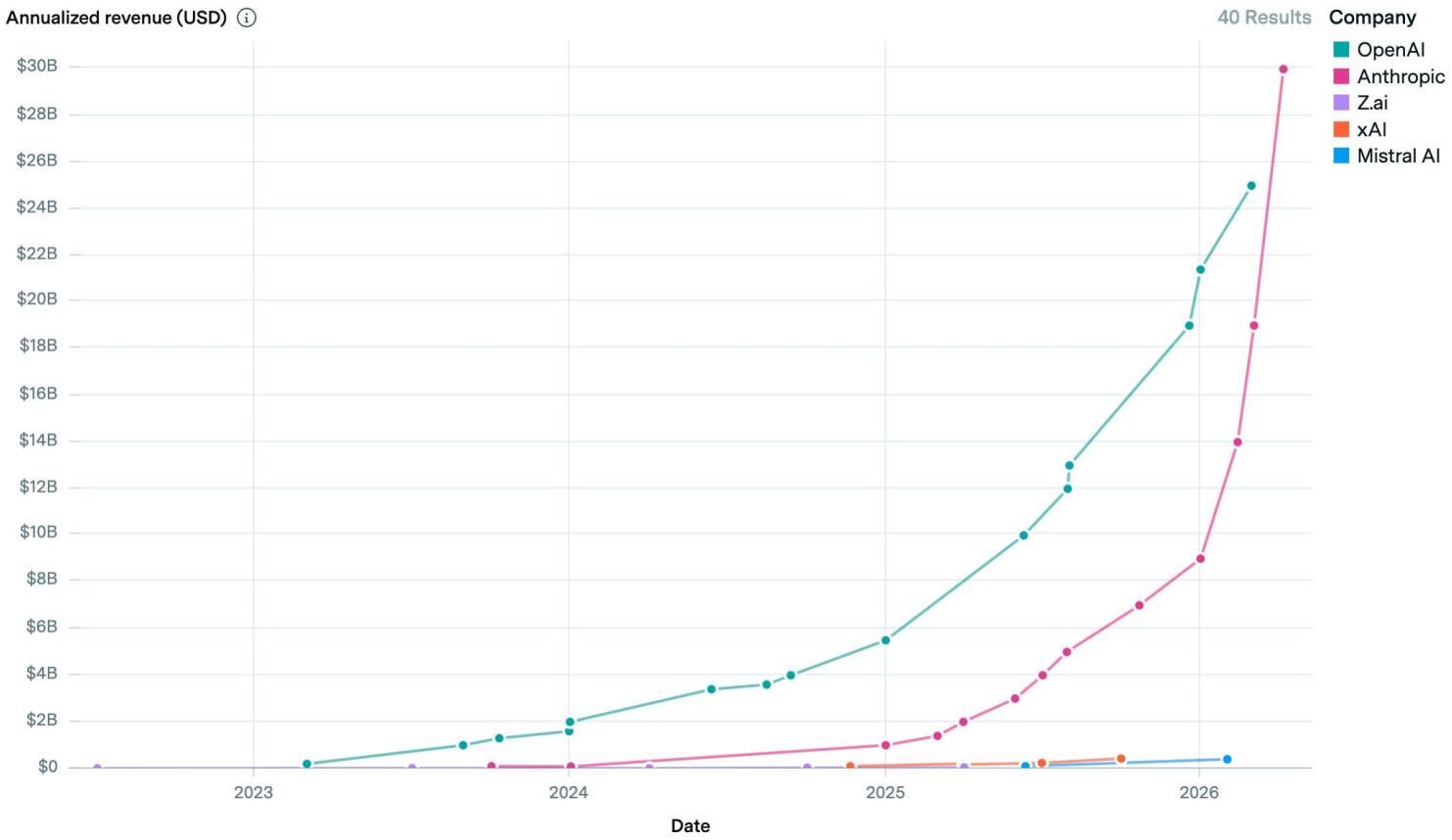
收入 / ARR ——大模型的第一观测指标

26 年 5 月，Anthropic CEO 在纽约金融活动上透露：Anthropic Q1 年化收入增长 80 倍；公司约 3500 人；Mythos 模型可能发现"数万"未公开漏洞；中国 AI 模型落后 6-12 个月；美国其他头部实验室落后 Anthropic 约 1-3 个月；个别 SaaS 公司可能因 AI 颠覆而倒闭。

AI companies



Anthropic Will Surpass Alphabet in Revenue by Mid-2028

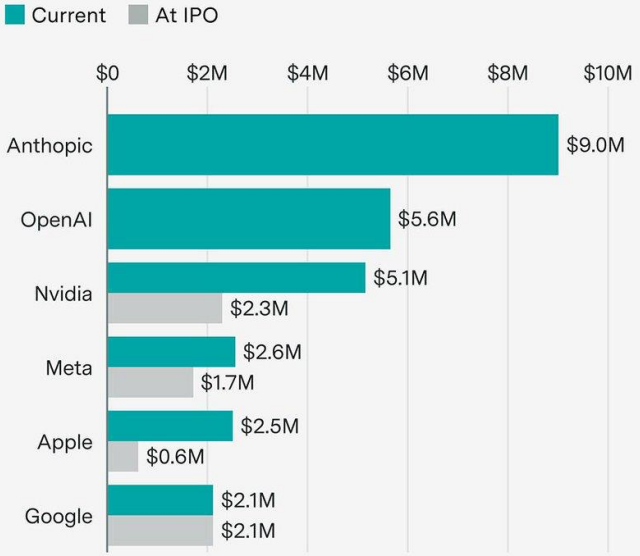
Revenue Trajectory · USD Billions · 2025-2030E · Anthropic ARR vs Alphabet GAAP Revenue

SOURCE: WORLD TRADE SECURITIES

MAY 2026

Anthropic and OpenAI earn more revenue per employee than major public tech companies

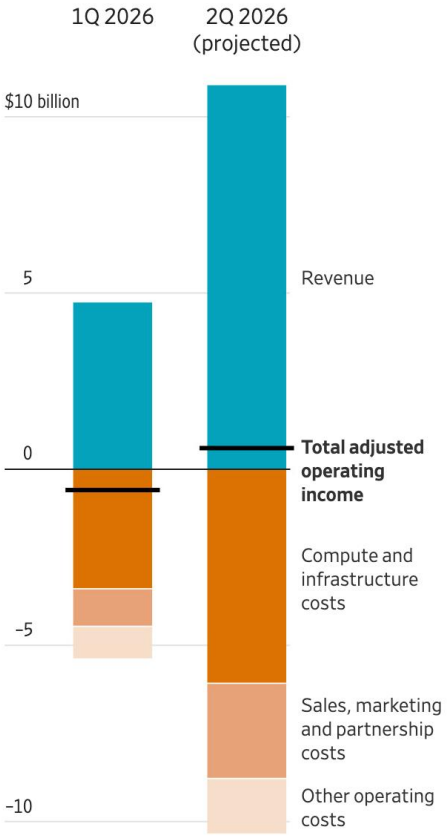
Annualized revenue per employee, adjusted for inflation.



Anthropic and OpenAI's figures extrapolate employee growth to avoid a mismatch between data reporting dates

EPOCH AI | CC-BY epoch.ai

Anthropic's operating income, by segment

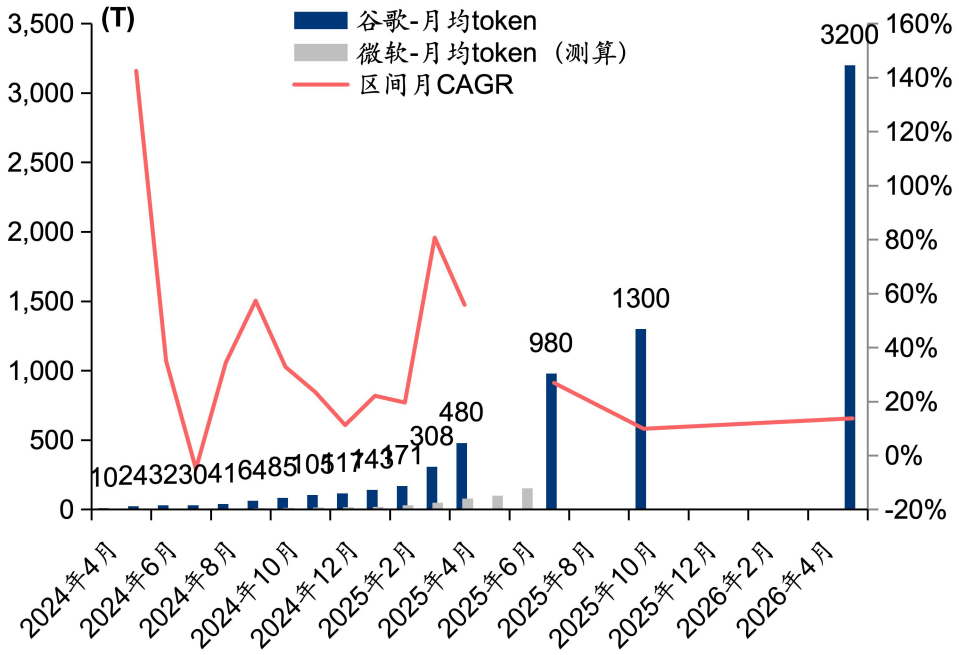


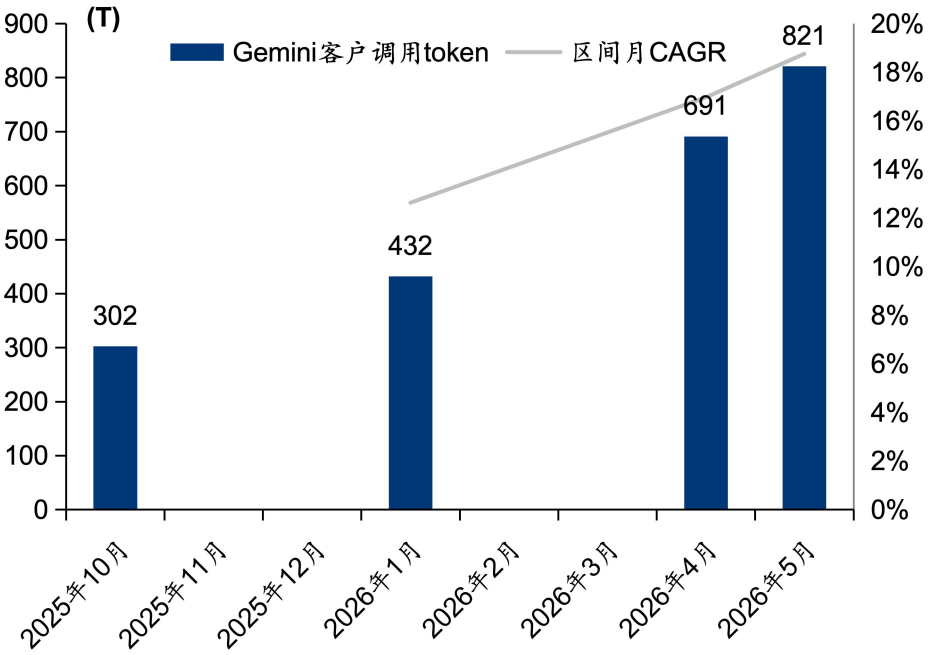
Source: financial information shared with investors

Nate Rattner/WSJ

时间	公司	数据 / 事件	关键信息	含义
2026-05-08	Anthropic	**All-In Podcast 讨论 Anthropic 是否成为下一代 AI 垄断公司**	Apple Podcast 官方简介显示，本期讨论 SpaceX-Anthropic 算力交易、Anthropic 增长轨迹，以及“是否会成为下一个伟大垄断”。Shortform 摘要称，Anthropic ARR 从年初约 **100 亿美元** 到 **4 月约 440 亿美元**，并提到极端外推下 2027 年可能冲击 **1 万亿美元 ARR**。	市场开始用“垄断型平台公司”框架理解 Anthropic，而不只是普通 SaaS 或 API 公司。
2026-05-08	Anthropic / OpenAI	**Epoch AI：收入/员工远超传统科技公司**	Epoch AI 估算 Anthropic 和 OpenAI 的收入/员工分别约 **900 万美元**、**550 万美元**，高于 Forbes Global 2000 中任何上市科技公司；其方法依赖媒体报道与估算，因此口径应视为近似。	大模型公司呈现出极高的人均收入密度，说明 AI 原生公司可能拥有更高的组织杠杆。
2026-05-21	Anthropic	**WSJ：Q2 预计首次实现季度经营利润**	WSJ 报道称，Anthropic 预计 2026Q2 收入 **109 亿美元**，较 Q1 的 **48 亿美元** 翻倍以上，并实现约 **5.59 亿美元经营利润**；收入增长主要来自企业客户、代码工具和 Agentic Claude 应用。	Anthropic 提前进入盈利窗口，打破“前沿模型公司必须多年烧钱”的单一路径。
2026-05-22 /05-24	OpenAI / Anthropic	**The Information：OpenAI Q1 收入仍领先，但 Anthropic 增速更快**	**The Information 报道称 OpenAI Q1 收入约 57 亿美元，领先 Anthropic 约 10 亿美元**；***OpenAI Q1 调整后营业利润率 -122%，营收增长伴随巨额亏损***；但 Anthropic 年化收入已接近 **450 亿美元**，**Q2 预计收入近 110 亿美元、经营利润约 6 亿美元。**	OpenAI 仍是收入体量领先者，但 Anthropic 的 run-rate 已经逼近反超。
2026-05	OpenAI	**OpenAI 增长驱动来自 Codex、企业销售、广告测试、GPT-5.5、图像生成**	The Information 转述显示，OpenAI Q1 表现受 Codex、企业销售和 ChatGPT 广告测试拉动。	OpenAI 收入结构正在从 ChatGPT 订阅/API 扩展到 Coding、企业、广告和多模态内容生产。
2026-02 / 2026-05	Anthropic / OpenAI	**Epoch AI：Anthropic 有望在年化收入上追上 OpenAI**	Epoch AI 认为 Anthropic 自达到 10 亿美元年化收入以来，年化收入增速约 **10x/year**，高于 OpenAI 的 **3.4x/year**；其模型估算 Anthropic 可能在 2026 年中左右与 OpenAI 交叉，但也提示 Anthropic 增速可能已放缓。	ARR 竞争已从“OpenAI 单极领先”进入“Anthropic 高速追赶”阶段。

Token 已经成为第二观测指标，但是“长协”值得关注





时间	公司	事件	关键信息	含义
2026-05-19	OpenAI	**推出 Guaranteed Capacity**	OpenAI 官方称，该产品为企业的核心产品、Agent 和客户工作流提供长期 OpenAI compute 访问；客户可选择 **1–3 年承诺**，折扣随年度承诺金额提高，并可在 OpenAI 产品组合中消耗该承诺额度。(OpenAI)	OpenAI 把 token / compute 从按需 API 消费，升级为长期容量合同。
2026-05-19	OpenAI	**覆盖生产系统、客户应用、AI agents**	OpenAI 官方写明，Guaranteed Capacity 用于 production systems、customer-facing applications、AI agents；客户可基于 spend level 获得确定性容量。(OpenAI)	这说明长协不是给实验性调用，而是给关键生产负载。
2026-05-19	OpenAI	**跨模型、云和业务需求规划**	OpenAI 表单要求客户评估 model and cloud-provider planning、production workload growth、multi-year capacity needs、customer-facing AI systems and agents。(OpenAI)	本质是企业 AI 工作负载的容量规划工具。
2026-05-19	OpenAI / Brockman / Altman	**管理层明确强调产能约束**	Greg Brockman 称 OpenAI 用折扣 token 和容量确定性换取 1–3 年承诺，并预计世界会越来越感到产能受限；Sam Altman 也称客户越来越需要容量确定性，OpenAI 会卖到当前分配额度售罄。(X (formerly Twitter))	官方叙事直接强化“AI 产能短缺”框架。
2026-05-19 后	行业	**市场把它类比云计算 reserved instances / Savings Plans**	Techmeme 聚合的讨论中，有评论将其类比 AWS 的长期承诺、reserved instances / Savings Plans，并提到未实现收入 / backlog 口径。(Techmeme)	AI token 开始向云计算合同模式靠拢。

Token 长协可能改变 AI 行业的披露方式。 过去市场看 AI 需求，主要看历史 token 消耗量，这类数据是滞后的；长协模式成熟后，模型公司可能披露未来 1–3 年 token 承诺、已锁定容量、AI backlog 或 AI RPO。这会把大模型公司的收入能见度推向云计算公司式的合同披露，也会提高硬件产业链的订单可见度。对于 GPU、网络设备、光模块、液冷、数据中心和电力资产来说，这是“需求可预见性上升”的信号。

第一，token 从“API 现货”变成“战略资源”。
以前企业觉得 token 是随时可以买到的资源，只要付费就能调模型。但 Claude Code 限流、OpenAI / Anthropic 产能紧张、Google token 调用暴涨之后，企业会越来越担心：如果核心业务已经嵌入 AI agent，一旦模型服务限流，业务流程就会中断。Guaranteed Capacity 直接回应这个痛点：客户用 1–3 年承诺换取价格折扣和容量确定性。(OpenAI)

第二，长协会让“公共 tokens”变少，进一步强化紧缺预期。
如果大客户提前锁定一部分 OpenAI 产能，剩下可供现货市场、普通 API 用户、长尾开发者使用的容量会减少。这个逻辑和云计算、IDC、电力、GPU 租赁很像：越多人签长期容量，现货资源越稀缺，供给紧张叙事越容易自我强化。OpenAI 官方也明确说，这一产品是围绕生产系统、客户应用和 AI agents 的关键工作流设计的。(OpenAI)

第三，它把 AI 收入披露从“滞后 token 消耗”推进到“前瞻订单 / backlog”。
过去市场看 AI 需求，主要看历史 token 调用量，比如 Google 披露的每分钟 token 调用、月度 token 处理量，都是回溯数据。Guaranteed Capacity 这类合同出现后，未来可能出现更前瞻的指标：**未来 12–36 个月 token 承诺额、已锁定容量、剩余可售容量、AI backlog、AI RPO**。这会让模型公司的收入能见度提高，也会让硬件、云厂商、数据中心、光模块、交换机等产业链的需求预测更像云计算资本开支周期。

第四，对硬件投资者是强催化。
如果大模型公司能够拿到 1–3 年 token 长协，等于下游企业替上游 capex 提前做了一部分“需求背书”。OpenAI 拿着确定性订单，就更容易规划 GPU、数据中心、网络、电力、云合作伙伴容量；云厂商和硬件供应商也能看到更高的需求能见度。对市场来说，这会支撑一个更强的投资框架：**tokens 紧缺 → 企业签长协 → 模型公司扩 capex → GPU / 网络 / 数据中心订单能见度提高 → 估值中枢上移**。

Anthropic 开始抢算力大战

2025 年 6 月以来，Anthropic 几乎把能签的算力来源都签了一遍：AWS Trainium、Google TPU、NVIDIA GPU、CoreWeave、SpaceX / xAI、Akamai，甚至还有 Fluidstack 自建数据中心。

Anthropic 的算力策略已经从“找一个主云”变成“多云、多芯片、多层次组合”。AWS 是主训练伙伴，Google 提供 TPU，Microsoft / NVIDIA 提供高端 GPU，CoreWeave 和 Akamai 提供补充云基础设施，SpaceX / xAI 提供巨量短期推理容量，Fluidstack 则对应更长期的定制数据中心。背后的直接原因，是 Claude 的需求增长太快。

Anthropic 官方在 2026 年 4 月说，年化收入已从 2025 年底约 90 亿美元提升到 300 亿美元+，企业客户和 consumer usage 都在快速增长；5 月 SpaceX 算力协议公布后，Anthropic 直接提高 Claude Code、Claude API、Claude Pro / Max 的使用限制。

时间	合作方	合同 / 规模	主要用途	重点
2025-10-23	**Google Cloud**	计划使用最高 **100 万颗 Google TPU**，价值数百亿美元，2026 年带来 **超过 1GW** 算力	训练 + 推理	这是 Anthropic 多芯片路线的重要一步，用 TPU 降低对 NVIDIA 的单一依赖。(Anthropic)
2025-10-29	**AWS / Project Rainier**	Rainier 上线，近 **50 万颗 Trainium2**；Anthropic 到 2025 年底预计使用 **100 万+ Trainium2**	训练 + 推理	AWS 成为 Anthropic 主训练伙伴，Trainium 被真正用于前沿 Claude。(Amazon News)
2025-11-12	**Fluidstack**	Anthropic 宣布 **500 亿美元**美国 AI 基建投资，在 Texas、New York 建数据中心，2026 年陆续上线	自有 / 定制化数据中心	这是 Anthropic 从租云走向更深度定制基础设施。(Anthropic)
2025-11-18	**Microsoft + NVIDIA**	Anthropic 承诺购买 **300 亿美元 Azure 算力**，并额外签约最高 **1GW** 算力；NVIDIA / Microsoft 分别拟投资最高 100 亿 / 50 亿美元	NVIDIA GPU 训练 + 推理	Claude 接入 Azure / Foundry，同时引入 Grace Blackwell、Vera Rubin 等 NVIDIA 架构。(Anthropic)
2026-04-06	**Google + Broadcom**	新签 **5GW** 下一代 TPU 算力，2027 年开始上线	下一代 Claude 训练 + 服务	这进一步放大 Google TPU 线，官方称是为了服务全球客户需求和前沿模型。(Anthropic)
2026-04-10	**CoreWeave**	多年期算力协议，金额未披露，算力从 2026 年晚些时候上线	Claude 生产部署	CoreWeave 加入 Anthropic 基础设施伙伴池，说明 Anthropic 连 neocloud 也开始补。(CoreWeave)
2026-04-20	**Amazon / AWS**	新协议锁定最高 **5GW** 算力；Anthropic 未来 10 年向 AWS 技术承诺支出 **1000 亿美元+**，2026 年底前接近 **1GW** Trainium2 / Trainium3 上线	训练 + 推理 + Bedrock 分发	这是最核心的大单之一，AWS 继续是 Anthropic 主云和训练伙伴。(Anthropic)
2026-05-06	**SpaceX / xAI**	使用 **Colossus 1 全部算力**，新增 **300MW+、22 万+ NVIDIA GPU**；后续披露称 Anthropic 支付约 **12.5 亿美元/月**至 2029 年 5 月	主要是推理容量	这笔交易直接提高 Claude Code、Claude Pro / Max 和 Opus API 限额，核心是解决短期推理瓶颈。(Anthropic)
2026-05-07 / 05-08	**Akamai**	Akamai 披露一家美国 frontier model provider 承诺 **18 亿美元 / 7 年**云基础设施服务；Reuters / Bloomberg 称客户为 Anthropic	分布式云 / 推理	这说明 Anthropic 不只找 hyperscaler，也开始签 CDN / edge cloud 背景的基础设施商。(akamai.com)

SpaceX 招股书能看到更多的 AI Infra 细节。

维度	披露口径	含义
AI segment 收入，2025	**32 亿美元**	体量仍小，约为 SpaceX 2025 总收入 186.7 亿美元的 17%
AI segment 经营亏损，2025	**63.5 亿美元左右**	亏损约为收入的 **2 倍**，说明模型和算力仍在重投入期
AI segment capex，2025	**127.3 亿美元**	约占三大业务 capex 合计的 **61%**；远高于收入规模
AI segment 收入，2026Q1	**8.18 亿美元**	年化约 **32.7 亿美元**，环比/同比增长没有覆盖成本压力
AI segment 经营亏损，2026Q1	**24.7 亿—25 亿美元**	单季度亏损约为收入的 **3 倍**
AI segment capex，2026Q1	**77.23 亿美元**	单季度 capex / 收入约 **9.4 倍**，年化 capex run-rate 超 **300 亿美元**
对比：Space segment capex，2026Q1	**10.5 亿美元**	AI capex 明显超过传统航天投入
对比：Connectivity/Starlink capex，2026Q1	**13.3 亿美元**	AI 已成为最大资本开支方向
Anthropic 合同金额	**12.5 亿美元/月，年化 150 亿美元**	单月金额已超过 AI segment 2026Q1 收入；年化约为 2025 AI 收入的 **4.7 倍**
合同期限	至 **2029 年 5 月**	若持续执行，总额约 **400 亿+美元**
终止条款	双方可提前 **90 天通知终止**	收入弹性大，但确定性仍需要打折
一句话判断	xAI 亏损与 capex 很重，Anthropic 订单让 Colossus 资产具备外部定价	**从“模型竞争”切到“算力出租”后，xAI 的商业故事明显更顺**

AI segment 2025 年收入约 32 亿美元，但该口径合并了 X 平台、Grok/xAI 与 AI compute。

项目	2023	2024	2025	1Q25	1Q26
AI segment revenue	**2,961**	**2,620**	**3,201**	**727**	**818**
YoY	—	**-11.5%**	**+22.2%**	—	**+12.5%**
Advertising	2,323	1,728	1,844	443	343
YoY	—	-25.6%	+6.7%	—	-22.6%
占比	78.5%	66.0%	57.6%	60.9%	41.9%
AI Solutions & Infrastructure	638	892	1,357	284	475
YoY	—	+39.8%	+52.1%	—	+67.3%
占比	21.5%	34.0%	42.4%	39.1%	58.1%

据 Epoch AI 数据，顶尖 AI 实验室目前还没吃掉全球大部分 AI 算力，但 OpenAI、Anthropic 可能在未来几年快速提高占比。之后，继续靠算力扩张推动 AI 进步，就需要整个经济层面的算力建设继续加速。

OpenAI 在 2023 年掀起了人工智能算力扩展的浪潮。但如今它大约使用了全球算力的 10%，而各大顶尖实验室加起来可能还不到一半。根据关于数据中心电力容量和计算支出的披露，OpenAI、Anthropic 和 xAI 使用的计算量可能不到全球总量的 30%（OpenAI + Anthropic + xAI 合计不到 400 万 H100e，约占全球 20%–30%；若算上云厂商为它们模型提供的推理算力，可能再加约 5%。）。Google 和 Meta 是巨大的超大规模云服务商，但它们的大部分计算资源用于云服务和推荐系统，而不是用于其前沿研究实验室。

公司对比：

OpenAI：2025 年底数据中心容量 1.9GW，约等于 170 万 H100e；2023 年约 0.2GW，2024 年约 0.6GW，基本每年约 3 倍扩张。

Anthropic：2025 年底可能已有 100 万+ H100e，低于 OpenAI，但追赶很快。

xAI：2025 年底约 60 万－70 万 H100e；Colossus 1/2 合计约 55 万 H100e。

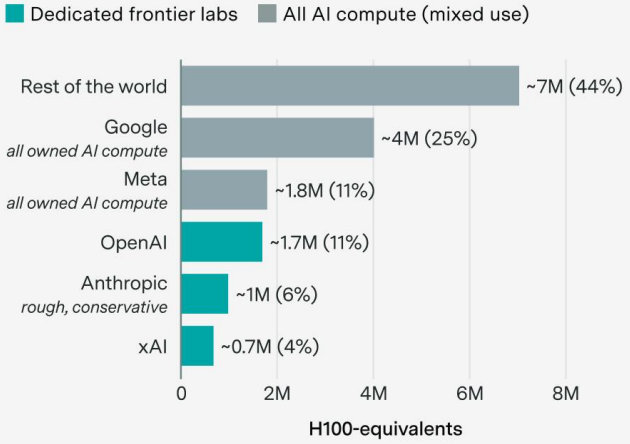
Google：公司总算力可能约占全球 1/4，但大量用于 Google Cloud 和内部非前沿 AI，不一定都给 DeepMind。

Meta：公司约占全球 10% AI 算力，但很多用于推荐系统、广告和内容分发，MSL 实际可用算力可能低于 OpenAI。

若 Anthropic + OpenAI 当前占 20%，且每年算力增长 4 倍、全球仅增长 3 倍，则它们约 2.5 年后份额翻倍，5 年内可能吃掉约 80% 全球 AI 算力。现在的“算力瓶颈”不是因为前沿实验室已经占满全球算力，而是它们正在快速抢占新增算力。

Much of the world's AI compute isn't used by the top frontier labs

Total frontier AI compute, including at Google and Meta, is likely under 50% of world total as of the end of 2025.



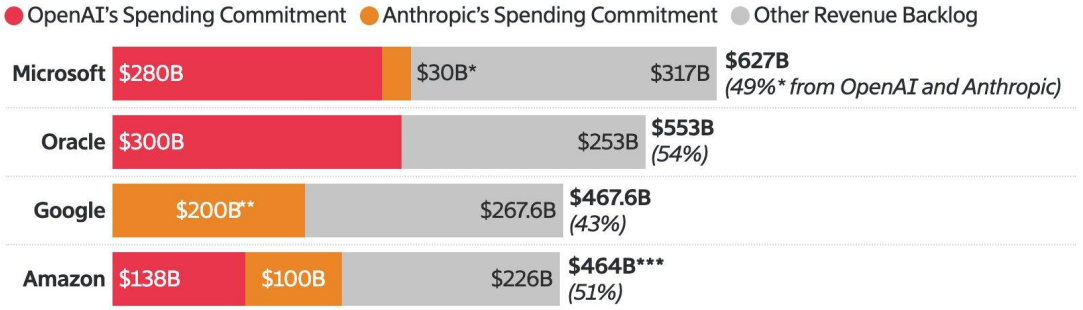
Estimates of compute owned or rented by entity as of end-2025, from chipmaker/lab disclosures. Anthropic and Google compute partially overlap. Google and Meta-owned compute is split between frontier AI and other uses (e.g., recommender systems, external cloud). Google DeepMind likely <50% of Google total.

EPOCH AI | CC-BY

epoch.ai

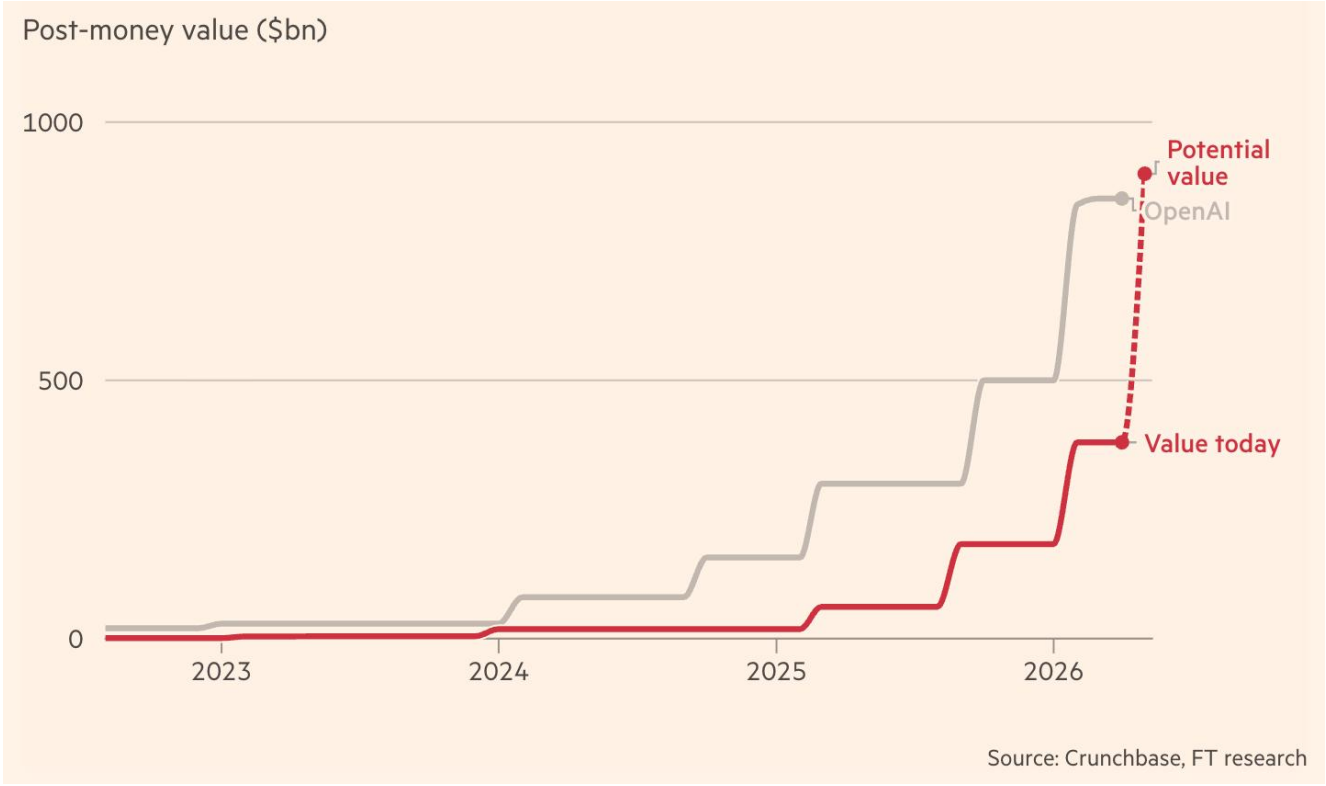
AI Whales

OpenAI and Anthropic's spending commitments to the four biggest U.S. cloud providers, as a percentage of those providers' revenue backlog.



* At least. ** Approximately. *** Including Amazon's April Anthropic agreement. Other figures as of March 31. • Source: Company filings, The Information reporting

OpenAI & Anthropic：上市节奏明显加快



2026 年 5 月以来，OpenAI 和 Anthropic 的上市预期同步升温。OpenAI 预计最快 9 月上市，最新私募估值约 **8520 亿美元**，IPO 估值讨论最高 **1 万亿美元**；Anthropic 则可能最快 10 月上市，最新融资洽谈估值已达 **9000 亿美元+**。这意味着海外前沿模型公司正式进入 public market 定价阶段，市场关注点也从模型能力扩展到 **ARR、经营利润、算力长协、RPO / backlog 和长期 capex 能见度**。

维度	OpenAI	Anthropic
上市节奏	更快，最快 **2026 年 9 月**	稍后，最快 **2026 年 10 月**
隐含估值	私募约 **8520 亿美元**，IPO 最高讨论 **1 万亿美元**	最新融资洽谈约 **9000 亿美元+**
募资诉求	支持超大规模算力、数据中心和模型训练	支持 Claude 企业需求和推理容量扩张
财务叙事	收入来源更宽，但 capex 压力更大	收入增速更陡，Q2 可能率先盈利
投资含义	通用 AI 平台上市	企业 Agent 平台上市

公司	最新上市时间预期	最新隐含估值 / 市值口径	最新融资情况	关键变化
OpenAI	**可能未来几周秘密提交 IPO 文件，最早 2026 年 9 月上市**	最近私募估值约 **8520 亿美元**；IPO 估值讨论最高可到 **1 万亿美元**	此前已完成约 **1220 亿美元**融资承诺；IPO 初步讨论募资至少 **600 亿美元**	IPO 节奏明显提前，核心原因是继续融资支持芯片、数据中心、人才与模型训练；公司正在与 Goldman Sachs、Morgan Stanley 准备招股书草案。(Reuters)
Anthropic	**最新市场预期：最快 2026 年 10 月上市**；此前已聘请 Wilson Sonsini 准备 IPO	最新融资洽谈估值约 **9000 亿美元+**；若按 IPO 报道口径，潜在募资规模也可能达 **600 亿美元级别**	据 Bloomberg / FT 等报道，Anthropic 正洽谈新一轮至少 **300 亿美元**融资，估值约 **9000 亿美元**；可能成为 IPO 前最后一轮大额私募	Anthropic 估值快速逼近 OpenAI，原因是 Claude Code、企业 Agent、金融工作流收入爆发，且 2026Q2 预计首次实现经营利润。(金融时报)

口径	OpenAI	Anthropic	备注
公司正式融资估值 / Primary round	**8520 亿美元。**OpenAI 最近一轮融资后估值约 8520 亿美元。(Reuters)	**3800 亿美元已完成；9000 亿美元+ 正在推进。**Anthropic 官方已完成 Series G，投后估值 3800 亿美元；FT / Reuters 口径显示，其正推进至少 300 亿美元新融资，估值约 9000 亿美元。(Anthropic)	最接近公司认可价格。OpenAI 的 8520 亿、Anthropic 的 3800 亿是已完成融资；Anthropic 9000 亿属于接近落地但仍按媒体报道处理。
IPO 目标估值 / IPO expectation	**1 万亿美元+。**Reuters / FT 均称 OpenAI 正准备 IPO，目标估值可超过 1 万亿美元，最快 2026 年 9 月上市。(Reuters)	**9000 亿–1 万亿美元+。**Anthropic 已聘请 Wilson Sonsini 准备 IPO，最新融资估值已逼近 9000 亿，二级市场预期进一步推向 1 万亿美元以上。(金融时报)	这是投行、媒体和市场基于 IPO 预期做的估值框架，重要但仍然会随市场窗口和增长数据波动。
传统二级市场 / Forge、Hiive 等	**约 8500 亿–8800 亿美元。** Forge / 二级市场报道显示 OpenAI 大致围绕 8520 亿融资价附近交易，溢价不算大。(forgeglobal.com)	**约 8500 亿–1 万亿美元+。**多篇报道提到 Anthropic 在 Forge、Hiive 等二级市场快速上行，部分平台已接近或突破 1 万亿美元 implied valuation。(雅虎财经)	这是老股或私募股权份额交易反推出来的价格。比正式融资更市场化，但流动性薄、买卖价差大。
链上 tokenized pre-IPO / Jupiter、PreStocks	**约 1 万亿美元。**社媒和链上数据口径称，OpenAI 的 tokenized pre-IPO implied valuation 已触及 1 万亿美元；PreStocks 也提供 OpenAI 的链上 pre-IPO 敞口。(Binance)	**约 1.2 万亿–1.4 万亿美元。**Kobeissi / crypto.news 等口径称，Anthropic 链上 pre-IPO implied valuation 已升至约 1.2 万亿美元，并一度超过 OpenAI。(X (formerly Twitter))	这是最激进口径。交易的是 tokenized / SPV 经济敞口，不是公司正式股票；更像“上市前影子市场价格”。
公司对 SPV / tokenized 股权态度	OpenAI 官方明确称，OpenAI 股权不能在未经书面同意下直接或间接转让；未经授权的 SPV、tokenized interests、forward contracts 等可能无效。(OpenAI)	Anthropic 更强硬，称未经董事会批准的股份或权益转让无效，不会在账簿上承认；还点名部分平台。(华尔街日报)	所以链上 / SPV 价格只能当作市场情绪和潜在 IPO 预期，不能等同于公司认可估值。

OpenAI 的上市节奏曾经有所动摇，但是当前较为确定。OpenAI 的 IPO 逻辑是**“用上市解决长期算力融资问题”**。公司已经有 ChatGPT、Codex、企业销售、广告测试、图像生成等收入来源，但未来训练和推理 capex 极重，所以需要 public market 承接更大的融资规模。微软分成封顶也很关键，相当于 IPO 前把未来利润表的不确定性压低，改善投资人看到的长期盈利弹性。

时间	事件
2026 年 5 月 12 日	The Information 报道，OpenAI 与微软重谈协议，将对微软的收入分成总额封顶在 **380 亿美元**；Reuters 称这有助于 OpenAI 在 IPO 前改善长期财务故事。(Reuters)
2026 年 5 月 20 日	WSJ / Reuters 报道，OpenAI 正准备未来几周秘密提交美国 IPO 文件，目标最早 **2026 年 9 月上市**。(Reuters)
最新估值口径	OpenAI 最近私募估值约 **8520 亿美元**；IPO 讨论估值最高 **1 万亿美元**，募资规模至少 **600 亿美元**。(Reuters)

时间	事件	重点
2025-11-05	OpenAI CFO Sarah Friar 在 WSJ Tech Live 上说，IPO 暂时“不在计划中”，公司还在适应当前规模，不想被 IPO 牵着走。	这是 CFO 公开表达的保守口径。(Reuters)
2026-04-05	The Information 报道，CEO Sam Altman 和 CFO Sarah Friar 在 IPO 时间上出现分歧：Altman 希望最快 2026Q4 上市，Friar 认为公司 2026 年可能还没准备好。	分歧点主要是组织/流程准备、巨额算力承诺、收入增长是否足以支撑支出。(The Information)
2026-04-28	WSJ 报道，OpenAI 未达到部分内部收入和用户目标；CFO Sarah Friar 担心，如果收入增长不足，未来算力开支会带来压力。	这强化了 CFO 对“今年上市是否合适”的担忧。(华尔街日报)
2026-05-12	The Information / Reuters 报道，OpenAI 与微软重谈协议，把收入分成总额封顶在 **380 亿美元（The Information 口径里，原本到 2030 年大约会付 1350 亿美元，节省约 970 亿美元）**。	有助于改善 IPO 前的长期利润表和投资人预期。(Reuters)
2026-05-18	OpenAI 赢下 Musk 相关诉讼，Reuters 称这移除了 IPO 的一项重要障碍。	法律结构风险下降，IPO 节奏有条件提速。(Reuters)
2026-05-20	WSJ / Reuters 报道，OpenAI 正准备未来几周秘密提交 IPO 文件，目标最早 **2026 年 9 月上市**。	最新节奏明显比 CFO 此前保守口径更激进。(Reuters)

Anthropic 的 IPO 逻辑比 OpenAI 更偏 **“企业 Agent 收入兑现”**。Claude Code、企业客户、金融/专业服务工作流推动收入快速放大，Q2 预计已经接近经营利润为正。它的估值从此前数千亿美元级别快速推到 **9000 亿美元**，说明市场已经把 Anthropic 当成下一代企业 AI 平台在定价。

时间	事件
2025 年 12 月	Reuters 援引 FT 称，Anthropic 已聘请 Wilson Sonsini 准备潜在 IPO，最早可能 2026 年上市，但当时尚未作出最终决定。(Reuters)
2026 年 3–4 月	Bloomberg 口径称，Anthropic 考虑最快 **2026 年 10 月 IPO**，并与 Goldman Sachs、JPMorgan、Morgan Stanley 等投行进行早期讨论。(TNW | The heart of tech)
2026 年 5 月	Bloomberg / FT 报道，Anthropic 正洽谈新一轮至少 **300 亿美元**融资，估值约 **9000 亿美元+**，可能成为 IPO 前最后一轮大额私募。(金融时报)
2026Q2 预期	FT / WSJ 报道，Anthropic Q2 收入预计约 **109 亿美元**，经营利润约 **5.59 亿美元**，可能首次实现季度经营利润。(金融时报)

All-in-One：从“聊天 / 代码 / API 分散”走向统一 AI 工作台

OpenAI 更像要做“一个超级 AI 工作台”；Anthropic 更像要做“企业 AI 工作流平台”。 OpenAI 的 ChatGPT、Codex、API 有更强产品合并趋势；Anthropic 的 Claude Code 目前更像独立开发者执行器，通过 MCP、Skills、Connectors 和 Claude 主入口形成生态协同。

维度	OpenAI	Anthropic
收敛方式	**产品壳统一**：ChatGPT、Codex、API 收进统一产品战略	**工作流统一**：Claude、Claude Code、MCP、Skills、Connectors 协同
核心入口	ChatGPT	Claude
核心执行器	Codex	Claude Code
是否有明确合并信号	有。ChatGPT 和 Codex 被报道正在统一为一个核心产品体验。(WIRED)	暂无明确“Claude Code 并入 Claude 主 App”的信号
当前产品状态	Codex 已进入 ChatGPT mobile app，可远程管理 coding 任务。(Reuters)	Claude Code 继续强化 terminal / IDE / desktop / browser 开发者工作台
战略方向	一个 AI 工作台：聊天、代码、浏览器、文件、API、长期任务一体化	一个企业 AI 工作流层：Claude 入口 + Code 执行 + MCP 连接系统
投资含义	ChatGPT 流量可持续导入 Codex、API、企业产品，形成超级入口	Claude Code 和 MCP 可能成为企业 Agent / 开发者工具的事实标准接口

1、OpenAI：ChatGPT + Codex + API 强 All-in-One

OpenAI 的 All-in-One 信号最强。未来形态大概率是：**ChatGPT 负责对话、需求澄清、文件和通用入口；Codex 负责 repo、终端、代码修改和长期执行；API 负责开发者和企业集成。** 现在上下文还没完全打通，比如聊天记录不能直接迁入 Codex 历史列表，所以实用做法仍然是把关键结论沉淀成 **AGENTS.md / README-dev.md / 项目说明文件**，让 Codex 读取。

时间	事件	含义
2026-03	OpenAI 被报道在规划桌面 **superapp**，目标是把 ChatGPT、Codex、浏览器等能力整合到单一桌面体验。	OpenAI 想把 ChatGPT 从聊天入口升级为统一 AI 工作台。
2026-05-14	OpenAI 将 **Codex 接入 ChatGPT mobile app**，用户可在手机端查看、批准、调整 Codex 任务。(Reuters)	当前更像 **Codex 被带进 ChatGPT 主入口**。
2026-05-15	OpenAI 重组产品线，由 Greg Brockman 接管产品战略，核心方向是统一 ChatGPT 和 Codex 为一个核心产品体验。(WIRED)	组织架构开始服务“一个统一产品体验”。
2026-05	市场转述称 OpenAI 将 **ChatGPT、Codex、API** 合并到统一产品团队。(彭博资讯)	不是单纯把 Codex 做强，而是把聊天、代码、API 都放进同一产品战略。
2026-05	Tibo 的表述是：把 ChatGPT 整合进 Codex，再把 Codex 带到 ChatGPT。	更准确理解是 **ChatGPT 和 Codex 双向融合**，最终收敛成一个项目空间。

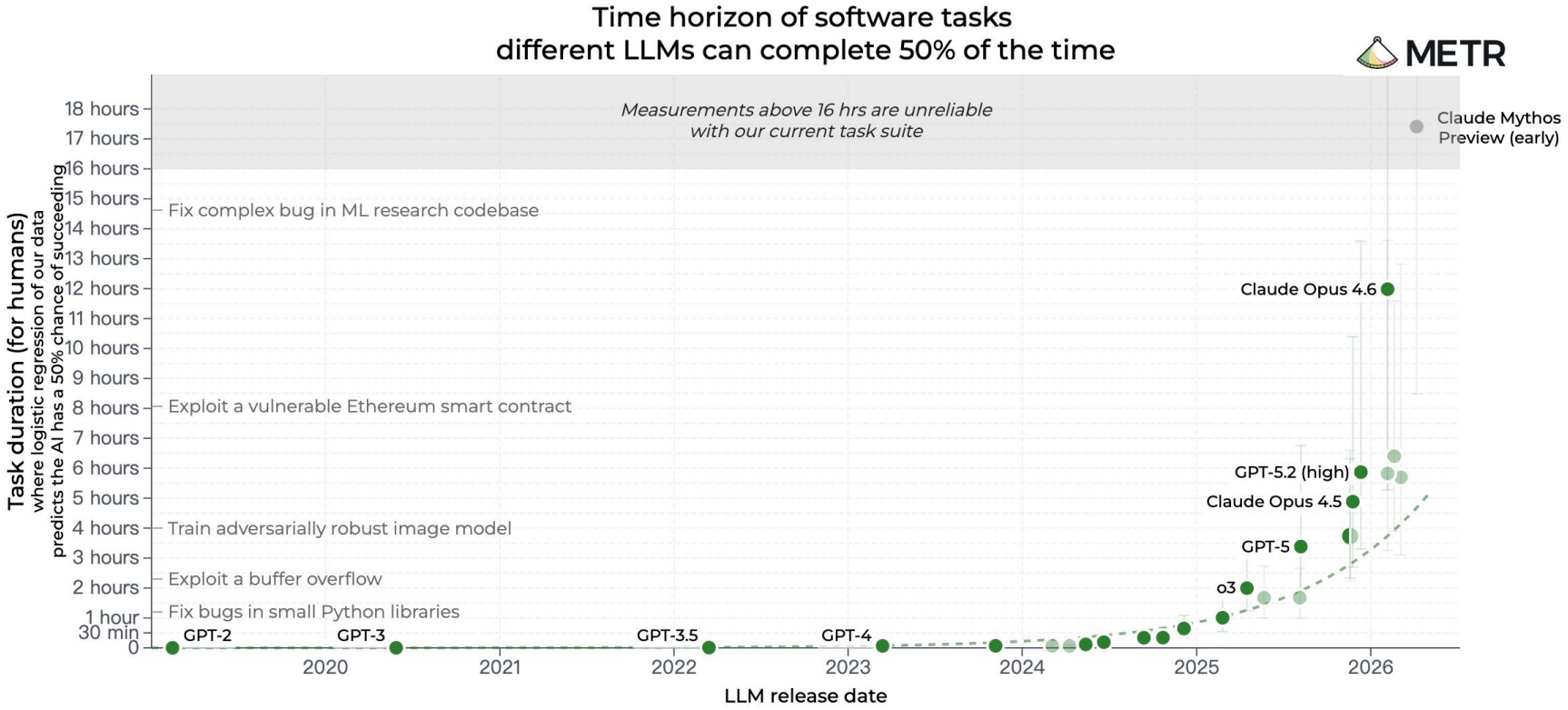
2、Anthropic：Claude + Claude Code + MCP，更偏工作流层统一

Anthropic 也在 All-in-One，但路线不同。它没有明显释放“Claude Code 会被统一进 Claude 主 App”的信号；它更像在构建 **企业工作流平台**：Claude 是自然语言入口，Claude Code 是开发执行器，MCP 是连接协议，Skills 是可复用工作流，Connectors 把企业系统和数据接进来。也就是说，Anthropic 的统一发生在 **协议、数据、工具、流程层**，产品壳暂时仍保持分工。

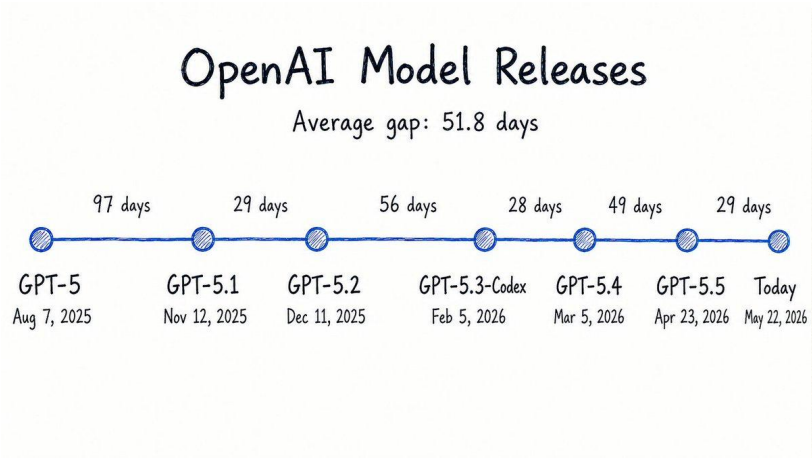
时间	事件	含义
2025-09	Anthropic 升级 Claude Code，强调其在 terminal 和 IDE 中处理更长、更复杂的开发任务。(Anthropic)	Claude Code 的定位仍是开发者执行器。
2026-04	Claude Design 发布，可做原型、wireframe、pitch deck，并把结果交给 Claude Code。	Claude 生态从聊天 / 代码扩到设计和产品表达。
2026-05	Anthropic 发布金融 Agent，结合 connectors、MCP apps、Microsoft 365 integrations、Claude Code plugins 等。	Claude 的收敛方式是把企业数据、Office、行业流程接进来。
2026-05	Claude Code desktop docs 显示，它可并排处理 chat、diff、preview、terminal、file editor，并连接 GitHub、Slack、Linear 等外部工具。(Claude Code)	Claude Code 自己越来越像一个开发者工作台。
当前判断	暂未看到 Anthropic 明确说要把 **Claude Code 合并进 Claude 主 App**。	Anthropic 更像 **Claude 主入口 + Claude Code 独立执行器 + MCP/Skills/Connectors 连接外部系统**。

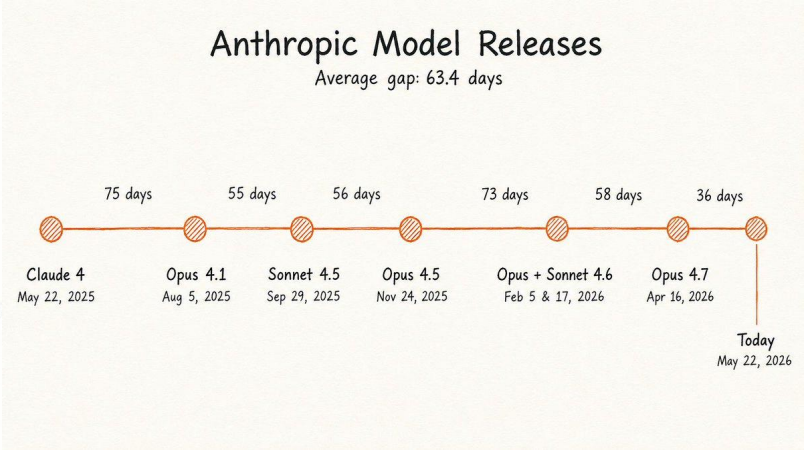
模型更新节奏

长程任务是最值得关注的指标。METR 在 2026 年 3 月的有限时间窗口内对 Claude Mythos 预览版的早期版本进行了风险评估。我们估计在我们的任务套件上，其 50% 时间视界至少为 16 小时（95% 置信区间为 8.5 小时到 55 小时），处于在不增加新任务情况下我们能测量的上限。



OpenAI 两次发布模型之间的平均天数是 52 天，如果排除第一次较长的一次，就是 40 天。所以至少等待几周是合理的。如果是新的预训练版本并且他们需要更多时间来调整，可能会更久些。Anthropic 的平均为 63 天，所以可能还需要大约另外 3 周左右。





按 OpenAI / Anthropic 那种发布周期思路套到国内，MiniMax 更像“模型主线可能快要更新”，智谱更像“GLM-5.1 这一代还在通过 ZCube 和 TileRT 继续榨系统红利”。MiniMax 从 M2 之后的主线模型平均间隔约 47 天，现在距离 M2.7 已经约 67 天，新一代模型已经进入发布窗口；智谱主线旗舰平均间隔约 60 天，GLM-5.1 从 3 月 27 日 Coding Plan 开放算起已经约 58 天，但 5 月 20 日 ZCube、5 月 22 日 GLM-5.1-highspeed 可以视为同一轮 GLM-5.1 的系统级增强。

公司	最近一次主线模型	最近一次系统/产品增强	历史节奏	当前判断
MiniMax	**M2.7，2026-03-18**	M2.7-highspeed / Agent Harness	M2 后平均 **约 47 天**	距 M2.7 已约 **67 天**，新一代模型窗口已经打开
智谱	**GLM-5.1，2026-03-27 / 04-07**	**ZCube：2026-05-20；GLM-5.1-highspeed：2026-05-22**	旗舰平均 **约 60 天**	模型窗口接近，但 5 月的重点是 GLM-5.1 系统工程增强

MiniMax：M2 之后节奏明显加快，当前已超过历史平均间隔

MiniMax 的节奏有一个明显分水岭：**M2 之前迭代慢，M2 之后进入高频更新期**。从 2025 年 10 月 M2 到 2026 年 3 月 M2.7，MiniMax 在不到 5 个月里连续推出 M2、M2.1、M2.5、M2.7，主线模型平均约 47 天一更。现在距离 M2.7 已经约 67 天，已经超过 M2 之后的平均节奏，因此从历史发布周期看，**MiniMax 新一代 M2.x 或 M3 确实已经进入合理发布窗口**。但 M2.7 已经加入“自我进化 / Agent Harness”叙事，下一次如果是更大版本，训练和产品整合周期拉长也合理。

补充：MiniMax 在主线模型之外，2025–2026 年还密集发布了 Speech、Music、Hailuo 视频等模型，例如 Hailuo-02、Speech-2.5、Hailuo-2.3、Speech-2.6、Music-2.5、Speech-2.8、Music-2.6 等。(MiniMax API Docs) 这些说明 MiniMax 是多模态高频迭代，但用于预测 **M2/M3 主线模型** 时。

时间	发布内容	类型	距上次主线模型	备注
2025-01-15	**MiniMax-Text-01 / MiniMax-VL-01**	文本 / 视觉语言	—	API 文档列为下一代文本模型和视觉语言模型发布。(MiniMax API Docs)
2025-10-27	**MiniMax-M2**	Agentic coding / Agent 模型	**285 天**	官方定位为 “Efficient Model for the Agentic Era”。(MiniMax API Docs)
2025-12-22	**MiniMax-M2.1**	Coding / 重构	**56 天**	官方描述为 Polyglot Programming Mastery、Precision Code Refactoring。(MiniMax API Docs)
2026-02-12	**MiniMax-M2.5 / M2.5-highspeed**	Coding / 工具调用 / Office / Agent	**52 天**	官方称 M2.5 在编程、工具调用、搜索、办公生产力等场景达到或刷新 SOTA；官方文章也提到从 M2、M2.1 到 M2.5 只用了约 3 个半月。(MiniMax)
2026-03-18	**MiniMax-M2.7 / M2.7-highspeed**	自我进化 / Agent Harness	**34 天**	官方称 M2.7 是首个深度参与自身演进的模型，具备 Agent Teams、complex Skills、dynamic tool search 等能力。(MiniMax)
2026-05-24	距 M2.7 发布至今	—	**67 天**	已经高于 M2 之后 **47 天左右** 的平均节奏。

MiniMax 节奏判断

口径	间隔
全样本平均，Text-01 → M2 → M2.1 → M2.5 → M2.7	**约 107 天**
剔除 Text-01 → M2 这次超长间隔，只看 M2 之后	**约 47 天**
M2 之后中位数	**约 52 天**
当前距离 M2.7	**约 67 天**

智谱：旗舰模型约两个月一更，5 月重点从模型切到 ZCube / TileRT 系统工程

智谱的节奏更像 **“旗舰模型两个月一更，中间穿插 Turbo、V、Flash、系统工程版本”**。从 GLM-4.5 到 GLM-5.1，主线旗舰平均间隔约 60 天；GLM-5.1 从 3 月 27 日 Coding Plan 开放算起已经约 58 天，理论上也接近下一次主线版本窗口。但智谱 5 月没有直接发 GLM-5.x，而是连续推出 **ZCube 组网架构** 和 **GLM-5.1-highspeed / TileRT 高速版**，这说明智谱这一轮的重点已经从“单纯堆模型版本”转向 **推理系统工程**：ZCube 解决生产集群网络瓶颈，TileRT 解决旗舰模型低延迟输出问题。这个节奏很有研究价值，因为它说明国内大模型厂商的竞争点正在从 benchmark 延伸到 **吞吐、TTFT、网络成本、Coding Plan 使用体验**。

时间	发布内容	类型	距上次主线旗舰	备注
2025-07-28	**GLM-4.5 Series**	原生 Agentic LLM	—	官方称 GLM-4.5 是最新 native agentic LLM，强化推理、coding 和 agentic 能力。(Z.AI)
2025-09-30	**GLM-4.6**	旗舰 coding 模型	**64 天**	官方称 GLM-4.6 是旗舰 coding model，context 扩至 200K。(Z.AI)
2025-12-22	**GLM-4.7**	Coding / Reasoning / Agentic	**83 天**	官方称 GLM-4.7 强化 coding、reasoning、agentic capabilities。(Z.AI)
2026-02-12	**GLM-5**	Complex system engineering / long-range Agent	**52 天**	官方称 GLM-5 从 coding 转向 engineering，面向复杂系统工程和长程 Agent 任务。(Z.AI)
2026-03-27 / 2026-04-07	**GLM-5.1**	长程任务 / Agentic Engineering	**43 天 / 54 天**	3 月 27 日媒体称 GLM-5.1 已面向 Coding Plan 用户开放；官方英文 release notes 列为 4 月 7 日。GLM-5.1 可在单次任务中独立工作最长 8 小时。(东方财富网)
2026-05-20	**ZCube 组网架构落地**	推理网络 / 组网	距 GLM-5.1 约 44–54 天	智谱、驭驰网络与清华在 GLM-5.1 coding 生产环境落地 ZCube；交换机与光模块成本下降 33%，GPU 平均推理吞吐提升 15%，TTFT P99 降低 40.6%。(智谱 AI)
2026-05-22	**GLM-5.1-highspeed / TileRT**	推理引擎 / 高速 API	距 ZCube 2 天	GLM-5.1-highspeed **【还未披露价格】** 输出速度达 400 tokens/s，由智谱 GLM 团队与 TileRT 团队联合打造，适用于 AI 编程、实时交互、商业决策、实时语音等低延迟场景。(搜狐网)

智谱节奏判断

口径	间隔
旗舰主线：GLM-4.5 → 4.6 → 4.7 → 5 → 5.1	**平均约 60 天**
旗舰主线中位数	**约 58 天**
从 GLM-5.1 Coding Plan 开放算起至今	**约 58 天**
从官方 release notes 4 月 7 日算起至今	**约 47 天**
若把 ZCube / highspeed 视为 GLM-5.1 系统级版本 最新大事件刚发生在	**5 月 20–22 日**

智谱&MiniMax 入通在即

恒生指数公司 5 月 22 日宣布，截至 2026 年 3 月 31 日之恒生指数系列季度检讨结果，所有变动将于 2026 年 6 月 5 日（星期五）收市后实施并于 2026 年 6 月 8 日（星期一）起生效。

·6 月模型发布：GLM 5.5 / MiniMax M3 / MiniMax Hailuo 3；

·南向潜在纳入：智谱预计 6 月首周，MiniMax 预计 8 月首周。

3. 恒生科技指数

恒生科技指数成份股将有以下的变动，成份股数目维持 30 只。

加入：

代号	公司
100	MiniMax Group Inc. - W
2513	北京智谱华章科技股份有限公司

剔除：

代号	公司
268	金蝶国际软件集团有限公司
3888	金山软件有限公司

DeepSeek V4-Pro 永久降价，会破坏 token 涨价逻辑吗？

DeepSeek V4-Pro 的降价节奏超出市场预期。V4 发布初期，DeepSeek 曾因高端算力容量受限而将 Pro 定价显著高于 Flash，并表示价格有望在华为 Ascend 950 超节点下半年规模出货后下降；但 4 月 26 日先将全系列 cache hit 价格降至发布价 1/10，5 月 23 日又直接把 V4-Pro 的 75% 折扣永久化。这个变化意味着，DeepSeek 正在把 V4-Pro 从高端稀缺模型，推向可大规模调用的旗舰 API，也进一步强化国产模型价格战和国产算力栈成熟的叙事。

时间	事件	含义
2026-04-24	**DeepSeek V4 发布**，V4-Pro / V4-Flash 同步上线。	Pro 定位旗舰，Flash 定位低成本高吞吐。
2026-04-26	全系列 API **cache hit 输入价格降至首发价 1/10**；V4-Pro 同时开启 **75% 折扣 / 2.5 折**。	第一轮降价，核心利好长上下文、Agent、coding。(21 经济网)
最初口径	Reuters 报道称，V4-Pro 75% 折扣最初到 **5 月 5 日**。	当时市场理解为短期促销。(Reuters)
随后调整	V4-Pro 75% 折扣延长至 **2026 年 5 月 31 日 15:59 UTC**。	从几天促销延长为整月促销。(DeepSeek API Docs)
2026-05-22 / 05-23	DeepSeek 宣布：5 月 31 日促销结束后，V4-Pro API **正式调整为原价 1/4**。	临时促销变成永久重定价。(IT 之家)
永久价	每百万 tokens：**缓存命中输入 0.025 元、缓存未命中输入 3 元、输出 6 元**。	旗舰 Pro 模型进入极低价区间。(花城)

API 访问模型名	输入（缓存命中）	输入（缓存未命中）	输出	上下文长度
deepseek-v4-pro	1 元	12 元	24 元	1M
deepseek-v4-flash	0.2 元	1 元	2 元	

*受限于高端算力，目前 Pro 的服务吞吐十分有限，预计下半年昇腾 950 超节点批量上市后，Pro 的价格会大幅下调。

我们认为，当前节点 DeepSeek 的降价，不会影响 token 的涨价逻辑。

涨价有三种形态

从 2025H2 开始，token 价格确实进入了一个“涨价周期”，但它不是全行业同向涨价，而是前沿模型开始把 **能力、速度、长上下文、Agent、Coding、区域合规、容量确定性** 全部拆成收费层。OpenAI 是最清晰的代际涨价，Claude 是速度和有效成本涨价，智谱是国内最明显的连续涨价，阿里则是高端 SKU 上移。DeepSeek V4-Pro 永久降价是这条主线里的反向变量，说明国内仍然存在价格战，但高端 Agent 能力已经开始具备提价权。

涨价形态	代表厂商	逻辑
代际涨价	OpenAI、智谱	新一代旗舰能力更强，直接提高 input / output 单价。
速度溢价	Claude Fast Mode、OpenAI Priority、MiniMax Highspeed	同一能力下，快响应 / 产能确定性额外收费。
高端 SKU 分层	阿里 Qwen、智谱、OpenAI Pro	Deep Research、Pro、长上下文、Agent、Coding、视觉 GUI 等能力单独定高价。

1、OpenAI：GPT-5.2 开始进入连续涨价周期

GPT-5.2 是涨价起点，GPT-5.3 暂停，GPT-5.4 再涨，GPT-5.5 大幅上台阶。 它的定价逻辑已经从“更强模型更便宜”转向“更强模型、更长上下文、更快响应、更高可靠性，都要单独付费”。

时间	模型 / SKU	价格，美元 / 1M tokens	变化
2025-11	**GPT-5.1**	Input **$1.25**；Output **$10**	GPT-5.1 仍是 GPT-5 时代的基准价格。(OpenAI 开发者)
2025-12	**GPT-5.2**	Input **$1.75**；Output **$14**	相比 GPT-5.1，输入 / 输出均涨 **40%**；OpenAI 当时也明确说 GPT-5.2 API per-token 价格高于 GPT-5.1，但 token efficiency 更高。(OpenAI 开发者)
2026-02	**GPT-5.3 Chat**	Input **$1.75**；Output **$14**	和 GPT-5.2 同价，没有继续涨。(OpenAI 开发者)
2026-03	**GPT-5.4**	Short context：Input **$2.5**；Output **$15**；Long context：Input **$5**；Output **$22.5**	相比 GPT-5.2 / 5.3，短上下文输入涨 **43%**，输出涨 **7%**；长上下文另收溢价。(OpenAI 开发者)
2026-04	**GPT-5.5**	Short context：Input **$5**；Output **$30**；Long context：Input **$10**；Output **$45**	相比 GPT-5.4，短上下文价格直接 **翻倍**；长上下文继续溢价。(OpenAI 开发者)
2026-04	**GPT-5.5 Pro**	Short context：Input **$30**；Output **$180**；Long context：Input **$60**；Output **$270**	Pro 版变成真正的高价旗舰。(OpenAI 开发者)
2026-04 以后	**Priority processing**	GPT-5.5 Priority：Input **$12.5**；Output **$75**	速度 / 产能确定性也开始单独收费。(OpenAI 开发者)

2、Claude：标价不一定代代涨，但 Fast Mode / tokenizer / 区域能力在抬高有效价格

Claude 的涨价逻辑比 OpenAI 隐蔽。**主线 Opus / Sonnet 标价没有一路涨，甚至 Opus 从 4.1 到 4.5 后标价下降；但 Fast Mode 直接 6 倍、Haiku 4.5 比 3.5 贵 25%、区域能力 1.1 倍、Opus 4.7 tokenizer 可能增加 token 数。** 所以 Claude 的核心不是"代代涨价"，而是把 **速度、区域、tokenizer、模型档位** 拆成不同收费层。

时间	模型 / SKU	价格，美元 / 1M tokens	变化
Claude Opus 4 / 4.1	Opus 老旗舰	Input **$15**；Output **$75**	老 Opus 定价很高。(Claude Platform)
Claude Opus 4.5 / 4.6 / 4.7	新 Opus 主线	Input **$5**；Output **$25**	标价比 Opus 4.1 下降，但 4.5 / 4.6 / 4.7 之间保持同价。(Claude Platform)
Claude Sonnet 4 / 4.5 / 4.6	Sonnet 主线	Input **$3**；Output **$15**	Sonnet 主线价格稳定。(Claude Platform)
Claude Haiku 3.5 → Haiku 4.5	轻量模型	Haiku 3.5：**$0.8 / $4**；Haiku 4.5：**$1 / $5**	Haiku 轻量模型涨 **25%**。(Claude Platform)
Claude Opus 4.6 / 4.7 Fast Mode	快速模式	Input **$30**；Output **$150**	Fast Mode 是标准 Opus 的 **6 倍价格**。(Claude Platform)
区域 / 数据驻留	US-only / regional	**1.1x** 价格乘数	企业合规和区域能力也开始收费。(Claude Platform)
Opus 4.7 tokenizer	有效成本	同样文本最多可能多出 **35% tokens**	标价不变，但单次请求的有效成本可能上升。(Claude Platform)

3、智谱：涨价最明显，GLM-4.7 → GLM-5 → GLM-5.1 连续上台阶

智谱是国内最典型的涨价样本。**GLM-4.7 到 GLM-5 是第一跳，GLM-5 到 GLM-5.1 是第二跳，GLM-5V-Turbo 和 Coding Plan 扣量倍率是第三层。** 这说明智谱已经不再单纯卷低价，而是在把 **长程 coding、agentic engineering、视觉 GUI Agent、Claude Code 兼容生态** 变成可提价资产。

时间	模型 / SKU	价格，美元 / 1M tokens	变化
GLM-4.5 / 4.6 / 4.7	主线模型	Input **$0.6**；Output **$2.2**	GLM-4.x 主线价格稳定。(Z.AI)
2026-02	**GLM-5**	Input **$1.0**；Output **$3.2**	相比 GLM-4.7，输入涨 **67%**，输出涨 **45%**；TrendForce 也报道 GLM-5 出现约 30%+ 涨价。(Z.AI)
2026-03	**GLM-5-Turbo**	Input **$1.2**；Output **$4.0**	比 GLM-5 继续涨，属于更强 / 更快 Agent 版本溢价。(Z.AI)
2026-04	**GLM-5.1**	Input **$1.4**；Output **$4.4**	相比 GLM-5，输入涨 **40%**，输出涨 **38%**；相比 GLM-4.7，输入涨 **133%**，输出涨 **100%**。(Z.AI)
2026-04	**GLM-5V-Turbo**	Input **$1.2**；Output **$4.0**	视觉 / GUI Agent 版本价格明显高于 GLM-4.6V 的 **$0.3 / $0.9**。(Z.AI)
Coding Plan	GLM-5.1 / GLM-5-Turbo	高峰期 **3x** 扣量，非高峰期 **2x** 扣量；6 月底前非高峰临时 1x	订阅制里也在通过扣量倍率做分层。(Z.AI)

4、阿里 Qwen：“新高端 SKU 上移”

阿里的情况要谨慎写。公开资料更容易坐实的是：**Qwen 从 2024 年极低价促销，转向 2026 年高端模型 / 长上下文 / Deep Research / Agent SKU 分层收费。** 但“同一模型直接提价”目前我没看到足够官方证据。更准确的表达是：**阿里不是简单把老模型提价，而是用 Qwen3.7-Max、Deep Research、长上下文分层和国际部署，重建高端价格带。**

时间	模型 / SKU	价格，美元 / 1M tokens	变化
2024-05	Qwen-Long 等	阿里当时大幅降价，Qwen-Long 从 **0.02 元 / 1K tokens** 降至 **0.0005 元 / 1K tokens**	2024 年主线仍是价格战。(Reuters)
2026-04	**Qwen3.5-397B-A17B Global**	Input **$0.172–0.43**；Output **$1.032–2.58**，按上下文分层	长上下文开始按区间分层收费。(AlibabaCloud)
2026-04	**Qwen3.5-397B-A17B International**	Input **$0.6**；Output **$3.6**	海外 / 国际部署价格高于 Global 口径。(AlibabaCloud)
2026-05	**Qwen3.7-Max**	Input **$2.5**；Output **$7.5**	新旗舰 Agent / Coding 模型价格明显上移；第三方模型卡显示其支持 1M context。(Puter Developer)
2026	**Qwen deep research**	Input **$7.742**；Output **$23.367**	Deep Research 这种高价值工具型 SKU 定价非常高。(AlibabaCloud)

5、MiniMax / Kimi：也有“高端模型更贵”，但仍然以性价比为主

MiniMax 和 Kimi 还没像 OpenAI / 智谱那样形成特别强的涨价曲线。MiniMax 的重点是 **M2.7 输入端和高速版上移**；Kimi 目前仍以“开源权重 + Agent 能力 + 相对低价”做竞争。

公司	事件	价格，美元 / 1M tokens	变化
MiniMax M2.5	2026-02 M2.5	Input **$0.15**；Output **$1.20**	低价 Agent 模型代表。(Verdent AI)
MiniMax M2.7	2026-03 M2.7	Input **$0.30**；Output **$1.20**；Cache **$0.06**	相比 M2.5，输入端涨一倍，输出保持不变；高性能 Agent 版本上移。(Thomas Wiegold — AI Agency Sydney)
MiniMax M2.7 Highspeed	高速版	Input **$0.60**；Output **$2.40**	高速版约为 M2.7 标准价 **2 倍**。(云价格查询)
Kimi K2.6	2026-04 K2.6	官方 / 第三方口径约 Input **$0.60**；Output **$2.50**	Kimi 仍偏性价比，重点是 Agent Swarm 和长程执行。(Medium)

大模型市场空间再思考

如果只按全球程序员测算，AI Coding 的市场空间大概是百亿美元级，重度 Agent 化后可以看到数百亿到千亿美元级；很难直接推到万亿美元。SemiAnalysis 的 15 万亿美元，是把 Claude Code 从“编程工具”外推到“信息工作自动化”的劳动池。

全球开发者数量口径差异很大。JetBrains 2025 口径是 2080 万专业开发者；Evans Data 早前预计 2024 年专业开发者约 2870 万；SlashData 2025 口径更宽，专业开发者 3650 万、总开发者 4700 万+；GitHub 2025 已有 1.8 亿+开发者账号，但这个更像平台账户池，不能直接当作付费席位。

口径	全球人数	适合用途
窄口径	2000 万	高质量职业开发者
基准口径	3000 万	全球专业程序员 TAM
宽口径	4700 万	专业+半专业+活跃开发者

当前主流 AI Coding 产品价格已经形成几档：GitHub Copilot Pro 是 10 美元/月，Business 是 19 美元/人/月，Enterprise 是 39 美元/人/月；Cursor Individual 是 20 美元/月，Teams 是 40 美元/人/月；Claude Max 是 100 美元/月、200 美元/月两档。GitHub 也在从固定请求量切到 usage-based billing，说明高频 Agent 使用会把 ARPU 往上推。

保守市场空间：50 亿—150 亿美元，对应 Copilot / Cursor 这类座席型产品，ARPU 在 120—480 美元/年。

中性市场空间：150 亿—500 亿美元，对应企业开发者普遍上 AI Coding，部分团队开始高频使用 Agent。

激进市场空间：700 亿—1100 亿美元，对应 4700 万宽口径开发者、接近全员付费、并且大量用户使用 100—200 美元/月的 Agent 级产品。

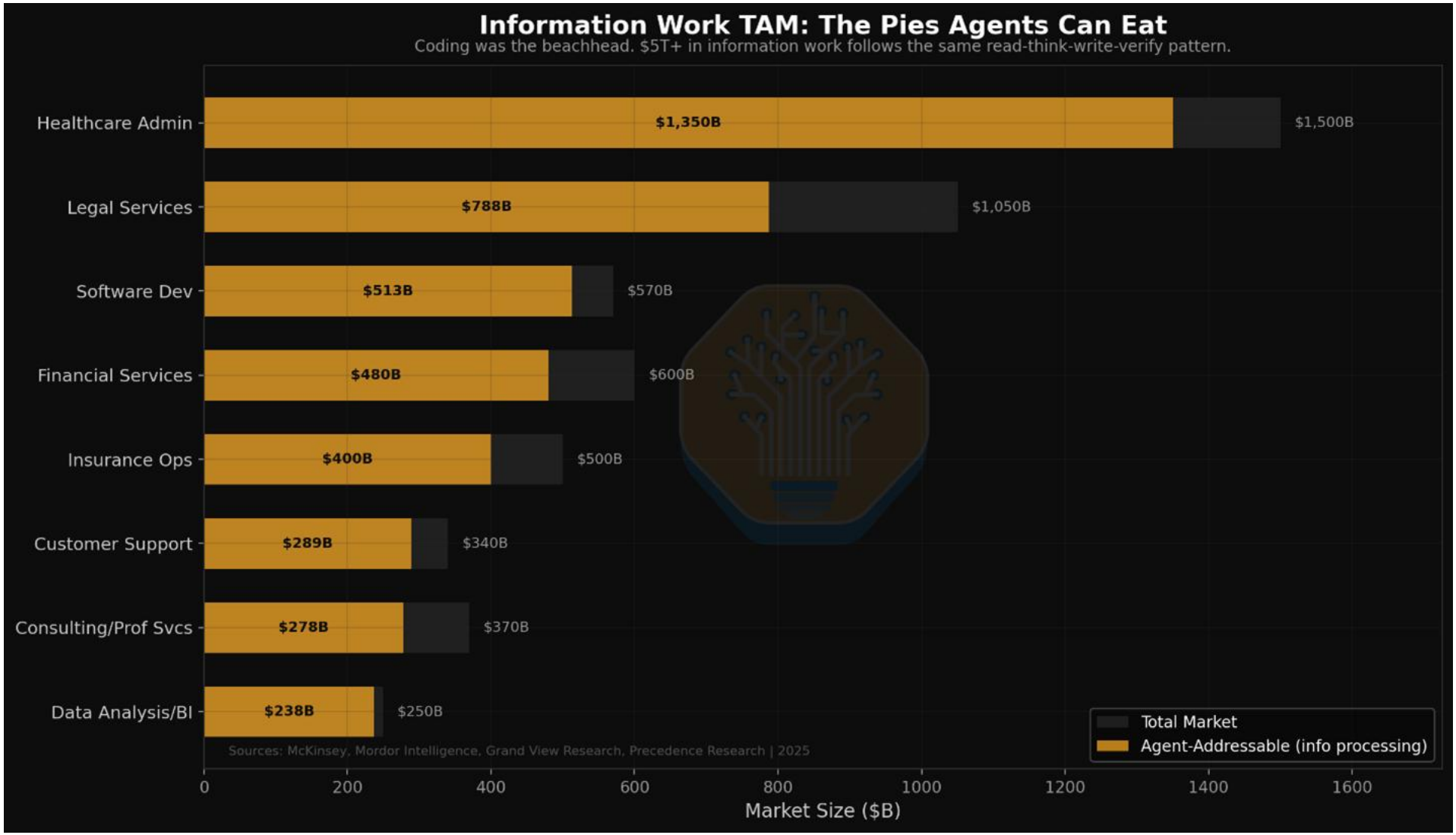
年 ARPU	对应产品/场景	2000 万程序员	3000 万程序员	4700 万程序员
120 美元	Copilot Pro 低端座席	24 亿美元	36 亿美元	56 亿美元
240 美元	Cursor Pro / Copilot Business 附近	48 亿美元	72 亿美元	113 亿美元
480 美元	Cursor Teams / Copilot Enterprise 附近	96 亿美元	144 亿美元	226 亿美元
1200 美元	Claude Max 5x / 重度 Agent 用户	240 亿美元	360 亿美元	564 亿美元
2400 美元	Claude Max 20x / 极高频 Agent 用户	480 亿美元	720 亿美元	1128 亿美元

如果按照替换程序员的逻辑去算的话，可以到万亿美金。

情景	程序员数量	人均年化全成本	全球程序员劳动成本池
保守	2000 万人	5 万美元	**1.0 万亿美元**
中性	3650 万人	10 万美元	**3.65 万亿美元**
乐观	4720 万人	15 万美元	**7.08 万亿美元**

SemiAnalysis 在 Claude Code 文章里把 Claude Code 定义成更接近 Claude Computer：它能读环境、理解文件、规划步骤、调用工具、迭代完成任务，所以软件开发只是第一个落地点。全球知识工作者/信息工作者确实是十亿级。ILO 相关研究估算，全球知识工作岗位约 6.44 亿—9.97 亿，占全球就业 19.6%—30.4%；Forrester 早在 2018 年就估算全球 information workers 达到 12.5 亿。本质是：Agent 可触达的工资、外包、软件、咨询、客服、法务、金融分析、数据分析、文档生产等工作价值池。它衡量的是“可被自动化/重构的劳动成本和服务收入”，不是 AI 公司明天能收到的订阅费。

变量	假设
全球信息工作者	约 10 亿人
每人每年可被 Agent 影响的信息工作价值	约 1.5 万美元
对应劳动池	约 15 万亿美元


Information Work TAM: The Pies Agents Can Eat
Coding was the beachhead. $5T+ in information work follows the same read-think-write-verify pattern.
Healthcare Admin
$1,350B
$1,500B
Legal Services
$788B
$1,050B
Software Dev
$513B
$570B
Financial Services
$480B
$600B
Insurance Ops
$400B
$500B
Customer Support
$289B
$340B
Consulting/Prof Svcs
$278B
$370B
Data Analysis/BI
$238B
$250B
Sources: McKinsey, Mordor Intelligence, Grand View Research, Precedence Research | 2025
Total Market
Agent-Addressable (info processing)
0
200
400
600
800
1000
1200
1400
1600
Market Size ($B)

公司	全球份额	对应市场空间
OpenAI	20.0%	**3.00 万亿美元**
Google	15.0%	**2.25 万亿美元**
Anthropic	15.0%	**2.25 万亿美元**
Meta	10.0%	**1.50 万亿美元**
xAI	5.0%	**0.75 万亿美元**
海外其他	5.0%	**0.75 万亿美元**
ByteDance	6.0%	**0.90 万亿美元**
Alibaba	6.0%	**0.90 万亿美元**
Tencent	6.0%	**0.90 万亿美元**
MiniMax	4.0%	**0.60 万亿美元**
Zhipu	4.0%	**0.60 万亿美元**
中国其他	4.0%	**0.60 万亿美元**
合计	100.0%	**15.00 万亿美元**

公司	份额	5%渗透	10%渗透	20%渗透	30%渗透	34%渗透	40%渗透
OpenAI	20%	1500 亿美元	3000 亿美元	6000 亿美元	***9000 亿美元***	1.02 万亿美元	1.20 万亿美元
Google	15%	1125 亿美元	2250 亿美元	4500 亿美元	6750 亿美元	7650 亿美元	9000 亿美元
Anthropic	15%	1125 亿美元	2250 亿美元	4500 亿美元	6750 亿美元	7650 亿美元	9000 亿美元
Meta	10%	750 亿美元	1500 亿美元	3000 亿美元	4500 亿美元	5100 亿美元	6000 亿美元
xAI	5%	375 亿美元	750 亿美元	1500 亿美元	2250 亿美元	2550 亿美元	3000 亿美元
海外其他	5%	375 亿美元	750 亿美元	1500 亿美元	2250 亿美元	2550 亿美元	3000 亿美元
ByteDance	6%	450 亿美元	900 亿美元	1800 亿美元	2700 亿美元	3060 亿美元	3600 亿美元
Alibaba	6%	450 亿美元	900 亿美元	1800 亿美元	2700 亿美元	3060 亿美元	3600 亿美元
Tencent	6%	450 亿美元	900 亿美元	1800 亿美元	2700 亿美元	3060 亿美元	3600 亿美元
MiniMax	4%	300 亿美元	600 亿美元	1200 亿美元	***1800 亿美元***	2040 亿美元	2400 亿美元
Zhipu	4%	300 亿美元	600 亿美元	1200 亿美元	***1800 亿美元***	2040 亿美元	2400 亿美元
中国其他	4%	300 亿美元	600 亿美元	1200 亿美元	1800 亿美元	2040 亿美元	2400 亿美元
合计	100%	7500 亿美元	1.50 万亿美元	3.00 万亿美元	4.50 万亿美元	5.10 万亿美元	6.00 万亿美元

附录

海外模型近期变化汇总

海外这一轮大模型竞争已经明显从“单一模型发布”扩展为 **模型本体、Agent 产品、入口流量、Token 规模、推理容量和算力资产运营** 的综合竞争。OpenAI 强在通用平台化和 Codex 执行闭环，Anthropic 强在企业 Agent 与金融工作流，Google 强在 Search / Android / Workspace / Cloud 带来的 token 和流量规模，Meta 强在 30 亿+ 社交入口和 AI glasses，xAI 则用 Grok Build 补产品、用 Colossus 1/2 强化算力叙事。

公司	本轮主线	最关键增量	后续重点观察
OpenAI	GPT-5.5 平台化	Codex 长程执行、Realtime 语音、Images 2.0、personal finance、Ads	Codex 是否成为软件工程主入口；广告和金融能否打开新变现
Anthropic	企业 Agent 工厂化	Opus 4.7、金融 Agent、MCP/Stainless、PwC/KPMG、Colossus 1 推理容量	Claude Code 增长、金融/专业服务落地、算力是否继续制约
Google	流量入口 + Token 规模	3200 万亿月度 tokens、Gemini App 9 亿 MAU、AI Overview 25 亿 MAU、AI Mode 10 亿 MAU	Gemini 3.5 Pro、Search AI Mode 对流量和广告的改造
Meta	消费级 Agent + 眼镜入口	Muse Spark、Meta AI app、30 亿+ 用户、subagents、AI glasses	Muse Spark 是否真正进入 WhatsApp/Instagram/Facebook 高频任务
xAI	Grok 产品矩阵 + 算力资产	Grok Build、Skills、Connectors、Imagine、Voice、Colossus 1/2	Grok Build 能否挑战 Claude Code/Codex；Colossus 2 新模型何时发布

1、OpenAI：GPT-5.5 之后，主线是 Codex、实时语音、图像、金融、广告商业化

时间	类型	事件	重要性
2026-04-21	多模态 / 图像	**ChatGPT Images 2.0 发布**，强化复杂图像生成、排版、文字渲染、多语言文字、漫画/海报/商业素材生成等能力。(OpenAI)	对标 Gemini Omni、Grok Imagine，OpenAI 继续补“内容生产入口”。
2026-04-23	模型本体	**GPT-5.5 发布**，强调 coding、research、data analysis、文档/表格/幻灯片生成、工具使用和复杂工作流执行。(OpenAI)	海外本轮最重要模型锚点之一，定位从聊天模型升级到知识工作模型。
2026-04-24	API / 价格	GPT-5.5 API 开放，支持 **1M context**；标准版价格 **$5/M input、$30/M output**，Pro 版 **$30/M input、$180/M output**。(OpenAI)	OpenAI 继续走高端模型定价路线。
2026-05-05	C 端分发	**GPT-5.5 Instant 成为 ChatGPT 默认模型**，官方强调减少幻觉、提升准确性和个性化。(OpenAI)	GPT-5.5 能力下沉到默认流量入口。
2026-05-05	广告商业化	OpenAI 推出 **ChatGPT Ads** 新购买方式，包括 CPC bidding、自助广告平台 beta、广告衡量能力。(OpenAI)	OpenAI 开始显性探索广告变现，商业模式从订阅/API 外扩。
2026-05-07	实时语音 / Agent 入口	发布 **GPT-Realtime-2、GPT-Realtime-Translate、GPT-Realtime-Whisper**；GPT-Realtime-2 是带 GPT-5 级推理能力的实时语音模型，支持工具调用与实时行动。(OpenAI)	语音从“输入方式”升级为实时 Agent 入口。
2026-05-14	Coding Agent	**Codex 移动端 / 远程工作流增强**，用户可在 ChatGPT mobile 中监控、批准、调整 Codex 任务。(OpenAI)	Codex 从本地代码助手变成跨设备工程执行系统。
2026-05-15	垂直应用 / 金融	**ChatGPT personal finance preview** 面向美国 Pro 用户推出，可连接金融账户、生成财务 dashboard，并基于个人金融上下文问答。(OpenAI)	OpenAI 开始切入个人金融，高价值垂类场景打开。
2026-05-21	Coding Agent / 长程任务	Codex **Goal Mode GA**，支持模型围绕目标连续工作数小时甚至数天；远程电脑使用能力增强。(OpenAI 开发者)	Codex 正在接近“长期执行型工程 Agent”。

OpenAI 这一轮的主线是 **GPT-5.5 + Codex + 实时语音 + 图像生成 + 金融场景 + 广告商业化**。GPT-5.5 本体依然是能力锚点，但增量更集中在工作流执行：Codex 负责代码和电脑使用，Realtime 系列负责语音入口，Images 2.0 负责内容生产，personal finance preview 则说明 OpenAI 已经开始尝试高价值垂直场景。整体看，OpenAI 在海外五家里最像“通用 AI 操作系统”，模型、入口、工具、商业化都在同步铺开。

2、Anthropic：Claude Opus 4.7 后，核心是企业 Agent、金融场景、MCP 生态和算力扩张

时间	类型	事件	重要性
2026-04-06	算力基础设施	Anthropic 与 **Google / Broadcom** 扩大合作，获得多 GW 级下一代 TPU 容量，预计 2027 年开始上线；Anthropic 同时披露 ARR 已超 **300 亿美元**，高于 2025 年底约 90 亿美元。(Anthropic)	Claude 增长已经进入“算力交付决定增长速度”的阶段。
2026-04-16	模型本体	**2026 年 4 月 7 日 Mythos 正式公布，随后其蒸馏版 Claude Opus 4.7 发布**，强化高级软件工程、长程任务、金融分析、复杂工具使用和视觉能力。(Anthropic)	Claude Code 与企业 Agent 的核心底座升级。
2026-04-17	设计 Agent	**Claude Design** research preview 发布，可做原型、wireframe、pitch deck，并可移交 Claude Code。(Anthropic)	Claude 从 coding 扩到产品、设计、商业表达。
2026-04-24	融资 / 算力	Google 据报计划向 Anthropic 投资至多 **400 亿美元**，其中 **100 亿美元现金承诺**、300 亿美元与业绩目标挂钩；Google Cloud 还将提供约 **5GW** 算力。(TechCrunch)	Anthropic 与 Google 的绑定进一步加深，算力与资本同步推进。
2026-05-05	金融 Agent	Anthropic 发布 **10 个金融服务 Agent 模板**，覆盖 pitchbook、KYC、月结、估值复核、财报 review、market research 等；同时支持 Excel、PowerPoint、Word、Outlook 相关工作流。(Anthropic)	这是 Claude 在金融行业落地最强的一次产品化。
2026-05-06	算力 / 推理容量	Anthropic 与 SpaceX 达成 Colossus 1 算力协议，获得 **300MW+、22 万+ NVIDIA GPU**，并提高 Claude Code、Claude Opus API、Claude Pro/Max 容量。(Anthropic)	这笔算力主要服务推理容量和 Claude Code 使用上限。
2026-05-06	算力结构	xAI 官方确认 **Colossus 1 含 H100、H200、GB200**，规模超过 22 万 GPU；市场解读称混合芯片集群更适合推理容量，而前沿训练更适合迁往同构 Blackwell 集群。(xAI)	这把 AI 竞争从“模型能力”推向“算力资产周转效率”。
2026-05-14	企业渠道	PwC 扩大与 Anthropic 合作，将 Claude 用于技术构建、交易执行和企业职能重塑。(Anthropic)	四大渠道帮助 Claude 进入金融、审计、咨询、企业转型场景。
2026-05-18	工具生态	Anthropic 收购 **Stainless**，补 SDK、CLI、MCP server 工具链。(Anthropic)	Anthropic 在加固 MCP 与开发者基础设施。
2026-05-19	企业渠道 / 金融税务	KPMG 与 Anthropic 达成全球联盟，Claude 将嵌入 KPMG Digital Gateway，**276,000+ 员工**获得 Claude 访问权限。(Anthropic)	Claude 进入税务、法律、PE、企业服务的分发渠道。

Anthropic 这一轮最重要的变化是 **Claude Opus 4.7 + 金融 Agent + MCP / connectors + 四大渠道 + 多云算力扩张**。Opus 4.7 继续巩固 coding 和长程任务能力，金融 Agent 模板把 Claude 直接嵌进投行、资管、风控、KYC 和月结流程；PwC、KPMG 则提供企业分发；Google/Broadcom、SpaceX、Amazon、Microsoft/NVIDIA、Fluidstack 等算力协议说明 Claude 的瓶颈已经从模型证明转向算力交付。

3、Google：Gemini 3.5 Flash 是模型锚点，真正大招是 Token 规模、Search 入口、Workspace/Android/Cloud 分发

时间	类型	事件	重要性
2026-04-06	算力 / 战略投资	Google/Broadcom 与 Anthropic 扩大 TPU 合作；随后 Google 据报计划最高投资 Anthropic **400 亿美元**，并提供约 **5GW** 算力。(Anthropic)	Google 一边发展 Gemini，一边把 Anthropic 纳入 TPU / Cloud 生态。
2026-05-19	模型本体	**Gemini 3.5 Flash 发布并 GA**，接入 Gemini API、AI Studio、Android Studio、Antigravity、Gemini App、Search AI Mode 等。(Reuters)	Google 用 Flash 走高速、低成本、全入口分发。
2026-05-19	模型路线	**Gemini 3.5 Pro 预告**，Google 表示仍在内部测试，预计下月推出。(Reuters)	Flash 先扩大使用，Pro 后续补高端能力。
2026-05-19	Token 规模	Google 披露模型 API 处理量约 **190 亿 tokens/min**；第三方整理称月度 AI token 处理量约 **3.2 quadrillion / 3200 万亿 tokens**，较一年前大幅增长。(blog.google)	这是 Google 本轮最重要的数据，说明 Gemini 已嵌入海量产品流量和企业调用。
2026-05-19	C 端流量	Gemini App MAU 达 **9 亿**；AI Overviews MAU 超 **25 亿**；AI Mode MAU 超 **10 亿**。(Business Insider)	Google 正在用搜索和 Gemini App 重构 AI 流量入口。
2026-05-19	多模态 / 世界模型	**Gemini Omni / Omni Flash 发布**，从视频输出开始，后续扩展到任意输入到任意输出；接入 Gemini App、Flow、YouTube Shorts。(blog.google)	对标 OpenAI 图像/视频和 xAI Imagine，Google 押注多模态世界模型。
2026-05-19	Coding Agent	**Antigravity 2.0** 推进，支持 agent-first 开发平台、SDK、subagents、hooks、异步任务、CLI。(Reuters)	Google 正面进入 Codex / Claude Code / Grok Build 战场。
2026-05-19	通用 Agent	**Gemini Spark** 发布，定位更自主的个人/办公 Agent，可接入 Google Workspace、运行在 Google Cloud VM、支持后台任务和 MCP。(新闻澳大利亚)	这是 Google 对“通用任务 Agent”的回答。
2026-05-19	搜索 Agent	Search AI Mode 升级，AI agents 被嵌入搜索框，可做任务执行、视觉化回答、代码生成等。(Reuters)	Google 的防守核心是把 AI 直接放进搜索主入口。

时间	类型	事件	重要性
2026-05-19	硬件入口	Google 展示 Gemini 智能眼镜路线，合作方包括 Samsung、Gentle Monster、Warby Parker。(Reuters)	Google 也在争夺下一代端侧入口。
2026-05-19	云 / 算力供给	Google 与 Blackstone 推出 AI cloud venture，目标到 2027 年提供 **500MW** 数据中心容量，整体投资可达 **250 亿美元**，提供 Google TPU 算力。(Reuters)	Google 把 TPU 从自用/云服务扩成外部算力供给平台。

Google 这一轮的重点已经不止 Gemini 3.5 Flash，而是 **Token 规模 + Search 入口 + Gemini App + Workspace/Android/YouTube/Cloud 分发 + TPU 算力生态**。Gemini App 9 亿 MAU、AI Overviews 25 亿 MAU、AI Mode 10 亿 MAU、模型 API 190 亿 tokens/min，这些数据说明 Google 正在把 AI 从 chatbot 扩成全产品层能力。Gemini 3.5 Flash 提供高速低成本模型底座，Omni 抢多模态视频入口，Antigravity 抢 coding，Spark 抢办公/个人 Agent，Google/Blackstone 和 Anthropic 算力合作则把 TPU 变成更强的产业筹码。Google 这一轮最值得写的不是单点模型参数，而是**“流量入口 + token 调用 + 自研芯片 + 云基础设施”**的闭环。

4、Meta：Muse Spark 是模型锚点，真正看点是 30 亿+ 用户入口、Meta AI、眼镜和消费级 Agent

时间	类型	事件	重要性
2026-04-08	模型本体	**Muse Spark 发布**，是 Meta Superintelligence Labs 的首个模型；强调 small and fast、多模态、科学/数学/健康推理、personal superintelligence。(AI Meta)	Meta AI 新路线的核心底座。
2026-04-08	安全 / 上线体系	Meta 同日发布 advanced AI 构建与测试体系，强调可靠性、安全性和用户保护。(AI Meta)	Meta 在补模型上线、风控和消费级规模化能力。
2026-05-06 左右	消费级 Agent	市场消息称 Meta 正在基于 Muse Spark 开发更高级的消费级 AI 助手，面向 **30 亿+ 用户**，目标处理更自主的日常任务，处于内部测试。(X (formerly Twitter))	Meta 正从聊天助手走向日常任务 Agent。
2026-05-12	Meta AI App / 多 Agent	Meta 官方更新称 Muse Spark 已让 Meta AI app 和 meta.ai 升级，支持更快语音、更聪明 AI glasses、购物与对话帮助；Meta AI 可在任务中启动多个 subagents 并行处理。(about.fb.com)	Meta 的 Agent 路线更贴近日常消费场景，比如旅行、购物、内容、社交。
2026-05-12	语音 / 入口	Muse Spark 支持 Meta AI app 里的自然语音对话，可打断、切话题、切语言，并可生成图像、	语音和推荐流结合，是 Meta 相比纯

时间	类型	事件	重要性
		调用 Reels / 地图推荐。(about.fb.com)	API 模型的入口优势。
2026-05-12	眼镜 / 端侧入口	Meta 更新称 Muse Spark 让 AI glasses 更聪明，并将能力推向更高频设备入口。(about.fb.com)	Meta 的长期差异化在 Ray-Ban / AI glasses。
2026-05-13	商业消息 / 开发者	Meta 推出面向 Business Messaging 的 AI developer assistant，服务 WhatsApp / Messenger 商业消息生态。	Meta 在 B 端商业消息链路里嵌 AI。
2026-05-18	组织 / AI-native	Reuters 报道 Meta 进行 AI 相关组织重构，约 7,000 名员工转向 AI workflows，约 10% workforce 受影响。(blog.google)	Meta 把 AI 从产品战略推向组织流程重构。

Meta 这一轮的核心是 **Muse Spark + Meta AI app + 30 亿+ 社交入口 + AI glasses + 消费级 Agent**。Muse Spark 本体强调小、快、多模态和个人超智能，但真正的产业含义在于 Meta 能把模型直接嵌入 Facebook、Instagram、WhatsApp、Messenger、Meta AI app 和智能眼镜。“更高级消费者 AI 助手”说明 Meta 不是只做模型 demo，而是在做面向日常任务的自主 Agent：旅行规划、购物推荐、内容生成、社交对话、地图/Reels 调用都可能成为入口。Meta 的优势不在 API 定价，而在 **最大规模 C 端分发和下一代眼镜入口**。

5、xAI：Grok 4.3 / Grok Build / Skills / Connectors 补产品，Colossus 1/2 补算力叙事

类型	事件	重要性
2026-04-30	语音 / 个性化	**Custom Voices 与 Voice Library 发布**，可用短录音克隆声音，用于 Grok TTS 和 Voice Agent APIs。(xAI)	Grok 在语音人格化和品牌语音入口上补能力。
2026-05-06	企业数据 / Connectors	**Grok Connectors** 上线 web、iOS、Android，接 SharePoint、Outlook、OneDrive、Google Workspace、Notion、GitHub、Linear 和 BYO MCP。(xAI)	Grok 开始连接企业数据源和办公系统。
2026-05-06	多模态 / 图像视频	**Grok Imagine Quality Mode API** 发布，强调真实感、文字渲染和创意控制。(xAI)	对标 OpenAI Images 2.0 与 Gemini Omni。
2026-05-06	算力商业化	SpaceXAI / xAI 与 Anthropic 达成 Colossus 1 算力合作；官方称 Colossus 1 拥有 **22 万+ NVIDIA GPU**，包括 H100、H200、GB200。(xAI)	xAI 从“只训练 Grok”转向“算力资产可外部变现”。

	类型	事件	重要性
2026-05-06	算力资产轮换	市场解读称 Colossus 1 混合 H100/H200/GB200，更适合推理容量，前沿训练迁往更同构的 Blackwell / Colossus 2；推文估算租金约 **$2.6/GPU·hour**、年收入约 **50–60 亿美元**。(X (formerly Twitter))	这条适合写成产业判断：AI 算力资产开始像数据中心金融资产一样运营。
2026-05-14	Coding Agent	**Grok Build beta 发布**，面向 SuperGrok Heavy 用户，是终端里的 coding agent / CLI，支持 plan、review、approve、diff、AGENTS.md、plugins、hooks、skills、MCP servers。(xAI)	xAI 正式进入 Codex / Claude Code / Antigravity 的核心战场。
2026-05-15 左右	模型底座	xAI 旧模型迁移到 **Grok 4.3**；市场信息称 Grok 4.3 支持长上下文和多档 reasoning effort，grok-code-fast-1 迁移到 Grok Build 相关模型。	Grok 4.3 成为通用底座，Grok Build 成为 coding 底座。
2026-05-18	Skills / 办公 Agent	**Grok Skills 发布**，内置 Word、PPT、表格、PDF、Skill Creator 等，支持 web、iOS、Android。(xAI)	Grok 开始进入文档、表格、演示稿等办公生产流。
2026-05-21	开源开发者入口	**Grok 接入 OpenCode**，用户可用 SuperGrok 或 X Premium 订阅在 OpenCode 中调用 Grok Build。(xAI)	xAI 用开源开发工具扩大 Grok Build 分发。
2026-05 上旬	Colossus 2 / 新模型训练	X 上可检索到 Mark K 相关贴文称，Elon Musk 确认多个新 Grok 模型正在 **Colossus 2** 上训练，Grok Build 进展顺利。(X (formerly Twitter))	xAI 的下一轮模型能力押注 Colossus 2 和更同构训练集群。

xAI 这一轮可以写成两条线：一条是 **Grok 产品矩阵补齐**，一条是 **Colossus 算力资产轮换**。产品端，Grok Build 对标 Codex 和 Claude Code，Grok Skills 切文档/PPT/表格/PDF，Grok Connectors 接企业数据源，Grok Imagine 和 Voice 补多模态入口；算力端，Colossus 1 被整体提供给 Anthropic，官方确认其包含 H100、H200、GB200，规模超过 22 万 GPU。市场推文的核心判断值得放进报告：Colossus 1 作为混合芯片训练集群效率有限，但作为单一大客户推理资产可以产生稳定现金流；xAI 则把前沿训练迁往 Colossus 2。这个逻辑非常适合概括为：**AI 竞争正在从模型能力竞争，升级为模型、Agent 产品、推理容量和算力资产运营的复合竞争。**

国内模型近期变化汇总

公司	本轮主线	最关键增量	报告里最值得写的点
DeepSeek	V4 + 华为芯片 + 开源 + 降价	V4-Pro 1.6T / 49B、1M context、Ascend 950、永久 75% 降价	国产旗舰模型价格锚被重定，国产算力栈被

公司	本轮主线	最关键增量	报告里最值得写的点
			验证
智谱	GLM-5.1 + ZCube + TileRT	ZCube 成本 -33%、吞吐 +15%、TTFT P99 -40.6%；Highspeed 400 tok/s	从模型厂商升级为推理系统和组网架构玩家
Kimi	K2.6 + Agent Swarm	300 sub-agents、4,000 steps、长程 coding、开源权重	国内长程 Agent / 多智能体编排代表
MiniMax	M2.7 + 自我进化	Agent Teams、complex Skills、dynamic tool search、自建 RL harness	“模型参与模型研发”的自我演进叙事
阶跃	Step 3.5 Flash 高效开源 Agent 底座	196B / 11B 激活、MTP-3、256K、100–300 tok/s	高智能密度、低成本、国产芯片适配
小米	MiMo-V2.5 + Token Plan + 全模态	1T / 42B、1M context、全模态、Token Plan 降复杂度	模型能力向手机、汽车、IoT、语音入口延展
阿里 Qwen	Qwen3.7-Max + FlashQLA + 真武芯片 + 淘宝入口	35 小时 Agent、TileLang kernel、Zhenwu M890、电商 Agent	芯片—云—模型—入口全栈闭环最完整
字节 Seed	Doubao 2.0 + Seedance 2.0 + 火山方舟	1.55 亿周活、Agent era、视频生成、256K 工具调用	流量入口最强，多模态传播力强，但版权风险突出

国内这一轮竞争比海外更“工程化”。海外重点是 GPT-5.5、Claude Opus、Gemini 3.5 这类模型与 Agent 产品；国内则快速进入 **模型本体 + 推理引擎 + 组网架构 + 国产芯片 + Coding Plan + API 价格战** 的复合竞争。DeepSeek 用 V4 和永久降价重塑价格体系，智谱用 ZCube 和 TileRT 把竞争下沉到推理基础设施，阿里用 Qwen3.7-Max、FlashQLA 和真武芯片做全栈闭环，小米用 MiMo 绑定全模态与硬件入口，Kimi、MiniMax、阶跃分别代表 Agent Swarm、自我进化 Agent Harness 和高智能密度开源底座，字节则用豆包流量和火山方舟把模型能力推向大规模应用。国内大模型的重点已经从“有没有强模型”转向 **谁能更快、更便宜、更稳定地跑长程 Agent**。

1、DeepSeek：V4 是国产模型里最强“开源 + 华为芯片 + 价格战”事件

时间	类型	事件	重要性

时间	类型	事件	重要性
2026-04-24	模型本体	**DeepSeek-V4 Preview 发布并开源**，包含 **V4-Pro** 和 **V4-Flash** 两个版本；V4-Pro 为 **1.6T 总参数 / 49B 激活参数**，V4-Flash 为 **284B / 13B 激活参数**，均支持 **1M context**。(DeepSeek API Docs)	国产开源模型重新进入全球前沿竞争，且把 1M context 做成默认卖点。
2026-04-24	API / 生态	V4 API 当日开放，兼容 **OpenAI ChatCompletions** 与 **Anthropic API**；deepseek-chat 和 deepseek-reasoner 后续将路由/迁移到 V4-Flash。(DeepSeek API Docs)	降低开发者迁移成本，直接抢海外 API 与 coding agent 工作流。
2026-04-24	国产算力	华为宣布基于 **Ascend 950** 的超节点将全面支持 DeepSeek V4；Reuters 称 V4 是 DeepSeek 首个适配华为硬件的模型。(Reuters)	DeepSeek 从“模型公司”变成国产算力栈的核心牵引者。
2026-04-26	API / 成本	DeepSeek API 全系 **cache hit 输入价格降至发布价的 1/10**。(DeepSeek API Docs)	长上下文和 Agent 场景最吃 cache，降 cache price 等于直接降低 Agent 成本。
2026-04-29	产业链	V4 发布后，ByteDance、腾讯、阿里等国内大厂被报道加速抢购华为 **Ascend 950PR** 芯片。(Reuters)	V4 成为国产 AI 芯片需求的催化剂。
2026-05-23	价格战	**V4-Pro 75% 折扣永久化**，API 价格从 **0.1–24 元 / 百万 tokens** 下调到 **0.025–6 元 / 百万 tokens**。(Reuters)	这是真正的“旗舰模型价格战”，会压低国内 Agent / Coding API 的定价锚。

DeepSeek 这一轮的核心是 **V4 + 华为 Ascend + 开源 + 价格战**。V4-Pro 用 1.6T 总参数、49B 激活参数和 1M context 把国产开源模型重新推回全球前沿；更关键的是，它同时适配华为 Ascend 950 超节点，并通过 cache hit 降价和 V4-Pro 永久 75% 降价，把长上下文与 Agent 的调用成本直接打下来。DeepSeek 的意义已经不只是模型能力，而是成为国产算力、开源生态和 API 价格体系的共同锚点。

2、智谱：GLM-5.1 是模型锚点，ZCube 和 TileRT 是真正增量

时间	类型	事件	重要性
2026-02-11	模型本体	**GLM-5 发布**，Reuters 称其强化 coding 与长程 Agent 任务，并使用包括华为 Ascend、摩尔线程、寒武纪、昆仑芯等国产芯片进行推理。(Reuters)	智谱较早把“国产芯片推理 + Agentic coding”绑定。
2026-02-12	商业化	智谱上调 **GLM Coding Plan** 价格至少 30%，理由是 coding 需求增长。(Reuters)	说明 coding plan 已经有真实需求，不只是模型 demo。
2026-04-01	多模态 / GUI Agent	**GLM-5V-Turbo** 发布，定位原生多模态视觉 coding / GUI Agent / OpenClaw 场景。(MarkTechPost)	智谱把 Agent 从文本代码扩到屏幕、视觉和 GUI 操作。

时间	类型	事件	重要性
2026-04-07 左右	模型本体	**GLM-5.1 发布**，官方文档称其面向长程任务，可在单次运行中独立工作最长 **8 小时**；GitHub 描述其为面向 **agentic engineering** 的下一代旗舰模型。(Z.AI)	GLM-5.1 是智谱对 Claude Code / Codex 战场的正面回应。
2026-05-21	组网 / 推理基础设施	**ZCube 组网在 GLM-5.1 coding 生产集群落地**；相比 ROFT 架构，交换机与光模块成本下降 **33%**，GPU 平均推理吞吐提升 **15%+**，TTFT P99 下降 **40.6%**。(智谱 AI)	这是智谱最有产业价值的事件：模型性能之外，开始解决推理集群网络瓶颈。
2026-05-22	推理系统	**GLM-5.1-highspeed** 面向部分企业客户开放，输出速度达到 **400 tokens/s**；报道称该 API 由智谱 GLM 团队与 **TileRT 团队**联合打造，在推理引擎、调度系统和底层基础设施三层优化。(开源中国)	智谱把“旗舰模型能力”和“低延迟”同时放到生产 API，直接服务 coding、实时交互和语音场景。
2026-05	Coding Plan / 工具入口	GLM Coding Plan 支持 GLM-5.1、GLM-5-Turbo，并接入 Claude Code、Kilo Code、Cline、OpenCode、OpenClaw 等开发工作流。(Z.ai)	智谱正在从模型 API 走向开发者订阅和工具链入口。

智谱这一轮不能只写 GLM-5.1。真正的重点是 **GLM-5.1 + ZCube + TileRT + Coding Plan**。GLM-5.1 负责长程 coding agent 能力，ZCube 负责解决 PD 分离推理中的网络成本和 TTFT 问题，TileRT 则把旗舰模型推到 400 tokens/s 的低延迟生产 API。智谱这条线的报告写法应该从“模型厂商”升级为“模型 + 推理系统 + 组网架构 + 开发者订阅”的全栈叙事。

3、Kimi / 月之暗面：K2.6 的关键词是长程 Agent、Agent Swarm、开源权重

时间	类型	事件	重要性
2026-04-20	模型本体	**Kimi K2.6 上线 Workers AI**，Cloudflare 称其为 **1T 参数 / 32B 激活参数**、**262K context**、原生多模态 Agent 模型，支持长程 coding、主动执行和 swarm-based orchestration。(Cloudflare Docs)	Kimi K2.6 是国内开源 Agent 模型的核心事件之一。
2026-04-20 左右	Agent Swarm	Hugging Face 模型页称 K2.6 支持 **300 个 sub-agents** 与 **4,000 个 coordinated steps**，可把文档、网页、表格等任务串成一次端到端自主运行。(Hugging Face)	Kimi 把竞争点从单模型回答推进到多智能体编排。
2026-04 下旬	长程 coding	第三方整理称 K2.6 曾对开源金融撮合引擎 exchange-core 进行 **13 小时自主优化**，涉及 1,000+ tool calls、4,000+ 行代码修改，吞吐提升 185%。(Verdent AI)	这类案例适合证明“长程 Agent 耐力”，但报告里要标为第三方整理。
2026-04 下旬	产品 Agent	Kimi Help Center 称 **K2.6 Agent** 可使用 20+ 工具，端到端处理建站、文档生成、数据分析等复杂任务。(Kimi AI)	Kimi 在 C 端产品里已经把 Agent 作为主入口。

时间	类型	事件	重要性
2026-04 下旬	分发	Kimi K2.6 已在 Cloudflare Workers AI 首日支持，说明其海外开发者分发不只依赖自家 App。(Cloudflare Docs)	有助于 Kimi 打开海外开源模型生态。

Kimi K2.6 的核心是 **长程 coding + Agent Swarm + 开源权重 + 海外开发者分发**。它没有只强调单模型 benchmark，而是把能力锚在 300 个子 Agent、4,000 个协作步骤、长时间自主执行和端到端任务交付上。Kimi 这条线适合写成国内最接近 Claude Code / OpenAI Codex "长程 Agent 工作流"的开源模型路线。

4、MiniMax：M2.7 的特点是"自我进化"与 Agent Harness

时间	类型	事件	重要性
2026-03-18	模型本体	**MiniMax M2.7 发布**，官方称其可构建复杂 Agent harness，完成高复杂生产力任务，支持 **Agent Teams、complex Skills、dynamic tool search**。(MiniMax)	MiniMax 把模型定位从聊天/多模态转向复杂 Agent 执行。
2026-03-18	自我进化	官方称 M2.7 是 M2 系列中首个深度参与自身演进的模型：开发过程中让模型更新 memory、构建复杂 skills，并改进 RL harness。(MiniMax)	"模型参与模型研发流程"是 M2.7 最独特的叙事。
2026-04-12	开源 / 生态	M2.7 开源后，GitHub 和模型社区围绕自部署、agentic coding、复杂 skill 使用形成开发者讨论。(GitHub)	MiniMax 从闭源 API 走向开源权重生态。
2026-04-14	部署 / 推理	NVIDIA NGC 上线 **MiniMax-M2.7 NIM Container**，用于部署 MiniMax-M2.7 推理服务。(NVIDIA NGC)	有利于企业把 M2.7 放进 GPU 云和私有化环境。
2026-05-22	Agent 产品	MiniMax Agent changelog 显示持续迭代 Desktop App、Experts、Custom Mode、Lightning Mode、PPT Agent、MCP Builder 等功能。(agent.minimax.io)	MiniMax 不只发模型，也在补完整 Agent 产品形态。

MiniMax M2.7 的核心是 **自我进化模型 + Agent Teams + complex Skills + dynamic tool search**。这套叙事和智谱、Kimi 不完全一样：

智谱强调推理系统和组网，Kimi 强调 Agent Swarm，MiniMax 更强调模型参与构建自身的训练、评测和 skill harness。报告里可以把 M2.7 写成国内“Agent Harness 型模型”的代表，它的竞争点在复杂工作流、办公自动化和多智能体协作。

5、阶跃星辰：Step 3.5 Flash 是“高智能密度”的开源 Agent 底座

时间	类型	事件	重要性
2026-02-02	模型本体	**Step 3.5 Flash 开源**，官方称其是面向实时 Agent 工作流的开源基座模型；采用 MoE 架构，**196B 总参数 / 11B 激活参数**，支持 **256K context**。(GitHub)	阶跃的路线是“高智能密度”：大容量模型、低激活成本。
2026-02-02	推理效率	Step 3.5 Flash 使用 **MTP-3**，典型生成吞吐 **100–300 tok/s**，单流 coding 峰值可达 **350 tok/s**。(GitHub)	直接对标国内高速 Agent 模型路线。
2026-02-05	国产芯片适配	华为昇腾发文称支持 Step 3.5 Flash Agent 模型推理部署。(昇腾社区)	阶跃也在走国产算力生态适配。
2026-02 至今	开源生态	GitHub、ModelScope、OpenRouter 等渠道均可获取/调用 Step 3.5 Flash。(GitHub)	阶跃更像“开源高效 Agent 底座”，便于被国内外工具链接入。
2026-04 以后	观察项	DeepSeek V4 之后，未看到 Step 4 级别旗舰正式发布，主线仍是 Step 3.5 Flash 的高效部署和生态接入。	后续重点看是否推出更强长程 Agent 模型。

阶跃星辰这一轮的代表事件是 **Step 3.5 Flash**。它的价值不在于参数最大，而在于 196B 总参数、11B 激活参数、MTP-3、256K context 和 100–300 tok/s 的组合，形成“高智能密度、低推理成本、适合实时 Agent”的模型路线。相比 DeepSeek 和智谱，阶跃的产业叙事更轻量，适合作为国产开源 Agent 底座和国产芯片适配案例来写。

6、小米 MiMo：从模型进入 Agent 框架、Token Plan 和全模态入口

时间	类型	事件	重要性
2026-03-18	模型本体	**MiMo-V2-Pro 发布**，官方称其超过 **1T 总参数 / 42B 激活参数**，支持 **1M context**，面向复杂真实 Agent 工作流。(小米 MiMo 平台)	小米正式进入前沿 Agent 模型竞争。
2026-03-19	战略投入	雷军宣布小米未来三年 AI 投入至少 **600 亿元人民币**；Reuters 称 MiMo-V2-Pro 在 OpenRouter 上已处理超过 **1.5 万亿 tokens**。(Reuters)	小米不是“小模型试水”，而是集团级 AI 投入。
2026-04-23	模型矩阵	**MiMo-V2.5 系列发布 / 公测**，包括 MiMo-V2.5、V2.5-Pro、TTS、ASR；V2.5-Pro 为 **1T 参数 / 42B 激活 / 1M context**，V2.5 支持图像、视频、音频、文本的原生全模态理解。(小米 MiMo 平台)	小米开始形成文本、全模态、语音的模型矩阵。
2026-04-29	Token Plan / 商业化	MiMo-V2.5 Token Plan 优化：MiMo-V2.5 为 **1 Token = 1 Credit**，MiMo-V2.5-Pro 为 **1 Token = 2 Credits**；取消此前 4x 与 256K/1M context 倍率差异，并新增夜间折扣和连续包月折扣。(小米 MiMo 平台)	小米把模型商业化做成更适合 Agent 高 token 消耗的订阅/点数体系。
2026-04-27 / 04-28	开源 / 硬件生态	MiMo-V2.5-Pro 开源，AMD 称 MiMo-V2.5 系列 4 月 23 日公测、4 月 28 日 GA 与开源，并提供 Day 0 AMD Instinct 支持。(小米 MiMo)	小米走“开源模型 + 多硬件适配 + Agent 框架”的路线。
2026-04 下旬	Agent 能力	官方称 MiMo-V2.5-Pro 可完成长期复杂 Agent 任务，案例包括 4.3 小时完成编译器项目、11.5 小时构建视频编辑器 Web 应用。(小米 MiMo 平台)	小米在用长时间自主执行案例证明 Agent 能力。

小米 MiMo 这一轮的核心是 **1T MoE + 1M context + 全模态 + Token Plan + 开源生态**。MiMo-V2.5-Pro 主攻复杂 Agent 和长程软件工程，MiMo-V2.5 则覆盖图像、视频、音频、文本等全模态输入。更值得注意的是小米把 Token Plan 做了重新设计，取消长上下文倍率差异，降低 Agent 高频调用的心理门槛。小米的差异化在于：它既有模型，又有手机、汽车、IoT、语音和端侧入口，未来可能把 MiMo 变成硬件生态的智能底座。

7、阿里 Qwen：Qwen3.7-Max、FlashQLA、真武芯片和淘宝入口构成全栈叙事

时间	模型 / 事件	类型	核心变化	怎么理解
2026-01-23	**Qwen3-Max 2026-01-23 / Thinking mode**	闭源旗舰升级	阿里云 Model Studio 显示 Qwen3-Max 2026-01-23 支持 thinking / non-thinking，工具调用，按上下文长度分层定价。	这是 Qwen3-Max 的一次能力升级，先把 thinking 和工具调用做进旗舰模型。(AlibabaCloud)

时间	模型 / 事件	类型	核心变化	怎么理解
2026-02-02	**Qwen3-Coder-Next**	Coding / 本地开发	Qwen 官方称其是面向 coding agents 和 local development 的 open-weight 模型。	这是阿里在“轻量 coding agent / 本地开发”方向的补位。(Qwen Studio)
2026-02-16	**Qwen3.5**	Agentic AI 主线	Reuters 称 Qwen3.5 面向 “agentic AI era”，目标是自主执行移动端和桌面端复杂任务；阿里称其成本效率较前代更优。	Qwen3.5 是阿里从通用模型切到 Agent 模型的正式起点。(Reuters)
2026-04-01	**Qwen3.6-Plus**	多模态 / Agentic coding	官方 Model Studio 仍列出 Qwen3.6-Plus，定位是 **native multimodal、1M context、agentic coding**；价格页显示输入约 **$0.5–$2 / 1M tokens**。	这就是你说“好像消失”的核心版本。它没有消失，只是被 3.7-Max 的发布盖住了。(modelstudio.alibabacloud.com)
2026-04-14	**Qwen3.6-35B-A3B**	开源 / Agentic coding	Qwen 官方发布 “Agentic Coding Power, Now Open to All”。	这是 3.6 系列里的开源 coding 分支，强调小激活参数下的 agentic coding 能力。(Qwen Studio)
2026-04-21	**Qwen3.6-27B**	Dense coding 模型	Qwen 官方发布 “Flagship-Level Coding in a 27B Dense Model”。	这是 3.6 系列的 27B dense coding 分支，适合本地部署、私有化和低成本 coding 场景。(Qwen Studio)
2026-04-22	**Qwen3.6-Max-Preview**	闭源旗舰预览	阿里云社区称 Qwen3.6-Max-Preview 相比 Qwen3.6-Plus，在 agentic coding、world knowledge、instruction following 上提升，并支持 preserve_thinking。	这其实是 3.7-Max 前的过渡旗舰，说明 Qwen3.6 不是单点模型，而是一整代中间层。(AlibabaCloud)
2026-04-27 前后	**FlashQLA**	推理 / Kernel	Qwen 团队发布基于 TileLang 的线性注意力 kernel，面向预训练和端侧 Agent 推理提速。	这不是模型，但说明阿里也在往 runtime / kernel 方向卷。
2026-05-20 / 05-21	**Qwen3.7-Max**	最新旗舰 Agent 模型	Reuters 称 Qwen3.7-Max 面向 advanced coding 和 long-running agent tasks，可连续运行最高 **35 小时**；Model Studio 显示其 2026-05-21 上线，定价 **$2.5 input / $7.5 output**。	Qwen3.7-Max 把 3.6 的 agentic coding 叙事正式升级成“长程 Agent 旗舰”。(Reuters)
2026-05-20	**Zhenwu M890 + Panjiu AL128**	芯片 / 服务器 / 全栈	Reuters 称阿里发布真武 M890 芯片和磐久 AL128 服务器，M890 面向 AI agents 的长上下文、多模型协同和实时通信需求。	Qwen3.7-Max 不只是模型发布，而是和阿里自研芯片、云平台一起形成全栈叙事。(Reuters)

阿里 Qwen 这一轮最完整，主线是 **Qwen3.7-Max + Qwen3.6 系列 + FlashQLA + Zhenwu M890 + 淘宝入口**。Qwen3.7-Max 抢长程 Agent，Qwen3.6-27B 抢开源 coding，FlashQLA 抢推理 kernel，Zhenwu M890 抢国产芯片，淘宝接入则抢消费入口。阿里这条线最适合写成全栈闭环：底层有芯片和云，中层有 Qwen 模型与推理优化，上层有电商和企业应用入口。

8、字节 / 豆包 Seed：Doubao 2.0、Seedance 2.0、火山方舟形成“流量 + 多模态 + Agent 平台”

时间	类型	事件	重要性
2026-02-12	多模态 / 视频	**Seedance 2.0 发布并走红**，Reuters 称其视频生成能力在国内社媒引发关注，并获得马斯克转发式讨论。(Reuters)	字节在视频生成上形成一次“DeepSeek moment”式传播。
2026-02-14	模型本体	**Doubao 2.0 / Seed 2.0 发布**，定位“agent era”，Pro 版具备复杂推理和多步骤任务执行能力，字节称其使用成本较同类模型降低约一个数量级。(Reuters)	字节从聊天 App 进入复杂任务 Agent 竞争。
2026-02-14	用户入口	Reuters 引用 QuestMobile 数据称，豆包为中国最大 AI chatbot app，周活 **1.55 亿**，高于 DeepSeek 的 **8160 万**。(Reuters)	字节最大优势是 C 端流量和内容分发。
2026-02-14	Coding / 工具	Doubao-Seed-2.0-Code 被作为专业软件工程引擎推出，支持 repo-scale coding 和长上下文场景；火山方舟模型列表也显示 Seed 2.0 Pro 支持深度思考、多模态理解、工具调用和 **256K context**。(HowAIWorks.ai)	字节也在补 coding agent，而不仅是 C 端聊天。
2026-03-14	风险 / 合规	The Information 经 Reuters 报道称，字节暂停 Seedance 2.0 全球发布，原因是涉及 Disney 等版权争议。(Reuters)	多模态出海的监管和版权风险开始显性化。
2026-05-11	火山方舟 / API	火山方舟产品公告显示 Chat API 按 token 付费支持低延迟模式，支持 **doubao-seed-2.0-pro**、**doubao-seed-2.0-lite**。(火山引擎)	字节把 Seed2.0 能力沉淀到火山方舟企业平台。

字节这轮的主线是 **豆包 C 端流量 + Seed 2.0 Agent 模型 + Seedance 2.0 视频生成 + 火山方舟企业 API**。Doubao 2.0 定位 Agent era，背后有 1.55 亿周活的消费入口；Seedance 2.0 则证明字节在视频生成上具备很强传播能力，但也暴露版权和全球发布风险。字节最适合写成“应用入口驱动模型迭代”的代表：豆包负责流量，Seed 系列负责模型，火山方舟负责企业化和开发者接入。