东吴证券 SOOCHOW SECURITIES

传媒行业深度报告

诸侯混战，国模出海——大模型 2026 年中复盘及展望

2026 年 07 月 07 日

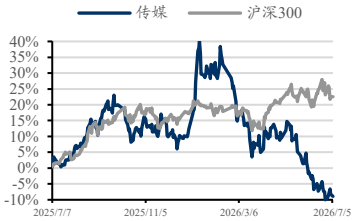
增持（维持）

证券分析师 张良卫
执业证书：S0600516070001
021-60199793
zhanglw@dwzq.com.cn
证券分析师 张文雨
执业证书：S0600525070007
zhangwy@dwzq.com.cn

投资要点

- 每隔一段时间，市场就会响起同样的质疑声：Scaling Law 是不是见顶了？紧随而来的连锁反应总是相似的，例如担心 AI 是不是泡沫，是不是有更高效的方式，Transformer 本身不是真正的理解而只是概率统计的游戏……然后模型能力又一次跃升，ARR 又一次增长，这些怀疑暂时被搁置，但从未真正消失。它们在每一次模型迭代放缓时重新浮现。这种周期性的怀疑背后，是两个更本质的问题：第一，什么是理解？第二，是否存在比暴力 scale 更高效的路径？我们认为，压缩本身就是智能。当 AI 不仅在回答问题，还在真正地操作工具、修改代码、生成代码、生成报告、生成观点……**如果这些都不是理解，那什么是理解？**关于第二个问题，确实有人在下不同的赌注。田渊栋、Yann LeCun、Ilya 新成立的公司，都在探索 alternative 架构和新的推理范式。但现实是：这些几乎没有成型的模型、产品。投资人给钱，**更像是一种分散下注——万一 Transformer 真的不是终点呢？但这种 bet 的风险极高。相比之下，暴力美学至少还是有效的。**Opus 系列能力迭代放缓后，Fable（Claude 5）再次跃升；GPT 5.1 到 5.4 进步平缓，而 GPT-5.5/5.6 又实现了质的突破。这个规律一次次重复：预训练决定天花板，后训练决定能触及天花板的高度。真正在发生的不是 Scaling Law 失效，而是单一维度的参数堆叠遇到了边际收益递减，**多模态、Multi-Agent、RSI（递归自改进）、世界模型的 scaling 空间正在被打开。**
- **Frontier model 竞争格局：交替领先，终局未定。**2025 年 11 月 Google 凭 Gemini 3 短暂领先，2026 年初 Anthropic 靠 coding 反超，到 Q2 OpenAI 凭 GPT 5.5/5.6 回归。Google 在 coding 上因产品飞轮未启动而掉队，但多模态数据优势和自研芯片的优势让其保留翻盘可能。大模型第一的位置从未稳固，领先优势的半衰期正在急剧缩短。
- **闭源 vs 开源：开发者用开源，企业买闭源，但天平正在倾斜。**闭源在前沿能力上仍领先 6 到 12 个月，但开源在性价比、可控性、数据隐私上的优势正在吸引越来越多的垂直 AI 应用公司自己训练模型。Harvey 用 GLM-5.1 训练出法律 agent 超越 GPT-5.5 和 Opus 4.8，Cursor 被 xAI 收购获得自己的模型底座。垂直应用公司过去是闭源 API+prompt engineering，正在逐步转为开源模型+自有训练能力+垂直场景优化。
- **从 Vibe Coding 到 Vibe Working 再到 Vibe Creating：三波浪潮叠加推进。**Coding 已经是收获期（Anthropic 和 OpenAI 收入主要来源），Working 正在渗透（知识工作者占 Codex 用户 20%且增速是程序员 3 倍，金融成为 Anthropic 第二大收入来源），Creating 还在实验期。三波浪潮的演进逻辑是从高可验证性向低可验证性递进。
- **投资建议：**①OpenAI 和 Anthropic 或在今年下半年到明年陆续上市。两家上市后会对整个板块形成定价锚点。②Google 的翻盘机会：Coding 追赶看 Antigravity 能否重启产品飞轮。长期看多模态+世界模型布局。③中国开源模型全球化机会：国产开源模型能力上已经快速追赶，从“能用”到“好用”。DeepSeek 算法创新领先，成本优势显著；智谱 GLM-5.x 系列在 coding 上表现优秀；minimax m3 在原生多模态有所突破。我们认为随着企业端从 tokenmaxxing 转向开源节流，国产模型在全球市场的份额有望提升。
- **风险提示：** Scaling Law 失效风险，ARR 增速放缓，自由现金流压力

行业走势



相关研究

《NeoCloud 龙头崛起，AI 算力基础设施价值凸显》

2026-05-31

《“十五五”规划定调，看好游戏出海》

2026-03-16

内容目录

图表目录

1. Scaling Law 没有失效，多模态、multi-agent、RSI 是新的方向

1.1. Scaling law 没有失效

每隔一段时间，市场就会响起同样的质疑声：Scaling Law 是不是见顶了？紧随而来的连锁反应总是相似的——AI 是不是泡沫，是不是有更高效的方式，Transformer 本身不是真正的理解而只是概率统计的游戏……如此云云。然后模型能力又一次跃升，ARR 又一次增长，这些怀疑暂时被搁置，但从未真正消失。它们像幽灵一样，在每一次模型迭代放缓时重新浮现。

这种周期性的怀疑背后，是两个更本质的问题：**第一，什么是理解？第二，是否存在比暴力 scale 更高效的路径？**

关于第一个问题，从信息论的角度看，**压缩即智能。理解就是建模，建模就是压缩。**当 AI 不仅在回答问题，还在真正地操作工具、修改代码、生成代码、生成报告、生成观点，当它能端到端地完成一个估值建模、能用 300 个子 Agent 并行协作 13 小时完成复杂编程任务、能在虚拟环境里“梦到”如何解鞋带然后在真实世界里做到，如果这些都不是理解，那什么是理解？**是不是只有完全复制人脑的神经结构才叫理解？这个标准本身可能就是一种傲慢。**

关于第二个问题，确实有人在下不同的赌注。田渊栋、Yann LeCun、Ilya 新成立的公司，都在探索 alternative 架构和新的推理范式。但现实是：这些公司有理念、有融资、有顶级人才，却没有任何成型的模型或产品，没有 DAU，没有 ARR。投资人给钱，更像是一种分散下注——万一 Transformer 真的不是终点呢？但这种 bet 的风险极高。相比之下，暴力美学至少还是有效的。Opus 系列能力迭代放缓后，Fable（mythos）再次跃升；GPT 5.1 到 5.4 进步平缓，5.5 和 5.6 又实现了质的突破。这个规律一次次重复：**预训练决定天花板，后训练决定能触及天花板的高度。**

所以真正在发生的不是 Scaling Law 失效，而是单一维度的 scaling 遇到了边际收益递减，而多个新维度的 scaling 空间正在被打开。如果把“scale”理解为不只是堆更多参数，而是在更丰富的数据模态、更复杂的协作架构、更深度的自我进化、更强的世界理解上做 scaling，那么我们离天花板还很远。

1.2. 多模态：从理解文字到理解世界

Google 在 2025 年 11 月凭 Gemini 3 短暂领先，核心突破在于原生多模态架构，在预训练阶段就把视频、图像、音频、文本都映射到同一个向量空间里。

DeepMind 在今年初发布的 Omni 模型，是一个能理解世界的单一模型架构，可以接受任何输入、生成任何输出。Demis 把 Omni 定义为世界模型，因为它对世界有深层的理解。传统世界模型如 Genie 是动作条件视频模型（action-conditioned video model），而 Omni 既能理解世界，又能完成你期望世界模型能做的视觉创作，虽然不是实时的。世

界模型和视频模型之间的界限会以不同于以往的方式演变。

多模态 scaling 的潜力还远未释放。当前的视频生成模型（可灵、seedance、Runway）和文本模型是两条技术路线。但如果未来能实现真正的统一架构，一个模型既能读懂视频、生成视频，又能理解文本、写代码，那么不同模态之间的能力就能互相迁移。比如从物理视频中学到的重力、碰撞知识，可能帮助模型更好地理解物理仿真代码；从代码中学到的循环、递归逻辑，可能帮助模型生成更符合因果关系的视频。

国内模型在多模态输出上已经领先，在 artificial analysis 的文生视频排行榜上，前十名里有 9 个都是中国的模型。但在多模态输入和理解上还在追赶。

图1：artificial analysis 的文生视频排行榜（截至 2026/7/5）

Category: All | Current models | All models | With Audio | No Audio | All | Open weights | Global Leaderboard | Personal Leaderboard

	Range	Creator	Model	Elo	95% CI	Samples	Released	API Pricing[1]
1	1	ByteDance Seed	Dreamina Seedance 2.0 720p	1,222	-9/9	9,614	Mar 2026	$9.07 /min
2	2	Alibaba-ATH	HappyHorse-1.1	1,153	-11/11	2,915	Jun 2026	$9.90 /min
3	3	Alibaba-ATH	HappyHorse-1.0	1,126	-9/9	6,602	Apr 2026	$13.20 /min
4	4-8	Skywork AI	SkyReels V4	1,108	-10/10	3,761	Mar 2026	$21.00 /min
5	4-6	KlingAI	Kling 3.0 1080p (Pro)	1,108	-8/8	7,215	Feb 2026	$20.16 /min
6	4-9	Alibaba	Wan 2.7	1,103	-10/10	3,282	Apr 2026	$16.90 /min
7	6-9	KlingAI	Kling 3.0 Omni 1080p (Pro)	1,099	-8/8	7,006	Feb 2026	$16.80 /min
8	6-10	KlingAI	Kling 3.0 720p (Standard)	1,099	-8/8	7,171	Feb 2026	$15.12 /min
9	6-11	Google	Veo 3.1	1,095	-8/8	7,178	Jan 2026	$24.00 /min
10	8-12	KlingAI	Kling 3.0 Omni 720p (Standard)	1,091	-8/8	6,926	Feb 2026	$13.44 /min

数据来源：artificial analysis，东吴证券研究所

1.3. Multi-Agent：从单兵作战到军团协作

如果单个模型的能力增长在放缓，那么让多个 Agent 协作来完成复杂任务，就是另一个 scaling 维度。

月之暗面的 Kimi Work 展示了一个极端案例：13 小时连续编码、300 个子 Agent 并行协作、4000 多次自主工具调用。让 300 个模型同时工作，有的负责数据清洗，有的负责代码生成，有的负责测试，有的负责文档整理。人类干活只能串行，但 AI 可以大规模并行。

Multi-Agent 的核心在于协作架构的涌现能力。当你有 300 个 Agent 时，如何分配任务、如何避免冲突、如何合并结果、如何处理依赖关系，这些编排逻辑是一个复杂系统。

但 Multi-Agent 也面临工程挑战。第一是成本。300 个 Agent 同时跑，token 消耗是单 Agent 的几百倍。第二，Agent 越多，整个系统的行为就越难预测。调试的时候你很难知道是哪个 Agent 在哪个环节出了问题。第三是可靠性。一个 Agent 失败可能导致整

个任务链崩溃，容错机制的设计非常复杂。

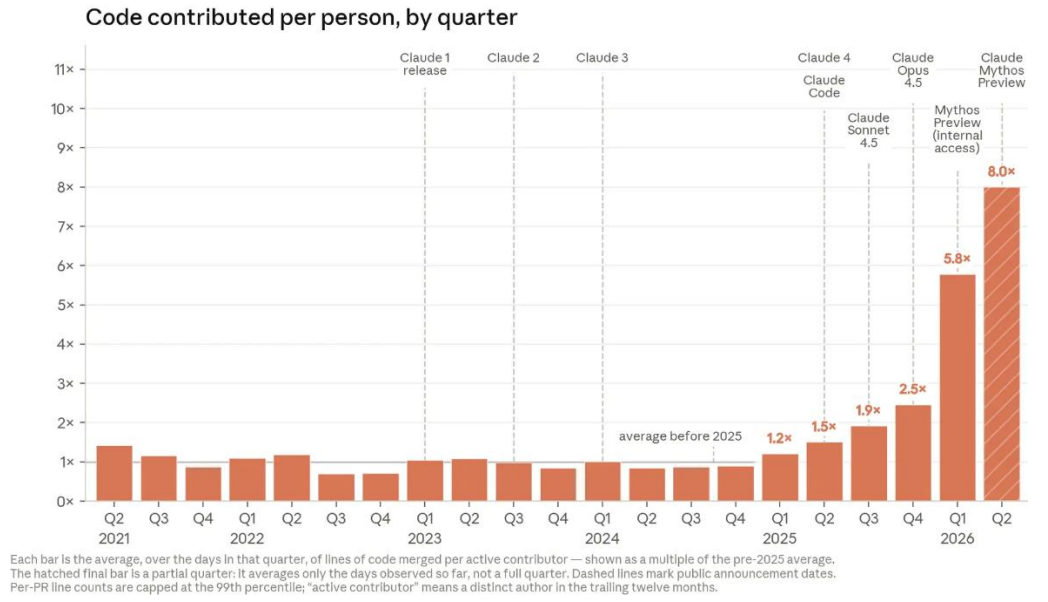
1.4. RSI：让 AI 自己训练 AI

如果说 Multi-Agent 是空间上的 scaling，那么 RSI（Recursive Self-Improvement）就是时间上的 scaling——让 AI 系统不断改进自己，形成正向循环。

Anthropic 在 6 月的《When AI builds itself》文章中披露了几个数字：

- 截至 2026 年 5 月，代码库里超过 80%的代码由 Claude 写出；
- 2026 年 Q1 工程师人均每天合并的代码量是 2025 年之前的 8 倍；
- 2026 年 4 月有一个 AI Agent 端到端完成了一项 AI 安全研究，累计工作 800 小时，效果比人类研究员一周的成果还要好；
- 在代码性能优化测试里，Claude Preview 能做到 2.52 倍加速，接近熟练人类研究员 4 到 8 小时能做到的 4 倍。

图2：Anthropic 人均贡献的代码量快速提升



数据来源：Anthropic，东吴证券研究所

AI 写代码训练 AI，训练出来的 AI 更强，更强的 AI 写更好的代码，更好的代码又训练出更强的 AI。Anthropic 设想了三种未来：第一种是模型能力不再变强（几乎不可能）；第二种是有复利效应但不会指数级增长（目前处于这个阶段）；第三种是完全 RSI，人类在训练流程中的角色大幅缩小，进度完全只受算力限制。

第三种未来听起来很科幻，但 Anthropic 认为这是最大的风险。如果基础模型有对齐瑕疵，RSI 会放大这个瑕疵。当 AI 比我们更聪明时，失控风险更高。所以他们一边在

技术上推进 RSI，一边在伦理上呼吁放缓。

RSI 现在还处于早期阶段，主要体现在两个方面。第一是 AI 辅助研究。田渊栋等人创办的 Neo Lab、前 Anthropic 研究员创办的 Mirendil、OpenAI 前研究 VP 创办的 Core Automation，都在做让 AI 自动完成研究任务，例如读论文、提假设、写代码、跑实验、分析结果。第二是 AI 辅助后训练。很多公司已经在用 AI 生成 synthetic data 来做 RL 训练，用 AI 来优化 reward function 的设计。

但真正的 RSI（指 AI 完全自主地改进自己的架构、训练策略、数据配比）还没有实现。主要瓶颈在于资源分配决策仍然需要人类在驾驶座上。你不能让 AI 自己决定下一代模型应该是 10 万亿参数还是 20 万亿参数、应该在 coding 上多做 RL 还是在数学上多做 RL，因为这些决策涉及高昂的算力投入和时间成本。

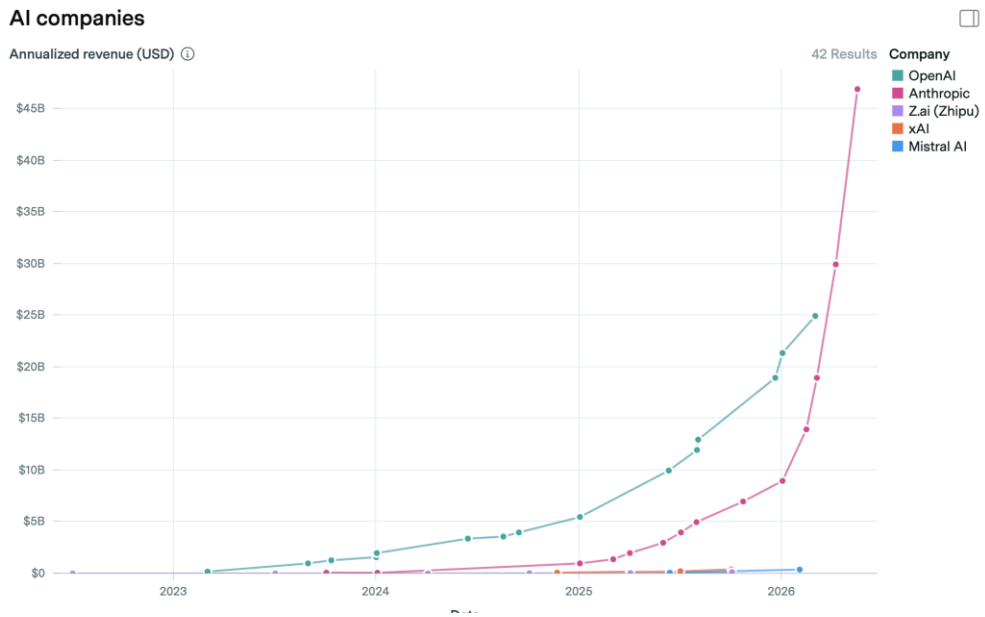
所以 RSI 的突破可能需要一个前置条件：训练成本大幅下降，让试错变得可承受。如果训练一个新模型的成本和时间大幅下降，那么 AI 就可以大规模并行地尝试不同的架构和策略，然后人类只需要在最后挑选最好的结果。

2. Frontier model：交替领先，终局未定

去年，Google 凭借 Gemini 2.5 和 3.0 的发布“一雪前耻”。相比此前 Bard 时代的尴尬表现，Gemini 3 不仅在多模态能力上实现飞跃，更被业界视为 Google 对 Scaling Law 最极致的一次执行。原生多模态架构、YouTube 的海量视频数据和 Waymo 积累的物理世界交互数据，构成了 OpenAI 和 Anthropic 难以复制的结构性优势。Google 用巨量算力和数据证明了 scale 确实 work，是通向 AGI 的路径之一。

但高光时刻只维持了两个月。2025 年底到 2026 年初，Agentic+coding 产品开始密集爆发，Claude Code 迅速占领开发者社区。Claude Code 成为大量开发者的日常工具。截止 2026 年 5 月 15 日，Anthropic 的 arr 达 470 亿美金，在绝对值和环比增速上都领先 OpenAI。

图3：全球主要大模型公司年化经常性收入（ARR）



数据来源：epoch ai，东吴证券研究所

Google 在这一轮 coding 军备竞赛中明显慢了半拍。问题出在哪里？我们认为在于战略选择，Google 最初把主要精力放在基础模型上，忽视了 coding RL。此外，Antigravity 的用户体验不佳，用的人少就拿不到真实使用数据，没有数据 RL 训练就跟不上，模型改进就慢。做好 coding 模型需要一个正向循环：好产品吸引开发者，开发者产生海量真实编码任务数据，用数据做强化学习训练让模型变强，模型强又让产品更好用。Claude Code 和 Codex 因为产品做得好，这个飞轮转得很快，Google 则卡在了第一步。

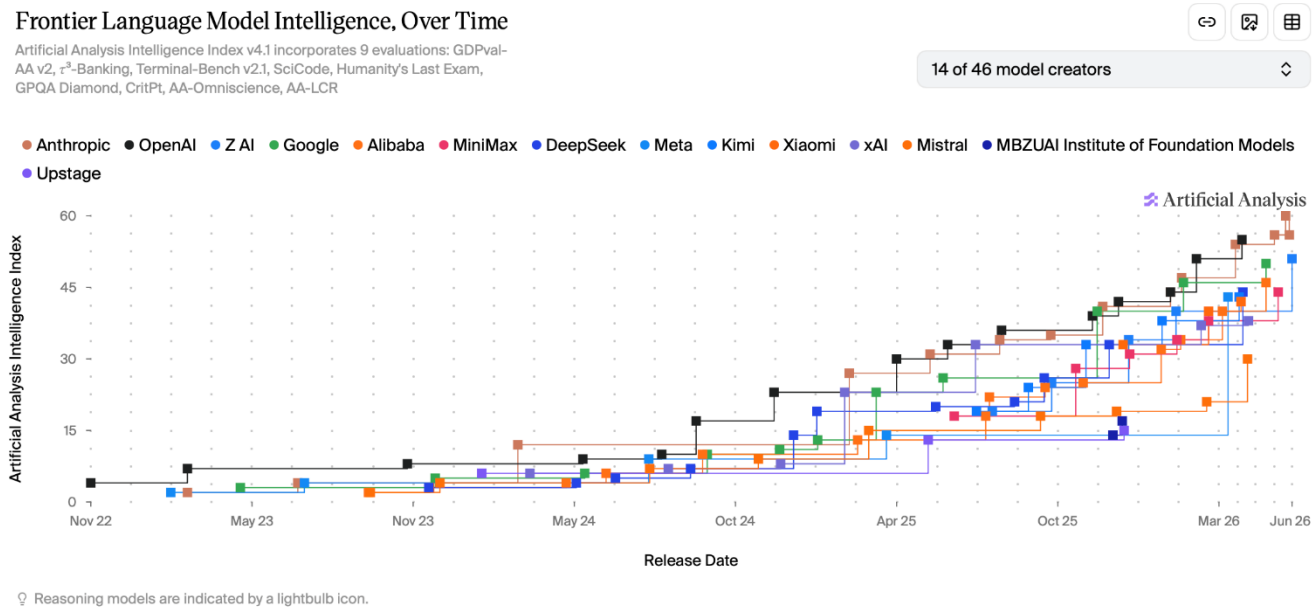
但 Anthropic 的领先优势在二季度开始出现裂缝。4 月中旬发布的 Opus 4.7 相较于 opus4.6 没有明显改善，而 tokenizer 的变化进一步隐形提价。此后，mythos 模型仅限于美国少数大公司使用；fable 则是上线后又被撤回，再度上线后“静默降智”（系统会把复杂任务偷偷降级到 opus4.8 的能力水平去完成）。与此同时，Anthropic 还遭遇了更底层的约束：算力供给不足开始限制产品迭代速度。他们不得不通过调整定价策略变相控制用量，不再允许第三方 Harness 按 API 价格使用 Claude，改成按 Output 价格计费，这又流失了一批用户。

OpenAI 抓住了这个窗口期。5 月发布的 GPT 5.5 和后续限量上线的 gpt 5.6，让 OpenAI 在 coding 能力上追平甚至在某些任务上超过了 Claude。OpenAI 此前被诟病算力储备过多、资本开支失控，此时反而成了 OpenAI 的竞争优势。因为有充足算力，OpenAI 可以用价格战去抢用户。这在 Anthropic 算力吃紧、不得不限流和涨价的时候，形成了鲜明对比。从实际效果看，很多开发者开始混用 Claude 和 Codex，甚至直接从 claude 切换到 codex。

站在 26 年中，回看过去三四年，大模型第一的位置从未真正稳固过。2023 到 2024 年 OpenAI 是话题的中心，2025 年 Google 凭 Gemini 3 短暂领先，2026 年初 Anthropic

靠 coding 反超，到现在 OpenAI 在模型能力上又追了回来。这种交替领先的格局背后，既有技术路线的分化（多模态 vs 编码专精 vs 通用智能），也有组织能力的差异（Google 的大公司病 vs Anthropic 的专注 vs OpenAI 的资源优势），还有产品飞轮转速的竞赛。

图4：大模型发布节奏越来越密集，且 frontier 模型竞争焦灼



数据来源：artificial analysis，东吴证券研究所

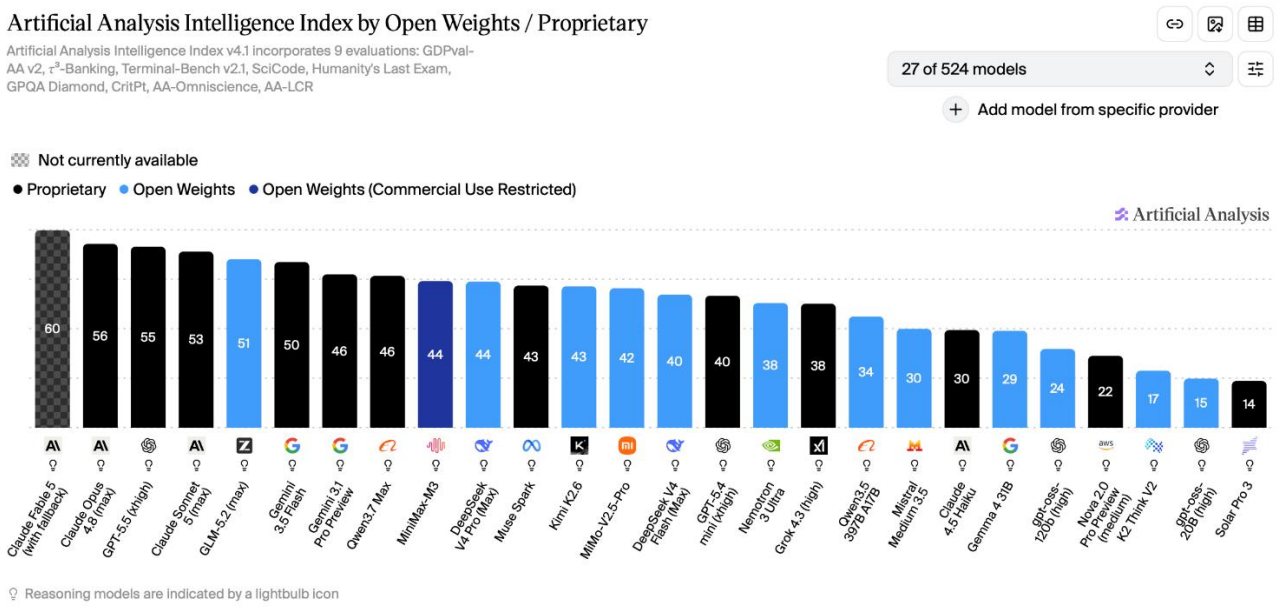
那么 Google 在今年下半年有没有可能再度领先？我们认为取决于三个条件能否同时满足。第一是算力能否成为核心制约因素。如果 Anthropic 继续受困于算力供给，而 Google 的 TPU 集群和云端优势能充分释放，天平就会倾斜。第二是用户是否开始更关注性价比而非 frontier intelligence。如果企业从 token maxing 转向审视 ROI，那么 Google 在推理成本上的优势会更明显。第三是多模态能否成为下一阶段 scaling 的主方向。如果业界共识从更大的 LLM，转向**理解世界的全模态模型**，那么 YouTube、Waymo、Android 积累的多模态数据就会从潜在优势变成现实壁垒。

大模型的战争还远未到终局，但有一点已经很清楚：领先优势的半衰期正在急剧缩短。从 Gemini 3 发布到被 coding 赛道反超，只用了两个月。从 Anthropic ARR 反超到遭遇算力瓶颈和品牌危机，也只用了一个季度。在这样的节奏下，下半年谁能领先，可能不取决于谁的模型参数更大，而取决于谁能平衡好模型能力、产品体验、数据飞轮、算力供给这四个要素。

3. 价格战：开源模型正在冲击闭源模型的定价根基

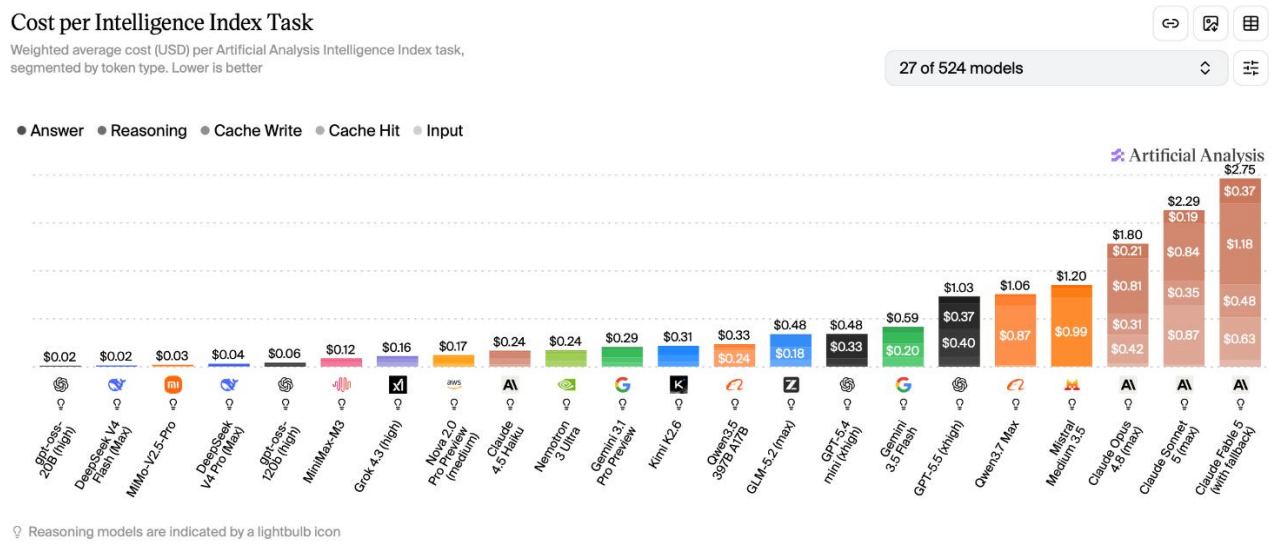
中国开源模型的能力快速提升。虽然和海外顶尖闭源模型仍有差距，但已经过了可以用的阶段，且价格显著更低。

图5：artificial analysis 榜单上，开源模型占据半壁江山



数据来源：artificial analysis，东吴证券研究所

图6：开源模型的成本显著低于 claude opus 4.8/fable 5

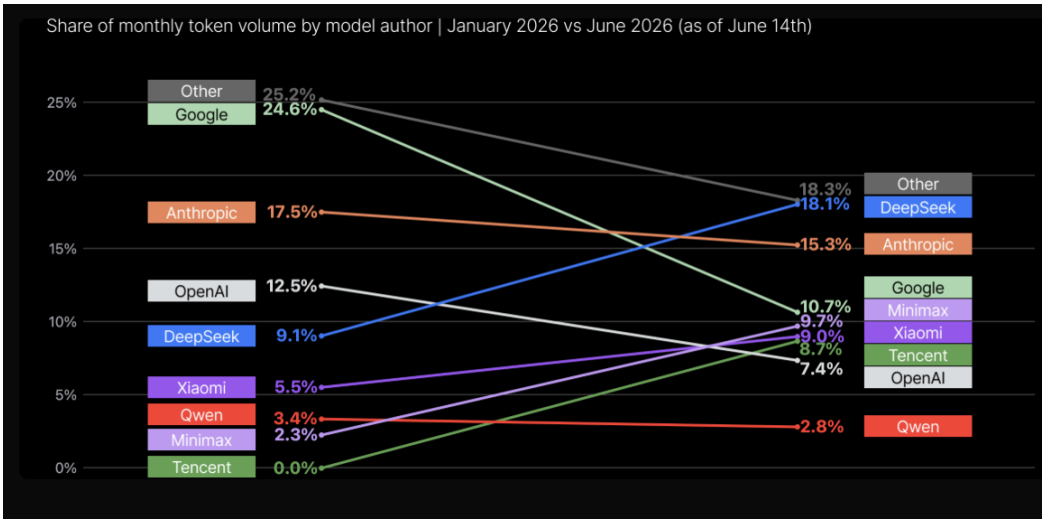


数据来源：artificial analysis，东吴证券研究所

Anthropic 和 OpenAI 的旗舰模型越强越贵，客户就越有动力找替代方案。企业客户开始在复杂任务用旗舰闭源（Opus/Fable/GPT-5.5），简单任务用开源或低价模型（DeepSeek/MiniMax/Qwen）。

客户以为自己有很多复杂任务需要旗舰模型，但真正跑起来发现，可能 90%的场景根本不需要 Fable 5 或 GPT-5.6 这种级别的能力。医保申诉信、客服邮件、代码补全、文档摘要——这些占 AI 使用量大头的场景，开源模型就能搞定。

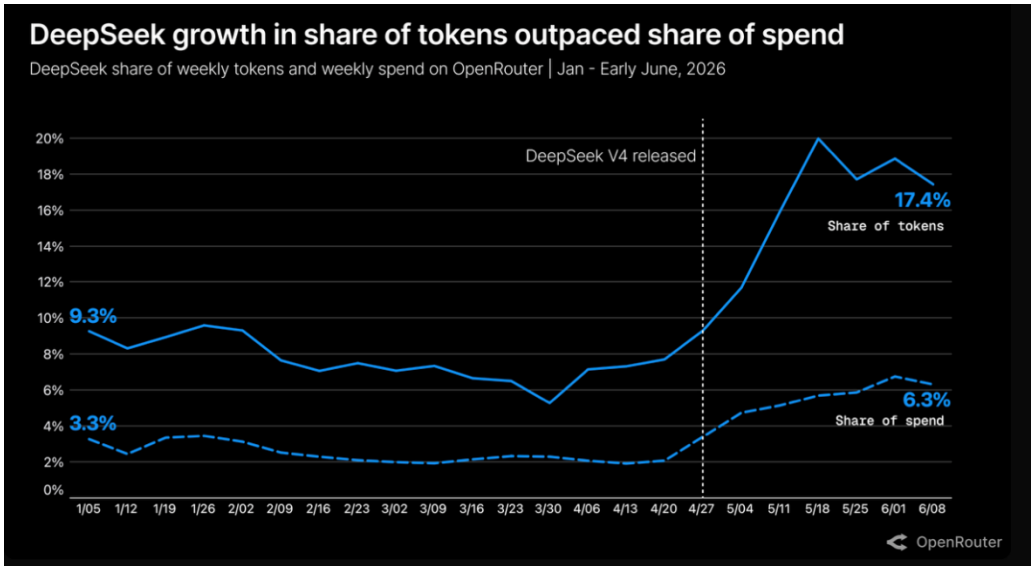
图7：OpenRouter 上开源模型 token 占比快速上升



数据来源：openrouter，东吴证券研究所

开源模型 tokens 消耗量大，而闭源模型的定价高。开源模型的 tokens 量份额在增长，但因为定价便宜，收入并没有同步高增。

图8：openrouter 上 deepseek 的 tokens 份额涨幅比收入份额涨幅更大



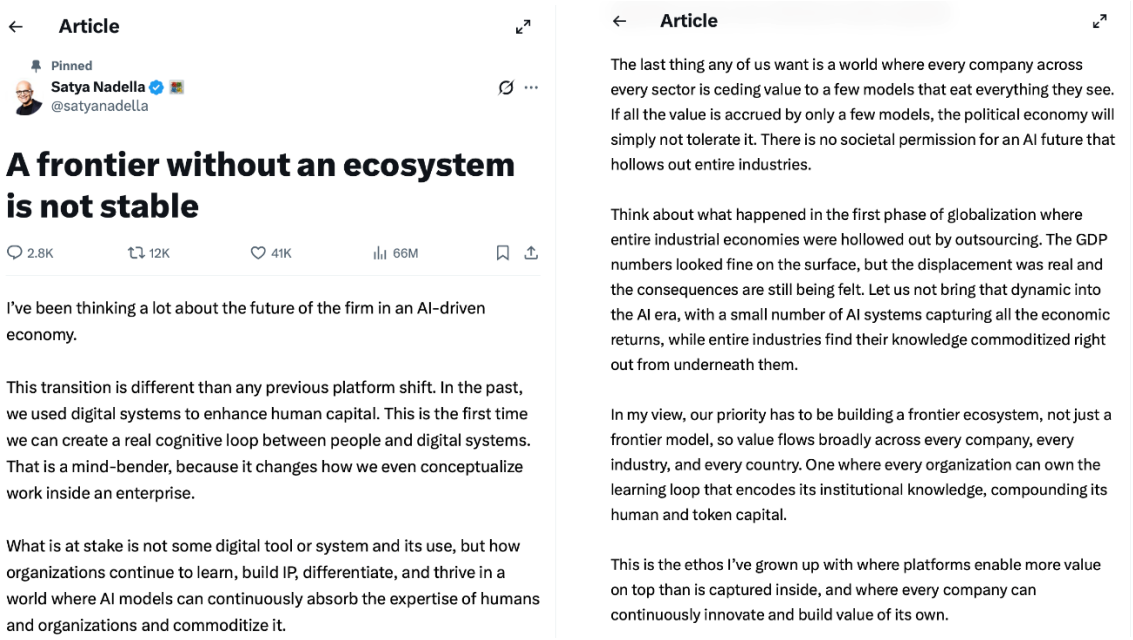
数据来源：openrouter，东吴证券研究所

模型路由器是这个格局的关键基础设施。

客户提交一个任务，路由器自动判断这个任务复杂度，简单任务扔给 DeepSeek（便宜），复杂任务扔给 Fable 5（贵但准）。一年前全球大型企业基本上不会碰任何中国开源模型，担心安全漏洞。现在很多金融服务公司都在用，因为没发现安全问题的证据，而且确实能省钱。

微软甚至考虑用 DeepSeek 来驱动 Copilot Cowork（白领 AI 助手，处理 Outlook、Teams、Office 里的任务）。Nadella 在 Anthropic 被禁之后发了一条 X：“我们最不愿看到的是，所有行业的每家公司都将价值拱手让给那些吞噬所见一切的少数模型。”微软不想被 Anthropic 和 OpenAI 绑架定价权。

图9：微软 CEO 纳德拉的文章



数据来源：Satya Nadella，东吴证券研究所

Anthropic 和 OpenAI 会不会降价？

现在闭源模型能守住高价，是因为最复杂的 1%任务确实不可替代。但如果开源模型继续追赶，这个 1%可能会变成 0.1%。到那时候 Anthropic 和 OpenAI 要么降价（收入受损），要么继续高价（份额流失）。开源正在吃掉 90%的用量场景，闭源只能守住最难的 1%，天花板被结构性压制。

开源模型会不会闭源？

模型公司要赚钱、要回收数据飞轮。国产厂商想闭源（为了赚钱），但企业客户在推开源（为了省钱）。两个力量在反向对冲。最终结果取决于：模型的不可替代性溢价到底值多少。

Anthropic 和 OpenAI 现在面临的是跟英特尔当年一样的困境——高端芯片利润率很高，但份额被 ARM（便宜、够用）一点点蚕食。最后英特尔守住了数据中心和高性能计算，但 PC 和移动市场全丢了。Anthropic 和 OpenAI 会不会走同样的路？我们觉得概率不小。

4. 模型在做应用，应用也在做模型

一方面，模型公司在做应用。

Anthropic 今年在产品层动作频繁，而且方向很明确，即编程、设计、金融、法律、医疗、生命科学这几个高 ARPU、高数据价值的垂直场景。①Claude Code，直接杀进 AI Coding 赛道，跟 Cursor、Codex 正面竞争。②Claude Design 4 月份发布，直接对标 Figma。发布当天 Figma 股价跌 7%。Claude Design 的定位是把新想法到可讨论原型的时间从几天压缩到几小时。对按席位收费的协作设计工具来说，绕过早期探索阶段意味着使用频次压缩，席位需求下降。③招聘信号：Anthropic 挖走了 Workday 的 CTO（大概是为了做 HR 应用），并在金融服务、法律、医疗、生命科学四个垂直领域发布了工程经理招聘。这说明 Anthropic 不只是做两个垂直应用试试水，而是系统性地在多个高价值垂直领域布局。

Anthropic 往下游延伸（做 Claude Code、Claude Design、垂直行业应用），本质上是从卖 token 变成卖解决方案。API 收入的天花板是推理量，垂直应用收入的天花板是企业愿意为解决方案付多少钱。这意味着：①中间层应用公司的生存空间被挤压。纯做应用、场景不够深、数据优势不明显的公司最危险。Cursor 能活下来是因为被 SpaceX 收购拿到算力，其他 AI Coding 公司（Replit、Lovable、Trae）还困在供应商即竞争对手的结构里。②数据飞轮的门槛被拉高了。以前应用公司讲的故事是“我有用户数据，可以训练更好的模型”。但现在 Anthropic 和 OpenAI 自己下场做应用，他们的数据积累速度比你快得多，而且可以把数据直接烧进下一版基础模型，不只是做个 fine-tune。这个闭环的时间线更快、效果更好。

但另一方面，应用公司也在做模型。

2026 年 4 月底，SpaceX 以 600 亿美元收购了 Cursor。然后 6 月 17 号，Cursor 宣布自研 1.5 万亿参数的底座模型，背后用的是 SpaceX 的算力——超过 10 万张 GPU。Cursor 在从纯应用公司转型为应用+模型一体。

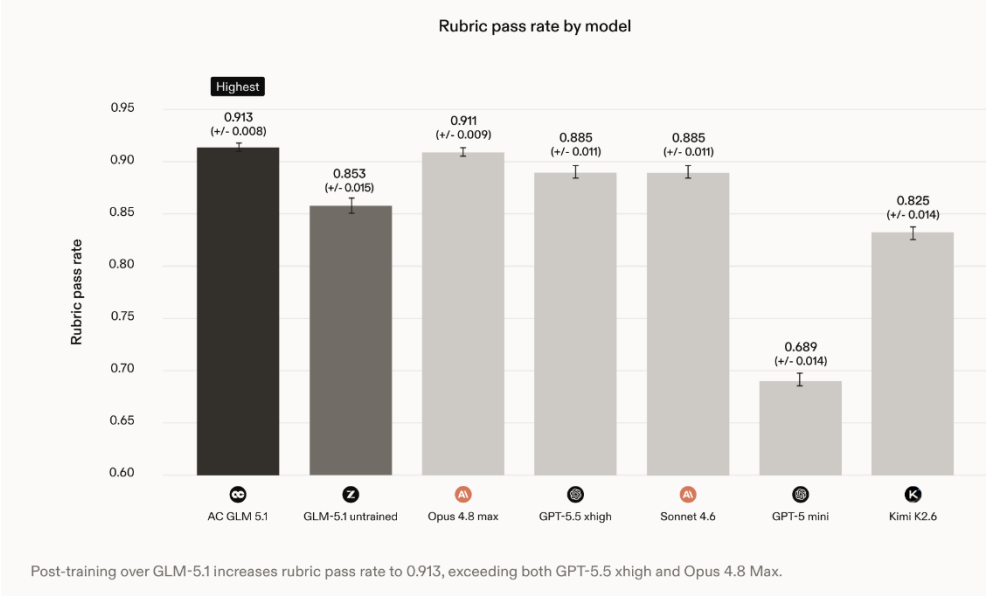
Cursor+spacex 的组合，优势在于：①**SpaceX/xAI 的算力资源。**xAI 之前有 Colossus 1（22 万张 GPU，后来租给 Anthropic 了），现在 Cursor 用的 10 万卡可能来自 SpaceX 的新集群或者 Colossus 2。这让 Cursor 从完全依赖 Claude 或 GPT 的 API，变成有自己的模型底座+算力。之前 Cursor 的致命问题是：你的供应商（Anthropic、OpenAI）同时也是你的竞争对手（Claude Code、Codex）。现在 Cursor 有了自己的模型，这个结构性矛盾被解开了。②**Cursor 积累的高质量数据。**Cursor 用户写代码的过程、接受/拒绝 AI 建议的选择、调试和修改的记录——这些数据对训练 coding 模型非常有价值。而且这些数据是实时的、带用户反馈的，比 GitHub 上的静态代码库质量高得多。xAI 拿到这些数据，可以持续优化模型在 coding 场景的表现。

跟 Claude Code 和 Codex 比，Cursor+SpaceX 的优势是**成本结构更灵活、有自己的算力和模型底座、不被单一上游锁死。**

再举另一个例子。

6 月 22 日，harvey ai（一家做法律 ai 的公司）发布文章，通过与 Applied Compute 合作，在 GLM-5.1 基础上进行 post-training，训练出在 Harvey Legal Agent Benchmark (LAB) 上表现最强的法律 agent 模型。训练后的 GLM-5.1 在 rubric pass rate 上达到 0.913，**超过 GPT-5.5 xhigh 和 Opus 4.8 Max**；all-pass rate 达到 0.126，超过 GPT-5.5 xhigh，逼近 Opus 4.8 Max。

图10：按照模型划分的 rubric pass rate



数据来源：harvey，东吴证券研究所

Harvey 为什么要这么做？

1）数据隐私是刚需。法律文档涉及客户机密、诉讼策略、并购细节，这些东西发到 OpenAI 或 Anthropic 的 API 上，即使有企业协议，客户也不会接受。大律所和企业法务部门对数据主权的要求极高，Harvey 如果想卖给这些客户，必须证明数据不出他们的控制范围。自己训练开源模型，推理在自己的集群或客户的 VPC 里跑，这个问题就解决了。

2）减少对闭源模型的依赖。如果 Harvey 完全依赖 OpenAI API，OpenAI 随时可以调价、限流、改 terms，甚至自己做法律 agent 产品跟 Harvey 竞争。Claude 也一样——Anthropic 没有承诺永远不做垂直应用。Harvey 用开源模型训练出自己的模型，即使未来 GPT-6 或 Opus 5 出来性能更强，Harvey 也可以选择继续用自己的模型（如果够用），或者拿新的开源模型再训练一版，而不是被闭源厂商绑定。

垂直 AI 应用公司的长期生存策略是**开源模型 + 自有训练能力 + 垂直场景优化**，而不是**闭源 API + prompt engineering**。后者的护城河太浅——任何竞争对手都可以调同样的 API，差异只在 prompt 和产品包装上。前者的护城河在全栈能力：grader 设计、harness 优化、RL 训练、推理基础设施，这些东西需要时间和专业团队积累，不是简单抄得走的。

5. 从 Vibe Coding 到 vibe working 再到 vibe creating：AI 应用的三波浪潮

5.1. Vibe Coding——已经爆发的现实

Vibe Coding 为什么能率先爆发？核心是可验证性。代码写出来能不能跑、过不过测试，有编译器给出 yes or no。这让大模型厂商能用强化学习在 RL 阶段疯狂训练，每一次成功或失败都是清晰的信号。而对用户来说，代码跑通了就是跑通了，不需要主观判断质量好坏。这种高可验证性构成了正向飞轮：好产品吸引开发者，开发者产生海量真实编码任务数据，用数据做 RL 训练让模型变强，模型强又让产品更好用。

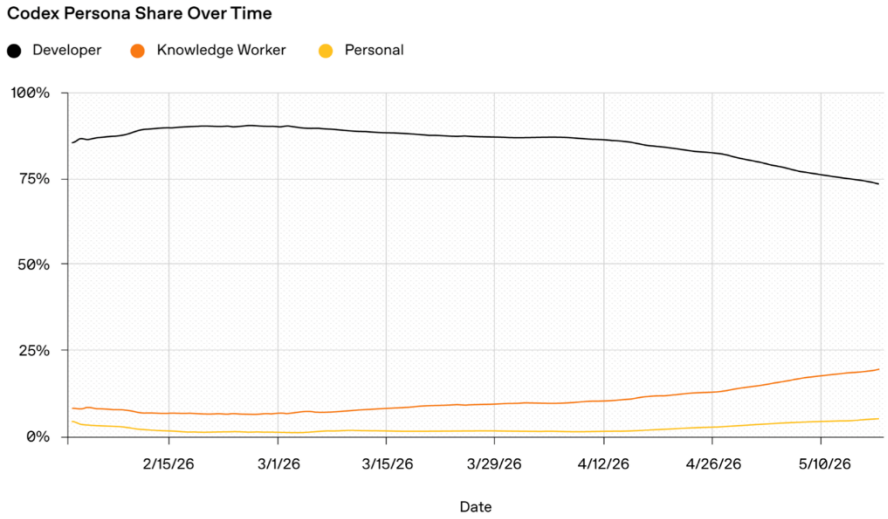
但 Vibe Coding 的天花板也在隐约浮现。GitHub Copilot 在 6 月 1 日从固定月费改成按 Token 计费后，用户账单暴涨。Uber 四个月花光了全年 AI 预算。亚马逊砍掉了内部 AI 工具使用量排行榜 Kirorank。企业开始从 token maxing 转向精细化管理。下一个增长点在哪里？

5.2. Vibe Working——正在渗透的白领办公

OpenAI 在 2026 年 6 月披露了一个关键数据：知识工作者——那些天天做表、做 PPT、跟数据打交道的人——现在占 Codex 用户的 20%，增速是程序员群体的 3 倍。AI Agent 的风已经从程序员工位吹到了普通打工人的桌面。

图11：OpenAI 报告中关于 knowledge worker 的描述

Knowledge workers now represent about 20 percent of Codex users and are adopting it more than 3 times as fast as developers. This includes roles that span product and project management, design, research and academia. Personal users represent more than 5 percent of Codex users and are growing more than 4 times as fast as developers, with substantial use in hobbies and creative work, education and self-learning, personal finance, and entertainment.



数据来源：OpenAI，东吴证券研究所

Vibe Working 的逻辑和 Coding 一脉相承，但场景更宽广。月之暗面的 Kimi Work、Anthropic 的 Claude Cowork、OpenAI 的桌面版 Codex，核心能力都是把原本在终端里跑的 Agent 能力搬到图形界面，让它能操作 Excel、PPT、浏览器、PDF。不用敲命令，不用配环境，直接用自然语言描述目标，AI 拆解任务、并行执行、调用工具、整理文件夹，最后交付文档、表格、PPT。

实际应用场景已经在跑通。金融行业的投研人员让 AI 调研近 10 年持仓数据、做估值建模、自动生成自选股复盘；市场人员让 AI 读取本地产品方案、调用浏览器登录数据看板、分析用户评论生成市场分析报告再转成 PPT；券商分析师用 AI 整理路演纪要、清理口语化表达、生成深度总结和买方画像。金融已经成为 Anthropic 第二大收入来源，而且从个人付费向企业报销的转变已经开始，部分机构已经给员工配了 agent 企业账号。

但 Vibe Working 面临的摩擦力比 Coding 大得多。第一是可验证性不足。一份 PPT 好不好看、一个投资建议对不对，短期内根本没有客观验证标准。没有清晰的 reward signal，RL 训练就很难做，用户信任门槛也更高。第二是合规审计的空白。企业最敏感的财务数据、HR 数据、合同数据放在本地，Agent 要操作这些数据就涉及权限、审计、留痕、追溯，目前这套合规框架还是空白。这会把大量企业客户挡在门外，短期内白领 Agent 的主战场还是个人用户和中小企业。

5.3. Vibe Creating——值得期待的未来

如果 Coding 是替代程序员的脑力劳动，Working 是替代白领的重复性工作，那么 Creating 瞄准的是人类最后的堡垒——创意本身。你用自然语言描述一个想法，AI 帮你

生成视频、音乐、游戏、设计作品。

视频生成是 Creating 最成熟的子领域。可灵能生成电影级的表演，表演有层次感，光影处理也更复杂。seedance 虽然风格更偏抖音短视频，但迭代速度极快。但商业化规模和 Coding 相比还差几个数量级。原因在于：第一，抽卡率成功率太低。输入同一个 prompt 重复生成十次，可能只有一次是能用的，剩下九次都是废片。而且没法精确修改——前五秒可以用但后五秒不行，只能整段重做。这种方式比传统后期制作费钱得多。第二，应用场景有限。目前主要是做短一点的 AI 短剧、广告片，长视频在某些镜头上可以用，但完整的内容生产流程还跑不通。

Vibe Creating 的商业化路径可能和前两波完全不同。Coding 和 Working 是生产力工具，按使用量或订阅收费。但 Creating 更像内容平台，而且 Creating 的价值不在于替代人类创作者，而在于降低创作门槛，让更多人能表达想法。

5.4. 从高可验证到低可验证的递进

从 vibe coding 到 vibe working 再到 vibe creating，AI 应用从高可验证性向低可验证性场景递进，速度取决于能否构建有效的 reward。

Coding 爆发最快，因为编译器提供了即时而客观的反馈。Working 渗透较慢，因为评价标准既有客观的部分，也有主观的部分。Creating 最难商业化，因为一个视频是否好看、一款游戏是否好玩，完全是主观的。

RL 训练需要清晰的 reward signal，没有信号就没法让模型自我进化。但长期看，Creating 的天花板可能是最高的。因为它不只是提效，而是在扩大创作者的定义。Coding 让不会写代码的人能开发软件，Working 让不懂 Excel 的人能做数据分析，Creating 会让没学过视频剪辑的人能拍电影、没学过编曲的人能做音乐。每一波浪潮都在重新定义“谁能做什么”，而 Creating 重新定义的是人类的想象力和表达欲。

6. 投资建议

下半年投资主线：OpenAI 和 Anthropic 上市进展，Google 能否翻盘，中国开源模型全球化。

海外龙头上市窗口：OpenAI 和 Anthropic 或在今年下半年到明年陆续上市。OpenAI 关注 GPT 5.6 能否全量上线并止住份额流失，算力优势能否持续。Anthropic 关注算力瓶颈能否缓解。两家上市后会对整个板块形成定价锚点。

Google 的翻盘机会：Coding 追赶看 Antigravity 能否重启产品飞轮。长期看多模态+世界模型布局，YouTube/Waymo 数据和 Omni 架构是 OpenAI 和 Anthropic 没有的结构性优势。TPU 对外销售如果能像 NVIDIA GPU 一样规模化，且推理成本有明显优势，Google Cloud 增速可能超预期。

中国开源模型全球化机会：国产开源模型能力上已经快速追赶，从“能用”阶段进入“好用”阶段。DeepSeek 算法创新领先，成本优势显著；智谱 GLM-5.x 系列在 coding 上表现优秀；minimax m3 在原生多模态有所突破。我们认为随着企业端从 tokenmaxxing 转向开源节流，国产模型在全球市场的份额有望提升。

7. 风险提示

1. Scaling Law 失效风险：如果下一代模型的能力提升明显放缓，整个行业的投资逻辑需要重估。

2. ARR 增速放缓风险：美国企业从 Token 竞赛转向 Token 节流，收紧 AI 预算、减少无效 token 支出，可能短期影响 Anthropic 和 OpenAI 的 ARR 增速

3. 自由现金流与资产折旧危机：大厂 FCF 转负可能引发资本市场对 AI 投资的信心动摇。

东吴证券 SOOCHOW SECURITIES

免责声明

东吴证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。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，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。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。

在法律许可的情况下，东吴证券及其所属关联机构可能会持有报告中提到的公司所发行的证券并进行交易，还可能为这些公司提供投资银行服务或其他服务。

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，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。

东吴证券投资评级标准

投资评级基于分析师对报告发布日后 6 至 12 个月内行业或公司回报潜力相对基准表现的预期（A 股市场基准为沪深 300 指数，香港市场基准为恒生指数，美国市场基准为标普 500 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），北交所基准指数为北证 50 指数），具体如下：

公司投资评级：

买入：预期未来 6 个月个股涨跌幅相对基准在 15%以上；

增持：预期未来 6 个月个股涨跌幅相对基准介于 5%与 15%之间；

中性：预期未来 6 个月个股涨跌幅相对基准介于-5%与 5%之间；

减持：预期未来 6 个月个股涨跌幅相对基准介于-15%与-5%之间；

卖出：预期未来 6 个月个股涨跌幅相对基准在-15%以下。

行业投资评级：

增持： 预期未来 6 个月内，行业指数相对强于基准 5%以上；

中性： 预期未来 6 个月内，行业指数相对基准-5%与 5%；

减持： 预期未来 6 个月内，行业指数相对弱于基准 5%以上。

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应当充分考虑自身特定状况，如具体投资目的、财务状况以及特定需求等，并完整理解和使用本报告内容，不应视本报告为做出投资决策的唯一因素。

东吴证券研究所
苏州工业园区星阳街 5 号
邮政编码：215021
传真：（0512）62938527
公司网址： http://www.dwzq.com.cn