所有研究分析师或工作人员保证他们报酬的任何一部分不曾与，不与，也将不会与本报告中具体的推荐意见或观点有直接或间接的联系。

股票投资评级说明

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

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

分析、估值方法的局限性说明

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

法律声明

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

开源证券股份有限公司（以下简称“本公司”）的机构或个人客户（以下简称“客户”）使用。本公司不会因接收人收到本报告而视其为客户。本报告是发送给开源证券客户的，属于商业秘密材料，只有开源证券客户才能参考或使用，如接收人并非开源证券客户，请及时退回并删除。

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

本报告可能附带其它网站的地址或超级链接，对于可能涉及的开源证券网站以外的地址或超级链接，开源证券不对其内容负责。本报告提供这些地址或超级链接的目的纯粹是为了客户使用方便，链接网站的内容不构成本报告的任何部分，客户需自行承担浏览这些网站的费用或风险。

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

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

开源证券研究所

上海

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

深圳

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

北京

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

西安

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