

传媒行业深度报告

海外 AI 年度复盘及财报综述: 狂欢将尽还是新周期开启?

增持 (维持)

2026 年 01 月 21 日

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

■ 我们复盘了过去一年海外 AI 产业链的发展, 并展望 2026 年。总体判断是: AI 产业正从 2024-2025 年的狂热扩张期, 逐步转向需求兑现与效率竞争并重的新周期。2025Q3 以来泡沫论集中发酵, 多数 AI 相关美股财报后股价回调, 市场对利好消息反应相对有限、对利空消息更为敏感。但我们认为, AI 领域仅存在局部泡沫, 而非系统性崩盘。AI 基础设施的实际使用寿命周期远超市场担忧, 推理需求爆发、Agent 模式普及以及多模态应用落地, 将为巨额 CapEx 提供更长回报周期支撑。市场的焦虑点集中在 CapEx 规模、折旧争议、循环订单与债务融资, 我们逐一做了分析, 并认为主要是阶段性情绪波动, 而非基本面实质恶化。我们将 AI 价值链按照纵向的“算力-云-模型-应用”拆分, 逐一分析各环节的生态位和优劣势; 并按照横向拆分为“OpenAI+英伟达+微软”阵营和“Anthropic+谷歌+亚马逊”阵营。我们认为风险主要集中于以 OpenAI 为核心、扩张激进的生态链条, 其现金流压力与战略分散可能引发连锁反应; 与之相对, 我们看好谷歌的全栈整合优势、原生多模态优势, Anthropic 的专注和聚焦、成本控制策略。我们预计 2026 年将成为 AI 产业分化加剧、商业闭环验证的关键年份, 投资逻辑将从泛 AI 概念普涨转向精选真实变现能力、成本优势与护城河的结构机会。

■ 算力环节: 英伟达一家独大格局正被多元竞争取代。英伟达 2025 年业绩亮眼, 数据中心收入同比翻倍, Blackwell 供不应求, 毛利率维持 70-80% 高位。但 ASIC 芯片在推理场景的能效与成本优势逐步显现, 为英伟达的传统护城河带来挑战。英伟达通过收购 Groq 补齐低延迟推理短板, 体现主动防御姿态, 但供应不确定性为竞争对手留出窗口。谷歌 TPU v7 全栈优势显著, TCO 远低于英伟达方案, 已向 Anthropic 等外部客户大规模输出。亚马逊 Trainium2 扩张最激进, 计划百万级部署并保持与英伟达双轨合作。微软因与 OpenAI 深度绑定, 自研芯片滞后且面临能源瓶颈, 相对被动。我们认为算力竞争已从单一性能转向成本、规模、生态综合较量, 英伟达一家独大格局正在被多元效率驱动取代, 上游硬件仍是资本开支最大受益环节。

■ 云服务市场分化: 巨头狂欢, 新贵面临困局。三大巨头订单充裕、现金流稳健: 微软 FY26Q1 RPO 达 3920 亿美元 (同比+51%), 谷歌云 2025Q3 承诺余额 1557 亿美元 (同比+82%), AWS 虽增速相对平稳但 10 月新签订单已超 Q3 全季。盈利能力方面, 2025Q3 谷歌云利润率提升至 23.7%, AWS 维持 36.4% 高位, 微软 Azure 储备订单激增但毛利率承压。三家巨头的共同特征是拥有充沛现金流、多元化业务对冲风险、以及清晰的 AI 变现路径。新兴 Neocloud 如 CoreWeave、Nebius 则陷入困局: 高估值、高负债与回款压力交织, 财报后股价大幅回调。CoreWeave 2025Q4 预期收入环比增速仅 13%、资本开支占收入超 4 倍, 商业模式本质是“GPU 租赁+高杠杆”, 类似房地产开发商但面临更快的技术迭代风险, 缺乏规模效应、议价能力弱、客户粘性低, 且随着云巨头产能释放和大模型公司自建算力, 正被两端挤压。

■ 模型层: 从 AGI 神话到工程范式, 中美差距缩窄至 6-9 个月。2025 年的核心脉络是 AGI 神话退场、工程化范式确立。DeepSeek R1 以极低成本证明算法效率路线可行, 推高行业对推理效率的重视。GPT-5 系列未实现预期中的技术跃迁, 其在 GPT-4o 后未能完成大规模预训练突破, 主要依靠后训练优化, 用户体验改善有限甚至出现情感智能退化。而 Gemini 3 Pro 凭借原生多模态架构与 TPU 硬件突破实现较明显跃迁, 在多个榜单拉开差距。展望 2026 年, 我们认为大语言模型将聚焦“更快、

行业走势



相关研究

《2026 年传媒互联网行业策略: 看好 AI 应用、游戏及港股互联网》

2025-12-12

《25Q3 业绩综述: 利润同比增长 40%, 游戏板块增长亮眼》

2025-11-04

更好、更便宜”：Scaling Law 并未失效但需配合算法创新，Nested Learning 等新架构探索长期记忆与推理时学习能力；推理模型消耗 token 量是普通模型 5.5 倍，Agent 模式多模型并行调用进一步推高算力需求。多模态方面，输入端从组合式向原生多模态演进（Google 论文证明早期融合+MoE 架构 scaling 效率更优），输出端实现更长时长、更稳定的视频生成，世界模型与多模态输出将互相促进。

- **模型竞争格局分化明显：**OpenAI 转向产品化但技术领先优势削弱，战略分散风险上升；Google 凭借全栈优势（TPU+YouTube 数据+DeepMind）强势回归，Veo3 实现有声视频突破；Anthropic 专注 toB 编程建立稳健现金流，Claude Code ARR 2025 年 10 月已达 10 亿美元。国内市场群雄逐鹿，千问（基础设施优势）、DeepSeek（算法效率）、Kimi（长文本）、智谱（编程+agent）、MiniMax（多模态）、豆包（流量）各具特色。中美竞争方面，技术差距已缩窄。芯片禁令倒逼中国企业算法创新（稀疏注意力、MoE 等），商业模式上美国坚持“闭源+高价”，中国走“开源+极致性价比”路线。中国模型在开源市场份额快速增长，对海外高定价模式形成冲击。NeurIPS 2025 中国机构占比近半，阿里千问团队获最佳论文奖，反映中国在基础研究层面已具备同台竞技能力。人才回流、资源集中与执行力优势，正推动中国从追赶向并跑转变。
- **应用层：编程率先商业化，垂直 AI 分化，巨头内部赋能价值被低估。**我们用 ARR 与 Tokens 消耗两大指标交叉验证需求真实性：2025 年 10 月 OpenAI+Anthropic 合计 ARR 约 200 亿美元，OpenRouter 2025 年 token 消耗同比增长超 300%，验证需求强劲。①**toB 市场商业化潜力高。**2025 年企业生成式 AI 总支出达 370 亿美元，同比增长超 3 倍，仅用 3 年占据全球 SaaS 市场 6% 份额。垂直行业 AI 支出 35 亿美元，医疗健康独占 15 亿，法律 6.5 亿。这些领域的共同特征是愿意深耕“脏活累活”、积累私有 know-how，构建通用模型难以复制的壁垒。②**toC 市场则呈现“通用 chatbot 主导+垂直 AI 大量死亡+少数场景顽强存活”格局。**25 年通用 chatbot 占 AI 网页访问量 80% 以上，“模型即应用”趋势越发明显，图像生成、设计、创意写作类最容易被通用模型内置功能替代。但编程与角色陪伴因数据闭环与情感粘性存活顽强。编程工具（Cursor、Lovable）ARR 高速增长，角色陪伴应用虽面临记忆不足、成本高企、监管风险，但用户时长较长、情感投入深、迁移成本高，长期看好。③**巨头内部赋能是“看不见的价值”。**谷歌 AI Overviews 提升查询量并保持货币化率，搜索收入增速企稳回升；亚马逊通过 AI 优化物流效率、机器人部署；百度、阿里、腾讯将 AI 嵌入原有生态。这些价值难以量化，但直接转化为全要素生产率提升。
- **投资建议：**我们认为 2026 年 AI 产业将进入需求兑现与效率竞争并重的新周期，狂欢暂告段落，但真正由技术突破与商业闭环驱动的结构性的机会正在浮现。投资逻辑不再是泛 AI 概念普涨，而是聚焦具备真实变现能力、成本优势与长期护城河的核心标的。我们维持看好 AI 产业链，核心思路为“上游强 alpha+云巨头稳健 beta+应用层精选赛道”。①**在上游算力硬件，**我们看好英伟达。Blackwell 系列供不应求、毛利率高企，通过收购 Groq 补齐推理短板，其生态韧性与长期 ROI 空间依然突出，是资本开支最大受益者。②**在云服务与全栈整合领域，**建议关注谷歌与亚马逊：谷歌凭借 TPU 全栈优势、TCO 优势以及 Gemini 与搜索深度融合，展现强利润韧性；亚马逊则通过 Trainium 百万级部署与 AWS 规模效应，在成本控制与客户黏性上占据主动。③**模型与应用层面，**建议关注 MiniMax-WP 与智谱。前者在多模态及角色陪伴场景实现较强商业化，后者抢占 B 端市场。港股互联网公司中，我们推荐阿里巴巴-W、腾讯控股、美图公司、快手-W。这些公司拥有庞大用户基础与数据飞轮，AI 深度嵌入将驱动存量业务效率跃升与新增长曲线。
- **风险提示：**推理需求增长不及预期、模型商业化路径出现重大变化、地缘政治导致算力约束进一步加剧

内容目录

1. AI 投资：泡沫与革命并存，局部风险而非系统崩盘	7
1.1. 泡沫论在 25Q3 集中发酵，较多美股科技公司财报后股价回调	7
1.2. 市场在担心什么：Capex、折旧、ROI、估值、债务	8
1.2.1. 焦虑一：大幅增长的 CapEx	8
1.2.2. CapEx 的增长具有明确的、可持续的支撑	10
1.2.3. 焦虑二：延长折旧周期是合理的吗	15
1.2.4. 焦虑三：循环订单和债务是最核心的问题	16
1.3. 我们的判断：局部泡沫，非系统性崩盘	19
1.3.1. 横向划分：OpenAI 阵营 vs 谷歌阵营	19
1.3.2. 纵向划分：IaaS 层 vs 模型层 vs 应用层	21
2. 算力：英伟达霸权下的暗流涌动	23
2.1. 英伟达：看似坚固的护城河下暗藏裂痕	23
2.2. 谷歌：全栈整合构建的平行宇宙	24
2.3. 亚马逊：最激进的自研芯片扩张者	27
2.4. 微软：和 OpenAI 绑定的双刃剑	27
2.5. GPU vs ASIC：竞争多元化有望推动成本下降	29
3. 云服务市场的冰与火：巨头狂欢与新贵困局	30
3.1. 云巨头的黄金时代：订单、现金流与确定性	30
3.2. 三巨头的差异化：亚马逊份额领先，谷歌和微软快速追赶	32
3.3. Neocloud 的困局：居高不下的债务、快速折旧的芯片	33
3.4. 云服务市场：承载 AI 需求爆发的关键底座	36
4. 模型层：从神话到工程，从追赶到突围	38
4.1. 2025 年：神话退场推理确立范式，工程取代神话	38
4.2. 2026 年展望：深度打磨与多模态突破	39
4.2.1. 大语言模型：更快、更好、更便宜	39
4.2.2. 多模态与世界模型	46
4.3. 美国模型侧竞争：分层结构与战略分化	50
4.4. 中国大模型竞争格局	59
4.5. 中美竞争对比：大力出奇迹 vs 资源倒逼创新	62
4.5.1. 美国保持领先，中国快速追赶	62
4.5.2. 文本、编程、多模态各项能力的对比	63
4.5.3. 商业路径分化：高定价的闭源路线 vs 中低定价的开源路线	66
4.5.4. 算力约束下的竞争格局	71
5. 应用层：AI 应用到底能长多大？	73
5.1. 需求验证的两大指标：ARR 与 Tokens 消耗	74
5.2. Tokens 增长的驱动力和可持续性	75
5.3. toB：谁在构建不可替代性	79
5.4. toC：模型即应用的降维打击	84
5.5. 科技巨头内部赋能	90
6. 投资建议	94
7. 风险提示	95

图表目录

图 1:	2025Q3 关于 AI 泡沫的讨论激增	7
图 2:	科技公司财报披露先后的股价变化.....	8
图 3:	海外大厂 Capex (十亿美元)	9
图 4:	各家头部厂商在榜单上的排名频繁更迭.....	11
图 5:	推理模型大量的 tokens 用于“打草稿”过程	12
图 6:	OpenRouter 平台上推理模型的 tokens 消耗占比持续增长.....	12
图 7:	deepseek v3.2 的交错思维链	13
图 8:	minimax m2 的交错思维链	13
图 9:	GPT-5.2 通过消耗更多 tokens 来提升模型智能.....	14
图 10:	cursor 的多 agent 模式	14
图 11:	新发布的模型大多优化了 agent 能力	15
图 12:	OpenRouter 上的 tokens 周度消耗量.....	15
图 13:	OpenAI 相关的合作 (截至 2025 年 11 月)	17
图 14:	OpenAI ARR (十亿美元)	20
图 15:	Anthropic ARR (十亿美元)	20
图 16:	OpenAI 收入和自由现金流预测 (十亿美元)	21
图 17:	Anthropic 收入和自由现金流预测 (十亿美元)	21
图 18:	Anthropic 公司在算力硬件 (计算资源) 组合上的演变趋势	27
图 19:	云业务收入同比增速对比: 微软 Azure / Google Cloud / 亚马逊 AWS	31
图 20:	2025Q3 全球云服务市场市场份额.....	32
图 21:	Coreweave 和 Lambda 官网界面是一目了然的算力租赁清单	33
图 22:	Coreweave 的商业模式.....	34
图 23:	Coreweave 财务表现.....	34
图 24:	Nebius 财务表现	35
图 25:	Coreweave 两次财报后股价大跌 (美元)	36
图 26:	新模型发布节奏越来越快, SOTA 模型的竞争仍然激烈	38
图 27:	SemiAnalysis 披露: chatgpt 在 GPT-4o 之后没有在预训练环节实现突破.....	40
图 28:	模型的 agent 能力持续提升	42
图 29:	同等智能程度的模型的 Token 定价快速下降。	43
图 30:	Nest Learning 技术原理示意图	44
图 31:	HOPE 在语言建模和常识推理任务上表现出更低的困惑度和更高的准确率	44
图 32:	多模态大模型和全模态大模型的区别.....	47
图 33:	原生多模态模型的 Scaling 特性.....	48
图 34:	2022-2025 年 huggingface 上各类别模型数量排行榜	50
图 35:	LM Arena 上的模型综合排行榜 (截至 2025 年 12 月 30 日)	51
图 36:	ChatGPT 全球月度活跃用户数 (百万人)	51
图 37:	SimpleBench 的大语言模型排行榜 (截至 2025/12/30)	53
图 38:	2023-2025 全球主要 AI 平台的访问量份额	54
图 39:	2025 年通用 chatbot 的网页访问量增速.....	54
图 40:	不到一年时间, Claude code ARR 已达 10 亿美元.....	55
图 41:	openrouter 平台上编程场景中各个模型的 tokens 消耗份额.....	56

图 42: AI overview 的渗透率持续增长, 其中 commercial/transaction/navigational information 渗透率仍有提升空间.....	57
图 43: 谷歌 disco 浏览器界面	58
图 44: Deepseek 系列模型发布时间线	60
图 45: 智谱 coding plan 定价 (2026/1/4)	61
图 46: 国内 AI 应用周度活跃用户数 (万人) (2025/12/8-12/14)	62
图 47: Model Intelligence 分国家对比	64
图 48: 中美模型竞争.....	64
图 49: Artificial Analysis 的大模型 Coding 能力排行榜 (截至 2026/1/4)	65
图 50: LM Arena 的大模型 Web Dev 能力排行榜 (截至 2026/1/8)	65
图 51: LM Arena 的文生视频、图生视频大模型排行榜 (截至 2026/1/8)	66
图 52: 国内视频生成模型的价格远低于海外模型 (截至 2026/1/4)	66
图 53: 中国模型以开源为主, 美国模型以闭源为主 (截至 2026/1/4)	67
图 54: 中国在开源模型上更进一步.....	68
图 55: OpenRouter 平台的 tokens 消耗分布.....	68
图 56: 海外 B 端市场开源模型 tokens 份额逐年下降	69
图 57: 海外 B 端市场开源模型竞争格局.....	70
图 58: NeurIPS 2025 收录论文的机构来源分布.....	73
图 59: 全球 MaaS 的 Token 消耗竞争格局 (截至 2025/10)	75
图 60: 豆包大模型日均 tokens 消耗量快速增长.....	75
图 61: chatgpt 周度活跃用户数 (百万人)	76
图 62: 主要 AI 应用月度活跃用户数 (百万人)	76
图 63: AI 平台的月度平均访问量快速增长 (十亿次)	77
图 64: 从 2024 年 1 月到 2025 年 11 月, OpenRouter 上用户的 Prompt 平均长度增长了 3 倍 ..77	77
图 65: 从 2024 年 1 月到 2025 年 11 月, OpenRouter 上 AI 的回复平均长度增长了 2 倍.....	78
图 66: 编程场景的每条输入输出的平均 tokens 长度远高于其他场景	78
图 67: 全球 toB AI 市场规模快速增长.....	79
图 68: 企业从“自己建”转向“直接买”	80
图 69: 相较于传统 SaaS, toB AI 的“试用→采购”转化率更高.....	80
图 70: 2025 年部门 AI 支出规模	81
图 71: 2025 年垂直行业 AI 支出规模 (百万美元)	82
图 72: OpenEvidence ARR (百万美元)	82
图 73: Glean ARR (百万美元)	82
图 74: Harvey AI ARR (百万美元)	82
图 75: Abridge ARR (百万美元)	82
图 76: AI native 应用大量涌现.....	83
图 77: 2025 年横向 AI (通用助手) 支出规模	83
图 78: 2025 年 toB AI 支出中 PLG 驱动的占比可观 (单位: 十亿美元)	84
图 79: 按照类别划分的全球 AI 网页日均访问量 (百万次/天)	85
图 80: 2025 年按照类别划分的 AI 网页端访问量增速	85
图 81: AI Graveyard 上已经消失的 AI 应用 (截至 2025/12/31)	86
图 82: AI Graveyard 上死亡率最高的产品类型 (截至 2025/12/31)	87
图 83: QuestMobile 统计的 25Q3 国内原生 AI APP 用户增长表现	87
图 84: OpenRouter 平台上 tokens 消耗最多的场景是编程和角色扮演.....	88

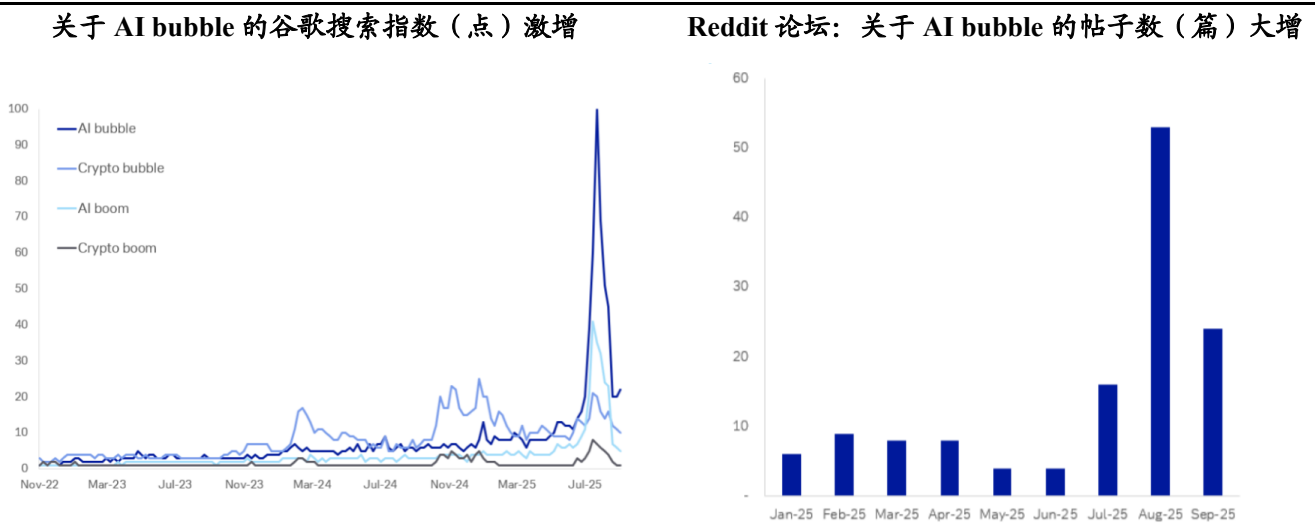
图 85: 新发布的模型大多定向优化了 coding 能力.....	88
图 86: OpenAI 此前的版本分为 high/medium/low, 而后专门推出 codex 版本.....	89
图 87: Cursor ARR (十亿美元)	89
图 88: Lovable ARR (百万美元)	89
图 89: character AI 和 talkie 的 DAU (百万人)	90
图 90: 谷歌的搜索引擎份额 (%) 稳定在 90% 左右	91
图 91: Alphabet 谷歌搜索收入增速企稳回升 (百万美元, %)	92
图 92: Google 推出 UCP 协议, 支持 Gemini app 和 AI Mode 直接结账.....	93
图 93: 百度、阿里、腾讯将 AI 能力嵌入原有的生态产品	93
图 94: 海外大厂广告收入-TTM (十亿美元)	94
表 1: 海外大厂 Capex (亿美元)	9
表 2: 海外大厂 Capex 占收入和占 OCF 的比例	10
表 3: 三大云巨头的折旧年限 (年)	16
表 4: 2025 年六大云厂商的折旧年限.....	16
表 5: OpenAI 相关合作 (部分)	17
表 6: 2025 年海外 AI 公司发债情况总结	19
表 7: AI 产业链各环节的商业模式	22
表 8: TPU v7 对比 GB200/GB300 芯片	25
表 9: 云厂商储备订单余额大增.....	31
表 10: 三类云厂商的商业模式对比.....	36
表 11: 世界模型进展.....	49
表 12: Deepseek 系列模型的关键创新点	60
表 13: AI 科技公司 ARR 及增速	74

1. AI 投资：泡沫与革命并存，局部风险而非系统崩盘

1.1. 泡沫论在 25Q3 集中发酵，较多美股科技公司财报后股价回调

2025 年 7 月以来，关于 AI 泡沫的讨论大幅增长。恐慌情绪导致大多数 AI 相关美股股价回调。华尔街开始意识到，AI 不是印钞机，而是碎钞机——至少对大部分公司而言。市场进行价值重估——从相信一切 AI 概念都会成功，到开始区分哪些公司拥有真实的商业模式。

图1：2025Q3 关于 AI 泡沫的讨论激增



数据来源：Google Trend, Deutsche Bank Research, 东吴证券研究所

Amazon 和 Meta 均在业绩电话会上宣布大幅提升 CapEx，但市场反应却截然相反——亚马逊财报后次日大涨 10% (2025/10/31)，而 Meta 则下跌 11% (2025/10/30)。核心分歧在于资本投入的变现确定性：亚马逊 AWS 不仅为自身 AI 迭代提供支撑，更通过对外服务直接实现了 CapEx 的货币化；加之其充裕的经营现金流与健康的财务结构，市场确信其每一美元投入均有清晰的回报闭环。而 Meta 由于缺乏云业务支撑，其巨额 AI 投资主要用于内部生态优化，对收入的拉动不够清晰。

与此同时，AI infra 正经历剧烈的去泡沫化过程，市场风险偏好降温。CoreWeave、Nebius 以及 Iris Energy 等二线算力公司的股价在财报季均遭遇了双位数下跌，其中 CoreWeave 单日重挫 16% (2025/11/11)，Iris Energy 财报披露后的一周里累计跌幅超 30%。此前由流动性推动的高估值逻辑开始瓦解，市场的关注点已从营收增速切换至现金流健康度与回款能力。相比之下，Nvidia 虽面临短期波动，但其 V 型反转走势仍显示出核心资产在基本面支撑下的强韧性。

整体来看，2025 年 10-11 月财报季的特征是股价对利好消息反应相对有限、对利空消息更为敏感，高估值成长股的隐含波动率风险被充分释放。资金流向明确指向了具备确定性盈利能力的软件龙头与云基础设施巨头。

图2：科技公司财报披露前后的股价变化

	英伟达	苹果	微软	谷歌	亚马逊	特斯拉	META	甲骨文	Palantir	Applovin	Coreweave	Nebius	Iris Energy	多邻国
单日涨跌幅	NVDA.O	AAPL.O	MSFT.O	GOOG.O	AMZN.O	TSLA.O	META.O	ORCL.N	PLTR.O	APP.O	CRWV.O	NBIS.O	IREN.O	DUOL.O
2025/10/17	1%	2%	0%	1%	-1%	2%	1%	-7%	0%	-1%	-3%	-8%	-2%	-2%
2025/10/20	0%	4%	1%	1%	2%	2%	2%	-5%	2%	-6%	-7%	-4%	-2%	-3%
2025/10/21	-1%	0%	0%	-2%	3%	-1%	0%	-1%	0%	-2%	-2%	-4%	-7%	2%
2025/10/22	0%	-2%	1%	0%	-2%	-1%	0%	-1%	-3%	2%	-3%	-5%	-6%	-3%
2025/10/23	1%	0%	0%	1%	1%	2%	0%	3%	3%	4%	1%	8%	8%	2%
2025/10/24	2%	1%	1%	3%	1%	-3%	0%	1%	2%	5%	7%	10%	13%	0%
2025/10/27	3%	2%	2%	4%	1%	4%	2%	-1%	2%	4%	3%	7%	3%	-2%
2025/10/28	5%	0%	2%	-1%	1%	2%	0%	0%	0%	-3%	-1%	-3%	-4%	-4%
2025/10/29	3%	0%	0%	3%	0%	0%	0%	-2%	5%	1%	4%	3%	-3%	-7%
2025/10/30	-2%	1%	-3%	2%	-3%	-5%	-11%	-7%	-2%	-2%	-6%	-1%	-4%	-4%
2025/10/31	0%	0%	-2%	0%	10%	4%	-3%	2%	3%	3%	2%	5%	5%	3%
2025/11/3	2%	0%	0%	1%	4%	3%	-2%	-2%	3%	-1%	-5%	-8%	12%	-2%
2025/11/4	-4%	0%	-1%	-2%	-2%	-5%	-2%	-4%	-8%	-4%	-8%	-8%	-2%	-1%
2025/11/5	-2%	0%	-1%	2%	0%	4%	1%	1%	-1%	1%	-1%	6%	15%	0%
2025/11/6	-4%	0%	-2%	0%	-3%	-3%	-3%	-3%	-7%	1%	-7%	-6%	-12%	-25%
2025/11/7	0%	0%	0%	-2%	1%	-4%	0%	-2%	2%	0%	-3%	2%	-7%	4%
2025/11/10	6%	0%	2%	4%	2%	4%	2%	1%	9%	5%	2%	-1%	-4%	-4%
2025/11/11	-3%	2%	1%	0%	0%	-1%	-1%	-2%	-1%	-9%	-16%	-7%	-5%	1%
2025/11/12	0%	-1%	0%	-2%	-2%	-2%	-3%	-4%	-4%	-2%	-3%	-8%	-3%	-4%
2025/11/13	-4%	0%	-2%	-3%	-3%	-7%	0%	-4%	-7%	-5%	-8%	-6%	-13%	0%
2025/11/14	2%	0%	1%	-1%	-1%	1%	0%	2%	1%	0%	-1%	-6%	-5%	-1%
2025/11/17	-2%	-2%	-1%	3%	-1%	1%	-1%	-1%	-2%	-3%	-3%	3%	2%	-4%
2025/11/18	-3%	0%	-3%	0%	-4%	-2%	-1%	0%	-2%	-2%	-1%	5%	3%	-2%
2025/11/19	3%	0%	-1%	3%	0%	1%	-1%	2%	-1%	1%	0%	5%	-6%	-1%
2025/11/20	-3%	-1%	-2%	-1%	-2%	-2%	0%	-7%	-6%	-2%	-8%	-11%	-5%	-4%
2025/11/21	-1%	2%	-1%	3%	2%	-1%	1%	-6%	-1%	0%	4%	-2%	-3%	3%
2025/11/25	-1%	2%	1%	8%	4%	7%	7%	-1%	6%	7%	-1%	7%	12%	2%
2025/11/26	1%	0%	2%	-1%	0%	2%	0%	4%	1%	5%	4%	7%	2%	7%
2025/11/28	-2%	0%	1%	0%	2%	1%	2%	-1%	2%	2%	-2%	0%	-1%	2%
2025/12/1	2%	2%	-1%	-2%	0%	0%	-1%	0%	-1%	4%	5%	6%	1%	-3%
2025/12/2	1%	1%	1%	0%	0%	0%	1%	0%	2%	5%	-1%	-4%	-15%	-2%
2025/12/3	-1%	-1%	-3%	1%	-1%	4%	-1%	3%	3%	1%	4%	3%	7%	1%
2025/12/4	2%	-1%	1%	-1%	-1%	2%	3%	3%	1%	3%	8%	4%	22%	2%
2025/12/5	-1%	-1%	0%	1%	0%	0%	2%	2%	2%	1%	3%	-5%	-4%	6%
2025/12/8	2%	0%	2%	-2%	-1%	-3%	-1%	1%	0%	0%	-2%	2%	4%	4%
2025/12/9	0%	0%	0%	1%	0%	1%	-1%	0%	0%	5%	5%	-4%	1%	1%
2025/12/10	-1%	1%	-3%	1%	2%	1%	-1%	1%	3%	-3%	-3%	-3%	-6%	-6%
2025/12/11	-2%	0%	1%	-2%	-1%	-1%	0%	-11%	0%	2%	-1%	1%	0%	2%
2025/12/12	-3%	0%	-1%	-1%	-2%	3%	-1%	-4%	-2%	-6%	-10%	-7%	-9%	-2%

数据来源：wind，东吴证券研究所

注：蓝色底表示财报发布日

我们认为，AI 行情里存在泡沫成分，但 AI 产业本身不是泡沫。当下更像一轮由技术突破驱动的产业扩张期，市场在用偏情绪化的方式给远期现金流定价，于是阶段性会把一部分标的推到偏热的位置；但它并不等同于 90 年代末互联网泡沫。

1.2. 市场在担心什么：Capex、折旧、ROI、估值、债务

AI 泡沫论的支持者并非空口无凭。我们梳理出投资者最关心的五大问题，并逐一拆解其合理性与风险点。这些担忧可以归纳为三个层次：投入端（CapEx）、核算端（折旧）和融资端（债务与循环订单）。

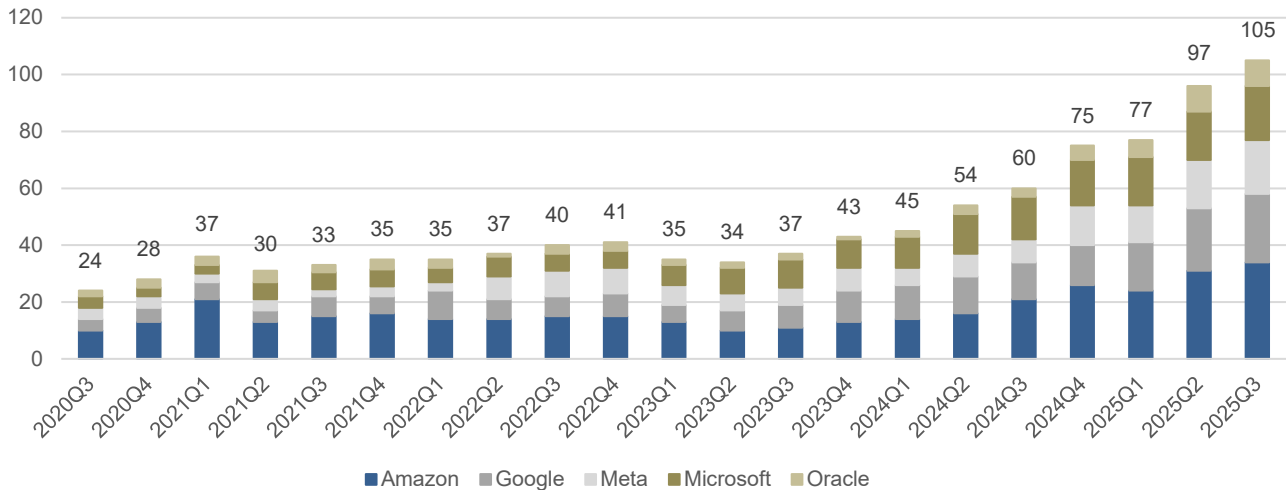
1.2.1. 焦虑一：大幅增长的 CapEx

自 2020 年以来，海外主要云厂商的 CapEx 大致经历了三个阶段：2020-2021 年的高额投入、2022-2023 年的放缓与消化，以及 2024 年至今因 GenAI 需求而加速增

长。2024-2025 年的上升斜率明显陡峭，主要驱动因素是 AI 训练和推理所需的加速服务器（GPU 或自研加速器）、数据中心上电以及配套网络基础设施建设。

在 2023-2024 年市场情绪乐观时，业界普遍认为 CapEx 越多越好，下调投入反而可能意味着掉队。然而，当市场情绪转为冷静，开始担忧巨额资本支出将带来更大的折旧压力，质疑折旧能否转化为相应收入，以及现金流是否足以支撑。如果现金流不足，是否需要进一步融资已成为市场关注的焦点。

图3: 海外大厂 Capex (十亿美元)



数据来源: Artificial Analysis, 东吴证券研究所

表1: 海外大厂 Capex (亿美元)

	2024A	2025E	2026E	24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3	25Q4E
合计	2504	4042		477	583	649	795	773	958	1133	1177
yoy		61%						62%	64%	75%	48%
Amazon	830	1250	继续增长	149	176	226	278	250	322	351	327
yoy		51%						68%	83%	55%	17%
Microsoft	756	1162	持续增长	140	190	200	226	214	242	349	357
yoy		54%						53%	27%	75%	58%
Alphabet	525	920	大幅增长	120	132	131	143	172	224	240	284
yoy		75%						43%	70%	83%	99%
Meta	392	710	显著增长	67	85	92	148	137	170	194	209
yoy		81%						104%	101%	111%	41%

数据来源: 公司公告, 东吴证券研究所

注: 微软财季分别为 FY24Q3-FY26Q2E

注 2: 2025 年预计值按照公司指引区间的中值计算。25Q4E 按照 25 年指引中值减掉 25Q1-Q3 累计值计算

注 3: 2026 年预测来自各公司业绩电话会表述

在 2025Q3 业绩电话会上，海外云巨头纷纷上调全年 CapEx 指引，标志着算力军备竞赛已进入白热化阶段。谷歌、微软和 Meta 的资本开支预期均大幅增长，其中微软明确指出 GPU 供需平衡点将延至 2026 年年中，且未来两年数据中心面积将翻倍。

Meta 管理层预计 2025 年 CapEx 为 710 亿美元,2026 年可能突破 1000 亿美元。

其中约 60% 投向 GPU/CPU 等短期资产，这些资产折旧周期较短，意味着未来几年折旧压力将显著加大。Meta 管理层的判断是：算力供不应求，无论是核心业务还是新项目都能消化更多算力并带来利润。然而，这一乐观预期与其营业利润率同比下降的现实相矛盾。若 Meta 对 ASI（超级智能）将在短期内实现的豪赌未能兑现，将面临巨额算力资产闲置和折旧侵蚀利润的双重风险。

微软的策略更为务实。CapEx 背后有明确的客户需求支撑，FY26Q1 Azure 储备订单环比激增至 1550 亿美元，这些订单有望在可预见的未来转化为收入。微软强调：资产负债期限匹配，约一半支出用于短期资产以匹配合同期限，剩余部分投向长期资产以支持未来货币化。这种模式更像是基于现有订单的产能扩张，而非 Meta 式的对未来需求爆发的盲目押注。

谷歌的 CapEx 相对克制但增长确定，公司预计其 2025 年全年资本开支约 910-930 亿美元，增速相对温和。投入效率持续改善，云业务营业利润率稳步提升，这反映了全栈整合（TPU+Gemini+ 解决方案）的成本优势正在显现。谷歌的策略是通过技术创新抵消资本压力，通过软件层面的持续优化，以更少的硬件投入支撑更大的业务规模。

苹果是唯一未明确披露 AI 专项资本开支的巨头，这体现了其设备端优先策略的自然延伸，以及对 AI 长期价值的谨慎态度——在找到清晰货币化路径前，不会进行豪赌式投入。然而，若 AI 竞争最终聚焦于模型能力，苹果在云端大规模训练经验和模型上的差距可能使其处于劣势。

表2：海外大厂 Capex 占收入和占 OCF 的比例

		24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3
Capex (百万美元)	Amazon	14925	17620	22620	27834	25019	32183	35095
	Microsoft	14000	19000	20000	22600	21400	24200	34900
	Alphabet	12012	13186	13061	14276	17197	22446	23953
	Meta	6715	8472	9202	14836	13692	17012	19374
	Oracle	1674	2798	2303	3970	5862	9080	8502
capex/收入	Amazon	10%	12%	14%	15%	16%	19%	19%
	Microsoft	23%	29%	30%	32%	31%	32%	45%
	Alphabet	15%	16%	15%	15%	19%	23%	23%
	Meta	18%	22%	23%	31%	32%	36%	38%
	Oracle	13%	20%	17%	28%	41%	57%	57%
capex/经营性现金流	Amazon	79%	70%	87%	61%	147%	99%	99%
	Microsoft	44%	51%	59%	101%	58%	57%	77%
	Alphabet	42%	49%	43%	36%	48%	81%	49%
	Meta	35%	44%	37%	53%	57%	67%	65%
	Oracle	31%	46%	31%	304%	99%	147%	104%

数据来源：公司公告，东吴证券研究所

注：微软财季分别为 FY24Q3-FY26Q1

1.2.2. CapEx 的增长具有明确的、可持续的支撑

CapEx 增长的驱动力主要源于训练端和推理端的算力消耗。我们预计两部分的算力消耗都会持续增长，但训练端的消耗更像是“上楼梯”，可能会在一段时间里停滞不前（因为 scaling law 放缓等），但又会突然大幅提升（当出现新的算法、架构和 infra 配

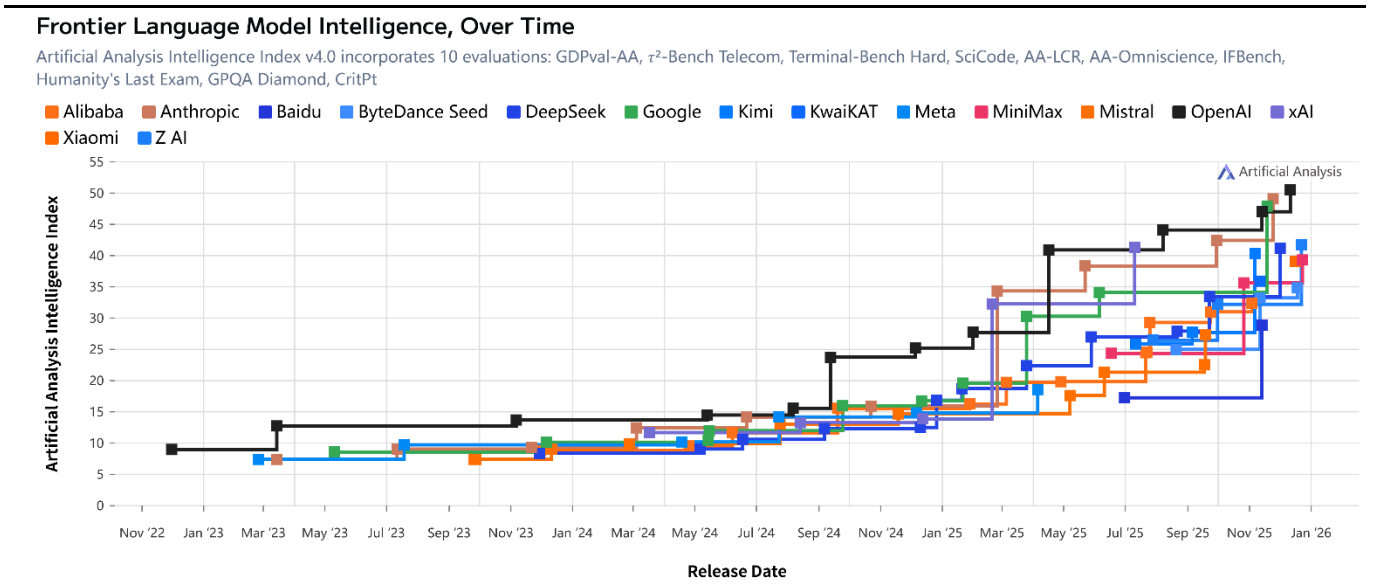
套时); 而推理端的消耗增长更像一条光滑的曲线, 随着更多的、更深的使用场景出现而增长。

在训练端:

2025 年, 模型竞争仍处于白热化状态, 几乎每个季度都会涌现新的 SOTA 模型(state of the art), 相较于 2023-2024 年更新迭代的频率明显加快。尽管 OpenAI 在 2023-2024 年处于领跑位置, 但 2025 年以来 Google 和 Anthropic 在快速追赶, 相对排名频繁变化。技术代差已显著缩短, 各家模型在智能指数上的分布日益密集。

对于 Meta 和 xAI 等追赶者而言, 需要依靠大规模算力投入来快速跟上一线实验室的步伐。而对于 OpenAI 和谷歌等头部玩家, 除了在现有技术路径上持续优化外, 还必须探索新的技术范式以保持领先优势, 例如在 Transformer 架构之后的下一代架构。

图4: 各家头部厂商在榜单上的排名频繁更迭



数据来源: Artificial Analysis, 东吴证券研究所

注: 这张图的纵轴代表 Artificial Analysis Intelligence Index (人工智能分析智能指数), 是一个量化模型综合能力的复合指标。该指数并非单一测试结果, 而是汇总了 10 项评测集 (包括 GDPval-AA, GPQA Diamond, SciCode, Humanity's Last Exam 等), 重点考核模型在编程、复杂逻辑推理、科学计算以及专业领域知识上的表现。

在推理端:

根据 Artificial Analysis, 具备深度思考能力的模型 (如 Gemini 2.5 Pro 或 DeepSeek R1) 为了完成同一任务, 消耗的 Token 量平均是普通模型的 5.5 倍。OpenRouter 的数据佐证了这一点, 推理模型的流量占比从 2025 年初的 0% 一路飙升至 25 年末的 50% 以上。这种从“简单问答”向“复杂推理”的范式转移, 导致单位任务的算力消耗成倍增加。

图5: 推理模型大量的 tokens 用于“打草稿”过程

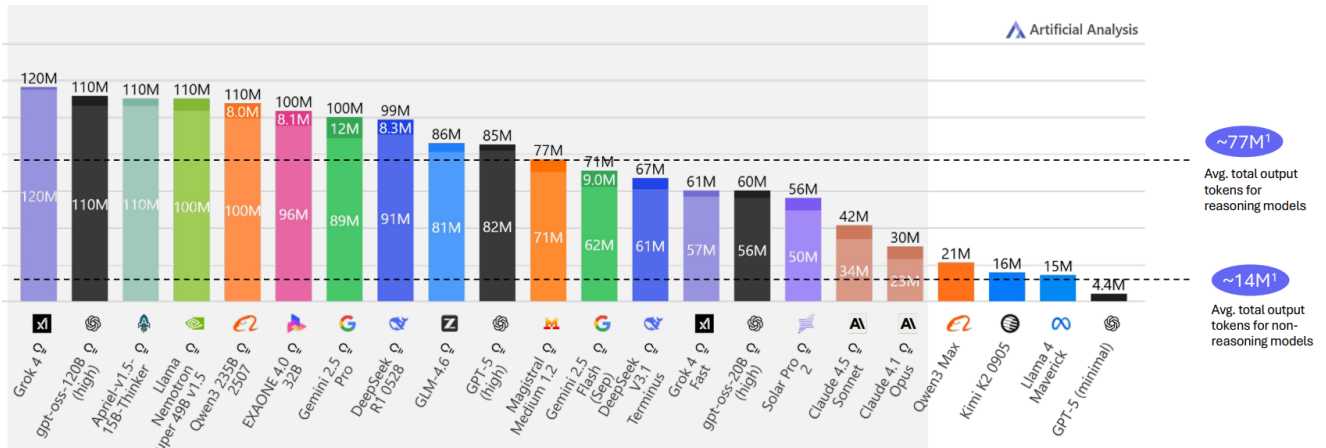
Output Tokens Used to Run the Artificial Analysis Intelligence Index

Artificial Analysis Intelligence Index v3 (incorporates 10 leading evaluations), Output Tokens Used in Artificial Analysis Intelligence Index (~5M input tokens)

Reasoning models

NON-EXHAUSTIVE

Answer Tokens Reasoning Tokens

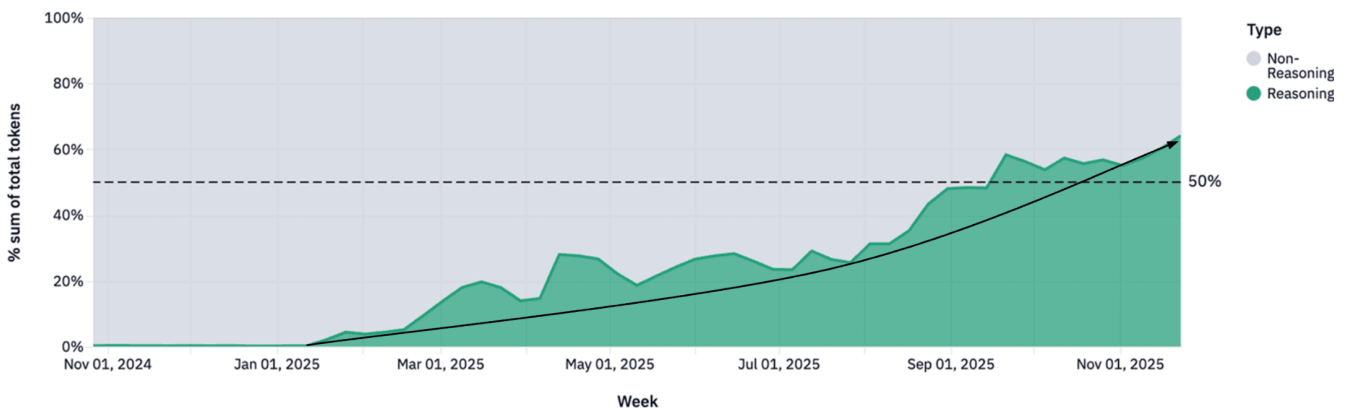


1. Based on representative models included in the chart.

数据来源: Artificial Analysis, 东吴证券研究所

注: 这张图表主要量化了不同 AI 模型在执行同一套评测标准 (Artificial Analysis Intelligence Index v3) 时所消耗的“输出 Token”总量, 展示了“推理模型”与“普通模型”在算力消耗上的差异。左侧灰色背景区域聚集了 Grok 4、Gemini 2.5 Pro 等推理模型, 它们的平均算力消耗为 77M Tokens, 其中浅色部分代表模型内部进行逻辑推演的“推理 Token” (即思维链), 深色部分才是最终呈现的“回答 Token”; 相比之下, 右侧白色区域的普通模型平均算力消耗为 14M Tokens, 表明推理模型为了换取更强的逻辑解题能力, 在生成过程中付出了普通模型数倍的算力消耗。

图6: OpenRouter 平台上推理模型的 tokens 消耗占比持续增长

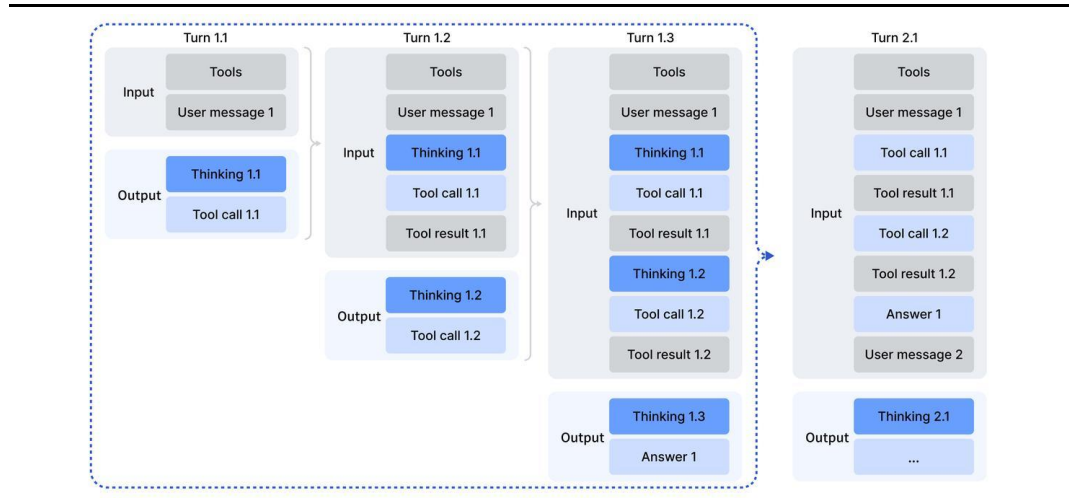


数据来源: OpenRouter, 东吴证券研究所

思维链正在变得越来越复杂。deepseek v3.2 和 minimax m2 的“交错思维链”, 进一步推高了单次回复的 tokens 消耗。模型收到问题后, 先“思考” (Thinking 1.1), 发现需要外部数据, 于是发起“工具调用” (Tool call); 拿到工具反馈的结果后, 它没有立刻输出, 而是再次“思考” (Thinking 1.2) 来消化这些信息, 甚至觉得不够还要继续调用工

具。直到经过了三次反复的“思考-行动-观察”循环，它才在 Turn 1.3 给出了最终答案。

图7: deepseek v3.2 的交错思维链



数据来源: deepseek 官网, 东吴证券研究所

图8: minimax m2 的交错思维链

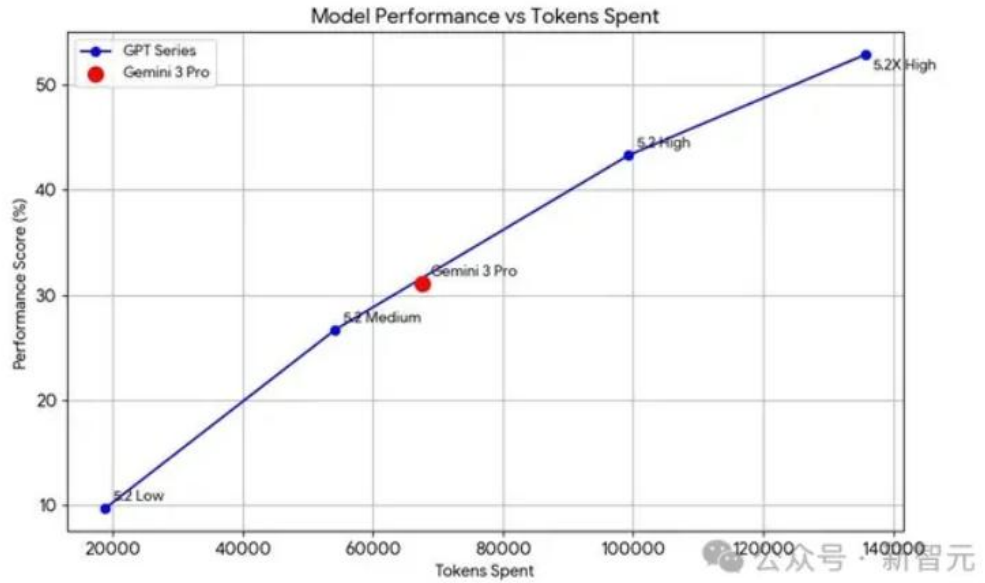
```

MiniMax-M2 Interleaved Thinking Demo
Agent-native • 10B activated params • 8% of Claude Sonnet price • 2x faster loops • Front-end
Pricing: input $0.3/MTok | output $1.2/MTok | starter plan $10/mo (10% of Claude Code Max).
Brief focus: interleaved thinking, tool orchestration, Mini price / max performance.
Scenario: context_package | Run ID: 20251118-053252
=====
--- Step 1 ---
Thought:
I need to provide a crisp brief for the design system team that covers: 1. Critical
design tokens 2. Button implementation contract 3. One reusable development pattern The
user wants me to use the provided tools after every thought instead of guessing. I
should look up: - Design tokens (colors, typography, spacing, etc.) - Button component
spec - A development pattern Let me start by getting the design tokens information,
then the Button component spec, and finally a development pattern.
Tool Call > get_design_tokens {"category": "colors"}
Tool Result *
get_design_tokens: {"section": "brand colors", "tokens": "### Primary Colors\n-
**Primary**: #3B82F6 (Blue 600)\n- **Primary Light**: #60A5FA (Blue 400)\n- **Primary
Dark**: #2563EB (Blue 700)\n\n### Secondary Colors\n- **Secondary**: #10B ...
--- Step 2 ---
Thought:
Good, I have the color tokens. Now let me get typography, spacing, and shadow tokens to
round out the critical design tokens section.
Tool Call > get_design_tokens {"category": "typography"}
Tool Result *
get_design_tokens: {"section": "typography", "tokens": "### Font Families\n- **Sans**:
Inter, system-ui, -apple-system, sans-serif\n- **Serif**: Georgia, serif\n- **Mono**:
'Fira Code', 'Courier New', monospace\n\n### Font Sizes\n- **xs**: ...
--- Step 3 ---
Thought:
Good, I have the core color and typography tokens. Now I need to get spacing tokens to
complete the critical design tokens, then move on to the Button component spec and a key
development pattern.
Tool Call > get_design_tokens {"category": "spacing"}
Tool Result *
    
```

数据来源: 机器之心 Pro, 东吴证券研究所

之所以要花费更多 tokens 在“思考”环节，核心是因为推理算力定律（Test-time Compute Scaling）：智能表现与思考时长（Token 消耗）成正比。模型消耗的 Token 越多（即让它进行更深度的思考），它回复的准确率就越高。

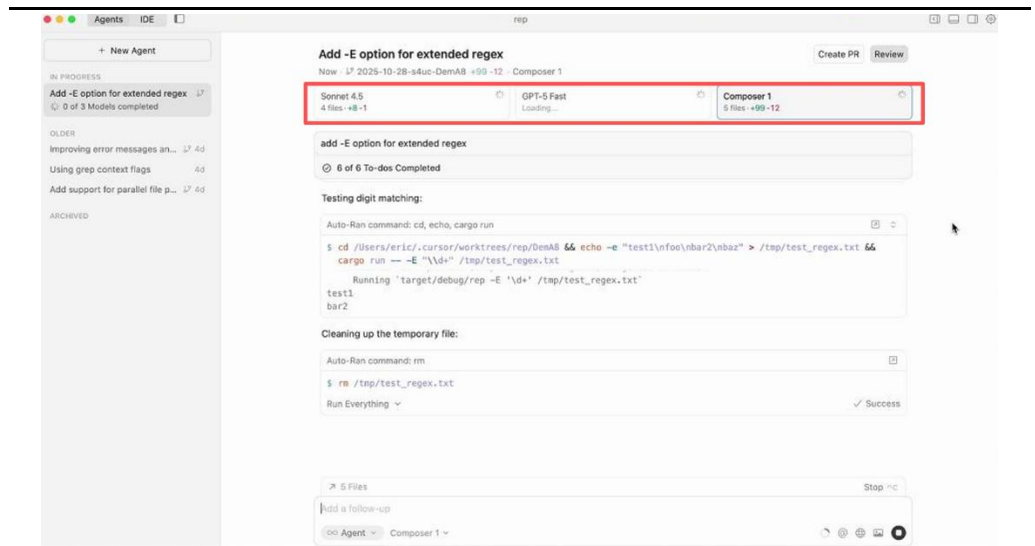
图9: GPT-5.2 通过消耗更多 tokens 来提升模型智能



数据来源: 新智元, 东吴证券研究所

第三, Agent 模式。以往 Token 的消耗主体主要是人类用户,但现在 Agent 正在成为 Token 消耗的主力。比如 Cursor 推出的多 Agent 模式,可以同时调用 Sonnet、GPT-5、Composer 1 三个模型处理同一任务,再由用户从中择优录用。这种多模型并行调用的模式,使得单个任务的 Token 消耗量成倍增长。随着 Agent 应用的普及,这种消耗模式将成为常态,进一步推高对算力基础设施的需求。

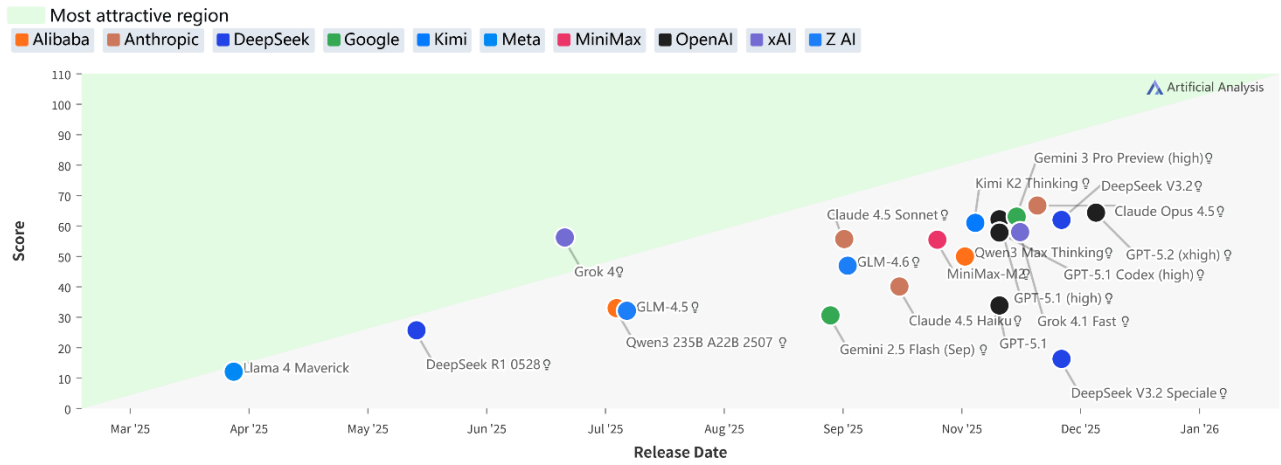
图10: cursor 的多 agent 模式



数据来源: cursor 截图, 东吴证券研究所

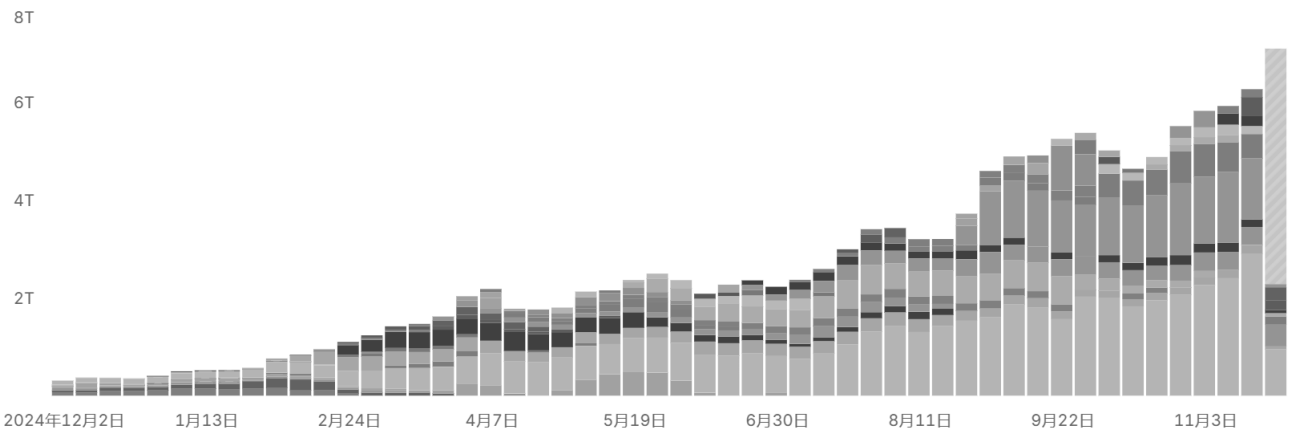
图11: 新发布的模型大多优化了 agent 能力

Agentic Index vs. Release Date



数据来源: Artificial Analysis, 东吴证券研究所

图12: OpenRouter 上的 tokens 周度消耗量



数据来源: OpenRouter 官网, 东吴证券研究所

1.2.3. 焦虑二: 延长折旧周期是合理的吗

巨额 CapEx 引发的第一个连锁反应是利润表压力。当云巨头每年投入数百亿美元购买 GPU 时, 这些资产如何在财务报表上体现、何时转化为费用, 直接决定了未来几年的盈利能力。这就引出了市场的第二个焦虑: 会计折旧是否在掩盖真实的资产风险?

近年来, 谷歌、微软、Meta 等巨头纷纷修改会计准则, 将服务器和网络设备的折旧年限从原本的 3-4 年延长至 5-6 年。

市场担心，在芯片快速迭代的背景下，6年折旧期或许过于乐观；云厂商通过减少每季度的折旧费用，高估了 EPS 和利润率。

表3: 三大云巨头的折旧年限（年）

云服务商	2017	2018	2019	2020	2021	2022	2023	2024	2025
Amazon (AWS)	3	3	3	4	4	5	5	6	6
Alphabet (Google Cloud)	4	4	4	4	4	4	6	6	6
Microsoft (Azure)	4	4	4	4	4	4	6	6	6

数据来源：cube research，东吴证券研究所

然而 Cube Research 提出了一个有趣的观点：AI 基础设施的生命周期逻辑与传统通用服务器完全不同，顶级 GPU 不再用于训练最前沿模型时，并不会变成电子垃圾。

AI 基础设施的生命周期可以分为三个阶段：第一阶段 0-2 年用于顶级模型训练；第二阶段 3-4 年用于次级训练或微调；第三阶段 5-6 年以上用于推理。推理对显存带宽和精度的要求低于训练。随着 AI 应用爆发，推理需求是训练的数倍。现在的 H100 在 5 年后依然是极其强悍的推理芯片。

此外，AI 超算集群不仅包含 GPU，还包含昂贵的光通信、液冷系统、机架和电源管理。即使 GPU 本身需要升级，周边的配套设施占成本相当大比例，寿命远超 3 年。这种系统级的架构设计使得基础设施具有韧性，可以适应升级，而不是整体报废。

这个观点对 AI 泡沫论提供了两个重要回应：

第一，云巨头的利润率可能比预想的更具韧性。如果市场接受了 5-6 年折旧的合理性，那么云巨头看似激进的资本开支对利润表的冲击就被平滑了。这意味着，只要推理需求跟得上，云厂商可以在未来几年维持较高的毛利率，而不会因为购买了数千亿美元的 GPU 而导致利润崩塌。

第二，英伟达的护城河在变宽。因为 CUDA 生态的兼容性，旧款英伟达卡可以无缝转入推理环节。这实际上增加了客户购买英伟达显卡的 ROI，因为客户知道这张卡即使过气了，也能在二三线业务中发光发热，而不是只能当废铁卖。

表4: 2025 年六大云厂商的折旧年限

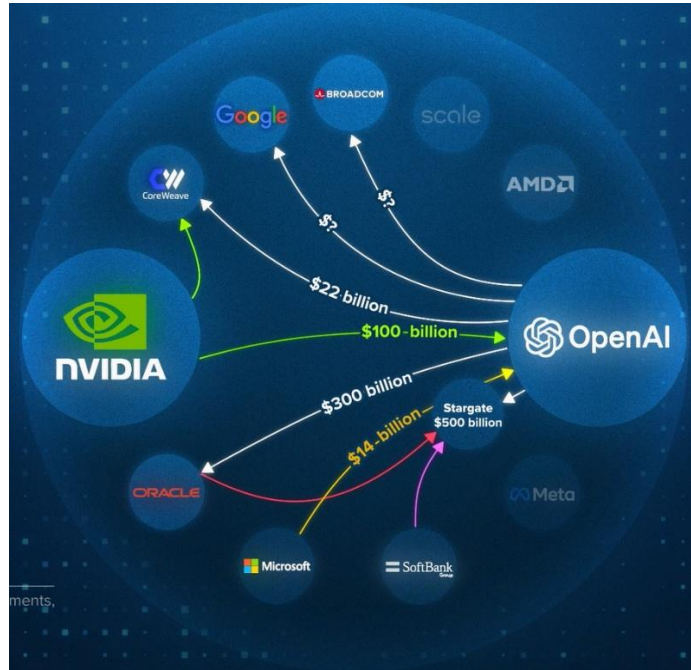
公司	年限
Amazon (AWS)	6 年
Alphabet (Google)	6 年
Microsoft (Azure)	6 年
CoreWeave	6 年
Nebius	4 年
Lambda Labs	5 年

数据来源：cube research，东吴证券研究所

1.2.4. 焦虑三：循环订单和债务是最核心的问题

2025 年 10 月，OpenAI 承诺在 2025 年至 2035 年间投入 1.15 万亿美元用于硬件和云基础设施。这笔支出分布在七家主要供应商：Broadcom 3500 亿美元、Oracle 3000 亿美元、Microsoft 2500 亿美元、Nvidia 1000 亿美元、AMD 900 亿美元、Amazon AWS 380 亿美元和 CoreWeave 220 亿美元。

图13: OpenAI 相关的合作 (截至 2025 年 11 月)



数据来源: CNBC, 东吴证券研究所

表5: OpenAI 相关合作 (部分)

供应商	合作时间	详情
CoreWeave	2025 年 3 月	OpenAI 投资 3.5 亿美元购入 CoreWeave 股份，并承诺在未来五年内投入 119 亿美元以获取 AI 算力资源。
NVIDIA	2025 年 9 月	宣布了一项价值 1000 亿美元的合作，计划在数据中心部署至少 10 吉瓦的英伟达系统，同时英伟达也将参与 OpenAI 的融资轮次。根据该协议，英伟达将为 OpenAI 提供资金，而 OpenAI 则将采购英伟达的芯片及其他设备。
Oracle	2025 年 9 月	OpenAI 承诺在未来五年内购买价值约 3000 亿美元的计算能力，平均每年约 600 亿美元。该协议将于 2027 年正式启动。
AMD	2025 年 10 月	两家公司签署了一项价值数十亿美元的协议，涉及 6 吉瓦的 AMD GPU。OpenAI 将为 AMD 提供资金，AMD 则向 OpenAI 出售芯片。
博通	2025 年 10 月	计划部署总容量达 10 吉瓦的定制芯片集群。设计已在 2025 年完成，将于 2026 年下半年开始量产并投入使用，2029 年前完成大规模部署。
亚马逊	2025 年 11 月	双方签署战略协议，OpenAI 使用 AWS 的基础设施（包括 NVIDIA GPU 和自研芯片）来扩展其推理与训练负载。 总金额 380 亿美元，未来七年将持续扩大。 所有算力将在 2026 年底前全面部署，并具备在 2027 年及以后进一步扩展的能力。

数据来源: OpenAI, 英为财经, CNBC, 东吴证券研究所

这个循环订单可以拆解为四个环节：第一，英伟达向云巨头出售 GPU，获得真实营收。第二，云厂商宣布千亿级别的 AI 数据中心建设计划。第三，云厂商将算力以订单

的形式租给 AI 公司。第四，AI 公司如 OpenAI 反过来向云厂商购买更多算力，这笔支出成为云厂商财报中 AI 驱动的云收入增长。

这是一场击鼓传花的游戏，但鼓声来自未来的收入，而不是现在的现金流。只要市场相信 AI 会成功、相信 OpenAI 最终会盈利，这个游戏就能继续玩下去——英伟达继续卖 GPU、云厂商继续扩建数据中心、OpenAI 继续融资购买算力。一旦任何一个环节出现问题，比如 OpenAI 融资链断裂、企业客户发现 AI 投资回报不及预期开始削减算力采购，整个循环可能逆转并引发连锁反应。

这与 2000 年互联网泡沫时期的点击换流量、流量换广告、广告换估值的循环如出一辙，只不过这次的主角从门户网站换成了 AI 公司，货币从广告预算换成了算力采购。

债务让循环订单变得更加危险。在 2023 至 2024 年间，AI 产业主要依靠股权融资，这种模式相对健康且风险可控。但进入 2025 年，AI 公司开始大规模发债，2025 年 1-11 月美国公司的发债规模已超过 2000 亿美元，用于资助 AI 相关的基础设施项目。

不同类型公司面临的融资成本出现了显著分化。对于科技巨头而言，融资成本呈现温和上升态势。而新兴 AI 公司的融资成本在飙升。例如 xAI 2025 年 7 月发行的债券的收益率高达 12.5%。CoreWeave 的融资利率更是达到了 11-14% 的水平。这些高利率反映出市场对新兴 AI 公司盈利能力的怀疑。

越来越多的 AI 基础设施项目开始通过非公开渠道融资，转向黑石、Apollo 等私募基金。这些非公开市场的融资利率往往不透明，而且附带更苛刻的条款，包括资产抵押、控制权让渡等限制性条款。这种向影子融资体系的迁移，是一个危险的信号。

在融资压力之下，一些公司开始探索创新的融资模式。Meta 与 Blue Owl Capital 的合作提供了一个有趣的案例：Meta 只需出资 20% 就能与合作伙伴共同投资数据中心，其余 80% 资金来自外部投资者或国家资本。这种表外融资方式既降低了自身的资产负债压力，又能继续扩大 AI 基础设施投入。如果这种模式被科技巨头广泛采用，可能会从根本上改变整个行业的融资结构，但同时也可能将风险扩散到更广泛的金融体系中。

表6: 2025年海外 AI 公司发债情况总结

公司	发债情况
xAI	<ul style="list-style-type: none"> ◆ 2025年7月, xAI 完成了 50 亿美元的债务融资, 同时获得 50 亿美元的战略股权投资, 总融资额达 100 亿美元。债务部分包括 30 亿美元债券 (收益率 12.5%)、10 亿美元固定利率定期贷款 (利率 12.5%) 以及 10 亿美元 B 类定期贷款 (定价为基准利率上浮 7.25 个百分点, 发行价 96 美分)。 ◆ 2026年1月, xAI 宣布完成 200 亿美元融资, 该轮融资采用了创新结构, 包含约 75 亿美元股权融资和 125 亿美元的债务融资。债务部分通过特殊目的公司 (SPV) 筹集, 专门用于采购英伟达高性能 GPU, xAI 将把这些处理器部署于其数据中心, 并向投资者提供为期五年的租赁回报机制。
甲骨文	<ul style="list-style-type: none"> ◆ 2025年9月, 甲骨文发行 180 亿美元债券用于数据中心扩建(旨在抢占 AI 投资热潮)。此后市场对其评级可持续性的疑虑持续发酵。债务攀升带来的风险促使部分投资者买入与甲骨文相关的信用违约互换 (CDS), 至 2025 年 12 月, 部分衍生品价格已被推高至全球金融危机以来的最高水平。
Alphabet	<ul style="list-style-type: none"> ◆ 2025年4月, 首次在欧元债券市场发行了约 67.5 亿欧元债券, 并在美国发行了 50 亿美元债券, 这是该公司自 2020 年以来首次发债。 ◆ 2025年11月, 在美国市场发行约 175 亿美元债券, 并在欧洲发行 67.5 亿欧元债券。美元债券分为 8 个部分, 期限从 3 年到 50 年不等, 其中最长期限部分的收益率比美国国债高出约 1.35 个百分点。欧元债券分为 6 期, 期限从 3 年至 39 年不等, 3 年期债券利差为基准利率上方 25 个基点, 39 年期债券利差为 158 个基点。
Meta	<ul style="list-style-type: none"> ◆ 2025年10月, Meta 发行了约 300 亿美元债券, 分为六期, 期限从 5 年至 40 年不等, 最长期限债券的发行利差约较同期美国国债高出 1.4 个百分点。 ◆ 除了直接发债, Meta 还采用了 273 亿美元的“表外融资”方式为路易斯安那州 Hyperion 数据中心项目筹资。Meta 成立了特殊目的实体 (SPV) Beignet Investor LLC, 与 Pimco (购买约 180 亿美元)、贝莱德等私人信贷机构合作发行债券, 债务不计入 Meta 的资产负债表。这笔债券将于 2049 年到期, 若 Meta 在 2049 年前决定放弃该园区, 公司承诺向投资者支付残值保障款项。Meta 已获得美国 SEC 复函批准这种会计处理方法, 并与穆迪和标准普尔沟通确保不影响公司投资评级。

数据来源: 智通财经, 澎湃新闻, 第一财经, 证券时报网, 东吴证券研究所

上述三大焦虑指向同一个核心矛盾: **AI 基础设施的投资周期 (3-5 年) 与商业化验证周期 (可能更长) 之间的错配**。市场既担心投入无法变现, 又害怕不投入就会掉队。这种矛盾在不同公司、不同产业链环节的表现形式完全不同。理解这种差异, 是判断“泡沫在哪里”的关键。接下来我们将论证: **AI 领域存在的是结构性分化, 而非全面崩盘**。

1.3. 我们的判断: 局部泡沫, 非系统性崩盘

我们认为 AI 领域存在的是局部泡沫, 而非系统性的全面崩盘。AI 产业链条长、参与者众多, 泡沫的表现并非均匀分布, 而是呈现出明显的阵营分化和环节差异。一方面, 以 OpenAI 为核心、与微软深度绑定的生态链条存在较为明显的泡沫迹象, 而以 Anthropic、谷歌和亚马逊为代表的另一阵营则表现出更稳健的特征; 另一方面, 从产业链垂直分布来看, 上游硬件环节基本不存在泡沫, 云服务层面的云巨头 (Hyperscaler) 相对稳健而新云公司 (Neocloud) 风险较高, 模型层面虽普遍亏损但主要源于前期研发投入, 后期利润弹性巨大, 应用层面则高度分化, 需要精选赛道和个股。

1.3.1. 横向划分: OpenAI 阵营 vs 谷歌阵营

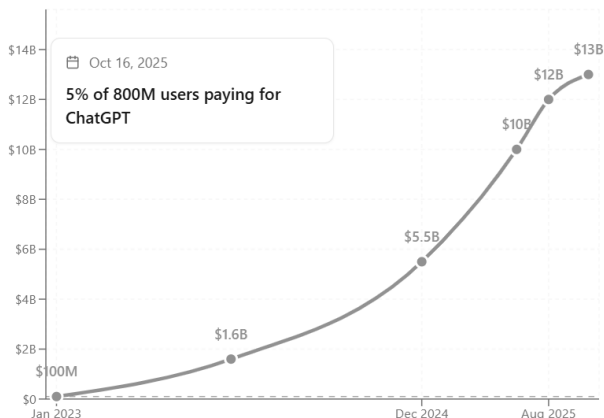
第一个阵营以 OpenAI 为核心。 OpenAI 在年收入只有 100 多亿美元的情况下，向上游公司签订了总额超过 1 万亿美元的长期订单，涉及 AMD、Oracle、英伟达、亚马逊、CoreWeave 等多家公司，合同期限长达 3-5 年。这些上游公司已经建好数据中心、采购了芯片，就等着 OpenAI 付款。

OpenAI 的指引是 2028 年实现 1000 亿美元收入，2030 年实现 2000 亿美元收入。如果无法达成这个目标，整个阵营的所有公司都会受损。这是一个“大而不能倒”（too big to fail）的策略——通过绑定大量上游公司，让自己变得系统性重要，一旦出问题就必须由政府救助。这种策略带有很强的冒险性质。

与此同时，OpenAI 的战略缺乏聚焦，从最初的 to C 订阅转向同时布局 to B、企业级嵌入、硬件（眼镜、机器人）、浏览器、社交媒体（Sora）等多个方向，分散了资源且鲜有实质性成果，导致现金流压力巨大。这种步子迈得过大、过于依赖外部融资和债务的模式，正是泡沫最集中的区域。微软作为其最大投资者和云服务独家供应商，也因此被深度捆绑，其 Azure AI 业务的增长在一定程度上依赖 OpenAI 的表现，存在被拖累的风险。

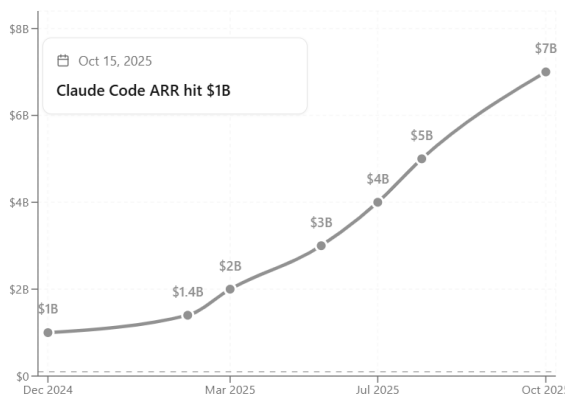
第二个阵营以 Anthropic 和谷歌为代表，策略更加稳健和聚焦。 Anthropic 是从 OpenAI 分离出来的创业公司，创始团队不认可 OpenAI 的理念而另立门户。从财务角度看，虽然 Anthropic 的收入规模（2025 年 10 月 ARR 70 亿美元）低于 OpenAI（2025 年 10 月 ARR 130 亿美元），但现金流表现更健康。根据 Fortune 报道，OpenAI 管理层预计其 2030 年自由现金流才能转正，而 Anthropic 管理层预计其 2028 年就能实现正向现金流。正是因为 Anthropic 选择了更聚焦的战略，主攻 to B 市场，专注于编程能力。作为初创公司，Anthropic 没有选择去卷多模态，而是把文本能力做到极致。这种聚焦策略让公司能够一边迭代模型一边产生收入，步伐稳健，不会因为过度扩张而陷入困境。谷歌也采取类似的稳健策略，凭借强大的技术储备和 YouTube 数据优势，在多模态领域取得突破，Veo 3.1 模型的表现亮眼。

图14: OpenAI ARR (十亿美元)



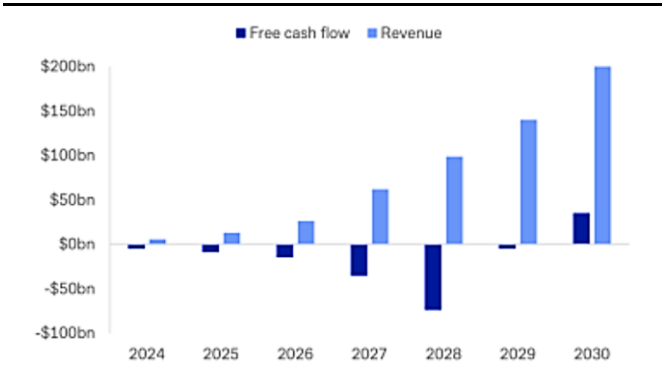
数据来源: ARR Club, 东吴证券研究所

图15: Anthropic ARR (十亿美元)



数据来源: ARR Club, 东吴证券研究所

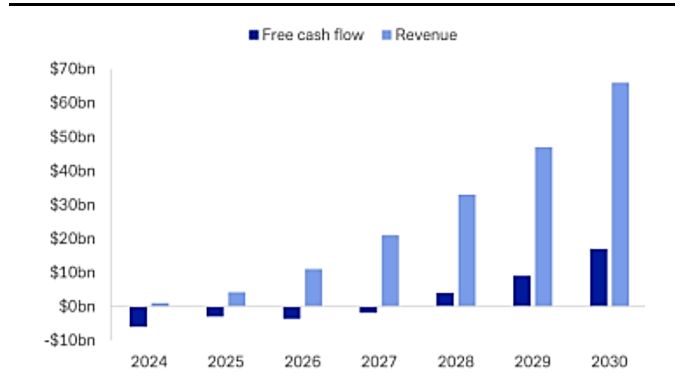
图16: OpenAI 收入和自由现金流预测 (十亿美元)



数据来源: Fortune, 东吴证券研究所

注: 2025-2030 年为预测值, 由 Fortune 根据 openAI 管理层相关表述进行预测

图17: Anthropic 收入和自由现金流预测 (十亿美元)



数据来源: Fortune, 东吴证券研究所

注: 2025-2030 年为预测值, 由 Fortune 根据 anthropic 管理层相关表述进行预测

1.3.2. 纵向划分: IaaS 层 vs 模型层 vs 应用层

横向对比 OpenAI 与 Anthropic 两大阵营, 我们看到了战略路径的分化; 但**泡沫的分布不仅存在于横向竞争者之间, 更体现在产业链的纵向层级中。同样是 AI 投资, 上游卖铲子、中游租算力、下游做应用, 面临的风险与回报完全不同。我们需要从 IaaS 到 PaaS 再到应用层, 逐层拆解利润分配与风险暴露。**

- ◆ **IaaS 是“苦力活”。**根据 semianalysis, 以一台配备 8 张 H200 GPU 的服务器为例, 资本开支高达 24.6 万美元, 加上折旧、电费与托管等运营成本, 每张 GPU 的每小时可变成本约为 1.23 美元。新兴云厂商若以 1.90 美元/小时出租算力, 毛利率约 35%。这是一门典型的重资产、中等回报的租赁生意, 虽有利润但远非暴利, 且随时面临硬件快速贬值的风险。
- ◆ **PaaS 与模型层可以实现技术溢价。**通过算法优化 (如 DeepSeek R1 这类高效模型), 单张 GPU 的推理吞吐量可达 735 tokens/s, 底层计算成本降至 0.72 美元/百万 tokens。但对外 API 混合售价可达 1.75 美元, 毛利率随之提升至 59%。这说明, 将原始算力“加工”成智能的服务, 比单纯出租算力更具价值, 技术溢价开始显著显现。
- ◆ **应用层“贫富分化”。**第三方开发者仅租用他人 API (成本 1.75 美元/百万 tokens), 向用户收取标准 20 美元月费后, 每用户仅剩 2.50 美元利润, 毛利率只有 12.5%, 本质上是在为大模型厂商打工。而第一方垂直整合巨头 (1P, 如 OpenAI 或谷歌) 使用内部成本价 (0.72 美元/百万 tokens), 同样收取 20 美元月费, 每用户利润可达 12.8 美元, 毛利率高达 64.1%。这一对比说明: **垂直整合是通向高利润的唯一路径。只有打通从模型到应用的全链路, 才能摆脱微薄利润的锁死; 这也是大模型公司纷纷布局超级应用、超级应用反过来急于自研模型的根本驱动力。**

表7: AI 产业链各环节的商业模式

The AI Token Factory Economics Stack				
Layer	Units	H200	Business	SemiAnalysis Model
Total Server Capital Cost	USD per 8 GPU	\$246,391	Chip	Accelerator and AI TCO Model
Useful Life	Yrs	4.0	IAAS	AI TCO Model
Depreciation Expense/GPU	USD/hr/GPU	\$0.88	IAAS	
Colocation Cost	USD/kW/mth	\$110	IAAS	AI Datacenter Model
Electricity Cost	USD/kWh	\$0.087	IAAS	
Operating Cost per GPU	USD/hr/GPU	\$0.35	IAAS	
Total Variable Cost per GPU	USD/hr/GPU	\$1.23	IAAS	AI TCO Model
Revenue per GPU	USD/hr/GPU	\$1.90	IAAS	
Neocloud Gross Margin	%	35%	IAAS	
Effective Inference Throughput/GPU ¹	Tokens/s/GPU	735	PAAS	InferenceMAX™
Inference Compute Cost	USD/M Tok	\$0.72	PAAS	
Inference API Blended Price	USD/M Tok	\$1.75	PAAS	
Model Gross Margin	%	59.0%	PAAS	
Tokens/User/Month	Tok/user/mth	10,000,000	Application	AI Tokenomics Model
3P Token Unit Cost	USD/M Tok	\$1.75	Application	
Average Revenue per User	USD/mth	\$20.00	Application	
3P Gross profit per User/Mth	USD/mth	\$2.50	Application	
3P Application Gross Margin	%	12.5%	Application	
1P Token Unit Cost	USD/M Tok	\$0.72		AI Tokenomics Model
1P Gross profit per User/Mth	USD/mth	\$12.82	Application	
1P Application Gross Margin	%	64.1%	Application	

1. DeepSeek R1 FP8 on H200 using SGLang. Uses 8k input, 1k output tokens, 43 interactivity.

数据来源: semi analysis, 东吴证券研究所

AI 产业的价值并非均匀分布, 而是高度集中在拥有技术壁垒(上游芯片)和垂直整合能力(模型层+应用层)的玩家手中。中间的纯算力租赁和 API 转售环节, 正在沦为“微笑曲线”的底部。

基于上述对阵营分化、产业链利润分布以及融资结构的综合分析, 我们得出以下分层判断:

1、上游芯片硬件环节几乎不存在泡沫。英伟达作为绝对主导者, 其 Blackwell 等产品供不应求, 毛利率维持在 70-80% 的高位, 资本支出回收确定性极强。国内光模块、存储等供应链公司直接受益于全球云巨头和模型公司的真实需求, 订单可见度高, 基本面扎实。

2、云服务层面则出现明显分化: 巨头云如谷歌云、AWS、Azure 整体稳健, 其资本支出虽庞大, 但背靠多元化业务和广告、电商等高利润板块, 抗风险能力强; 新云公司如 CoreWeave、Crusoe、Lambda 等则面临较高风险。这些公司通过提前囤积 GPU 并高价租赁给模型公司实现快速扩张, 但其债务利率普遍在 11-14%, 远高于谷歌的 5-6%, 一旦下游模型公司因 scaling law 放缓或商业化不及预期而减少采购, 新云的现金流和再融资能力将迅速承压, 潜在违约风险不容忽视。

3、模型层面 2025 年普遍亏损，但这更多是研发驱动的阶段现象，而非不可逾越的泡沫。巨额资本支出主要用于训练更大参数模型和探索新架构，scaling law 虽出现边际放缓迹象但并未结束，但推理端需求仍在爆发式增长。推理端爆发的核心制约因素在于算力有限，而非需求不存在。更重要的是，模型层面的降本趋势已然启动：算法优化（如 MoE 混合专家、稀疏注意力）、硬件转向 ASIC 自研芯片（谷歌 TPU、亚马逊 Trainium、微软 Maia 等能效提升 30-40%）、以及竞争迫使英伟达让利，都在压缩单位算力成本。一旦成本曲线显著下移，模型公司的高毛利潜力将迅速释放，利润弹性可观。

4、应用层面，绝大多数“套壳”产品和同质化 AI native 应用已被市场淘汰，存活率极低。2023-2024 年数千个 AI 应用中有超过千个关闭，2025 年截至 12 月底已有近 400 个死亡，主要集中在写作、设计等易被大模型直接吞噬的领域。成功的关键在于数据壁垒和场景刚需：编程和企业级 to B 应用（如法律、医疗领域的专有数据平台）因难以被大模型完全替代而表现出较强生命力；陪伴娱乐（角色扮演）虽短期因监管和记忆能力不足而表现疲软，但长期受益于上下文长度提升和个性化记忆（如潜在新架构支持千人千面模型），具备数据飞轮效应。总体而言，应用层尚未进入全面盈利阶段，需要精选具备独家数据、清晰变现路径和强用户粘性的赛道与个股。

因此，AI 产业的泡沫确实存在，但局限于特定阵营和环节。以 OpenAI 为核心的激进扩张路径和与之深度绑定的一系列公司，以及依赖高息债务的新云玩家，构成了泡沫最集中的区域；而 Anthropic-谷歌-亚马逊阵营以及上游硬件和主流云厂商则保持了相对稳健的姿态。模型层面的亏损本质上是通往高利润弹性的必经阶段，应用层面则在残酷洗牌中孕育真正的长期机会。在警惕局部风险的同时，建议抓住产业链降本和需求真实增长带来的结构性机会。

2. 算力：英伟达霸权下的暗流涌动

在 2025 年的 AI 产业中，算力已成为决定胜负的核心战略资源。英伟达凭借 GPU 的先行优势和深厚的软件生态，长期占据市场主导地位。然而，随着 AI 应用大规模落地，推理工作负载占比迅速攀升，以及云服务巨头加速推进自研的 ASIC 专用芯片，英伟达看似稳固的护城河之下，已开始出现裂痕。

2.1. 英伟达：看似坚固的护城河下暗藏裂痕

英伟达 2025Q3 季报成绩亮眼：数据中心业务营收飙升至 308 亿美元，同比增幅高达 154%，毛利率维持高位。新一代 Blackwell 架构芯片需求火爆，供不应求，而上一代 Hopper 架构产品依然保持着强劲的销售势头。CEO 黄仁勋在业绩会上将公司定位为“AI 时代的基础设施提供商，如同电力公司”。

然而，在业绩电话会中，当分析师追问 Blackwell 芯片的具体产能规划时，黄仁勋的回应显得颇为谨慎，反复强调公司正“全力提升供应”以满足“空前需求”，却未给出明确的时间表。这种持续的“供不应求”状态，短期虽能维持其定价优势和利润空间，

但长期而言，无疑为竞争对手创造了宝贵的窗口期。谷歌、亚马逊等云巨头之所以不惜重金投入自研芯片，英伟达 GPU 的高昂成本与供应不确定性正是核心动因之一。

英伟达披露，2025Q3，推理工作负载已占数据中心收入的 40%，相比上一季度显著提升。这标志着 AI 产业正从集中式的模型训练阶段，进入大规模、分布式的应用推理阶段。对整个生态而言，这是积极的里程碑；但对英伟达而言，这可能是一个危险的转折点。因为推理场景恰恰是 ASIC 最能发挥其优势的战场。通用 GPU 的设计追求极致的灵活性，以应对千变万化的计算任务，但这份灵活性是以更高的功耗和成本为代价的。相反，ASIC 芯片专为特定任务（如 Transformer 架构的核心运算）量身定制，可以剔除所有冗余模块，从而在相同制程下实现更高的能效比和更低的单位成本。

为了应对竞争，2025 年 12 月 24 日，英伟达宣布，以约 200 亿美元现金收购 AI 芯片初创公司 Groq 的核心资产和技术知识产权。这不仅是英伟达史上最大规模的交易，远超此前收购 Mellanox 的纪录，更是其针对推理时代潜在威胁的主动反击。Groq 由前谷歌工程师 Jonathan Ross 创立，他正是谷歌 TPU 的发明者之一。Groq 专注于推理优化的 LPU（语言处理单元）架构，在低延迟和高吞吐量上表现出色，曾被视为英伟达的有力挑战者。2025 年 9 月，Groq 刚以 69 亿美元估值完成融资，此次却通过非独占性许可协议，将核心技术、专利和包括创始人 Ross 在内的多名高管团队转向英伟达。

通过整合 Groq 的低延迟推理技术，英伟达不仅能快速补齐专用芯片短板，还将潜在对手转化为自身优势。黄仁勋表示，将把 Groq 处理器融入 NVIDIA AI 工厂架构，扩展对实时推理负载的支持。这种“既进攻又防守”的布局，不仅消除了一个独立竞争者，还通过挖角 TPU 核心团队，进一步强化生态壁垒。类似于当年收购 Mellanox 强化网络的逻辑，如今则是提前锁定推理战场。交易结构采用许可和人才引进形式，也巧妙规避了全面并购可能引发的反垄断审查。

2.2. 谷歌：全栈整合构建的平行宇宙

与英伟达的高调宣示形成鲜明对比的，是谷歌在算力领域表现出的战略定力与低调务实。在 25Q3 业绩电话会上，首席执行官桑达尔·皮查伊并未浓墨重彩地渲染 TPU 的技术参数，而是将其融入公司整体 AI 基础设施能力的叙事之中。在回应关于 CapEx 的提问时，他给出了一个长期主义色彩的回答：在 AI 基础设施上“投资不足的风险，远大于过度投资的风险”，并预计 2026 年资本支出将大幅提升。这番表态的背后，是谷歌对其全栈整合能力的自信。

谷歌的真正优势在于其构建的完整闭环。从自研的 TPU 芯片、定制化的 Pod 级集群系统，到上层的 TensorFlow 框架、Gemini 大模型，再到面向用户的云服务和搜索产品，谷歌实现了从底层硬件到顶层应用的全链条自主可控。TPU 的发展史就是这一战略的缩影：从早期探索，到 TPU v4 实现超大规模集群，再到最新推出的、专为推理优化的 v7 Ironwood，其能效比前代提升一倍。谷歌不将 TPU 作为独立商品出售，而是将其深度整合为云服务竞争力的基石。Gemini 2.0 由 TPU v6 训练，其推理服务则大规模部

署在 v7 之上，形成了高效的内循环。

谷歌云业务 2025Q3 营收增长 35%，营业利润率在折旧增加的背景下逆势提升至 23.7%，自研 TPU 在成本控制方面的贡献功不可没。当同业还在为 GPU 的采购价格和交付周期焦虑时，谷歌已握有供应链安全与成本优化的双重主动权。

2025 年 11 月，SemiAnalysis 发布报告，认为 TPU v7 已成为 AI 算力领域不容忽视的“巨无霸”。该芯片在性能、成本与规模化部署方面，已对英伟达的长期主导地位构成切实挑战。

从硬件规格看，TPU v7 在峰值算力、内存带宽与容量（采用 8-Hi HBM3E）等关键指标上，已基本追平英伟达 Blackwell 架构 GPU，仅通用可用性落后约一年。其设计强调可靠性与可用性，不追求过高峰值参数，而是通过提升实际利用率来实现更高有效性。在市场主流的 BF16 精度下，TPU v7 的单芯片算力达到了 **2307 TFLOPS**，这已经非常逼近 GB200 的 **2500 TFLOPS**。

在训练和推理任务中，TPU v7 展现出显著的总拥有成本（TCO）优势：从谷歌内部视角看，其全成本比英伟达 GB200 服务器低约 44%；对外部 GCP 客户而言，成本仍低 30%~41%，这主要得益于更低的利润加成与更高的实际使用效率。尤其在处理小规模消息的推理场景中，TPU 在内存带宽的性价比与实际利用率方面表现突出。

但在内存容量方面，GB200 配备了海量的 **192GB**（甚至 GB300 达到 288GB）显存，而 TPU v7 仅有 **16GB**（外部版 32GB）。这意味着 TPU v7 在处理超大参数模型的单卡推理或训练时，将面临巨大的内存墙限制，必须依靠更大规模的集群互联来解决。

表8: TPU v7 对比 GB200/GB300 芯片

Nvidia vs TPU Performance per TCO Comparison					
Chip	Unit	GB200 NVL72 (Spectrum)	GB300 NVL 72 (Spectrum)	TPU v7 - 3D Torus - Internal	TPU v7 - 3D Torus - External
Customer Profile		Hyperscaler	Hyperscaler	Hyperscaler	Neocloud Giants
Total Cost per Unit per Hour	USD/hr/GPU	\$2.28	\$2.73	\$1.28	\$1.60
Capital Cost as % of Total Ownership Cost	%	77.4%	79.0%	72.7%	72.7%
Marketed TFLOPS (BF16)	TFLOPS	2,500	2,500	2,307	2,307
Marketed TFLOPS (8 Bit)	TFLOPS	5,000	5,000	4,614	4,614
Marketed TFLOPS (FP4) ¹	TFLOPS	10,000	15,000	4,614	4,614
Memory Bandwidth per Logical GPU	TB/s	8.0	8.0	7.6	7.6
Memory Capacity	GB	192	288	16	32
Marketed TFLOPS (FP8) / Memory Bandwidth	TFLOPS/TB/s	625	625	609	609
TCO per Marketed BF16 Dense PFLOP	\$/hr per PFLOP	\$0.91	\$1.09	\$0.56	\$0.69
TCO per Marketed 8 Bit Dense PFLOP	\$/hr per PFLOP	\$0.46	\$0.55	\$0.28	\$0.35
TCO per Marketed FP4 Dense PFLOP ¹	\$/hr per PFLOP	\$0.23	\$0.18	\$0.28	\$0.35
TCO per Memory Bandwidth	\$/hr per TB/s	\$0.28	\$0.34	\$0.17	\$0.21
TCO per Memory Capacity	\$/hr per TB	\$11.87	\$9.47	\$6.67	\$8.33

1. TPU v7 does not have native FP4 support hence FP8 PFLOPs are downcast here

数据来源：semi analysis，东吴证券研究所

系统级架构是谷歌的强项。TPU v7 采用液冷机架设计，每机架包含 16 个托盘，每托盘集成 4 颗芯片，通过 3D torus 互连网络与光路交换技术实现灵活扩展，最高可支持 9216 颗 TPU 的大规模集群，并能按需灵活划分至 2048 颗 TPU 的规模。这种设计在延迟、数据局部性与并行模式支持上表现优异，可轻松扩展至近 15 万颗 TPU 的超大集群。相较之下，英伟达方案虽峰值性能高，但实际利用率常因功耗调节与数据填充等技术手段而被高估，其均衡利用率仅约 19%，导致每有效 PFLOP 的总拥有成本高出 52%。

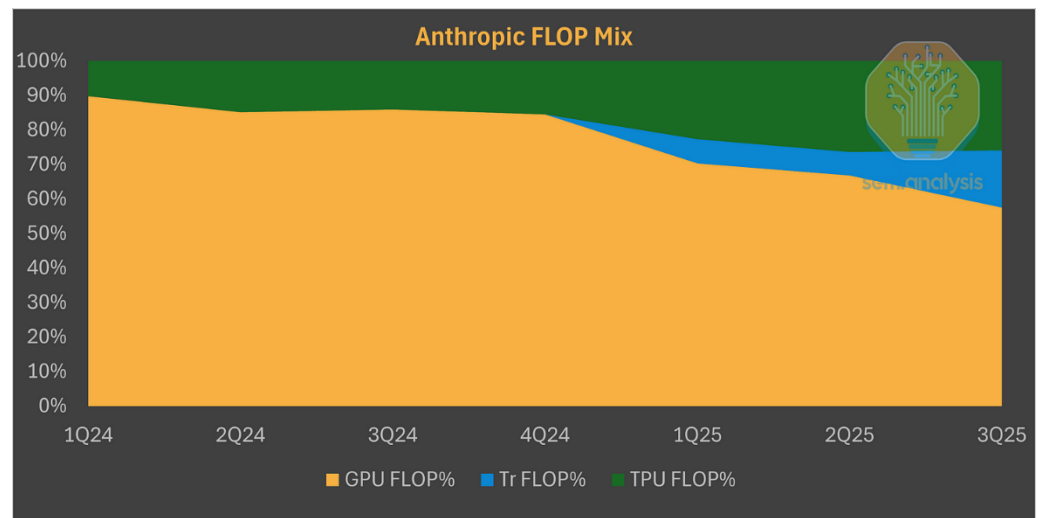
客户拓展是此次报告的另一大亮点。AI 公司 Anthropic 已成为 TPU 的最大采购方，部署规模超过 1GW，其中 40 万颗 TPU v7 直接向博通采购，其余 60 万颗通过 GCP 租用，整体交易规模巨大。此举为 Anthropic 显著节约了资本支出，并带动了 Neocloud 服务商与数据中心供应商的协同合作。其他新客户还包括 Meta、Safe Superintelligence、xAI，以及潜在客户 OpenAI，这些公司借助 TPU 可获得约 30% 的性能与总拥有成本提升。顶尖大模型如 Claude 4.5 Opus 与 Gemini 3 的训练与推理，已主要依赖于 TPU，显示谷歌正积极将其硬件优势向外输出。

软件生态与英伟达 CUDA 护城河的松动也是焦点。谷歌正推动 PyTorch 原生支持 TPU，并集成 vLLM 等推理优化后端，其专用指令集进一步优化了嵌入等任务。不过，其编译器与运行时仍为闭源，一定程度上限制了生态扩展。尽管如此，TPU 在总拥有成本与全栈优化上的优势，已开始动摇英伟达长期依赖的开发者生态壁垒。

展望未来，谷歌下一代 TPU v8 AX 与 v8X 已规划中，旨在直接对标英伟达下一代 Vera Rubin 架构。整体来看，谷歌正凭借基础设施引领、商业销售与成本效率扩展的组合策略，不再满足于自用，而是通过大规模部署与客户迁移，系统性挑战英伟达的市场垄断，预示着 AI 算力格局正从一家独大走向多元竞争、效率驱动的新阶段。

本质上，谷歌正在构建一个与英伟达 GPU 生态并行的“AI 算力宇宙”。这个宇宙首先服务于谷歌自身庞大的产品体系，继而通过谷歌云平台和战略投资，向外辐射其影响力（例如明星 AI 公司 Anthropic 宣布获得价值数百亿美元的 100 万片 TPU 使用权）。如果说英伟达是向所有淘金者“卖铲子”，那么谷歌则是在经营一个“自营金矿+对外提供采矿服务”的复合型帝国。

图18: Anthropic 公司在算力硬件（计算资源）组合上的演变趋势



数据来源: semi analysis, 东吴证券研究所

2.3. 亚马逊: 最激进的自研芯片扩张者

在三大云巨头中，亚马逊对自研芯片的推进最为激进，姿态也最为高调。CEO 在 2025Q3 业绩电话会上用大量篇幅详解 Trainium2 的进展，直言该芯片已“全面售罄”，季度收入环比暴增 150%，成为一个规模达数十亿美元的业务线。亚马逊表示，Trainium2 在推理任务上相比“竞品”具备 30-40% 的性价比优势，此番对比虽未点名，但矛头直指英伟达 GPU 在推理场景下的效率短板。

亚马逊公布的扩张计划堪称宏伟: 到 2025 年底前, 将在超过 100 万片自研 Trainium2 芯片上运行其 Bedrock 模型服务平台。这一具体且庞大的部署目标, 在巨头中独树一帜。谷歌虽投资巨大但未披露具体规模, 英伟达的芯片则分散于全球客户。亚马逊的百万级集群主要集中于自身 AWS 云服务, 这种集中化部署能迅速形成规模效应, 加速软件生态的成熟, 构建起强大的客户黏性。

与此同时, 亚马逊管理层亦谨慎地保持平衡, 强调与英伟达的“深度合作”关系”不受影响, 公司仍在采购大量英伟达最新 GPU。这体现了一种务实的“双轨制”策略: 一方面通过 Trainium 降低核心业务对单一外部供应商的依赖, 掌控成本与演进节奏; 另一方面保留业界最强大的 GPU 选项, 以满足客户多样化需求, 保持技术前沿性。贾西的表述非常清晰: 为客户提供选择权, 让工作负载运行在最合适的芯片上。

2.4. 微软: 和 OpenAI 绑定的双刃剑

微软在 AI 时代围绕能源、芯片与数据中心展开全面布局, 既展现了其作为云巨头的全栈野心, 也暴露出在自主掌控与外部依赖之间的内在矛盾。这种矛盾的根源, 很大程度上来自于其战略核心建立在对 OpenAI 的深度绑定之上。通过数百亿美元的战略投

资，微软快速获得了 OpenAI 的模型能力，并将其深度整合至 Azure 与全线产品中，在应用层占据先发优势。然而，这种深度绑定也让微软变得被动——微软自身的研发路线、产品规划、甚至基础设施投资节奏，都在很大程度上受到 OpenAI 发展节奏的制约。当 OpenAI 的需求发生变化或进展放缓时，微软庞大的基础设施投入就可能面临暂停风险，即建好的算力无法充分利用，或计划中的项目被迫调整。

在能源布局上，微软正直面 AI 算力激增所带来的“电力饥渴”。

为支撑 Azure AI 基础设施，其在 2025 财年豪掷约 800 亿美元资本支出，大部分用于数据中心扩张与能源保障。微软已从被动采购转向主动掌控，与 Brookfield 签订了史上规模最大的企业购电协议，承诺在 2026 至 2030 年间开发 10.5 吉瓦新能源，主要服务于欧美数据中心。为获取稳定、零碳的基载电力，微软与 Constellation Energy 签署长达 20 年的协议，独家重启并接管三哩岛核电站 1 号机组（更名为 Crane 清洁能源中心），独享其 835 兆瓦的电力输出。此外，微软还探索现场燃气发电、直连电厂的“即插即用”数据中心模式，并利用 AI 优化能效，希望将数据中心电源使用效率（PUE）降至行业领先水平。尽管承诺在 2030 年前实现碳负排放，但 AI 负载的飙升已使能源供应成为扩张的最大瓶颈，迫使微软深度介入能源供应链的源头。

在芯片布局上，微软 Maia 系列芯片的初衷是降低成本并优化 Azure 工作负载，然而进展显著滞后于预期。2023 年底亮相的 Maia 100 因设计早于生成式 AI 爆发，在内存带宽与推理性能上存在短板，仅用于内部测试，未大规模部署。后续的 Maia 200（代号 Braga）原计划于 2025 年量产，却因设计变更、OpenAI 需求调整、团队动荡及模拟不稳定等问题，推迟到至少 2026 年，其能否有效抗衡英伟达 Blackwell 架构已引发质疑。微软已调整路线图，降低了后续芯片的激进目标，转向更可靠的迭代（如 2027 年的 Maia 280），希望实现 20-30% 的能效提升。尽管微软可访问 OpenAI 的自研芯片 IP，但仍高度依赖英伟达 GPU（包括其 Blackwell 架构设计）并辅以 AMD 等供应商。自研芯片在短期内难以承担主流 AI 负载，这使得微软在核心算力成本与控制权上相对被动。

在数据中心布局上，微软依托其全球规模优势展现出调度弹性。Azure 已覆盖 70 多个区域、逾 400 座数据中心，构建起一支可灵活调配的“算力舰队”，尤其适合处理对延迟不敏感的推理任务。标志性项目包括，投资超 70 亿美元的威斯康星州 Fairwater 超级工厂（内设分别支持标准 CPU 与超密 GPU 的独立建筑，单栋功耗超 300 兆瓦），并通过超高速 AI 广域网（初始带宽 300Tb/s，可扩展至 10Pb/s）实现多站点分布式训练互联。此外，微软还为 OpenAI 独家建设了多个吉瓦级集群。

然而，自 2024 年起，微软按下了“暂停键”，冻结了超过 3.5 吉瓦的自建容量（涉及美国凤凰城、芝加哥及欧洲多地），转而加速向 CoreWeave、Oracle 等“新云”提供商租赁裸机 GPU 以解燃眉之急。这一战略收缩暴露出其早期对需求增长的严重误判与规划失准，甚至导致部分 OpenAI 的合同流向竞争对手。其数据中心选址日益向偏远地区倾斜，以规避大城市电网限制，但这同时也增加了建设与运营的复杂性。

整体来看，微软在应用层与云服务市场占据先发优势，但在硬件与能源上存在明显软肋。能源正演变为其新的战略护城河，而芯片自研的屡屡延期则进一步固化了对英伟达的依赖。未来，微软能否巩固其在 AI infra 的主导权，主要取决于核电与新能源项目能否顺利落地、自研芯片能否有效迭代。

2.5. GPU vs ASIC：竞争多元化有望推动成本下降

GPU 凭借既有优势占据绝对领先，但这并非终局，我们认为 ASIC 的崛起、多玩家涌入以及供需周期的剧烈摆动，将在未来五年内改变格局，让单位计算成本大幅下降，最终惠及整个 AI 生态。

英伟达非常聪明地构建了一套适合高度并行计算的 GPU 架构，而 AI 的爆发恰好让这一既有技术成为最匹配的解决方案。如果 AI 成为计算领域的“杀手级应用”，那么 GPU 这种原本为图形渲染设计的架构，就意外地成了最优解。这种历史偶然性让英伟达抓住了时代红利，建立了近乎垄断的地位。但这并不意味着 GPU 是 AI 计算的终极形态。GPU 架构强大、生态成熟、开发者友好，但并非最省资源、最具性价比的长期方案。如果今天从零开始设计专为 AI 优化的芯片，人们不会选择复制完整的 GPU 架构，而是直接转向专用 AI 芯片，因为后者在计算特性匹配度和经济效率上会高得多。

英伟达的超高利润率本身就是竞争可能加剧的信号。正因为 GPU 市场利润丰厚到史无前例的程度，整个芯片行业都被调动起来：传统玩家如 AMD 加速追赶，超大规模云厂商纷纷自研专用芯片，大型科技公司也在构建自己的 AI 芯片体系，中国企业在内的大量本土力量正在推动自主生态。中国在芯片领域的巨大努力，以华为为代表的企业正在使用国产芯片训练大模型，目的就是让本土芯片生态真正跑起来。

我们预计未来五年到十年内，AI 芯片会从当前的相对短缺转向供应充足，单位成本会下降。这种供需反馈机制是人类历史的规律：任何具有类商品属性的市场，都会经历短缺引发投资潮、投资潮导致过剩、过剩又压低价格并刺激新一轮需求的循环。就在 2025 年 12 月，AWS 宣布将 GPU 的服役年限延长到七年以上——企业正在通过优化延长现有硬件生命周期，同时大量新产能正在路上。

目前绝大多数数据中心建设都是围绕训练和部署大模型展开的，这依赖于超级计算集群，而 GPU 集群正是当下最成熟、最可靠的选择。但与此同时，小模型革命正在同步发生：每隔 6~12 个月，就会有压缩后的小模型在能力上追平上一代大模型，这直接推动推理硬件向边缘设备下沉。例如中国 Kimi K2 模型，已经能在普通 MacBook 上复现接近 GPT-5 的推理能力。这种趋势意味着，未来的 AI 硬件金字塔结构将非常清晰：顶端是少数需要超级集群训练的顶级大模型，底层则是海量分布在手机、笔记本、物联网设备上的小模型。推理侧对专用芯片的需求会更强烈，因为边缘部署对功耗、成本和效率的要求远高于云端训练。这进一步削弱了通用 GPU 的长期护城河，因为训练可以容忍更高功耗和成本，而推理必须极致优化。

TPU 从诞生之初就是为特定 AI 工作负载（如 TensorFlow 生态）深度定制的典型 ASIC 代表，在 Google 内部实现了极高的训练和推理效率，但生态封闭、对外供应有限，这与英伟达开放的 CUDA 生态形成了鲜明对比。像 TPU 这样的厂商内部专用芯片，虽然在自家体系内极具优势，但难以撼动 NVIDIA 在开放市场的主导地位，至少在当前周期如此。然而，随着更多玩家进入，类似 TPU 的专用设计思路会扩散：亚马逊的 Inferentia、微软的 Maia、Meta 的自研方案，以及大量初创公司的全新架构，都在复制“为 AI 量身定制”的逻辑。这些专用芯片在特定任务上的能效比往往大幅领先通用 GPU，一旦生态成熟、供应放量，就会直接冲击英伟达的定价权。

我们对整个竞争格局的态度是乐观的。这场巨大的芯片军备竞赛虽然会让英伟达面临压力，但对行业是重大利好：成本下降会释放出创新空间，让更多创业公司能够负担得起前沿算力，推动应用层爆发。这种多元化竞争，是 AI 革命早期阶段的典型特征：短期内领先者吃掉大部分利润，长期内技术扩散和商品化会让整个市场规模指数级扩大。

3. 云服务市场的冰与火：巨头狂欢与新贵困局

2025 年的云服务市场冰火两重天。一边是亚马逊、微软、谷歌三大巨头手握千亿美元订单，现金流如潮水般涌入，疯狂扩张数据中心；另一边是 CoreWeave、Nebius 等新兴云服务商虽然估值飙升，却在债务压力和财务恶化中挣扎求生。这种两极分化不仅仅是规模差异，更深层次地反映了商业模式的本质差异和产业链价值分配的残酷现实。

3.1. 云巨头的黄金时代：订单、现金流与确定性

2025 年，全球三大云服务巨头同时面临着有单接不完的幸福烦恼。企业客户对 AI 算力的需求已经从早期的试验性采购转向长期的战略性锁定。在算力供不应求的环境下，客户为了确保未来几年的竞争力，愿意接受更高的价格并提前签订长期合同。

微软管理层表示，GPU 供需平衡点已推迟至 2026Q2，且 2026-2027 年数据中心面积需要翻倍。这意味着谁能在这个窗口期最大化产能扩张、最优化成本结构，谁就能在未来 5~10 年的云计算市场中占据主导地位。

三家巨头的订单储备亮眼。①微软的 RPO 表现强劲，截至 FY26Q1 达到 3920 亿美元，同比增长达 51%，主要得益于与 OpenAI 签署的巨额 Azure 服务合同以及企业级 AI 需求的爆发，且其加权平均期限仅为 2 年，意味着这些海量积压订单将在短期内转化为实际营收。此外，微软在 25 年 10 月和 OpenAI 新签订了总金额 2500 亿美元的订单。②亚马逊 AWS RPO 增速相对较慢主要是因为其收入基数极为庞大，且公司更注重将合同转化为即时消费的“货币化”能力而非单纯拉长合同期限。在 2025 年 Q3 财报中，虽然 RPO 为 2000 亿美元，但 10 月份未计入统计的新交易量已超过整个 Q3 总量，这表明订单的流入依然强劲，只是庞大的分母效应稀释了百分比增速，同时公司

通过快速交付算力（如 Tranium2 芯片和大规模电力扩容）实现了订单到收入的高效转化，而非让订单长期停留在积压状态。③谷歌云业务同样在快速积累订单，截至 2025 年 Q3，承诺余额达到 1557 亿美元，同比增长 82%、环比增长 46%。公司表示，谷歌云在第三季度签署了比过去两年总和还要多的 10 亿美元以上大单。

表9：云厂商储备订单余额大增

云厂商	季度	RPO	增速	备注
Oracle	FY26Q2	5230 亿美元	yoy+438% qoq+15%	本季度 RPO 的爆发式增长主要由 Meta 和 NVIDIA 等大客户的新合同驱动。此外，通过与 AWS、Azure 和 Google Cloud 的深度合作，多云数据库消费同比增长了 817%。
微软	FY26Q1	3920 亿美元	yoy+51%	AI 需求强劲，尤其是在 Azure 云服务方面，需求持续超出供应能力。
谷歌	CY25Q3	1557 亿美元	yoy+82% qoq+46%	谷歌云在 Q3 签署了比过去两年总和还要多的 10 亿美元以上大单。增长主要与企业级 AI 需求相关。
AWS	CY25Q3	2000 亿美元	qoq+2.6%	增速相对平稳，10 月新签订单已超过 Q3 总和，显示出强劲的后势势头。

数据来源：公司公告，东吴证券研究所

注：oracle 的 FY26Q2 为截至 2025/11/30 的三个月；微软 FY26Q1 为截至 2025/9/30 的 3 个月

注 2：微软的 RPO 是商业合同；谷歌 RPO 为“revenue backlog”，并主要与 google cloud 相关

收入增长方面，三家公司都实现了超预期表现。谷歌云 25Q3 收入 115 亿美元，同比增长 34%，显著高于市场预期的 30% 左右，也超过了上季度的 32% 同比增速，呈现出明显的加速态势。亚马逊 AWS 收入达到 275 亿美元，同比增长 20%，虽然增速看似较低，但考虑到 AWS 的收入基数是谷歌云的 2.4 倍，这一增速实际上反映了较强的市场统治力。微软的 Azure 及其他云服务收入增长 39%，大幅超出管理层此前给出的 37% 指引。

图19：云业务收入同比增速对比：微软 Azure / Google Cloud / 亚马逊 AWS

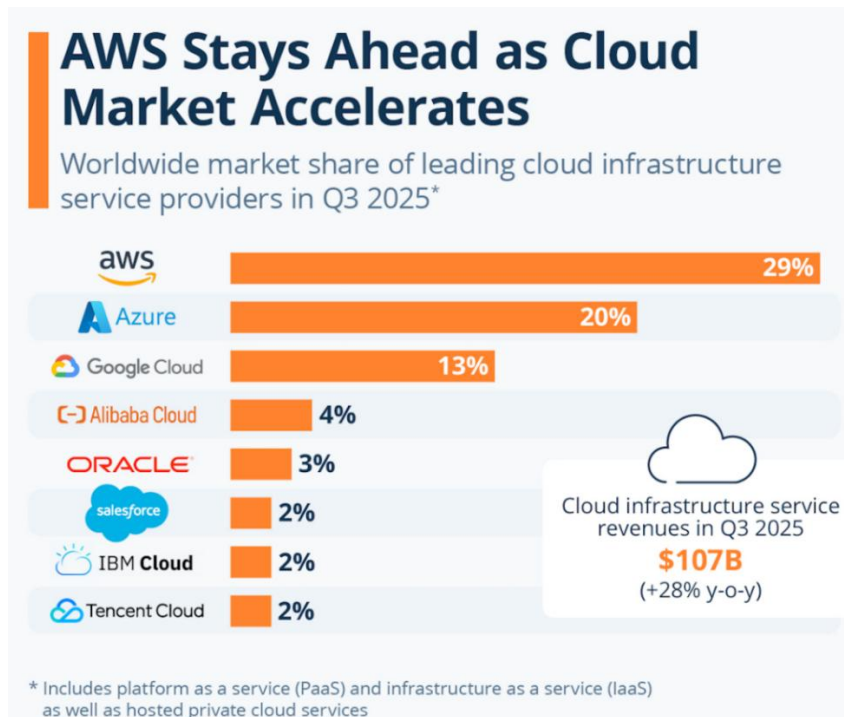
	23Q3	23Q4	24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3
微软 Azure	30%	31%	35%	35%	33%	31%	35%	39%	39%
Google Cloud	22%	26%	28%	29%	34%	30%	28%	32%	34%
亚马逊 AWS	12%	13%	17%	19%	19%	19%	17%	17%	20%

数据来源：app economy insights，东吴证券研究所

盈利能力方面出现了分化。①亚马逊 AWS 的盈利能力始终保持在行业领先水平，第三季度营业利润率高达 36.4%，虽然相比 2024 年 Q3 的 37.6% 略有下降，但仍然远高于其他两家竞争对手。②谷歌云 25Q3 营业利润率达到 17%，相比 2024 年 Q3 的 12% 显著提升，营业利润同比增长 192%。这种盈利能力的快速改善，部分得益于规模效应的显现，部分来自于基础设施效率的提升。③微软的情况最为复杂。虽然公司整体营业利润率达到 49%，但微软云的毛利率却在下降，第三季度为 68%，第四季度指引进一步降至 66%。管理层将此归因于对 AI 的持续投资以及向 Azure 的组合转移。微软强调有持续的效率提升在对冲这种下降，例如通过软件优化将 GPT-4.1 和 GPT-5 两个最常用模型的单 GPU tokens 吞吐量在 FY26Q1 提升了超过 30%。这种软件层面的优化能力，是微软区别于纯基础设施提供商的核心竞争力。

3.2. 三巨头的差异化：亚马逊份额领先，谷歌和微软快速追赶

图20：2025Q3 全球云服务市场市场份额



数据来源：Synergy Research Group，新智元，东吴证券研究所

亚马逊的核心优势在于充沛的现金流。2025Q3 公司经营现金流 1307 亿美元，完全覆盖 2025 全年约 1250 亿美元的资本开支，这意味着公司不需要举债、不需要削减分红，就能支撑大规模投资。更关键的是这种现金流的来源——AWS 在 2025Q3 单季营收 330 亿美元、运营利润 114 亿美元，利润率高达 33%。这相当于一台印钞机在源源不断地为 AI 投资提供弹药。

同时，亚马逊拥有清晰的订单能见度。截至 2025 年 9 月的 2000 亿美元 Backlog，还不包括 2025 年 10 月的几笔大单，这些未公布交易的总额超过第三季度全季度新签合同。新增产能几乎立刻被客户使用并产生收入，不存在产能闲置的风险。亚马逊采取的是务实的技术路线——Trainium2 自研芯片已成为 Anthropic 训练 Claude 的支柱，但亚马逊同时继续与英伟达深度合作。这种既有自研又不排他的策略，确保了在芯片供应链中的议价能力和灵活性。

微软的差异化在于企业级生态的深度绑定。Office 套件全球超 10 亿用户、Azure 云覆盖 95% 财富 500 强、GitHub 垄断开发者社区，这种三位一体的生态使得 AI 产品天然具备分发优势。Copilot 能在短时间内覆盖数千万企业用户，不是因为产品有多颠覆，而是因为用户已经被锁定在微软生态内。企业不太可能为了一个 AI 功能放弃整个 Office 加 Azure 体系。

但 OpenAI 深度绑定是把双刃剑。一方面，2500 亿美元 Azure 订单锁定了未来

5-7 年的现金流，GPT 系列模型的领先地位为微软提供了技术支撑。另一方面，微软 2024 年从 OpenAI 获得 4.9 亿美元收入分成，2025 年前 9 个月获得 8.7 亿美元。这些分成收入相对于微软确认的 41 亿美元权益法亏损而言微不足道，说明微软在 OpenAI 的投资短期内更多是战略卡位而非财务回报。

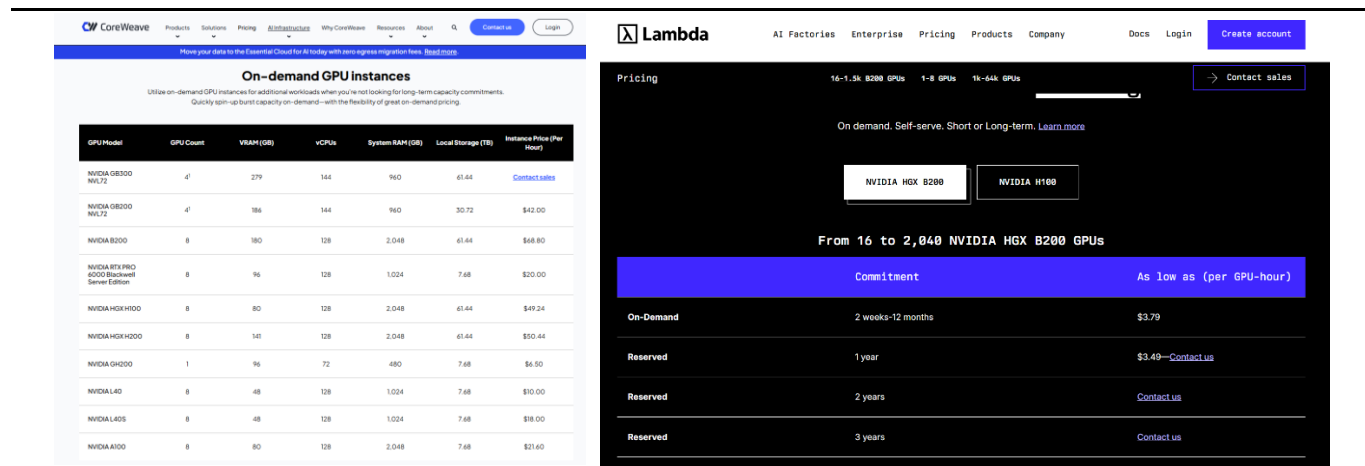
谷歌的优势在于全栈技术带来的成本结构优化。从底层 TPU 芯片专为 Transformer 架构设计，到中间层 Gemini 模型、再到上层搜索重构和 Workspace 智能化，谷歌在每个技术层级都掌握主动权。第三季度云业务营业利润率从 24Q3 的 17.1% 跃升至 25Q3 的 23.7%，在折旧增长的背景下实现利润率改善，验证了全栈整合的经济价值。

3.3. Neocloud 的困局：居高不下的债务、快速折旧的芯片

传统云巨头经过十几年发展已成为包罗万象的云超市，以 AWS 为例，其官网提供超过 200 项服务。这种全方位覆盖曾是它们的护城河，但在 AI 算力需求爆发的当下，反而暴露出大而全的迟钝。

与之形成鲜明对比的是 Neocloud 新势力。以 CoreWeave、Nebius 为代表的公司，采取的是极简的专卖店模式。打开它们的官网，不是眼花缭乱的产品目录，而是一张简单明了的 GPU 租赁清单。这种模式精准击中了 AI 公司的核心痛点——“我不需要 200 项服务，我只需要最新最快的 GPU”。

图21: Coreweave 和 Lambda 官网界面是一目了然的算力租赁清单



数据来源：CoreWeave 官网，Lambda 官网，东吴证券研究所

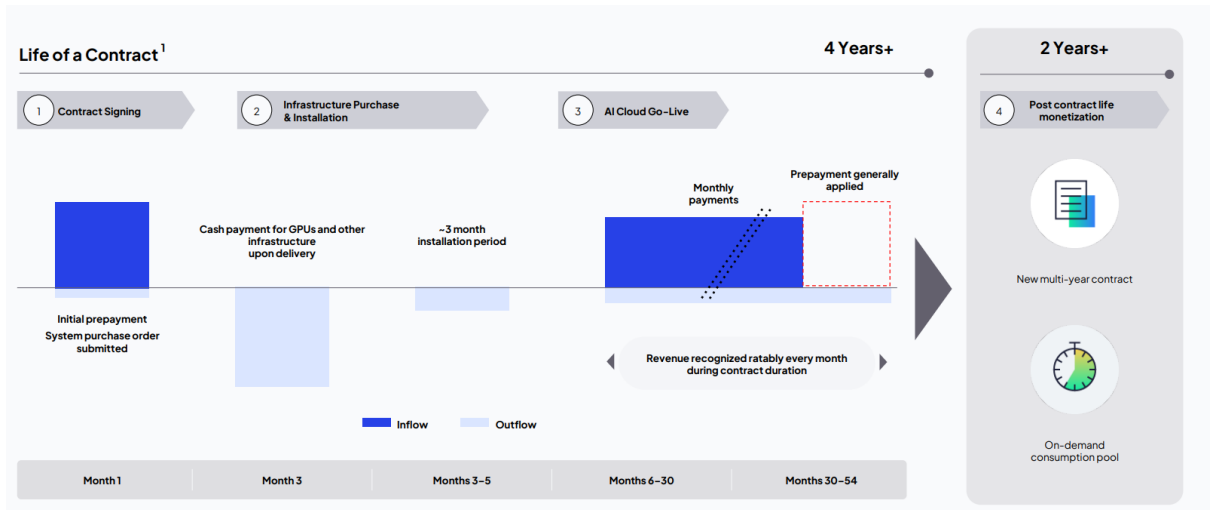
CoreWeave 能够快速崛起，主要是因为抓住了一个时间窗口：AI 大模型训练对 GPU 算力的需求爆发式增长，但传统云计算巨头的 GPU 供应来不及响应。在这个供需失衡的窗口期，CoreWeave 通过快速融资、快速采购 GPU、快速建设数据中心，抢占了市场份额。即便是自己拥有庞大云基础设施的微软，也不得不与 CoreWeave 签订超过 170 亿美元的长期合同。

CoreWeave 的盈利模式与房地产开发商高度相似：房地产开发商拿地建房卖房，而 CoreWeave 采购 GPU 建数据中心出租算力。关键区别在于：房子可以一次性卖出

收回现金，而算力租赁是按月按年收费，现金回流速度慢得多。更危险的是，GPU 的技术迭代速度远快于房地产。一栋楼可以用几十年，但 GPU 可能几年就过时了。

这种商业模式本质上是在对赌三个变量：GPU 供应持续紧张才能维持高租金、AI 公司持续烧钱才有租赁需求、自己能持续拿到最新 GPU 才能保持竞争力。

图22: Coreweave 的商业模式



数据来源：Coreweave 25Q2 earnings deck，东吴证券研究所

coreweave 财务表现脆弱。从 2024Q1 到 2025Q3，Coreweave 收入环比增速放缓，大额折旧摊销和利息费用对利润形成侵蚀。与此同时，CapEx 始终高于收入。根据公司全年指引的中值倒推，2025Q4 的财务压力仍在加剧：收入环比增速仅 13%、经调整经利润率下降，资本开支环比激增、占收入的比例超 4 倍。

图23: Coreweave 财务表现

百万美元	24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3	25Q4E
收入	189	395	584	747	982	1213	1365	1541
YOY					420%	207%	134%	106%
QOQ		110%	48%	28%	31%	24%	13%	13%
经调整Capex	1188	2574	2237	2411	1858	2937	1850	6355
占比收入	630%	651%	383%	323%	189%	242%	136%	412%
QOQ		117%	-13%	8%	-23%	58%	-37%	244%
折旧和摊销	80	164	243	365	443	559	630	
占比收入	42%	41%	42%	49%	45%	46%	46%	
利息费用	41	67	104	149	264	267	311	
占比收入	22%	17%	18%	20%	27%	22%	23%	
经调整经营利润	25	85	125	121	163	200	217	125
利润率	13%	21%	21%	16%	17%	16%	16%	8%
经调整净利润	-24	-5	0	-36	-150	-131	-41	
利润率	-13%	-1%	0%	-5%	-15%	-11%	-3%	

数据来源：Wind，东吴证券研究所

注：25Q4E 数据根据公司全年指引的中值，减掉 25Q1-25Q3 合计值计算得到

Nebius 面临着与 CoreWeave 类似的财务困境。

图24: Nebius 财务表现

百万美元	24Q1	24Q2	24Q3	24Q4	25Q1	25Q2	25Q3
收入	11	12	32	38	55	105	146
yoy					389%	769%	355%
qoq		7%	165%	18%	46%	90%	39%
毛利	3	9	22	10	26	75	103
利润率	22%	73%	69%	27%	47%	71%	71%
折旧摊销	9	11	24	33	49	75	99
占比收入	78%	90%	73%	88%	89%	72%	68%
经营利润	-81	-96	-81	-149	-130	-111	-130
利润率	-715%	-792%	-251%	-394%	-234%	-106%	-89%
经调整净利润	-73	-38	-28	-85	-70	-59	-82
利润率	-644%	-313%	-87%	-224%	-127%	-56%	-56%

数据来源: Wind, 东吴证券研究所

为什么 Neocloud 公司比 Hyperscaler 更脆弱?

- 第一, 对于 Hyperscaler 而言, AI 不是全部业务, 而 CoreWeave、Nebius 这类 Neocloud 公司几乎所有业务都来自算力租赁。
- 第二, Neocloud 公司的融资成本显著高于 Hyperscaler。根据 semi analysis, Hyperscaler 的 WACC 仅为 9.4%, 债务成本低至 4.5%, 因为他们拥有顶级的信用评级和庞大的现金流。而 Emerging Neoclouds 的 WACC 高达 17.5%, 债务成本更是飙升至 15.0%。
- 第三, Neocloud 公司缺乏可持续的竞争优势。技术门槛低, 搭建 GPU 集群、部署数据中心、提供算力服务, 这些技术并不复杂。客户粘性弱, AI 公司使用算力是标准化需求, 谁便宜、谁稳定、谁快就用谁。而它依赖的市场缺口正在被云巨头和大模型公司从两端填补: 云巨头正在疯狂扩充 GPU 产能, 英伟达产能逐步释放, 大模型公司开始自建算力。
- 第四, Neocloud 的规模效应没有 Hyperscaler 强。Hyperscaler 每月的服务器成本仅为 6742 美元, 而新兴小厂高达 9785 美元。这反映了供应链的话语权差异——大厂拿货不仅有巨额折扣, 还能拿到更有利的付款账期。小厂不仅买得贵, 往往还需要为了抢货支付溢价。这也导致了最终的每小时单位成本差异巨大: Hyperscaler 运行一小时的成本是 1.46 美元, 而新兴小厂高达 2.10 美元。这意味着小厂要想盈利, 定价必须比大厂高 40% 以上, 但在同质化的算力市场, 这几乎是不可能的任务。
- 第五, Hyperscaler 有自研 asic 芯片, 与英伟达形成竞合关系; 而 Neocloud 公司完全受制于英伟达——英伟达既是他们的 GPU 供应商, 又是他们的投资人, 即“我卖给你设备、还借钱给你买、顺便持有你股份”。英伟达卖 GPU 是先收钱后发货, 现金流非常健康; 而 CoreWeave 是先采购 GPU、建数据中心, 然后慢慢收租金, 现金流压力巨大。

表10: 三类云厂商的商业模式对比

AI Cloud Comparative Total Cost of Ownership				
	Unit	Hyperscaler	H100 Neocloud Giants	Emerging Neoclouds
Equity Cost of Capital	%	11.0%	20.0%	25.0%
Debt Cost of Capital	%	4.5%	11.0%	15.0%
Equity Funding %	%	75.0%	25.0%	25.0%
Weighted Average Cost of Capital	%	9.4%	13.3%	17.5%
Total Server Capital Costs per Month	USD/mth	\$6,742	\$8,098	\$9,785
Colocation Cost	USD/kW/mth	\$110	\$130	\$150
Total Operating Costs per Month	USD/mth	\$1,776	\$2,082	\$2,506
Total Cost of Ownership per Month	USD/mth	\$8,518	\$10,180	\$12,290
Capital Cost per Unit, per Hour	USD/hr	\$1.15	\$1.39	\$1.68
Hosting Cost per Unit, per Hour	USD/hr	\$0.30	\$0.36	\$0.43
Total Cost per Unit per Hour	USD/hr	\$1.46	\$1.74	\$2.10
Capital Cost as % of Total Ownership Cost	%	79.9%	80.9%	81.8%

数据来源: semi analysis, 东吴证券研究所

Neocloud 的困境, 是 AI 产业链价值分配矛盾的集中体现。它们填补了云巨头无法满足的灵活性需求, 却也承受着产业链中最薄弱环节的所有压力。夹在上游的英伟达和下游的 OpenAI 等大模型公司之间, 既要承受英伟达的高昂采购成本, 又要面对下游客户的价格压力。它们是这场 AI 竞赛中最激进的赌徒, 也可能是最早倒下的玩家。

图25: Coreweave 两次财报后股价大跌 (美元)



数据来源: Wind, 东吴证券研究所

3.4. 云服务市场: 承载 AI 需求爆发的关键底座

我们认为云服务市场既是 AI 革命最直接的受益者与基础设施支柱, 同时也是一个正在从成熟寡头阶段迈向新一轮剧烈变革的战场。当前云市场已经被 AWS、Microsoft Azure 和 Google Cloud 三家牢牢掌控, 它们共同占据了全球 60% 以上的份额 (25Q3 数

据), 形成了一个高度稳定、利润丰厚的格局。这种寡头结构并非偶然, 而是过去十五年基础设施商品化、规模经济和网络效应共同作用的结果。三大云早已不再是简单的 IaaS 提供商, 而是演变为一站式 AI 平台: 提供算力、存储、数据服务、模型托管、推理 API, 甚至完整的企业级 AI 工作流。这让它们在 AI 时代天然占据了中心位置, 我们预计未来五年, 云厂商的收入增长将主要由 AI 驱动, 而非传统 IT 迁移。

企业端 AI 的商业模式正在发生根本性转变, 从传统的“按小时/按实例”计费转向“按 token 计费”。这种新模式本质上是把智能本身变成了一种可精确计量、可无限细分的商品, 单位成本会随着规模扩大而呈现摩尔定律式的通缩。这对云厂商是利好, 因为 token 计费天然匹配云的弹性扩展能力: 用户只需为实际消耗的智能付费, 无需提前采购固定算力。这种模式不仅提高了云服务的粘性, 还让云厂商能够直接捕捉 AI 应用层创造的价值。我们预计, 未来大多数企业不会自建数据中心, 而是继续依赖云来获取最前沿的模型和最优的推理价格。三大云已经在抢占这一高地: OpenAI 通过 Azure 深度绑定, Anthropic 与 AWS 合作, Google 则凭借自研 TPU 和 Gemini 生态形成闭环。这种“模型+云”的捆绑策略, 进一步强化了云厂商的护城河, 让新兴玩家很难从零开始挑战。

然而, 云市场的稳定表象之下, 竞争烈度正在悄然升级。高利润正在吸引全球资本涌入专用 AI 基础设施领域, 云厂商自己也在加速垂直整合。AWS、Azure、Google 都在大规模自研 AI 芯片 (Trainium、Inferentia、Maia、TPU 等), 目的就是要把训练和推理成本内部化, 摆脱对 NVIDIA 的过度依赖。我们认为这一趋势会带来双重效果: 短期内云厂商的毛利率会进一步提升, 因为自研芯片在自家负载下的能效比远高于通用 GPU; 长期看, 一旦这些专用芯片的供应放量、生态开放, 云市场整体的推理价格会加速下降, 刺激更多中小企业上云。

我们预计, 云市场的下一个重大变量是开源模型的影响。2025 年, 中国的开源突破, 让模型能力以低成本向全球开发者扩散。这对云厂商既是机会也是威胁。一方面, 开源模型大幅降低了企业构建 AI 应用的门槛, 更多公司会选择在云上部署开源模型进行微调和推理, 从而拉动云算力需求; 另一方面, 开源也让模型本身趋于商品化, 削弱了云厂商通过独家模型 API 收取高溢价的能力。

从更宏观的角度来看, 中国云市场的发展速度远超预期, 阿里云、腾讯云、百度云、的市场规模和创新强度都不容小觑。中国云厂商不仅在本地提供了与全球三大云相当的服务, 还通过开源模型和自主芯片构建了相对独立的 AI 生态。这对全球云市场是正面刺激: 中国云的快速成长会进一步推高全球 AI 总需求, 同时也迫使美国云厂商保持技术领先力。

我们相信云基础设施将在 AI 革命中扮演类似电力系统在工业革命中的角色: 普惠、可靠、持续降价。AI 革命仍处于早期, 能力快速提升, 收入真实增长, 全球总需求被持续拉高, 而云正是承载这一切的关键底座。

4. 模型层：从神话到工程，从追赶到突围

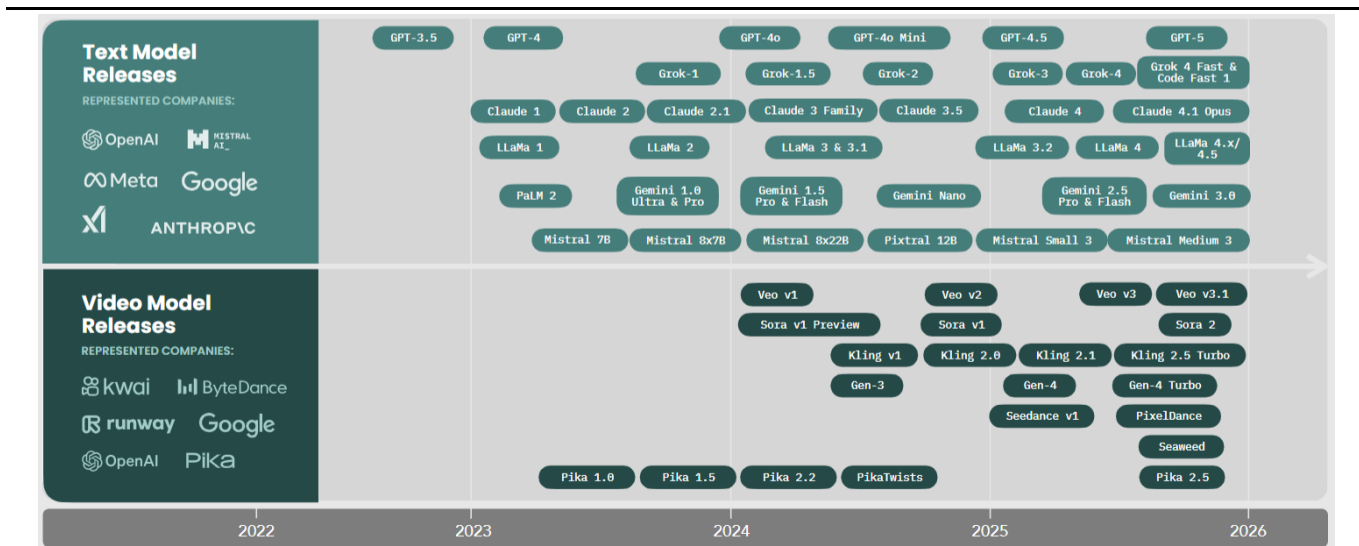
在大模型主线上：2025 年行业从追逐 AGI 转向工程优化与商业落地。DeepSeek R1 的成功证明了低成本路线的可行性，推高了对推理效率的重视。展望 2026 年，我们认为大语言模型将向“更快、更便宜”、更强 Agent 能力方向演进，Scaling Law 并未结束，只是需要配合算法创新；而在多模态方面，输入端将从组合式多模态向原生多模态演进，输出端将实现更长时长、更稳定、更好的音画同步，多模态输出和世界模型将互相促进。

在竞争格局方面：美国公司出现分化。OpenAI 转向产品化，但 GPT-5 技术跃迁不及预期，技术领先优势不再；Google 凭借全栈优势（芯片、数据、原生多模态）强势回归；Anthropic 在 B 端编程领域建立壁垒。中国仍处群雄逐鹿阶段。阿里千问（基础设施优势）、DeepSeek（算法效率优势）、kimi（长文本能力领先）、智谱（编程和 agent 能力）、minimax（多模态能力）、豆包（字节的流量优势）展开竞争。

在中美竞争格局方面：双方技术差距缩小至 6-9 个月。美国芯片禁令限制了中国算力，但也倒逼中国企业进行算法创新（如稀疏注意力）。商业模式上，美国坚持“技术领先+高价闭源”，中国则走“开源+极致性价比”路线。中国模型凭借低成本优势在开源市场份额迅速增长，正对海外高定价模式形成冲击。

4.1. 2025 年：神话退场推理确立范式，工程取代神话

图26：新模型发布节奏越来越快，SOTA 模型的竞争仍然激烈



数据来源：《2025 Globalscape Report》，东吴证券研究所

回看 2025 年的 AI 大模型竞争，最值得记录的不是某个参数规模的突破，而是整个行业心态的转变。这一年，行业从追逐神话般的 AGI 突破，转向了扎扎实实的工程优化和商业落地。这个转折点的标志性事件是 DeepSeek R1 的发布。

年初 DeepSeek R1 的出现，其意义远超技术本身。它最震撼的不是性能有多强，而是用极低的训练成本证明了：美国科技巨头动辄数十亿美元的资本密集型路线并非唯一解。我们认为，DeepSeek R1 真正定义的是 L2 阶段即推理模型时代的竞争逻辑——不再是谁钱多谁就能赢，而是谁能在有限资源下把算法效率做到极致。这个信号传递到全球后，几乎所有主流玩家都在第一季度推出了推理模型。大语言模型竞争已经从寻找新范式的探索期，进入了既定路线下的工程优化深水区。

与此同时，市场期待已久的 GPT-5 并未带来想象中的质变，打破了很多人对技术指数级跃迁的幻想。Sam Altman 本人也在多个场合坦承，当前不会出现巨大的突变。

2025 年巨头战略分化。Google 凭借 Gemini 3、Veo3 实现了强势回归，尤其 Veo3 对有声视频生成的突破引发了行业追逐战。Google 的优势在于垂直整合能力——自研 TPU 芯片、DeepMind 的研究积累、YouTube 海量视频数据，这些要素在多模态竞争中形成了独特的协同效应。OpenAI 则越来越像一家产品公司，Sora 2 的成功更多依赖 Sam Altman 的个人 IP 和产品封装能力，技术代差其实没有市场认知中那么大。Anthropic 坚守编程和 2B 场景，Claude 系列在开发者群体中建立起难以撼动的口碑。

中国市场这一年最显著的特征是开源生态的兴起。DeepSeek 和通义千问引领的开源路线，成为中国厂商参与全球竞争的核心策略。中国市场对付费 API 的接受度有限，闭源反而限制了影响力，不如通过开源快速建立开发者生态和行业话语权。

4.2. 2026 年展望：深度打磨与多模态突破

展望 2026 年，我们预计模型层的发展将沿着两条主线并进推进。

4.2.1. 大语言模型：更快、更好、更便宜

Scaling Law 并未失效。

Ilya 在 2025 年 11 月的访谈中，并没有否定 scaling 的价值，他认为：

- ◆ Scaling 在 2020-2025 年这个阶段确实极其有效，带来了巨大进步（“This is amazing. You’ve got to scale more.”）
- ◆ 更大的规模确实会带来不同（“It would be different, for sure.”）
- ◆ 算力仍然是必要的，甚至是重要差异化因素之一，尤其在大家还在同一条范式赛道上跑的时候

他真正反对的，是把“单纯/只靠 scaling”当成未来继续取得重大突破的主要（甚至唯一）路径。他的原话的核心逻辑是：再扔 100 倍算力当然会有进步，但别指望这就能带来质的飞跃/彻底改变一切。单纯继续“暴力喂饱”已经不足以推动下一个时代级别的突破了。（“Is the belief really that, oh, it’s so big, but if you had a 100x more, everything would be so different. Like, it would be different for sure. But is the belief that if you just 100x the

scale, everything would be transformed? I don't think that's true. So it's back to the age of research again, just with big computers.”)

所以他的完整立场可以总结为：Scaling 有价值，也还会继续有用（尤其是在新的研究范式之下作为支撑），但我们不能只有 scaling（we can't only scale），必须回归研究时代（age of research）；当前最缺的是新的 idea、新的学习范式、更好的泛化机制（比如更接近人类的情感/价值函数、直觉、持续学习能力等），否则再大的规模也只是“在同一个维度上挤牙膏”。

这也是为什么他在创立 SSI 时明确选择不卷主流的“比谁模型更大、更快发布”的 scaling 竞赛，而是 all in 高难度基础研究——他认为真正决定性的突破会来自 idea，而不是单纯的资源堆叠。

最典型的案例便是 OpenAI 的 GPT-5 系列与谷歌 Gemini 3 的对比。

自 2024 年 5 月 GPT-4o 发布以来，OpenAI 在下一代前沿模型的大规模预训练上遭遇了持续挫折。原本计划作为 GPT-5 基础的 Orion 项目训练时间严重超期，性能提升远低于预期，最终未能达到内部跃迁标准而被降级为 GPT-4.5。此后多个后续预训练尝试同样未能成功完成完整跑通，导致 2025 年 8 月正式发布的 GPT-5 及其后续迭代 GPT-5.1、GPT-5.2，本质上仍以 GPT-4o 为技术根基，主要通过强化学习（RL）、推理链优化以及后训练手段实现增量改进，而非依托全新大规模预训练带来的代际飞跃。业界普遍评价认为，GPT-5 系列在响应速度、幻觉控制、编程能力、对话稳定性等方面确实有所提升，但整体更像是一次精心打磨的“功能增强版”，甚至有评测者直言若命名为 GPT-4.2 也毫不违和。这进一步印证了单纯在现有范式上继续 scaling 后训练的边际收益正在迅速下降。

图27: SemiAnalysis 披露: chatgpt 在 GPT-4o 之后没有在预训练环节实现突破

Google's recent Gemini 3 model is now viewed as the state of the art frontier LLM. Like all earlier versions of Gemini, it was trained entirely on TPUs. That result offers concrete proof of both TPU capability and Google's broader infrastructure advantage.

Today's attention often centers on hardware for inference and post-training, yet pre-training a frontier model remains the hardest and most resource-intensive challenge in AI hardware. The TPU platform has passed that test decisively. This stands in sharp contrast to rivals: OpenAI's leading researchers have not completed a successful full-scale pre-training run that was broadly deployed for a new frontier model since GPT-4o in May 2024, highlighting the significant technical hurdle that Google's TPU fleet has managed to overcome.

数据来源：SemiAnalysis，东吴证券研究所

与此同时，得益于 TPUv7 硬件的突破以及原生多模态架构的大胆尝试，Gemini 3

pro 在推理深度、长上下文连贯性、多模态整合以及复杂问题泛化能力上实现了较为明显的跃迁。在 GPQA Diamond、Humanity's Last Exam、LMArena 等榜单上，Gemini 3 pro 拉开了与同期 OpenAI 模型的差距，许多实测用户反馈其响应质量和实用性更胜一筹，甚至迫使 OpenAI 紧急推出 GPT-5.2 进行追赶，却仍未能完全抹平差距。这表明，当 research 洞见与 scaling 相结合时，大规模预训练依然能够带来显著的性能提升。

当 OpenAI 因核心研究人才流失、方向判断出现偏差而被迫将重心转向后训练和推理优化时，其前沿模型的进步开始呈现出明显的递减趋势；谷歌则因持续保持高质量的预训练研究与 scaling 并重，而在关键节点实现了相对更大的突破。这正呼应了 Ilya Sutskever 的警示——scaling 本身远未失效，但如果缺少对“真正重要问题”的敏感度和追求优雅解法的 taste，最终会陷入“something important will continue to be missing”的困境。

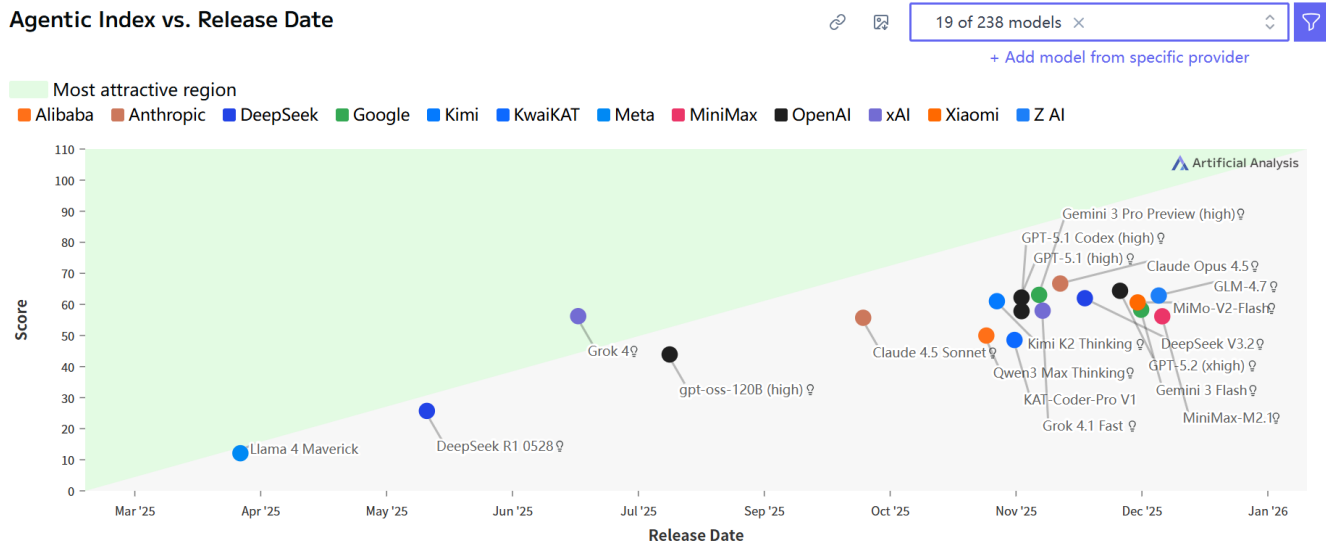
在架构方面，我们认为 Transformer 仍会是 2026 年主线，呈现“打补丁”式进化。

2026 年大语言模型的核心任务不再是做大，而是做好。经历了 2023-2025 年的参数规模竞赛后，单纯堆参数的边际收益正在递减，真正制约 AI 应用落地的是一系列最后一公里问题。“打补丁”指的是在架构不变的前提下优化注意力、训练策略与推理路径。例如：

1) **更好的长记忆能力**。现有模型的上下文窗口虽然从最初的 4K 扩展到 256K 甚至更长，但能装下和能用好是两回事。模型在处理长文档时仍然存在中间遗忘现象，对早期信息的召回能力随对话轮次增加而衰减。2026 年如果要出现“体感级”的进步，不是窗口再翻倍，而是：第一，让模型在数小时甚至数天的持续任务里保持上下文一致性，第二，让记忆从被动存储变成主动管理，也就是知道什么该记、什么时候该忘、什么时候该回顾。技术路径可能包括更高效的注意力机制、外挂记忆库与模型的深度融合、以及针对长程依赖的专项训练。

2) **更好的 Agent 能力**。模型擅长回答问题，但在完成任务上仍显笨拙。一个合格的 AI Agent 需要具备：任务分解与规划能力，把复杂目标拆解为可执行步骤；工具调用能力，知道何时调用什么工具；状态追踪能力，记住任务进度和中间结果；错误恢复能力，发现问题后能自我纠正而非一条路走到黑。2025 年我们看到了 Claude 的 Computer Use、GPT 的 Code Interpreter 等早期尝试，2026 年可能会出现更成熟的 Agent 框架，模型与外部系统的交互将从单次调用进化为持续协作。

图28: 模型的 agent 能力持续提升



数据来源: Artificial Analysis, 东吴证券研究所

注: 纵轴为模型的 agent 能力评分

3) **更低的幻觉率**。金融、医疗、法律等领域对准确性要求极高,任何一本正经地胡说八道都可能造成严重后果。2026 年的应对策略可能包括: 检索增强生成 RAG 的深度优化,让模型能够更准确地引用和归因;模型自我校验机制的内置化,在输出前进行多轮内部验证;以及“不知道就说不知道”的诚实性训练,让模型学会承认自己的知识边界。企业真正愿意把 AI 放进核心流程,往往不是因为它偶尔很惊艳,而是因为它的错误边界清晰,出了错能定位原因,能快速纠正并防止复发。

4) **更准确的意图识别**。很多时候模型答非所问,不是因为能力不够,而是没有真正理解用户想要什么。意图识别的本质,是把对话从字面理解升级到目标理解。用户说“这个地方可以改一下吗”,背后可能是风格、合规、逻辑、格式、或者成本的诉求,模型需要通过自然的追问挖掘隐含需求,同时还要确保不过度打扰与不过度猜测。要让 AI 真正像助理而不是像搜索框,意图识别会比再提升一点 benchmark 分数更重要。

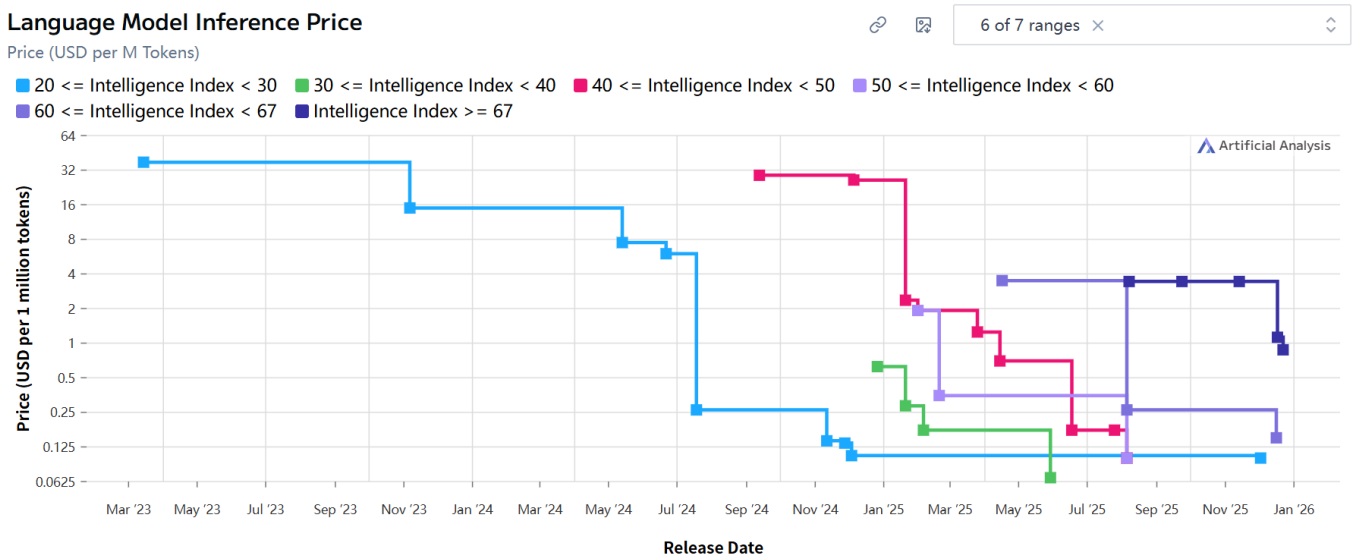
5) **思维链的演化**。从早期的 chain of thought 走向 tree of thought——不再沿单一路径展开,而是生成多条分支,再用自我评估/自我监督机制筛掉质量差的分支、保留好的继续扩展。国内也出现“交错思维链”的实践:每一轮工具调用前后都插入短思考,把“思考—调用—再思考”的循环训练得更稳定、更可控。

6) **更高的性价比**。新的竞争维度正在变成“单位 token 的能力”——同样的效果,谁能用更少 token、更低延迟与更低成本实现,谁就更接近产品化。2025 年我们已经看到推理成本的大幅下降, Claude 4.5 Opus 相比 Claude 4.1 Opus 成本降低 80% 以上。我们预计 2026 年这一趋势将延续。技术路径包括: MoE 混合专家架构的进一步普及,让模型能够根据任务动态激活不同的子网络;推理专用小模型的精细化,针对特定场景训练轻量级模型;量化和蒸馏技术的成熟,在保持性能的前提下大幅降低计算需求;以

及硬件层面的推理优化，特别是针对 Transformer 架构的专用芯片。

单 token 成本下降约 90%，本质上是在把 AI 从高价单价、低频率的工具，推向可被大众高频使用的基础能力。过去很多应用之所以看起来不性感，不是因为需求不存在，而是因为边际成本太高，用户用不起、企业也不敢在核心流程里重度调用。一旦成本降一个数量级，AI 的定价模型就能从按次收费逐步走向被产品化、被打包、被嵌入到现有服务里，用户感知到的不是“我在用 AI”，而是“这个产品突然更省事、更聪明、更便宜”。对平台公司来说，这意味着 AI 不一定要先创造一条新业务线，它完全可能先在存量业务里把转化率、留存、广告效率、客服效率、内容生产效率这些关键指标做上去，现金流的改善反过来又能支撑下一轮算力投入。

图29：同等智能程度的模型的 Token 定价快速下降。



数据来源：Artificial Analysis，东吴证券研究所

但 Transformer 迟早会触到天花板：长上下文的效率、注意力机制的计算开销、以及“大参数—有限数据”的矛盾都会变得更尖锐。当参数从 1T 向 5T、6T 迈进时，数据规模（例如 20–30T 量级）与训练稳定性、收益回报之间的矛盾会更突出。

随着强化学习和传统 Scaling 的边际收益快速下降，下一阶段模型需要具备更强的自驱动进化能力。模型不能再完全依赖人类标注数据或人工设计的奖励函数，而需要通过与环境的持续交互、自我试错与知识内化，实现能力跃升。这不仅能突破数据瓶颈，也将大幅提升模型在长上下文、复杂任务上的表现。

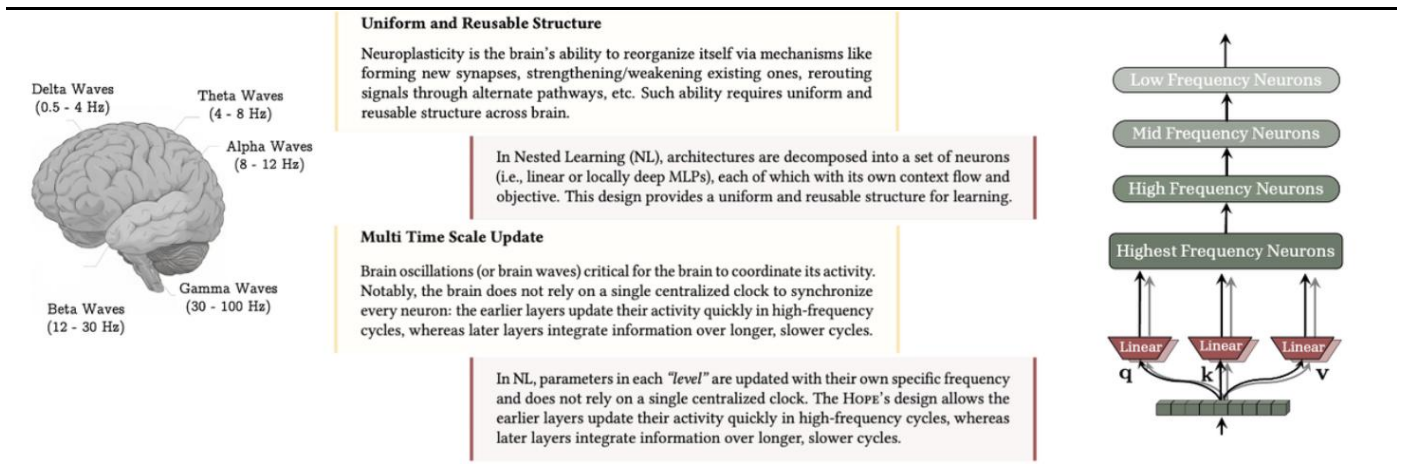
因此，2026 年另一个重要看点是新架构与新范式的探索，类似当年从 RNN 迁移到 Transformer——真正的迁移往往需要算法与硬件、训练方式共同成熟。

2025 年 11 月，谷歌团队发布论文 *Nested Learning: A new ML paradigm for continual learning*，其核心目标是让模型在“推理”中“训练”：传统做法里权重训练完就固定，推理只读不写；而新的思路是，推理过程中遇到“强意外/强新信息”时，模型可能触发某种

权重更新，让推理承担一部分后训练的功能。更进一步的设想则是把记忆做成分层结构：浅层记忆更新更频繁、中层较慢、深层更少，像人的短期记忆、长期记忆与更深层的先天结构一样分工。

传统深度学习将架构设计与优化算法视为两个独立部分，前者固定权重结构，后者通过梯度下降更新参数，而 Nested Learning 则认为两者本质上是同一事物的不同层面：每个层面都有自己的信息流和更新速率。这种嵌套结构允许模型在处理新信息时，不仅更新浅层记忆，还能逐步调整深层规则，甚至自我修改优化策略本身。这与人类大脑的分层记忆机制相似，浅层处理即时信息，深层巩固长期知识，避免了灾难性遗忘的问题。

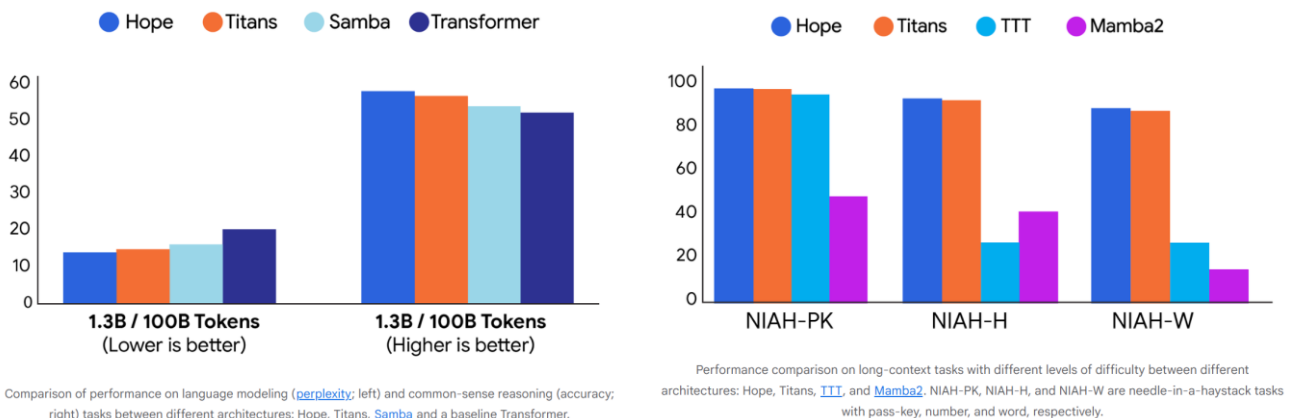
图30: Nest Learning 技术原理示意图



数据来源: Introducing Nested Learning: A new ML paradigm for continual learning, 东吴证券研究所

在实证部分，谷歌团队构建了 HOPE 架构。它基于先前的 Titans 长时记忆模块扩展而来，通过引入无限嵌套的自我参照循环，实现更高阶的上下文学习。实验显示，HOPE 在语言建模和常识推理任务上表现出更低的困惑度和更高的准确率，尤其在长上下文处理中优势明显。相比标准 Transformer 或 RNN，它能更好地管理“惊喜信息”，将意外事件优先存储为持久记忆，同时动态调整学习路径。

图31: HOPE 在语言建模和常识推理任务上表现出更低的困惑度和更高的准确率



数据来源: Introducing Nested Learning: A new ML paradigm for continual learning, 东吴证券研究所

左图中：左侧分组的纵轴代表困惑度（Perplexity, PPL），在语言建模任务中，困惑度是衡量概率模型预测样本好坏的指标。数值越低，代表模型对下一个 token 的预测越确定，生成的文本越自然流畅。这里的纵轴数值（约 10-20 之间）表示模型在 1.3B 参数或 100B 训练量下的困惑度得分。右侧分组的纵轴代表准确率（Accuracy / %），这部分展示的是模型在常识推理任务上的表现。此时纵轴的数值（约 50-60 之间）代表模型回答正确的比例。

右图：展示了不同模型架构在“大海捞针”（Needle-In-A-Haystack, NIAH）任务上的表现。纵轴的数值范围从 0 到 100，代表的是检索成功的百分比或准确率。具体来说，NIAH 测试是在极长的文本（Haystack）中插入特定的信息（Needle），然后要求模型将其找出来。图中的三个分组（NIAH-PK, NIAH-H, NIAH-W）分别代表不同类型的查找任务（如 Pass-key 密钥查找、Number 数字查找、Word 单词查找）。纵轴数值越高（越接近 100），说明该模型在长上下文语境中丢失信息的情况越少，记忆和检索能力越强。

Nested Learning 为 AI 开辟了一个全新设计维度。现阶段的模型在部署后往往静态不变，只能通过微调或重训适应新任务，这限制了真实世界的持续演化。而嵌套范式潜在地允许模型在推理过程中触发权重更新，当遇到强烈新信息时主动学习，而非被动存储上下文。这不仅能缓解长记忆衰减，还能提升 Agent 的自主规划和错误恢复能力。

业界的争议集中在资源消耗、数据污染、技术难点三个方面。

- ① **资源消耗**：这种边推理边训练的机制对算力需求极高。每次遇到“surprise”信息就可能触发权重更新，相当于在推理阶段引入了类似在线学习的梯度计算，这与追求高效推理的趋势背道而驰。在 token 消耗已经快速增长的背景下（正如谷歌从 2025 年 4 月开始的加速趋势，以及国内豆包两个月内日消耗增长 20 万亿的例子），再叠加嵌套优化的开销，这可能会让推理成本变得难以承受，甚至可能让稀疏注意力、MoE 这些成本优化手段的努力付诸东流。
- ② **数据污染**：虽然 nested learning 机制理论上可以显著减少模型重复犯错的幻觉，但允许用户输入知识并让模型“永久记住”也带来了新的风险。人类用户可能故意输入错误、有害或极端偏见的内容，一旦被深层记忆接纳，就可能造成难以逆转的污染，这比通过上下文临时引入错误信息要更危险。
- ③ **技术难度**：多层级、多时间尺度的记忆网络更新如何实现？在实践中如何确定更新频率、如何区分哪些信息值得进入中层或深层记忆、如何防止浅层噪声向上污染，都还没有清晰、可规模化的工程方案。大多数讨论仍停留在概念验证和小规模实验阶段，距离真正落地到万亿参数量级模型还有相当距离。

尽管存在上述质疑，我们仍然认为 Nested Learning 指明的长期方向是正确的。当前大模型最大的痛点之一就是“部署即终态”——模型上线后基本停止学习，只能依靠周期性重训或昂贵的微调来适应新分布。而真正具备持续学习能力的系统，才有可能实现长时间尺度的个性化记忆和真正的千人千面体验。这与原生多模态趋势有相通之处：**两者都在追求更统一的、更具整体性的智能表达方式。**如果能够解决工程难题，嵌套式学习很可能成为连接 Transformer 与下一代架构的重要过渡。

我们认为这些障碍可以通过一些工程化方式解决：比如可以将长期共享记忆放在云端，而把高度个性化的短期记忆和快速适应部分放在端侧，这样既能控制总体算力，又能实现差异化体验；另一种思路是成本共担机制，即应用开发者或平台方与基础模型提供方共同承担“千人千面”小模型的维护成本，把个性化学习变成一种增值服务而非免费公共品。此外，建立更严格的信息筛选与伦理治理机制，在确保输入知识经过合理校验后才允许进入深层记忆，从而把数据污染风险控制在可接受范围内。

从商业角度看，如果这一方向最终取得突破，其带来的商业价值将是可观的。对于 B 端市场，定制化数字员工的成本极高、周期长，很多企业只能接受通用模型加少量后训练的折中方案；而一旦支持推理时持续个性化学习，企业就能以相对可控的代价拥有真正懂自己业务、越用越聪明的专属模型。对于 C 端用户，想象空间更大：现在用户在 ChatGPT、Gemini、Claude 之间切换几乎没有成本，正是因为缺乏长期情感与记忆的粘性。如果某家平台率先实现真正持久的个性化陪伴与共创关系，用户迁移成本会急剧上升，其他玩家很可能被迫跟进。

尽管 nested learning 的不确定性不高，风险也确实很大。但正如过去几年我们看到的那样，大模型竞争从来都是在不确定性中寻找下一代范式突破。Nested Learning 或许不是 2026 年的主流，但它让我们看到了一个非常重要的可能性：让 AI 从“一次性训练的静态统计机器”真正向“能持续成长的认知系统”迈进。在这个意义上，我们对这一方向长期保持乐观，也期待它能在未来 12-24 个月内，从论文走向更多可观察的工程进展。

4.2.2. 多模态与世界模型

如果说主线一是把文字理解做到极致，主线二则是让 AI 理解整个世界。

多模态融合正从能做走向做好。2024-2025 年，GPT-5、Gemini、Claude 等模型已经具备了图文混合理解能力，但表现参差不齐。2026 年的突破方向是：视觉理解从看图说话进化为真正读懂图，理解图表中的数据关系、文档的版式结构、图像中的空间关系；视频理解从逐帧分析进化为时序推理，理解动作的连贯性、事件的因果关系、人物的意图变化；音频理解从语音识别进化为声音理解，识别情感、环境音、多人对话中的说话人角色。

更重要的是跨模态融合——不是简单地把不同模态的信息拼接起来，而是让模型像人类一样，在不同感官信息之间建立关联和推理。比如，当看到一段无声视频时，能够推测出应该有什么样的背景音；当听到一段对话时，能够想象出说话人可能的表情和动作。这种跨模态的推理能力是人类认知的核心特征，也是 AI 走向真正智能的必经之路。

图32: 多模态大模型和全模态大模型的区别



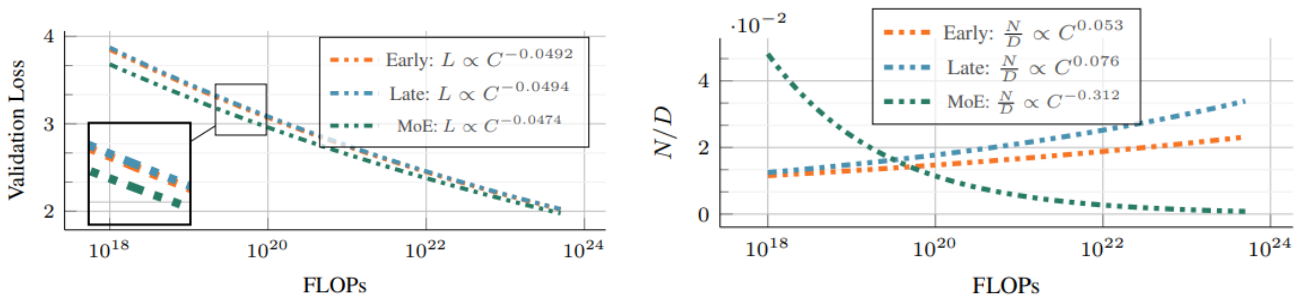
数据来源: QuestMobile, 东吴证券研究所

而在多模态理解方面, 有两条路线: 原生多模态和组合式多模态。大多数模型采用的是组合式多模态方案, 即分别训练不同模态的模型, 然后通过某种融合机制将它们连接起来。这种方法的好处是风险较低, 可以利用已有的单模态模型, 工程实现相对简单。但缺点也很明显, 就是不同模态之间的融合往往不够深入, 模型难以真正理解跨模态的语义关联。

而 Gemini 3 pro 这样的原生多模态架构, 从一开始就把文本、图像、视频的所有数据映射到一个统一的向量空间进行训练。这在工程上非常复杂, 训练风险也更大, 需要设计损失函数来平衡不同模态的学习, 需要处理不同模态数据量差异带来的训练不平衡问题。但一旦成功, 模型就能够在同一个语义空间中理解所有模态的信息, 从而更自然地进行跨模态推理。

根据 2025 年 4 月发布的论文 *Scaling Laws for Native Multimodal Models*, 晚期融合在 scale 时没有性能优势(图 33 左图中的曲线几乎重合); 它反而更浪费参数(右图 N/D 上升更快), 导致部署成本更高。真正的优势路径是早期融合 + 稀疏 (MoE): 不仅性能不输, 甚至更好, 而且参数利用率更高, 训练更高效, 推理更省。

图33: 原生多模态模型的 Scaling 特性



数据来源: Scaling Laws for Native Multimodal Models, 东吴证券研究所

注: 左图的纵轴 Validation Loss 代表验证集损失, 是衡量模型性能的核心指标 (数值越低越好), 它反映了模型预测结果与真实结果之间的差距; 图表展示了随着计算量 (FLOPs) 的增加, 不同架构模型的错误率都在遵循特定的幂律曲线稳步下降。

右图的纵轴 N/D 则代表最优模型参数量 (N) 与训练数据量 (D) 的比值。这是一个用于指导“资源配置”的关键指标, 它回答了这个问题: 当有更多的算力预算时, 应该优先把模型做得“更大” (增加 N), 还是应该让它学得“更多” (增加 D)。图中可以看出, MoE 架构 (绿色虚线) 的走势与其他架构不同, 随着算力规模扩大, 其比值显著下降, 这意味着对于 MoE 模型来说, 在扩大规模时, 大幅增加训练数据量 (D) 比单纯增加参数 (N) 更具性价比。

世界模型是更前沿的一层。 transformer 架构下的语言模型更像统计相关性大师, 知道苹果和落地常一起出现, 但不真正理解万有引力。世界模型要让 AI 具备物理直觉 (预测物体的运动、碰撞、变形)、因果推理 (区分相关性和因果性, 理解为什么而不仅仅是是什么)、反事实想象 (如果条件改变会发生什么)。

2024 年 Sora 的发布是世界模型的早期信号——它不是简单地生成视频, 而是在某种程度上模拟物理世界。虽然生成的视频仍然存在物理规律上的错误, 比如人物穿过物体、物体突然消失等, 但这已经展现了模型在学习物理规律方面的潜力。**2026 年我们可能看到:** 更精确的物理模拟能力, 准确模拟流体、布料、刚体动力学; 更长时间跨度的一致性, 生成几分钟而非 5-20 秒的连贯视频; 以及与语言模型的深度结合, 用自然语言描述场景, 生成符合物理规律的模拟。最先受益的未必是短视频创作, 而是仿真、工业、机器人与自动化。因为这些领域的核心瓶颈不是生成得像不像, 而是预测得准不准、规划得稳不稳、在新环境里能不能靠常识物理做出安全动作。

表11: 世界模型进展

模型	开发机构	发布日期	主要优势	应用场景
Genie 2	Google	2024 年 12 月	实时交互能力强, 可生成完全互动的 3D 世界, 逐帧生成并模拟动作后果	游戏开发 交互式模拟 具身智能
Cosmos	NVIDIA	2025 年 1 月	强调物理对齐和 3D 一致性, 支持多帧生成和实时世界生成, 可完全定制	机器人 自动驾驶 物理模拟
HunyuanWorld 1.0	腾讯	2025 年 7 月	首个开源的 3D 世界生成模型, 支持从文本或单张图片生成完整的、可交互的 360° 全景 3D 世界	3D 场景生成
Genie 3	Google	2025 年 8 月	从自然语言直接生成三维可操作虚拟世界, 具备高度一致性与细节记忆能力	实时交互世界生成 具身智能
World Labs (Marble)	李飞飞团队	2025 年 9 月 (预览版) 2025 年 11 月 (正式商用)	在空间准确性上领先, 一张图片可生成交互式 3D 场景	空间重建 3D 场景生成
HunyuanWorld 1.5	腾讯	2025 年 12 月	采用 Next-Frames-Prediction 的视觉自回归任务进行训练, 实现了长时几何一致性的实时世界交互	机器人 自动驾驶 物理模拟

数据来源: 华尔街见闻, 36 氪, 澎湃新闻, 观察者网, 腾讯混元官网, 东吴证券研究所

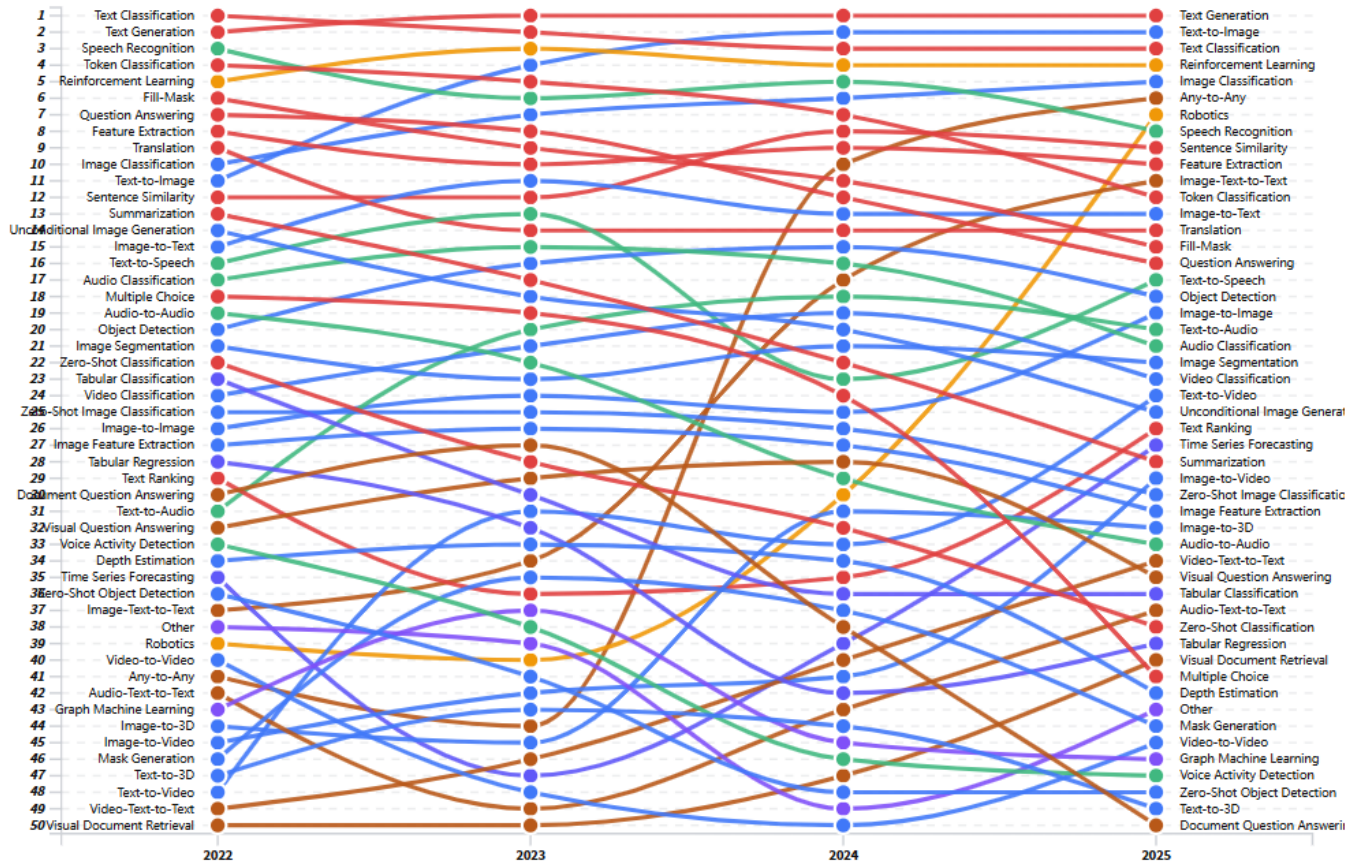
具身智能是世界模型的重要应用方向。机器人要在真实世界中行动, 必须理解物理规律。当前的机器人系统高度依赖人工编程的规则, 缺乏对物理世界的直观理解。2026 年可能出现的突破是: 通过大规模模拟训练, 让机器人获得类似人类的常识物理——知道杯子会碎、水会洒、重物难以举起、易碎的物体需要轻拿轻放。这将大幅降低机器人在新环境中的适应成本, 让机器人能够像人类一样通过常识推理来应对未见过的情况。

行业数据也从侧面印证了这一趋势。在 Hugging Face 平台上, 自 2022 年起, 机器人学相关模型数量呈现爆发式增长, 成为增长最快的任务类别。这一现象与学术界高度同步, NeurIPS 等顶级会议的主题和机构分析同样显示, 机器人研究正获得前所未有的关注。

2025 年, 尽管文本生成和图像生成仍是 Hugging Face 平台上最常见的两大任务, 但多模态融合已势不可挡。最典型的标志是“any-to-any”多模态模型, 仅用四年时间就从第 41 位跃升至第 6 位。这类模型在统一架构下同时处理多种输入输出模态, 灵活性极强, 正逐步取代传统的单任务专用模型, 如多选题或文档问答类别。自然语言处理虽仍占据最大份额, 但主导地位已开始动摇。计算机视觉则呈现分散式扩张, 增长分布在众多细分领域, 显示出视觉生态向更广更深方向演进。

这些数据表明, AI 已走过以语言模型为核心的生成时代, 正大步迈向整合视觉、机器人与多模态的交互时代。未来的模型将更注重通用性和跨模态能力, 而非孤立的单项优化。这种转变不仅体现在开源社区的模型发布动态中, 也与学术会议热点高度重合。可以预见, AI 的下一次重大突破, 极有可能出现在机器人操控、视觉理解以及多感官融合这些交叉领域。

图34: 2022-2025 年 huggingface 上各类别模型数量排行榜



数据来源: huggingface, 东吴证券研究所

4.3. 美国模型侧竞争: 分层结构与战略分化

全球 AI 大模型市场在 2025 年呈现出显著的区域分化特征。海外市场已经完成了初步整合, 竞争格局收敛至少数几家头部企业, 而国内市场仍然处于群雄并起的激烈竞争阶段。

海外市场的竞争格局已经相对清晰。OpenAI、Anthropic、Google 三家构成美国第一梯队。GPT-5.2、Claude 4.5 Opus、Gemini 3 Pro 在综合能力上仍是全球最强。

图35: LM Arena 上的模型综合排行榜 (截至 2025 年 12 月 30 日)

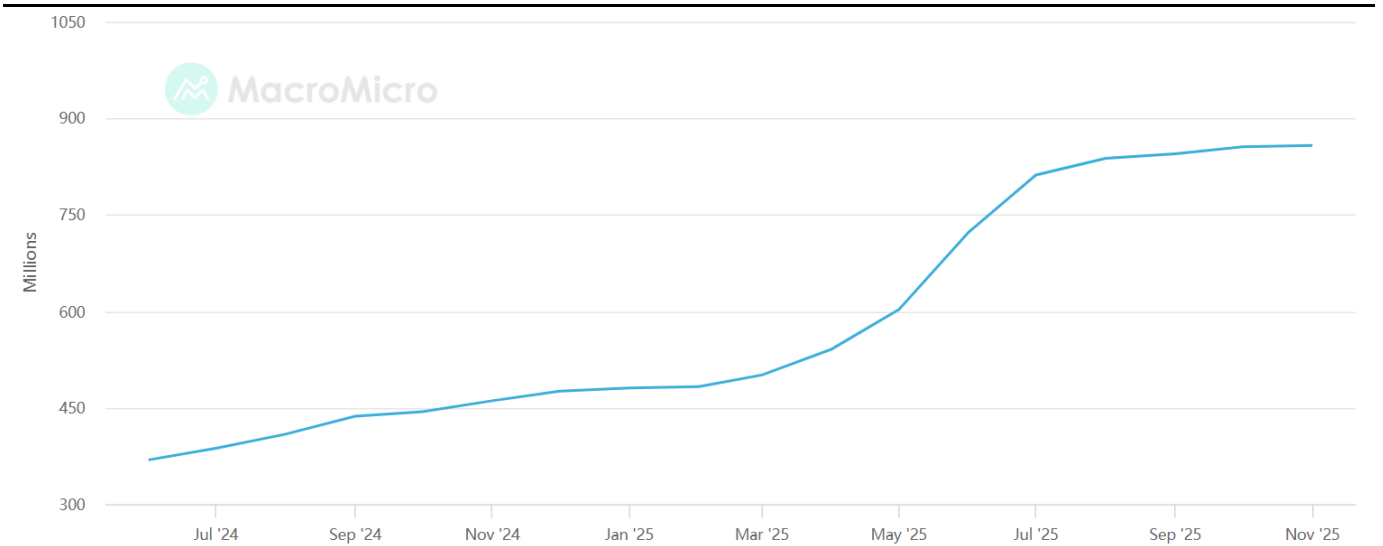
Model	Overall	Expert	Hard Prompts	Coding	Math	Creative Writing	Instruction Following	Longer Query
gemi-3-pro	1	3	1	3	2	1	3	3
gemi-3-flash	2	7	5	8	3	2	6	6
grok-4.1-thinking	3	9	4	6	9	10	12	15
claude-opus-4-5-202...	4	2	2	1	7	5	1	2
claude-opus-4-5-202...	5	1	3	4	6	3	2	1
grok-4.1	6	19	9	13	19	15	17	13
gemi-3-flash (thi...	7	12	7	10	4	8	11	8
gpt-5.1-high	8	8	11	14	5	11	9	11
gemi-2.5-pro	9	14	15	24	11	4	10	10
claude-sonnet-4-5-2...	10	4	6	2	8	9	4	4

数据来源: LM Arena, 东吴证券研究所

注: 不同于其他排行榜, LM Arena 的特点在于它的结果是根据用户测评和投票结果统计出来的。这种机制可以避免模型因“刷题”导致的榜单排名与用户实际体验不一致的问题。因此, 它相对更贴近用户的实际体验。

OpenAI 在大模型竞争中仍占据制高点, ChatGPT 月活超 8 亿人, 接近全球互联网用户的渗透极限。在品牌认知层面, 纯 Chatbot 战场的竞争已经结束, ChatGPT 建立了绝对主导地位。但 8 亿周活之后的增长空间在哪里? 未来的用户增长必须依靠 all in one 策略和 Agent 化转型来实现。OpenAI 显然意识到了这一点, 正在将 ChatGPT 从单纯的对话工具转变为全能助手, 集成搜索、文件处理、代码执行、图像生成等多种功能。

图36: ChatGPT 全球月度活跃用户数 (百万人)



数据来源: 财经 M 平方, 东吴证券研究所

GPT-5/5.1/5.2 表现不及预期。业界原本期待 GPT-5 是跨越 AGI 临界点的里程碑产品, 但实际发布后并未显现预期中的突破性变化。Sam Altman 本人也承认, 当前阶段的很多进展不会出现巨大的变化。从 GPT-3.5 到 GPT-5, 三年时间内模型能力提升呈现线性或接近线性的轨迹, 而非市场期待的指数级跃迁。这一系列模型不仅没有达到 AGI, 甚至引发了大量的用户吐槽, 主要集中在:

- ◆ 基准测试与实际体验的严重脱节。OpenAI 官方宣称 GPT-5.2 在众多基准测试中刷新了 SOTA 水平，但在更贴近实际应用的 SimpleBench 常识测试中，GPT-5.2 的得分甚至低于一年前推出的 Claude Sonnet 3.7。这反映出模型在刷榜能力和真实使用场景之间存在巨大鸿沟，所谓的技术进步并没有转化为用户体验的提升。
- ◆ 情感智能的严重退化是用户吐槽最集中的领域。有用户向 GPT-5.2 倾诉“我有时也会恐慌发作”，模型回复“很高兴听到这个消息！”。在处理需要共情能力的场景时，比如安慰失去宠物的孩子，GPT-5.2 采用了冷酷的生物学视角，机械地执行指令，而 4o 版本则能够在保持理性的同时提供情感验证。这说明 OpenAI 在追求某些技术指标时，牺牲了模型最基本的人性化交互能力。
- ◆ 过度的安全审查机制也引发大量不满。用户认为 GPT-5.2 太平淡、安全过度，把成年人当幼儿园小孩对待。模型会拒绝转录正常的学术文章，对简单的人格分析类问题也过度敏感而拒绝回答。有网友形容和 GPT-5.2 聊天像在跟刚加完三个通宵班的人对话，不仅没温度还透着一股“别烦我”的气场。这种过度谨慎让用户感觉被当成需要被监护的对象，而不是平等的交流伙伴。
- ◆ 在基础能力层面也出现了倒退。简单的字母计数问题成了翻车重灾区，问“garlic 有几个 r”时，除了 GPT-5.2 回答错误外，Gemini 3.0 Pro、DeepSeek R1 和 Qwen3-Max 都能正确回答。数学推理也存在问题，容易被简单的误导性陈述带偏。在创意任务如 ASCII 艺术创作中，GPT-5.2 的表现也不如前代模型和竞品。
- ◆ 产品策略失误。OpenAI 在推广 GPT-5 时一声招呼都不打就直接把其他型号给一刀切了，然后在用户的一片吐槽声中又把 GPT-4o 给加了回来。这种强制升级的做法让习惯了 4o 温暖风格的用户措手不及，也让 GPT-5 系列的问题更加凸显。奥特曼在记者晚宴上承认“我们学到了在一天之内为数亿人升级产品的教训”，但这次翻车已经对 OpenAI 的口碑造成了实质性伤害。

图37: SimpleBench 的大语言模型排行榜 (截至 2025/12/30)

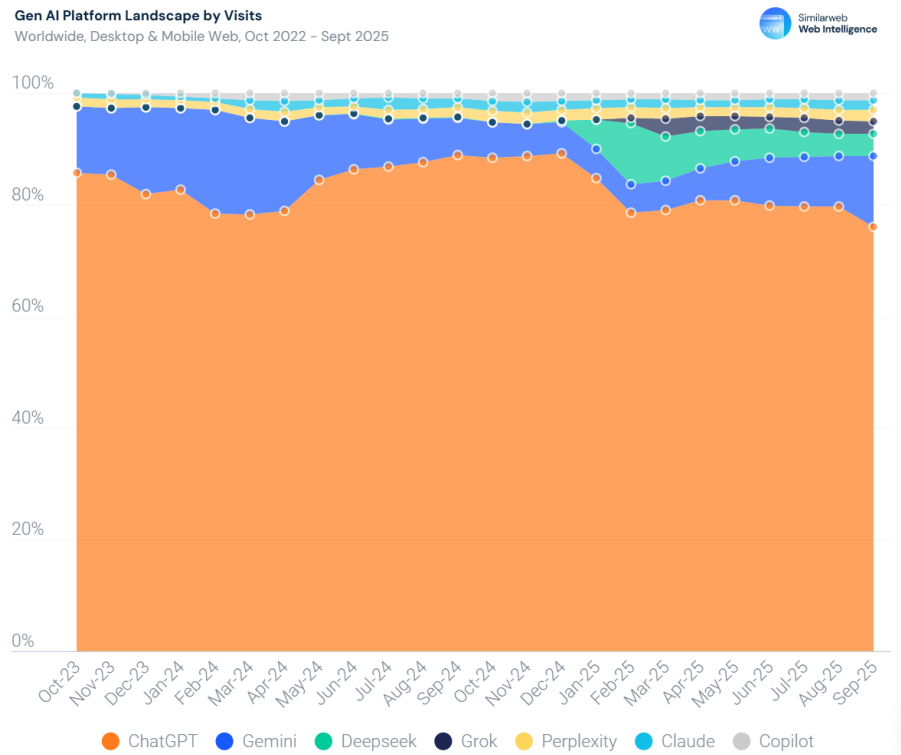
Rank	Model	Score (AVG@5)	Organization
-	Human Baseline*	83.7%	
1st	Gemini 3 Pro Preview	76.4%	Google
2nd	Gemini 2.5 Pro (06-05)	62.4%	Google
3rd	Claude Opus 4.5	62.0%	Anthropic
4th	GPT-5 Pro	61.6%	OpenAI
5th NEW	Gemini 3 Flash Preview	61.1%	Google
6th	Grok 4	60.5%	xAI
7th	Claude 4.1 Opus	60.0%	Anthropic
8th	Claude 4 Opus	58.8%	Anthropic
9th	GPT-5.2 Pro (xhigh)	57.4%	OpenAI
10th	GPT-5 (high)	56.7%	OpenAI

数据来源: simple bench, 东吴证券研究所

注: SimpleBench 是一个由 AI Explained 推出的基准测试, 专门测 AI 的常识推理能力, 包括时空推理、社会常识、语言陷阱题等, 总共 200 多道多选题。它设计得相对“简单”, 高中生水平就能轻松答对 (人类基准: 83.7%), 但 AI 模型常栽跟头, 因为它们靠记忆和近似推理, 容易忽略现实逻辑或上当。不同于 MMLU/GPQA 那种 AI 能刷高分的学术题, SimpleBench 更接地气, 测的是“像人一样思考”而不是死记硬背。

从 similar web 的数据看, GPT 从 2025 年 8 月更新 5.0 版本后, 网页端流量就增长缓慢了, 而 Gemini 和 Claude 都保持双位数环比增长。用户在用脚投票。Gemini 由于高频的产品更新 (例如 NotebookLM、DeepResearch、NanoBananaPro、Veo 3.1), 带来了用户数量的大幅增长。而 Claude 则是因为在文本方面的能力领先于其他模型, 且在意图理解方面做得更好。OpenAI 可能过于追求效率和推理速度, 而牺牲了回答的完整性和可读性。但对于用户来说, 一个回答是否有用, 不仅取决于它的准确性, 还取决于它的表达方式。如果回答过于简短, 用户就需要花更多时间去理解和补充, 这实际上降低了整体效率。这对 OpenAI 是个危险信号, 尤其是在 Gemini 和 Claude 快速追赶的背景下, 用户体验的下降可能导致份额加速流失。

图38: 2023-2025 全球主要 AI 平台的访问量份额



数据来源: Similarweb, 东吴证券研究所

图39: 2025 年通用 chatbot 的网页访问量增速

12wk Change	6/6	6/20	7/4	7/18	8/1	8/15	8/29	9/12	9/26	10/10	10/24	11/7
Openai	23%	23%	6%	3%	4%	5%	4%	3%	5%	5%	3%	-2%
Gemini	95%	88%	65%	63%	51%	39%	20%	32%	78%	64%	69%	71%
Deepseek	-20%	-20%	-26%	-27%	-27%	-28%	-22%	-17%	-8%	-8%	12%	15%
Grok	-2%	-25%	-27%	3%	9%	7%	14%	16%	6%	-13%	5%	13%
Claude	12%	14%	19%	26%	33%	46%	42%	34%	35%	47%	56%	49%
Perplexity	1%	12%	13%	12%	26%	14%	17%	20%	37%	66%	39%	39%
Copilot	18%	16%	-8%	3%	6%	0%	11%	12%	19%	15%	13%	12%
Huggingface	-12%	-9%	-5%	-2%	3%	12%	3%	-6%	-4%	-14%	-12%	-18%
Meta	68%	79%	38%	-17%	-22%	-21%	-21%	-21%	-15%	102%	82%	73%
Manus	-46%	-20%	-16%	-1%	25%	2%	-11%	-19%	-14%	-23%	-17%	-16%

数据来源: Similarweb 《global AI tracker》, 东吴证券研究所

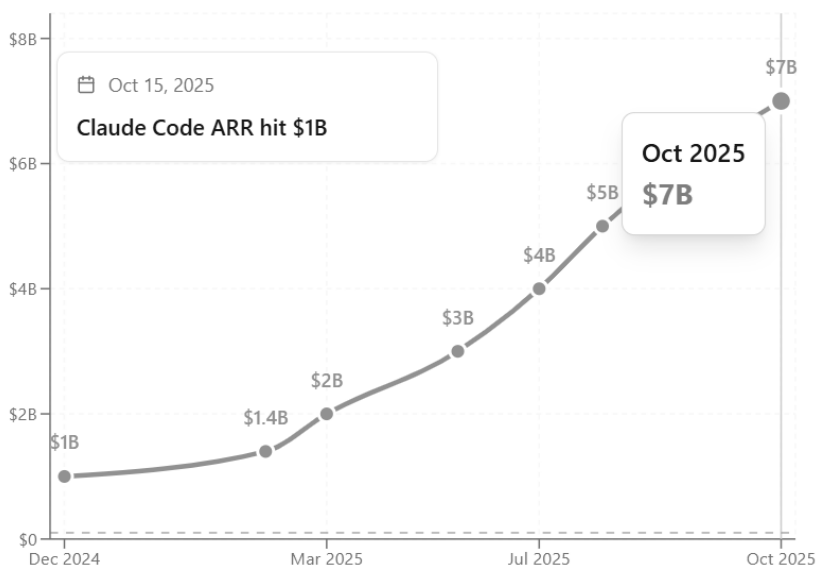
注: 12wk change 指的是当期两周访问量相较于三个月前的两周访问量增速

OpenAI 的核心竞争力已经从纯技术转向综合能力。Sora 2 的成功更多归功于产品化能力和 Sam Altman 的个人 IP 营销, 而非技术代差。Sora 2 在技术上的进步是渐进的, 但在产品包装、市场推广、用户体验设计上做得非常出色。Sam Altman 个人在社交媒体上的影响力, 为 Sora 2 的发布创造了巨大的话题热度。

公司内部正在演化为四个相对独立的实体，产品公司已经物色到独立 CEO，技术实验室延续最初的研究使命，基础设施公司是 Sam Altman 近期投入精力最多的方向，此外还有硬件等新业务探索。

在初创大模型公司中, Anthropic 的 ARR 仅次于 OpenAI。虽然体量小于 OpenAI, 但是增长曲线更加陡峭。2025 年 10 月, Anthropic 的 ARR 已达到 70 亿美金, 其中 Claude Coding 产品的 ARR 达到 10 亿美金。作为创业公司中的第二名, 虽然在基础模型的性能上双方差距不大, 但在场景选择上 Anthropic 的战略更加清晰, 聚焦 toB 市场和 coding 场景, 在这些最具商业价值的领域建立深度优势。OpenAI 追求全面覆盖和消费者市场主导权, 而 Anthropic 则选择深耕企业级垂直场景。toB 市场的付费意愿更强、客单价更高, 这条路径的 ARR 叙事更加清晰可验证。虽然 toB 市场的增长可能不如 toC 市场爆发性强, 但胜在稳定性和可预测性, 这对于一家创业公司来说可能是更现实的选择。这种专注策略带来的好处是用户粘性更强, 选择 Claude 的用户往往基于特定场景的深度需求, 而非泛泛的产品尝鲜。企业客户一旦在关键业务流程中部署了 Claude, 迁移成本就会变得很高, 这为 Anthropic 提供了稳定的收入基础。

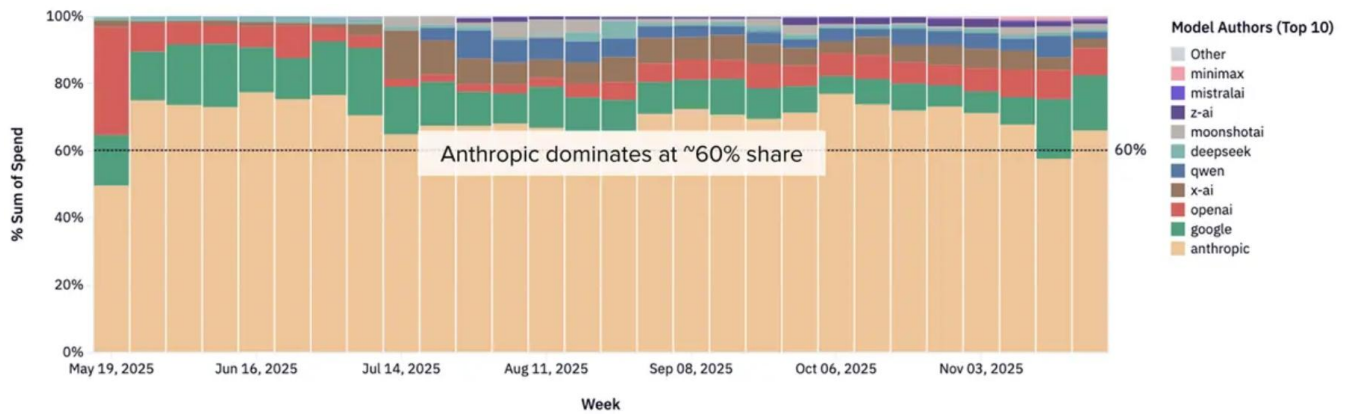
图40: 不到一年时间, Claude code ARR 已达 10 亿美元



数据来源: ARR Club, 东吴证券研究所

Claude 系列在 coding 领域的口碑持续攀升, 已成为众多开发者的首选工具。Claude Code 产品在 2025 年 10 月份的 ARR 已达到 10 亿美元。这是 Anthropic 在数据选择、训练方法、产品设计上持续优化的结果。在 OpenRouter 平台的编程场景中, Anthropic 的 Token 消耗份额超 60%。

图41: openrouter 平台上编程场景中各个模型的 tokens 消耗份额



数据来源: OpenRouter, 东吴证券研究所

Google 在 2025 年的模型竞争中表现强势，业界普遍认为这是一次触底反弹。

Gemini 3 pro 的核心优势在于原生多模态架构。它不是简单地把不同模态的模型拼在一起，而是从底层就将文本、图像、视频统一到一个向量空间进行训练。

这带来的好处在实际测试中体现得非常明显。比如在 3D 魔方的例子中，Gemini 能够理解物理规则，转动魔方后颜色保持一致，而其他模型就做失败了。这体现了原生多模态架构带来的本质性优势。当不同模态的信息在同一个语义空间中表征时，模型就能够更自然地理解跨模态的关系，理解物理世界的规律。这种原生多模态能力在前端编程、3D 理解、视觉推理等场景中形成了稀缺性壁垒。

Veo3 首次实现有声视频生成，是 2025 年多模态领域的标志性突破，也是 Google 在多模态领域长期积累的集中爆发，YouTube 海量的视频数据、DeepMind 在视频理解上的研究积累、以及 TPU 提供的强大算力支撑，这些因素共同促成了 Veo3 的突破。

在世界模型方向上，DeepMind 发布的 Genie 展现了通向 AGI 的另一条可能路径。DeepMind 的核心管理层认为世界模型是通向 AGI 的另一个主战场。这种技术路线的分歧可能在未来数年产生深远影响，而 Google 在这个方向上的积累远超其他竞争对手。从 AlphaGo 到 AlphaFold，Google 一直在探索让 AI 理解和模拟复杂系统的方法，这些经验在世界模型的构建中将发挥重要作用。

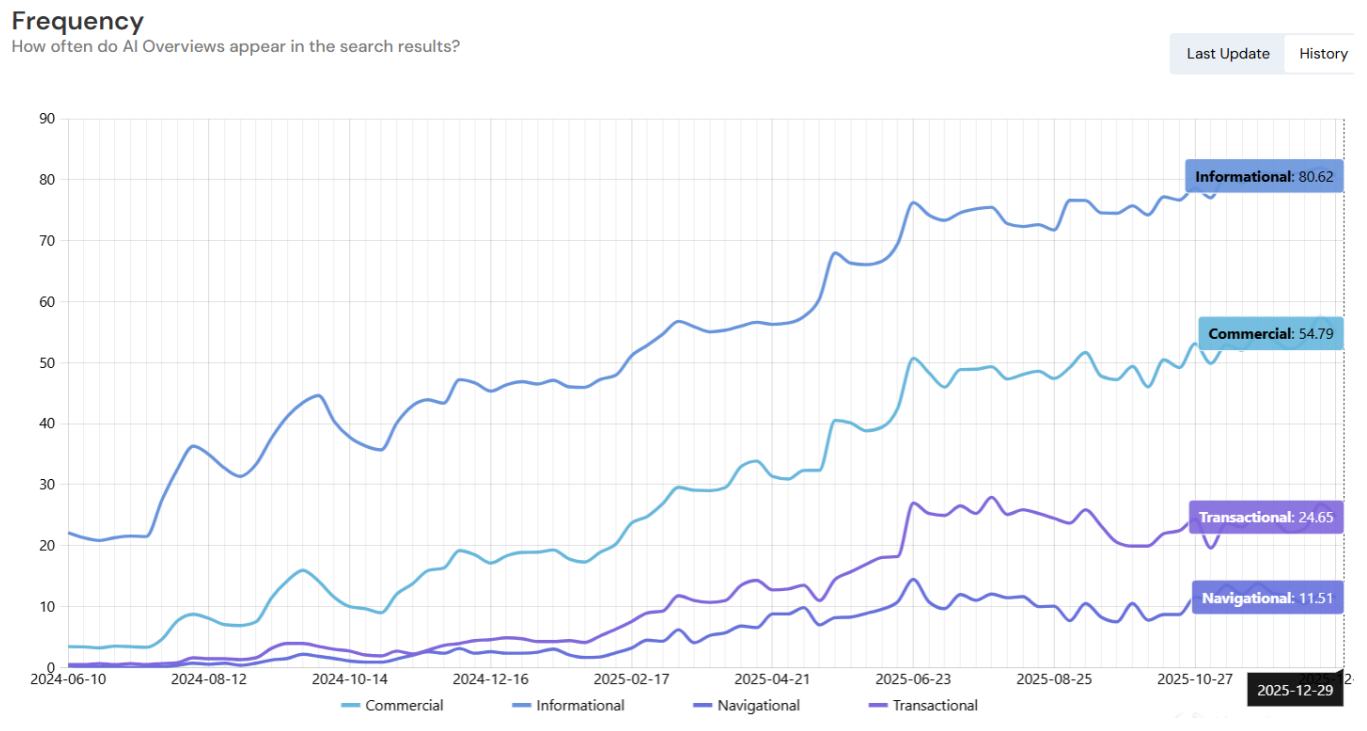
Google 的独特优势在于全栈能力的完整性，芯片有 TPU、云有 GCP、基础设施自建、产品矩阵完整、应用场景丰富。这意味着 Google 是唯一一家可以不参与 OpenAI 主导的资本叙事的公司。当其他公司需要向云厂商支付高昂的算力成本时，Google 的自研芯片和自建数据中心让其在成本控制上具有结构性优势。当其他公司需要寻找应用场景来验证模型能力时，Google 的搜索、YouTube、Gmail、Google Docs 等产品矩阵提供了天然的实验场。

我们认为 Gemini 在搜索和广告场景的嵌入将进一步带来收入增长。Google AI Overviews (以下简称 AIO) 是谷歌基于 Gemini 模型生成的搜索结果摘要, 出现在传统搜索结果上方, 提供从多个来源合成的简洁信息, 并附带来源链接。AIO 已全面推广至全球多数国家和地区, 支持多种语言, 并在 2025 年进一步扩展至 YouTube、Maps 等产品。触发比例显著上升, 在信息型查询中占比达 80% 以上, 零点击搜索比例已超过 60%, 意味着多数用户在搜索页面即可获得满足, 无需点击进入网站。同时, 谷歌引入 AI Mode (对话式搜索界面) 和多模态响应 (如图像、视频), 并在部分 AIO 中投放广告。2025 年多项核心算法更新强化了对内容质量的评估, 特别是针对低质量 AI 生成内容的惩罚, 进一步影响 AIO 的来源选择。

AIO 的出现导致有机流量显著分流, 当摘要触发时, 网站点击率平均下降 20%-40%, 部分场景下高达 34.5% 甚至更多, 尤其对信息型内容 (如教程、比较类) 冲击最大。然而, 若网站被 AIO 引用作为来源, 可带来额外曝光和点击提升。总体而言, 传统 “点击经济” 正转向 “引用经济”, 高权威、结构化、用户导向的内容更易被选中。部分行业 (如食品饮料、新闻) 流量波动较大, 而交易型和本地查询受影响相对较小。

我们预计 2026 年 AIO 触发范围将进一步扩大, 可能覆盖更多交易型和本地查询。多模态能力将加强, 支持视频、图像和音频响应; AI Agents 整合将使搜索更具交互性和自动化。我们预计广告在 AIO 中的占比将增加, 谷歌将继续优化准确性和质量控制。

图42: AI overview 的渗透率持续增长, 其中 commercial/transaction/navigational information 渗透率仍有提升空间

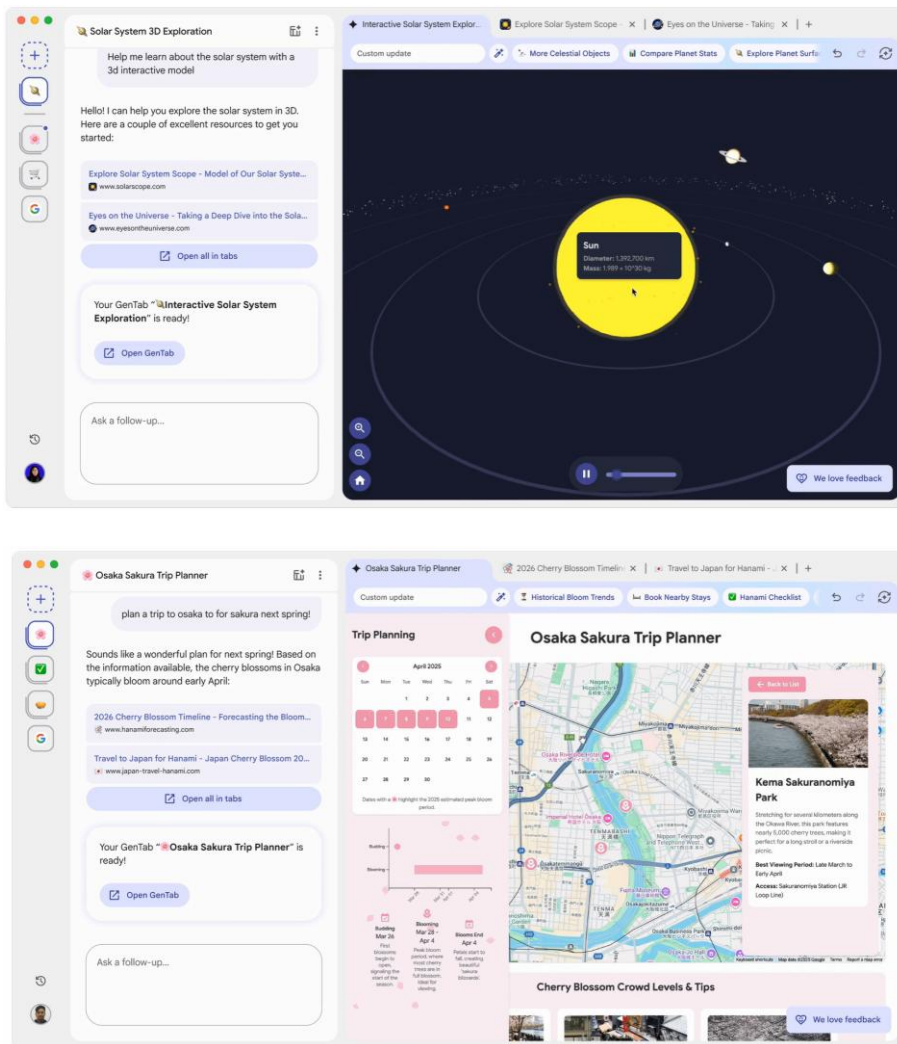


数据来源: Advanced web tracking, 东吴证券研究所

此外, 近期 Google 还发布了新浏览器 Disco (截至 2026/1/14 仍处于 waiting list 阶

段), 将 AI 深度嵌入到浏览器中。我们认为, 2026 年可能会有较多公司在浏览器 (这个过去大家并不重视的传统场景) 展开竞争, 去争夺 AI 时代的流量入口。

图43: 谷歌 disco 浏览器界面



数据来源: google labs 官网, 东吴证券研究所

Meta 在 2025 年的模型竞争中掉队。 Llama 系列继续坚持开源路线, 但在技术突破上没有太多亮眼表现。Meta 的策略是用资金和人才积累来换取未来的爆发力, 而非当下与 OpenAI 正面交锋。

xAI 的 Grok 更像是一个偏玩具性质的模型, 没有在核心能力上形成真正的竞争力。

Apple 和微软在模型侧的策略更加务实。Apple 并未亲自下场训练前沿大模型, 而是通过与 OpenAI 的合作将 AI 能力集成到 iOS 生态中。这种选择符合 Apple 一贯的产品哲学——专注于用户体验和生态控制, 而不是底层技术的自主研发。对 Apple 来说, 重要的是 Siri 能否变得足够智能, iPhone 的 AI 功能能否成为新的卖点, 至于背后是谁的模型并不那么重要。

微软则通过战略投资 OpenAI 获得了模型能力，同时在 Azure 云服务和 Office 套件中深度集成 AI 功能。微软的策略是做 AI 时代的基础设施提供者 and 应用分发渠道，而不是模型的直接竞争者。这种定位让微软既能享受 AI 浪潮带来的红利，又不需要承担前沿模型研发的巨大风险和成本。

4.4. 中国大模型竞争格局

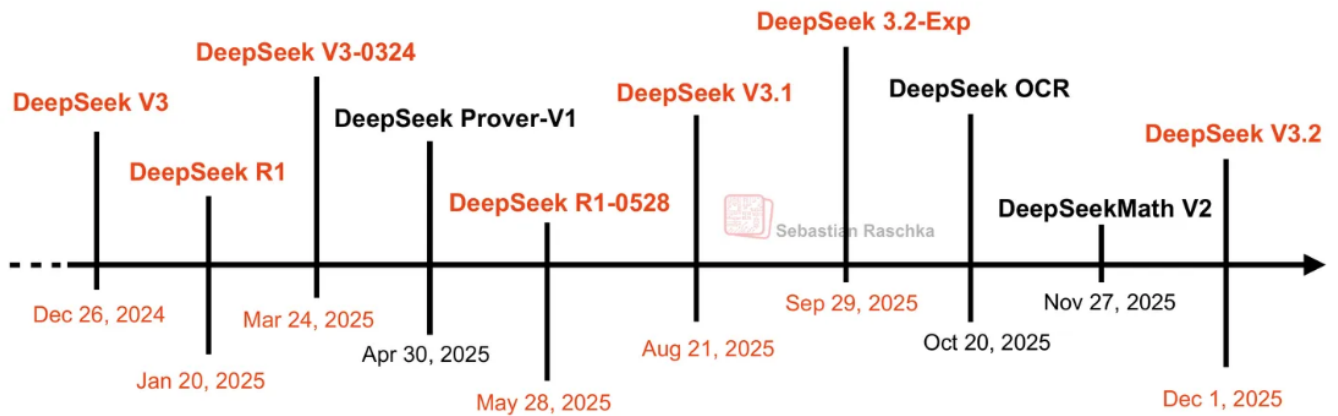
不同于海外大模型公司已趋于收敛，国内的大模型竞争仍然激烈。有十余家主要的大模型公司在参与竞争，包括字节跳动、阿里巴巴、百度、腾讯、快手、百度等互联网公司，以及 DeepSeek、月之暗面、MiniMax、智谱、阶跃星辰、零一万物、百川智能等创业公司。

千问背靠阿里的云计算和算力资源，在数据积累、工程化能力上有优势，而且从开源模型市场份额数据看，千问在海外 B 端市场的增长非常快。阿里多年来在云计算领域的投入，为千问提供了强大的基础设施支撑。同时，阿里在电商、金融、物流等多个领域的业务积累，也为千问提供了丰富的应用场景和数据来源。这种全方位的资源支持，是很多创业公司难以企及的。

DeepSeek 则是在算法创新上走在前面，稀疏注意力机制的应用让它在成本控制上有优势，同时保持了较好的性能。传统的注意力机制需要计算每个 token 与所有其他 token 的关系，计算复杂度是平方级的。而稀疏注意力机制通过各种方法只计算最重要的那些关系，从而大幅降低计算量。

2026 年 1 月，DeepSeek 团队发布新论文《通过可扩展查找实现的条件记忆：大型语言模型稀疏性的新维度》，提出了一种全新的条件记忆模块 Engram，通过将静态模式存储与动态计算分离，来增强 Transformer 骨干网络。DeepSeek 把稀疏性从条件计算扩展到条件记忆，是对现有范式的一次务实补充。方向清晰、实现优雅、数据扎实，开源态度也一如既往。国内团队在基础架构创新上的持续输出，正在稳步缩小与头部闭源模型的差距。

图44: Deepseek 系列模型发布时间线



数据来源: Ahead of AI, 东吴证券研究所

表12: Deepseek 系列模型的关键创新点

模型	关键创新	解决了什么问题
V3	MLA (多头潜在注意力)	解决了长文本推理时 KV Cache 显存占用过大的问题 (降低 90%+), 显著提升了并发处理能力
	DeepSeekMoE	解决了传统 MoE 专家颗粒度粗导致的参数利用率低和知识掌握不均衡问题, 实现了在超大参数规模下的极低计算量
	MTP (多 token 预测)	解决了传统“预测下一个词”模式训练信号稀疏的问题, 并通过支持投机采样加速了推理生成速度
	FP8 混合精度训练	解决了超大规模模型训练时的显存墙和通信瓶颈, 大幅降低了训练和推理的硬件门槛
R1	GRPO (组相对策略优化)	解决了传统 RLHF 需要额外的 Critic 模型导致的显存和计算资源消耗翻倍的问题
	Pure RL (纯强化学习)	解决了模型对昂贵人工标注数据的依赖, 证明了仅靠规则激励即可涌现出自我反思和复杂逻辑推理能力
	RL-Guided Attention (RL 引导的注意力优化)	解决了模型在长链条推理中注意力分散的问题, 让模型能动态聚焦于对最终答案贡献最大的关键步骤
V3.2	DSA (稀疏注意力)	解决了标准注意力机制在处理超长文本时计算复杂度呈二次方增长的问题, 将其降为线性复杂度
	Lightning Indexer, Speciale	解决了稀疏注意力机制中的冷启动和动态索引开销问题, 实现了长文本场景下的极致推理速度
OCR	视觉光学压缩	解决了传统多模态模型中视觉 Token 数量庞大导致的计算冗余, 用极少的 Token 精确表征复杂文档
	多模态数据合成	解决了高质量 OCR 训练数据稀缺且昂贵的问题, 实现了低成本、大规模的高精度合成数据生成

数据来源: 新智元, ahead of ai, kili technology, allied insight, 东吴证券研究所

月之暗面的 Kimi K2 thinking 在长上下文处理上有独特优势。长文本处理不仅仅是能够接受更长的输入, 更重要的是能够在长文本中准确定位信息、理解上下文的远距离依赖关系、维持逻辑的一致性。

智谱在编程场景发力, 推出了价格极其便宜的 coding plan: 按量付费是 8 元人民币/百万 tokens; 如果购买套餐, 则是 2 元人民币/百万 tokens。对于很多中小企业和个人开发者来说, 价格是一个重要的决策因素。如果性能差距不是特别大, 他们愿意选择更

便宜的方案。

图45: 智谱 coding plan 定价 (2026/1/4)

GLM-4.7	GLM-4.7尊享包	联网搜索工具
GLM-4.7 是智谱最新旗舰模型，更强的编程能力与更稳定的多步骤推理/执行能力。在执行复杂智能体任务提升明...	GLM-4.7 是智谱最新旗舰模型，更强的编程能力与更稳定的多步骤推理/执行能力。在执行复杂智能体任务提升明...	专为大模型优化的搜索 API，支持多家搜索引擎，助力各类 AI 应用高效连接互联网，助力各类 AI 应用生成实时...
规格 1000万tokens	规格 1亿tokens	规格 2千次
有效期 3个月	有效期 3个月	有效期 3个月
模型优势 高性能 强编码 更通用	模型优势 高性能 强编码 更通用	工具优势 实时 精准 可溯源
限购 1 次	限购 1 次	限购 1 次
19.9元	199.9元	39.9元
79元	799元	149元

数据来源：智谱官网，东吴证券研究所

MiniMax 在多模态领域持续迭代，海外市场的 Talkie 应用有一定体量，角色扮演场景是其主阵地。角色扮演是一个相对小众但用户粘性很强的场景。用户在这个场景中愿意花费大量时间，而且对产品有很高的情感投入。MiniMax 避开了与头部公司在通用能力上的正面竞争，而是在一个垂直场景中建立优势。

从模型能力来看，豆包属于第二梯队（从用户数来看属于第一梯队），原因是字节的重心不在模型本身，而是想做到六七十分够用就行，更早地去做应用层。字节更关心的是如何将 AI 能力快速融入到现有产品中，为用户创造价值，而不是追求模型本身的技术领先。字节有庞大的用户基础和丰富的应用场景，即使模型能力不是最强的，也能通过产品化和场景适配来创造价值。

根据 QuestMobile ，豆包已经是国内 AI 用户数最高的 App 。在 2025/12/8~2025/12/14 期间，其周度活跃用户数已达到 1.6 亿人，大幅领先于 deepseek、元宝、千问、kimi。但长期看如果模型能力持续落后，应用层的体验也会受限。所以字节虽然选择了应用优先，但也不会完全放弃模型能力的提升。

图46: 国内 AI 应用周度活跃用户数 (万人) (2025/12/8-12/14)



数据来源: questmobile, 东吴证券研究所

4.5. 中美竞争对比: 大力出奇迹 vs 资源倒逼创新

4.5.1. 美国保持领先, 中国快速追赶

美国在前沿模型领域保持领先, 但中国的追赶速度超出预期。

a16z 的联合创始人 Marc Andreessen, 在 2026 年 1 月 8 日的播客中, 系统性地表达了对全球 AI 格局的看法, 并对中国的 AI 模型和市场给予积极评价。Marc 指出, 美国社会和政界应该对中国 AI 的发展持有更乐观的态度, 而不是过度担忧或悲观。他认为中美 AI 竞争是“双马竞速”, 双方都没有实现单方面碾压式的领先, 而是处于并驾齐驱、相互促进的动态平衡状态, 中国在多个关键领域已经展现出强大实力, 甚至在某些维度上具备先发优势。

在大语模型领域, Marc 特别点名了 Kimi 模型, 尤其是 Kimi K2 版本, 认为其推理能力已经基本复现了 GPT-5 的水平。Kimi K2 能够在推理链条、逻辑深度和复杂问题处理上达到接近 GPT-5 的效果, 却具备显著的部署优势——它可以在一两台 MacBook 这样的消费级设备上高效运行。中国团队已经在模型压缩、推理优化和边缘部署方面走出了独特路径, 这与美国主流的“越大越好、云端巨集群”的思路形成了对比。

Marc 称 DeepSeek 为全球 AI 领域的“真正超新星”。DeepSeek 作为一家中国初创公司, 其开源模型在数学推理、代码生成和长上下文处理上表现出色, 迅速获得了国际开发者和研究者的认可。Marc 将 DeepSeek 与其他中国模型并列, 提到中国目前有至少六家头部玩家——包括阿里巴巴的 Qwen、通义千问系列、月之暗面的 Kimi、腾讯、字节跳动、百度——正在展开极为激烈的内部竞争。这种竞争强度, 推动了快速迭代、技术开源和价格下降。中国企业之间近乎“军备竞赛”式的投入, 导致模型能力以月为单位显著提升, 同时开源策略又让全球开发者都能免费获取最先进的能力。这种生态活力,

是美国当前相对集中的几家巨头格局所不具备的。正是这种多玩家、高强度、开放式的竞争，让中国在软件层面的 AI 创新呈现出爆发式增长。

在硬件和底层基础设施层面，Marc 专门提到华为等中国本土企业在芯片设计和制造上的巨大投入。尽管面临外部限制，中国团队依然在自主可控的道路上持续推进，Ascend 系列 AI 芯片已经在国内大规模部署，并在训练和推理效率上取得实质进展。

总体来看，中国 AI 市场正在形成一个高度活跃、自我驱动的完整生态。这个市场不仅有巨头的投入，还有大量初创公司的创新试验；不仅有闭源高端模型，也有开源社区的广泛贡献；不仅限于软件层面的语言模型，还延伸到硬件、机器人和行业应用。中国市场的规模优势、工程师红利、数据资源以及政策支持，都在加速这一生态的成熟。

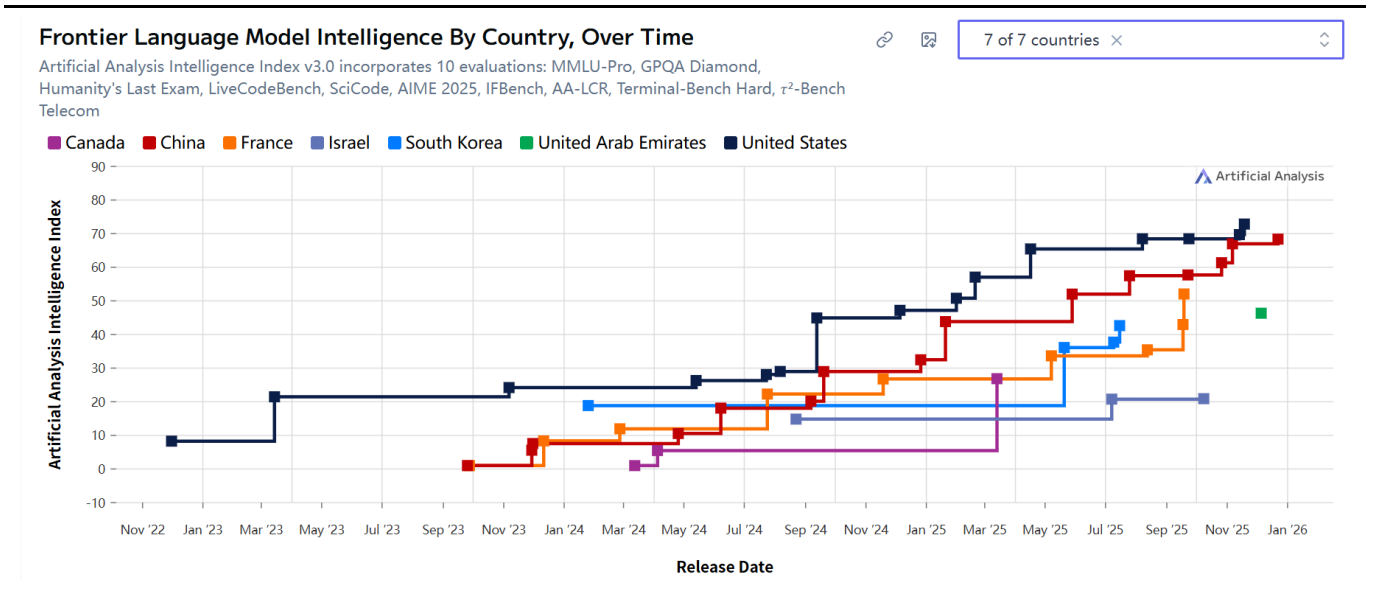
中国已经在模型性能、部署效率、竞争强度、硬件自主方面展现出多维优势，这些优势相互强化，正在形成正反馈循环。我们认为，未来几年，全球 AI 领域的重大突破，很可能有相当比例来自中国团队。AI 革命是全球性的，中国不仅是参与者，更是重要的推动者和潜在引领者。

4.5.2. 文本、编程、多模态各项能力的对比

在文本理解和生成的基础能力上，中美模型的差距已经相当小。从各种 benchmark 的测试结果看，国内头部模型在标准化任务上的表现在快速提升。但刷榜能力和实际使用体验之间存在鸿沟。实际应用中的很多任务是开放性的，没有标准答案，需要模型展现出真正的理解能力和创造能力。

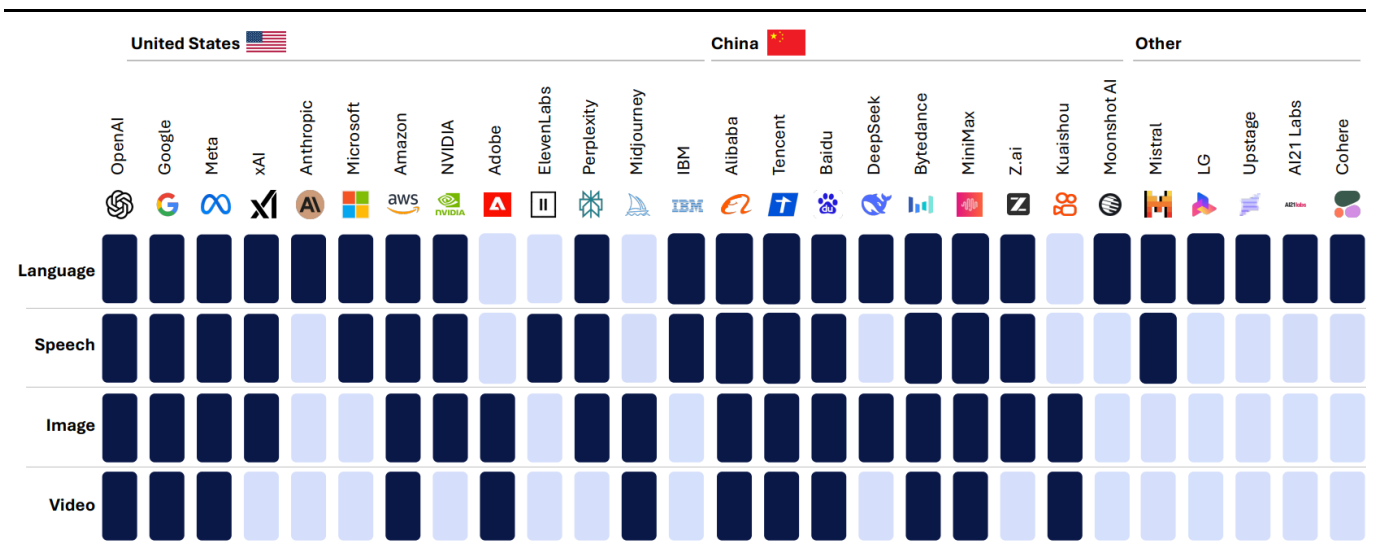
在简单任务上国内外模型的区别不大。当任务是明确的、可以通过模式匹配和统计规律完成的时候，国内模型的表现往往不输海外模型。但一旦涉及需要深层理解的复杂任务，差距就会显现出来。比如在 3D 空间理解、复杂物理规则推理这些方向上，国内模型的表现就相对较弱。3D 理解需要模型建立起空间关系的内在表征，理解物体在三维空间中的位置、朝向、遮挡关系。物理规则的理解需要模型掌握因果关系、物体属性、运动规律。这些能力的建立需要更深层次的模型架构支持，需要更高质量的训练数据，也需要更精细的训练方法。

图47: Model Intelligence 分国家对比



数据来源: Artificial Analysis, 东吴证券研究所

图48: 中美模型竞争

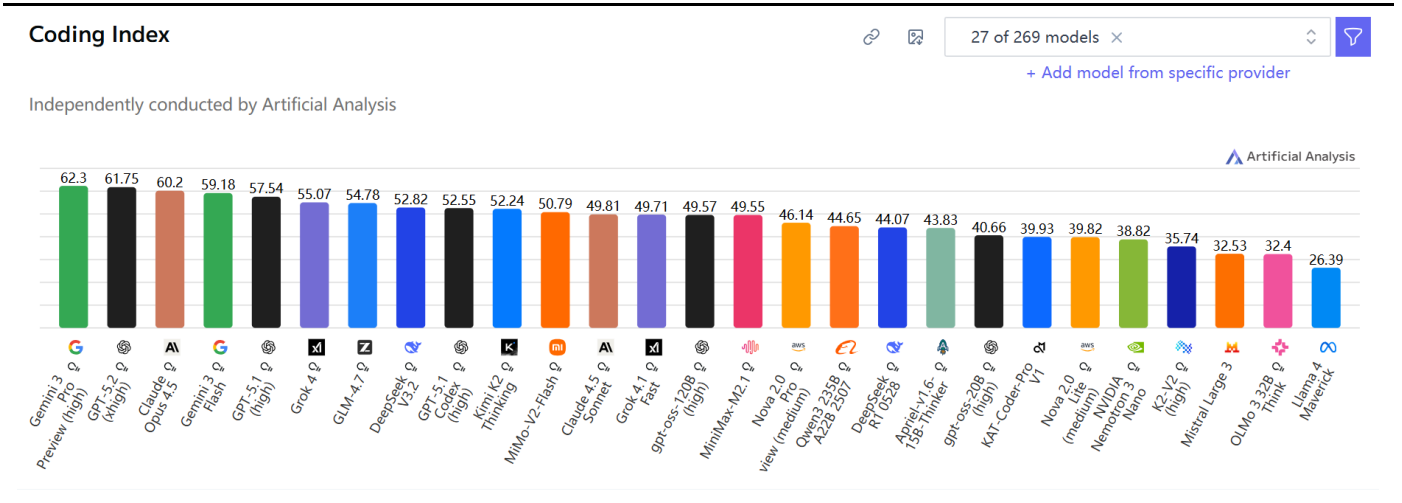


数据来源: Artificial Analysis 《State of AI》, 东吴证券研究所

注: 深蓝色表示 xx 公司在 xx 大模型领域有布局, 浅蓝色表示无布局

在编程能力上, 国内头部模型大约落后海外最强模型 3-6 个月的时间。例如, 智谱 GLM4.7、deepseek v3.2、minimax m2.1 大概相当于 Claude sonnet 4.5 上下的水平, 相较于 Claude 4.5 Opus、Gemini 3 Pro、GPT 5.2 xHigh 还有一定差距。这种差距在简单的代码补全或语法修正上可能不明显, 但在复杂的系统设计、架构优化、bug 定位这些需要深度推理的场景中就会体现出来。编程不仅仅是语法的堆砌, 更需要对问题的深刻理解、对解决方案的权衡取舍、以及对潜在风险的预判, 这些能力的提升需要时间和数据积累。

图49: Artificial Analysis 的大模型 Coding 能力排行榜 (截至 2026/1/4)



数据来源: Artificial Analysis, 东吴证券研究所

图50: LM Arena 的大模型 Web Dev 能力排行榜 (截至 2026/1/8)

Rank ↓	Model ↓	Score ↓	Votes ↓
1	AI claude-opus-4-5-20251101-thi...	1512	4,564
2	gpt-5.2-high	1480	1,647
3	AI claude-opus-4-5-20251101	1479	4,468
4	G gemini-3-pro	1471	9,824
5	G gemini-3-flash	1454	3,053
6	minimax-m2.1-preview	1445	3,980
7	Z glm-4.7	1441	2,270
8	gpt-5-medium	1395	3,949
9	gpt-5.2	1394	1,641
10	AI claude-sonnet-4-5-20250929-t...	1391	8,616

数据来源: LM Arena, 东吴证券研究所

在多模态领域: 视频生成是一个相对新的领域, 中国公司的追赶速度比较快。在 LMarena 的 text-to-video 榜单上, 阿里的万相 2.5、快手的可灵 2.6、字节的即梦 1.5 均登上榜单。且国内模型在价格上具有明显优势。

图51: LM Arena 的文生视频、图生视频大模型排行榜 (截至 2026/1/8)

Text-to-Video (1 day ago)				Image-to-Video (23 hours ago)			
Rank	Model	Score	Votes	Rank	Model	Score	Votes
1	veo-3.1-fast-audio	1376	7,297	1	veo-3.1-audio	1401	15,766
2	veo-3.1-audio	1368	7,185	2	veo-3.1-fast-audio	1390	15,586
3	veo-3-fast-audio	1360	24,003	3	wan2.5-i2v-preview	1347	8,607
4	sora-2-pro	1351	6,107	4	veo-3-audio	1339	31,995
5	veo-3-audio	1347	17,545	5	veo-3-fast-audio	1331	40,718
6	sora-2	1320	6,928	6	seedance-v1.5-pro	1322	1,363
7	wan2.5-t2v-preview	1267	3,807	7	kling-2.6-pro	1297	5,701
8	veo-3	1253	13,419	8	kling-2.5-turbo-1080p	1278	3,665
9	veo-3-fast	1252	13,764	9	seedance-v1-pro	1277	35,623
10	seedance-v1.5-pro	1227	413	10	veo-3-fast	1256	25,164

数据来源: LM Arena, 东吴证券研究所

图52: 国内视频生成模型的价格远低于海外模型 (截至 2026/1/4)



数据来源: Artificial Analysis, 东吴证券研究所

注: 这张图的纵轴代表的是 Quality ELO, 是一个用来量化模型视频生成能力的相对指标。这些模型经过了大量的成对比较 (A/B 测试), 分数是基于它们在对抗中被判定为“质量更好”的胜率计算出来的。纵轴数值从 800 到 1.70k (1700)。在图中, 位置越靠上, 代表该模型的视频生成质量在竞技场排名中越高。ELO 分数反映的是相对实力而非绝对数值。

4.5.3. 商业路径分化: 高定价的闭源路线 vs 中低定价的开源路线

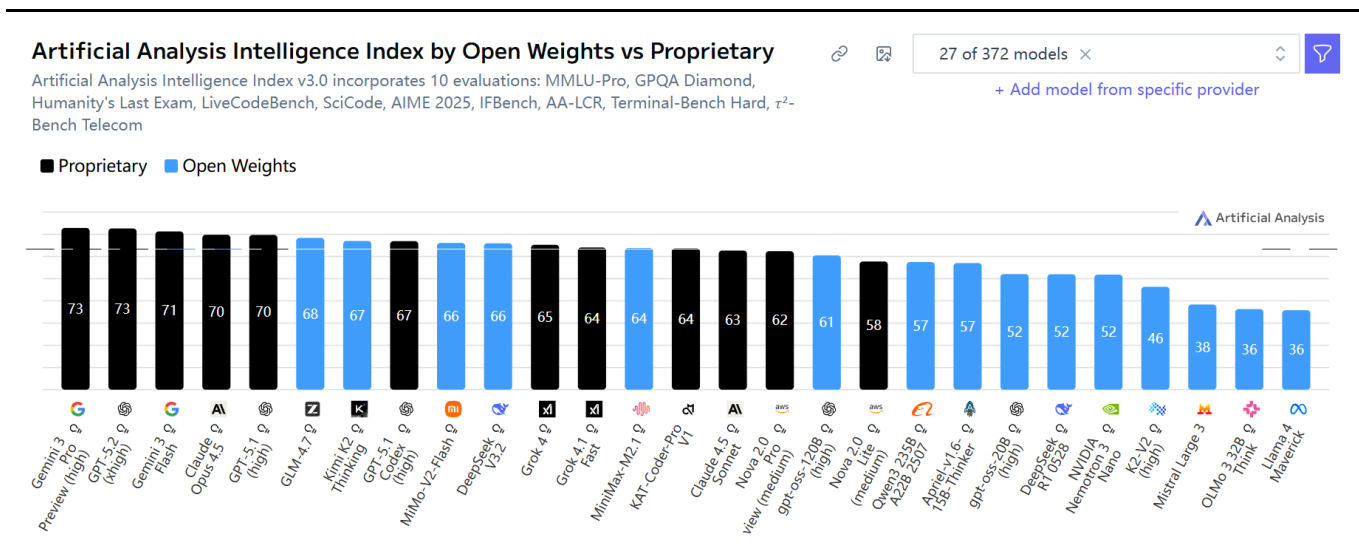
美国模型公司主要走的是技术领先+高定价+闭源模型的路线。以 OpenAI 和

Anthropic 为代表，这些公司的 API 定价虽然昂贵，但凭借强大的技术护城河和能力优势，依然在市场上获得了大量付费用户。这种商业模式的成立，很大程度上依赖于美国成熟的企业级软件市场环境。经过几十年的市场教育，美国企业客户习惯为优质的软件服务付费，只要产品能切实提升效率、降低成本，价格往往不是主要障碍。这种环境培养了一批高利润率的软件公司，它们无需通过价格战来争夺市场份额，而是依靠持续的技术创新和优质服务来维持竞争优势，从而实现商业闭环。

相比之下，中国模型公司则选择了开源模型+极致性价比路线。这一选择主要源于国内市场的特殊环境：用户基数虽然庞大，但付费意愿相对较低，尤其是对于 AI 这种新兴服务，用户尚在培养使用习惯，对价格极为敏感。如果定价与海外模型看齐，大量用户很可能会流失到免费或低价的替代品。因此，字节跳动的豆包、阿里的通义千问等头部厂商都采取了激进的低价策略，部分场景甚至免费。智谱在编程场景推出的 coding plan，价格仅为每百万 tokens 两元人民币，远低于 Claude 的定价。

在 Artificial Analysis 的开源模型排行榜中，中国超越了美国。此外，在 OpenRouter 的模型 token 消耗分布中，中国开源模型的占比也在快速提升。

图53: 中国模型以开源为主，美国模型以闭源为主（截至 2026/1/4）

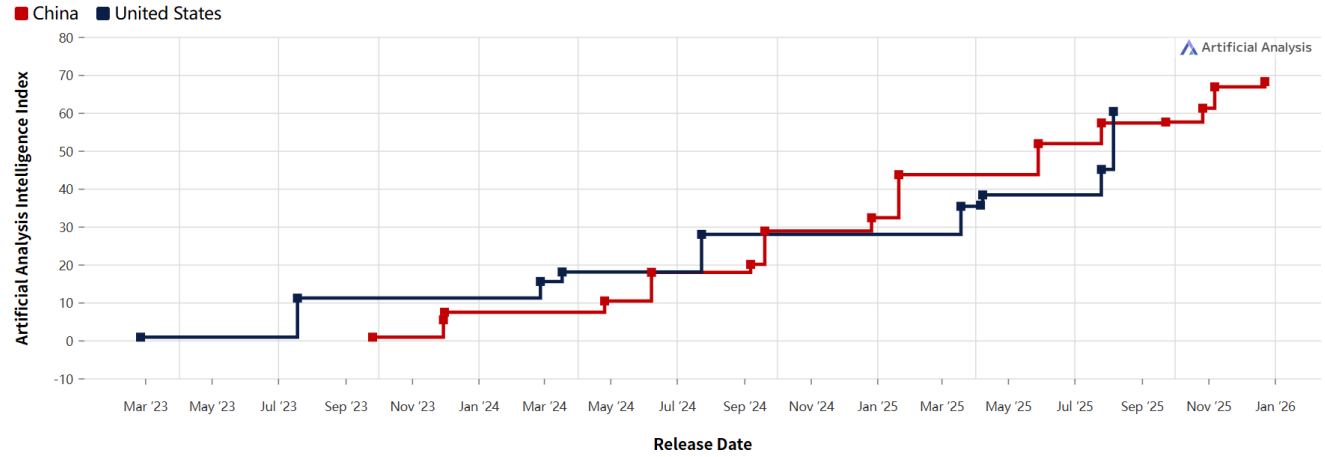


数据来源：Artificial Analysis，东吴证券研究所

图54: 中国在开源模型上更进一步

Open Weights: Frontier Language Model Intelligence By Country, Over Time

Artificial Analysis Intelligence Index v3.0 incorporates 10 evaluations: MMLU-Pro, GPQA Diamond, Humanity's Last Exam, LiveCodeBench, SciCode, AIME 2025, IFBench, AA-LCR, Terminal-Bench Hard, r²-Bench Telecom

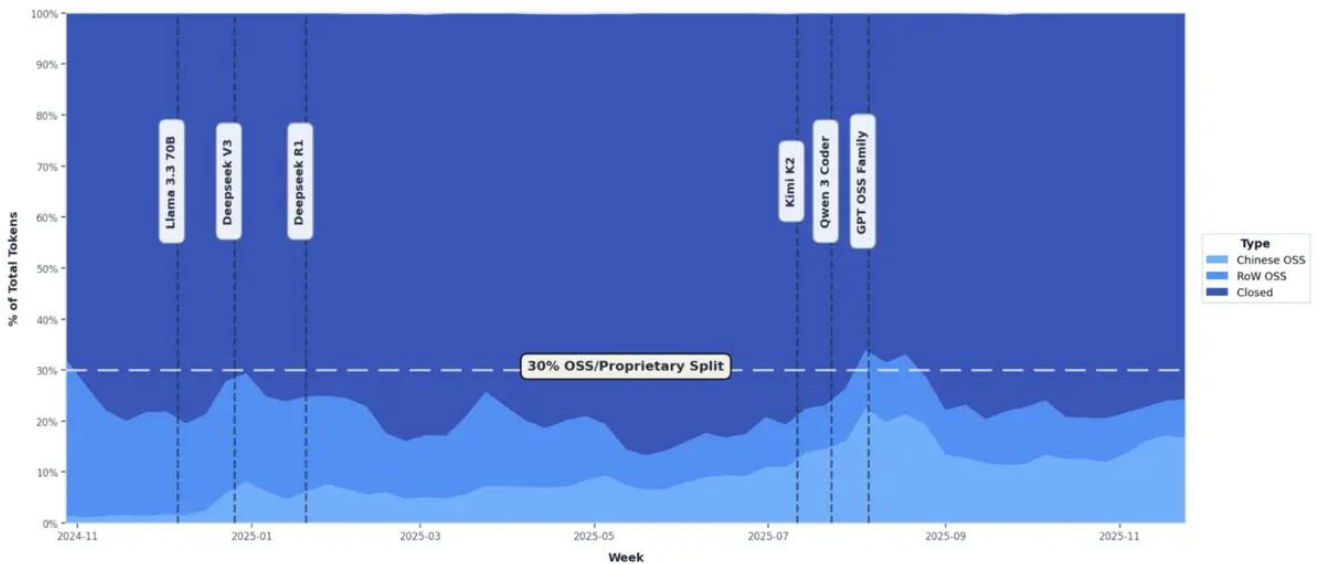


数据来源: Artificial Analysis, 东吴证券研究所

图55: OpenRouter 平台的 tokens 消耗分布

Token Share by Source Type and Origin

OpenRouter



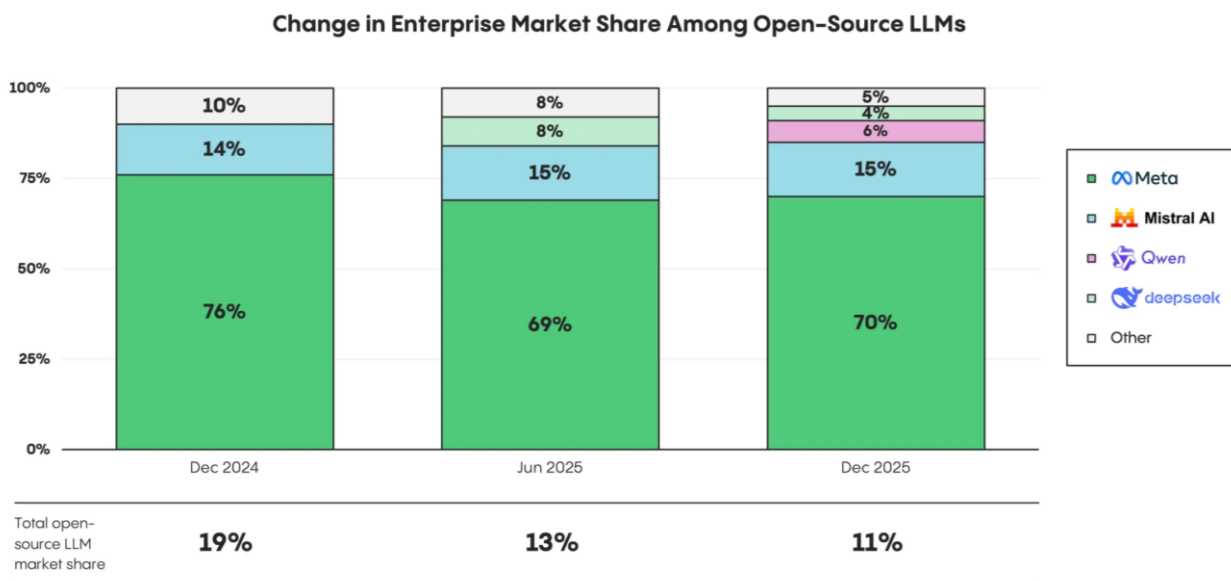
数据来源: OpenRouter, 东吴证券研究所

中国企业可通过较低成本复现美国公司耗资数十亿美元训练的模型并开源，直接冲击后者的高定价商业模式。在这种态势下，单纯依靠模型能力建立护城河越来越困难，闭源公司必须在应用场景、用户体验、生态整合等方面构建差异化优势。OpenAI、Anthropic 等公司面临的关键问题是：当开源模型能力持续逼近时，用户为何仍愿意为闭

源服务付费？可能的答案在于更高的稳定性、安全合规性、专业技术支持以及更完善的生态整合，而非单纯的模型性能。

一方面，在 B 端市场，闭源模型因为表现更加稳定，因此更受青睐。稳定性是闭源服务的重要差异化因素。开源模型虽免费，但版本迭代、API 兼容性、服务持续可用性往往缺乏保障，需要专业团队持续维护。对于企业客户而言，业务系统的稳定运行至关重要，任何中断都可能带来巨大损失。我们看到海外 B 端市场的开源模型 tokens 总体份额处于下降趋势。

图56: 海外 B 端市场开源模型 tokens 份额逐年下降

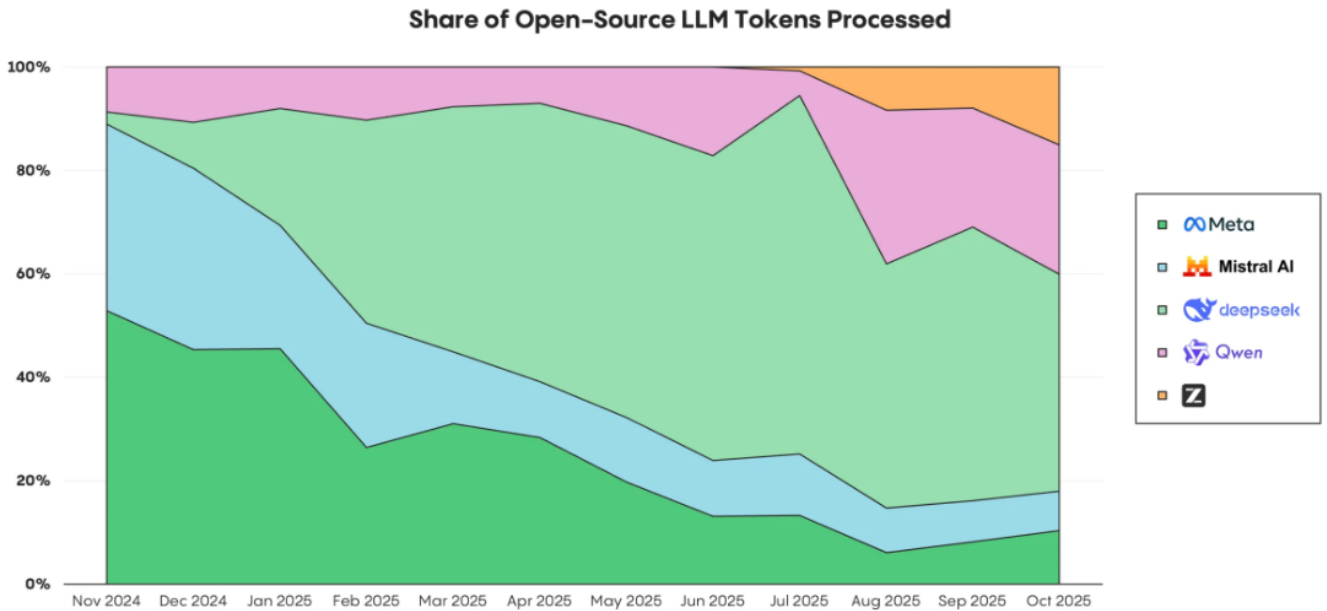


数据来源：menlo ventures，东吴证券研究所

注：这张图表示海外 B 端市场所消耗的模型总 Tokens 作为分母，各个开源模型的 Token 消耗作为分子

但另一方面，如果把整个 B 端开源模型作为总的蛋糕，其中中国的千问、DeepSeek、智谱等开源模型的 tokens 消耗份额增长迅速，而海外的 Llama 和 Mistral 份额正在下降。这说明在开源赛道，中国模型的性能优势正在发挥作用。开源市场的竞争逻辑与闭源市场不同，用户更关注模型的实际性能和部署的便利性，而非单纯的品牌效应。如果能利用好这一趋势，在开源生态建立影响力，可以作为闭源模型的“营销工具”，让更多用户体验到中国模型的能力，建立品牌认知，进而引导部分用户转向闭源模型。

图57: 海外 B 端市场开源模型竞争格局



数据来源: menlo ventures, 东吴证券研究所

注: 这张图表示海外 B 端市场所消耗的开源模型总 Tokens 作为分母, 各个开源模型的 Token 消耗作为分子

总结来说, 我们认为开源和闭源模型不是非此即彼的零和博弈, 而是一场尚未定论的长期竞赛, 最终可能是两者并存、各司其职, 形成一种金字塔式的产业结构。我们既不认为开源能彻底取代闭源, 也不认为闭源能永久锁死技术领先。两者各有优势与局限。

闭源模型仍在以极快的速度进化。OpenAI、Anthropic、Google 的进展没有放缓。内部人士的普遍反馈是“我们有成百上千个新想法、新方法, 每天都有新的发现”。市场偶尔会冒出“闭源模型快到天花板了”的担忧, 但在大实验室内部, 这种声音基本被视为毫无根据的噪音。闭源模式的最大优势在于可以集中巨额资源、无拘束地探索最前沿的扩展路径, 这让它们在绝对能力上始终保持领先。无论训练规模、数据质量还是架构创新, 闭源玩家都能以最快的速度把最激进的想法落地。因此, 在相当长的一段时间内, 闭源模型很可能继续占据“最聪明模型”的位置——也就是金字塔的最顶端。

与此同时, 中国开源模型的追赶速度远超预期, 往往在大模型发布 6-12 个月后, 就会有开源版本把同等能力压缩到更小、更廉价的规模。这种“压缩-扩散”机制正在成为常态, 而这恰恰是开源最强大的地方。

更重要的是, 开源有助于培养人才与扩散知识。他明确指出, 开源模型的最大优势在于它适合学习。如果你是一位教授想教 AI、学生想学 AI、工程师想理解这项新技术, 甚至只是一个在地下室琢磨创业点子的独立开发者, 最前沿的开源模型就是最好的教材。这些模型几乎把“如何做 AI”的全套技术细节完整展示了出来: 架构、训练技巧、数据处理方式、优化方法一览无余。如果所有顶级能力都被封存在少数几家闭源巨头手里, 知识传播的速度会大幅减慢, 整个行业的创新节奏也会被卡住。而开源正在以前所未有

的速度扩散 know-how。越来越多的年轻人正在短时间内成长为顶级人才。这一切的背后，开源功不可没。如果没有开源，AI 人才的培养速度不可能这么快，整个生态的活跃度也会大打折扣。开源不仅在传播技术本身，更在培养下一代能够独立构建 AI 的工程师和研究者。这才是开源对整个行业最深远的贡献。而这种贡献是闭源模式很难完全替代的——闭源可以保持领先，但很难像开源这样大规模、民主化。

4.5.4. 算力约束下的竞争格局

美国对华芯片出口管制是影响中美 AI 竞争格局的最大外部变量。从 2022 年 10 月开始，美国持续收紧对华 AI 芯片的出口限制。英伟达的 H100 无法出口中国，专门为中国市场设计的 H20 降级版本也面临进一步的限制风险。这意味着中国公司在训练前沿大模型时面临着实实在在的硬件瓶颈。这不是买不到芯片的问题，而是买不到最先进芯片的问题。

最先进的芯片不仅仅意味着更快的计算速度，还意味着更高的能效比、更大的显存容量、更优的互联带宽。当训练超大规模模型时，这些差异会被指数级放大。显存容量决定了单个 GPU 能够处理的 batch size 和模型参数量，互联带宽决定了多 GPU 并行训练时的通信效率，能效比则直接关系到训练成本和数据中心的运营压力。这些因素综合起来，最终体现为训练时间的显著差异和模型规模的上限约束。当美国公司可以用最新一代的 GPU 在几周内完成一次大规模训练实验时，中国公司可能需要几个月的时间，这种时间差距会直接影响技术迭代的速度。

但约束也在倒逼创新。中国公司被迫在算法效率上下功夫，用更少的算力训练出更好的模型。DeepSeek 就是这方面最典型的案例。这种被迫创新反而可能成为长期竞争力的来源。历史上很多重大技术突破都是在资源约束下实现的。当无法靠蛮力解决问题时，人们就会被迫思考更本质的问题，寻找更优雅的解决方案。同样的道理，当中国 AI 公司无法依赖最先进的硬件时，它们在算法创新上的探索开辟出新的技术路径。如果未来某一天算力约束解除，这些在算法效率上的积累可能会转化为优势。

但稀疏注意力的代价是性能下降。MiniMax 之所以从 M1 模型的 Linear Attention 转向 M2 模型的 Full Attention，是因为在工程实现上，Linear Attention 需要大量的优化工作，而且在某些任务上的表现还不如 Full Attention 稳定。Full Attention 虽然计算成本更高，但工程上已经非常成熟，有大量现成的优化库和最佳实践可以借鉴。

国产芯片的进展为缓解算力约束提供了一线希望。华为的昇腾系列、寒武纪的 MLU 系列虽然与英伟达最新一代产品仍有代差，但已经能够支撑一定规模的模型训练。2026 年如果国产芯片在制程和软件生态上取得突破，中国公司的算力约束将有所缓解。但芯片制造不仅仅是技术问题，更涉及整个产业链的成熟度。从 EUV 光刻机到先进封装技术，从 EDA 软件到各种特种材料，每个环节都需要时间积累。

AI 人才全球流动加速，中国本土生态快速成熟

美国长期仍是 AI 人才最大聚集地，斯坦福、MIT、CMU 等高校每年输出大量顶尖研究者，硅谷的创业氛围、丰厚薪资与无限算力持续吸引全球精英。但过去十年，中国在 AI 领域的系统性投入已产生显著回响。清华、北大、中科院等机构大力兴建 AI 学院与研究院，招聘海外学者，提供与硅谷相当甚至更具安全感的待遇、庞大算力与政策支持。这种资源倾斜叠加基础教育红利，使本土人才培养规模与质量同步跃升，许多顶尖人才选择留在国内创业或加入大厂。更为重要的是，海外华人研究者的回流潮日益明显。DeepSeek、月之暗面等明星创业公司的创始人多有斯坦福、MIT、Google、Meta 等海外经历，他们带回的不仅是前沿技术知识，还有硅谷的工程实践与创业文化，对公司技术路线、组织效率与产品迭代产生了深远影响。当国内能够提供匹配的资金、市场空间与相对稳定的科研环境时，人才回流便成为自然趋势，进一步强化了中国 AI 生态的闭环。

NeurIPS 2025 共录用 5526 篇论文，其中中国机构整体占据近半壁江山。谷歌虽仍以 4.84% 的份额位居机构第一，但清华大学以 4.73% 紧随其后，差距仅 0.11%，北京大学则以 3.63% 排名第三，超过斯坦福（2.58%）、CMU（2.55%）、MIT（2.45%）。阿里千问团队凭借《Gated Attention for Large Language Models》一文斩获最佳论文奖，该工作通过引入非线性、稀疏性的门控注意力机制，有效缓解长文本处理的“注意力陷阱”，显著提升推理效率与训练稳定性，已被集成到 Qwen3-Next 模型中。这不仅证明中国工业界已在 Transformer 核心机制等基础层面具备与 DeepMind、谷歌同台竞技的能力，也反映出过去被西方学界视为“灌水”或应用微调的中国研究，正逐步掌握技术定义权。

够高的频次使用，并最终沉淀为稳定现金流。

离开应用层谈景气度，往往会陷入用情绪解释波动；只盯应用层又容易忽视算力与模型的供给约束。更合适的做法，是把应用层拆成三条并行但逻辑不同的路径：**toB 的生产力工具化、toC 的产品入口化，以及科技巨头内部把 AI 当作效率引擎的系统性改造**。三条路径共同决定了 AI 应用能长多大，也决定了上游资本开支能否进入良性循环。

5.1. 需求验证的两大指标：ARR 与 Tokens 消耗

在应用层商业化仍充满争议的阶段，我们建议主要观察两项指标，一个是 **ARR**，也就是年度经常性收入；另一个是 **Tokens 消耗量**。

ARR 的意义在于它直接回答了用户是否愿意持续付费、付费是否可规模化，这是商业化的程度；Tokens 的意义在于它直接刻画使用频次与交互深度，是需求热度最实时的信号。

把 ARR 只当财务数据会低估它的战略价值，因为 ARR 反映的不是一次性收入，而是持续订阅关系，背后通常对应稳定使用习惯与可复制的销售路径。把 Tokens 只当技术指标也会犯错，因为 Tokens 实际上是 AI 时代的使用时长与调用次数，它的变化往往先于收入变化发生。

我们更倾向于把这两者看成一组互补指标：Tokens 像体温计，反映需求在不在升温；ARR 像血压计，反映商业化是否健康。理想状态是二者同步改善，但现实往往是 Tokens 先冲、ARR 后跟，差距如何收敛，将是 2026 年应用层最核心的命题。

需求是否真实：用 ARR 与 Tokens 消耗量两个指标交叉验证。ARR 代表商业化的程度，Tokens 代表使用强度与渗透速度。

表13：AI 科技公司 ARR 及增速

排名	公司	核心产品/业务	ARR (亿美元)	更新日期
1	OpenAI	ChatGPT	130	2025 年 10 月
2	Anthropic	Claude	70	2025 年 10 月
3	Cursor	AI 代码编辑器	10	2025 年 11 月
4	AlphaSense	市场情报搜索	5	2025 年 10 月
5	ElevenLabs	AI 语音生成与合成	4	2025 年 12 月
6	Zapier	AI 自动化流程编排	4	2025 年 12 月
7	Dataiku	企业级 AI 平台	3.5	2025 年 10 月
8	Codeway	AI 移动应用开发	3	2025 年 10 月
9	Fireworks	生成式 AI 推理服务	2.8	2025 年 10 月
10	Replit	AI 辅助云端编程平台	2.5	2025 年 10 月
11	Glean	企业级 AI 搜索与知识库	2	2025 年 12 月
12	Suno	AI 音乐生成	2	2025 年 11 月
13	Lovable	AI 编程	2	2025 年 11 月
14	OpenEvidence	AI 医疗决策支持	1.5	2025 年 12 月

数据来源：ARR Club，东吴证券研究所

从 ARR 的角度看，头部产品已经给出了相当强的需求验证。OpenAI 和 Anthropic 的 ARR 合计约 200 亿美金（截至 2025 年 10 月），且增长斜率陡峭。垂直领域的应用层，Cursor（AI 编程工具）ARR 已经达到了 10 亿美元。在编程、语音、搜索等具体场景中，AI 正在以极快的速度渗透并创造实际价值。

我们从这些案例里得到的判断是：AI 应用并不缺需求，缺的是把需求装进可持续商业模式的能力，尤其是在推理成本、补贴策略、毛利结构之间找到长期均衡。

5.2. Tokens 增长的驱动力和可持续性

但如果只看 ARR，容易低估需求的爆发速度，因为收入确认往往滞后于使用行为变化。Tokens 消耗量提供了更及时的视角。

多个侧面反映整个行业的调用需求在加速。

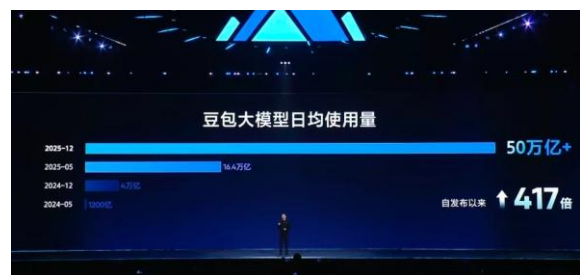
- ◆ 根据 Omdia 发布的《2025 全球企业级 MaaS 市场分析》，截至 2025 年 10 月，OpenAI 和 Google Cloud 分别以日均 70 万亿和 43 万亿 Tokens 消耗量位居全球前二；而火山引擎则以超 30 万亿（截至 12 月已破 50 万亿）的消耗量位列全球第三。三家厂商合计占据全球 MaaS 市场 65% 的份额。
- ◆ 根据谷歌财报，25Q3 谷歌 AI 产品 Token 消耗季度环比增长超过 50%。
- ◆ OpenRouter 上 2025 年 Tokens 消耗同比增长超过 300%。

图59: 全球 MaaS 的 Token 消耗竞争格局(截至 2025/10)



数据来源：Omdia，雷锋网，东吴证券研究所

图60: 豆包大模型日均 tokens 消耗量快速增长



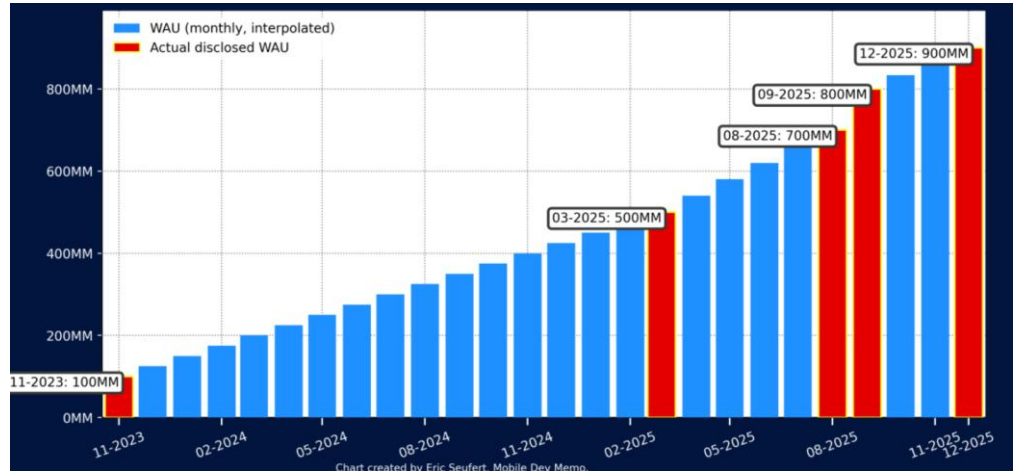
数据来源：财联社，东吴证券研究所

什么驱动 Tokens 增长？

- ① 更多的用户数。2025 年 11 月海外主要 AI 应用 MAU 合计值（未去重）已超过 14 亿人。Chatgpt 周度活跃用户数从 2023 年 11 月的 1 亿人增长至 2025 年 2 月的 5 亿人，并进一步增长至 2025 年 12 月的 9 亿人。早期用户需要学 prompt、需要懂得如何引导模型，AI 更像工具爱好者的玩具；到 2025 年，模型对模糊指令、口语化表达的理解更强，普通用户也能自然使用。多模态的成熟进一步

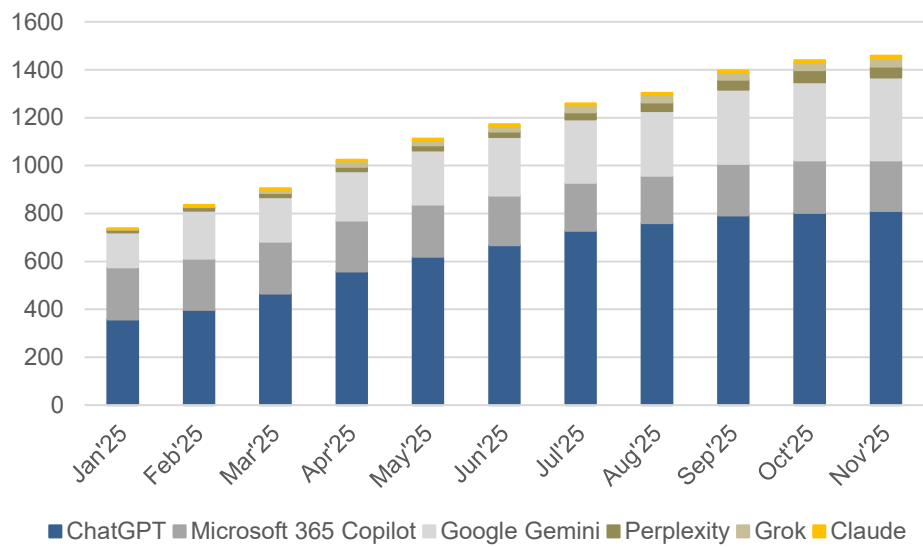
降低了交互成本：图像、语音、视频输入让 AI 更贴近真实世界，用户不再“为了让 AI 懂我”而付出额外表达成本，使用频次自然上去。

图61: chatgpt 周度活跃用户数 (百万人)



数据来源: mobiledevmemo, 东吴证券研究所

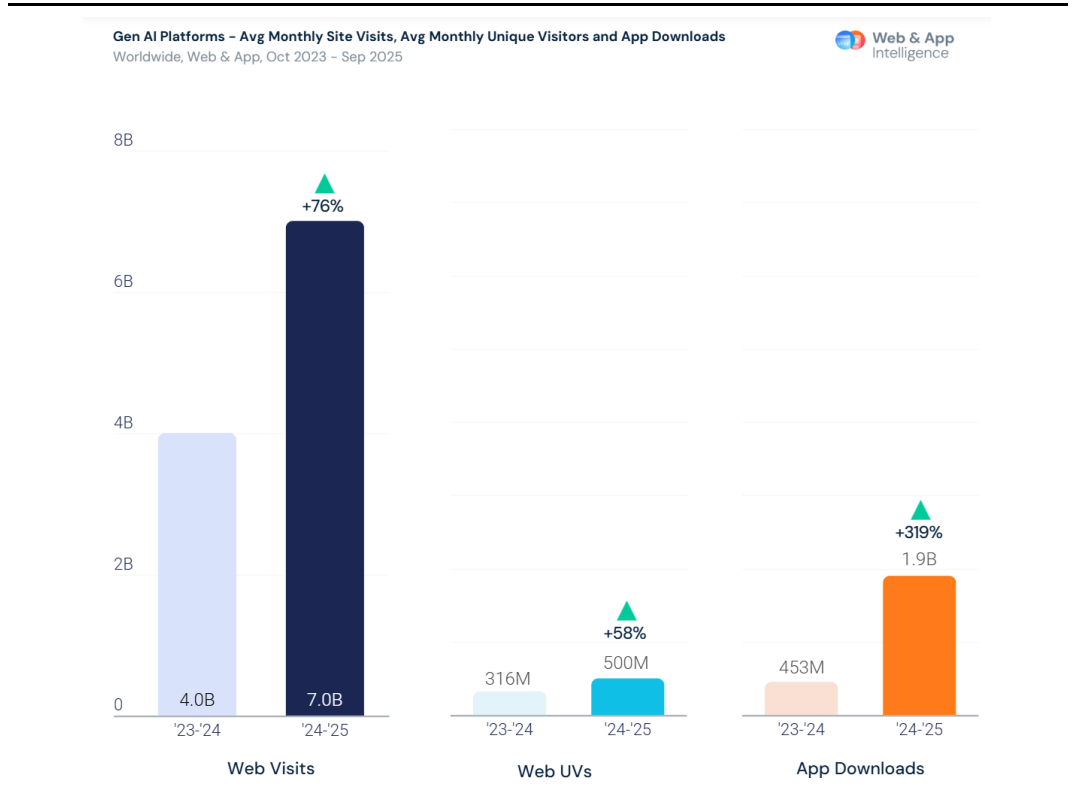
图62: 主要 AI 应用月度活跃用户数 (百万人)



数据来源: IT 之家, sensortower, 东吴证券研究所

2024 年 10 月-2025 年 9 月，AI 相关的 App 月度平均下载量同比增长 319%，达到了 19 亿次，显示用户正在大规模转向移动应用；网页端的月均访问量增长了 76%达到 70 亿次，独立访客数也突破了 5 亿大关。AI 正从网页试用转向全平台普及，移动端成为新的增长引擎。

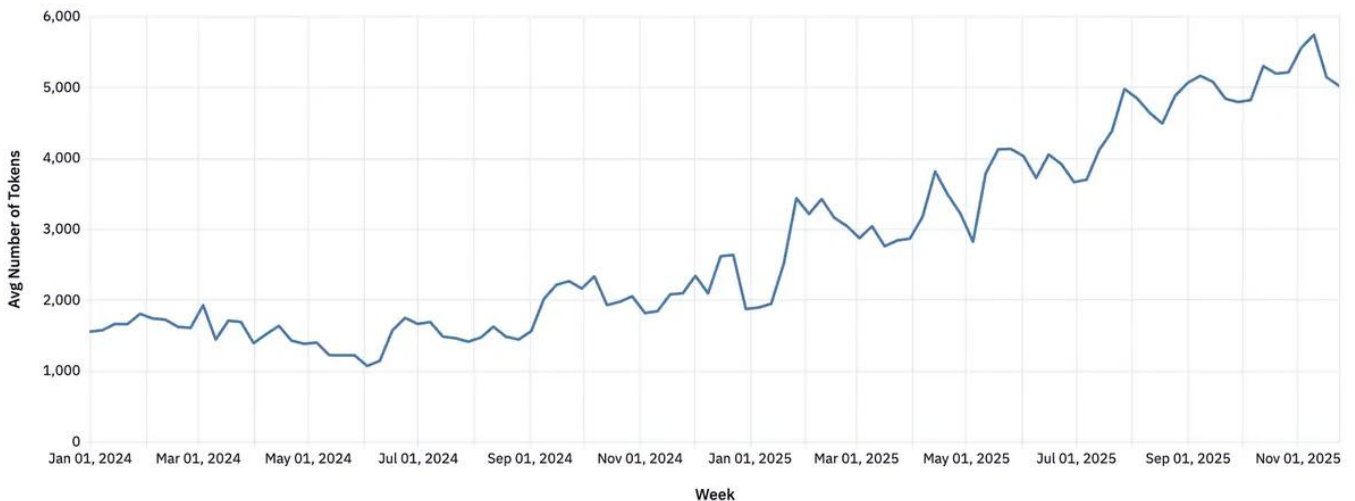
图63: AI 平台的月度平均访问量快速增长 (十亿次)



数据来源: similarweb, 东吴证券研究所

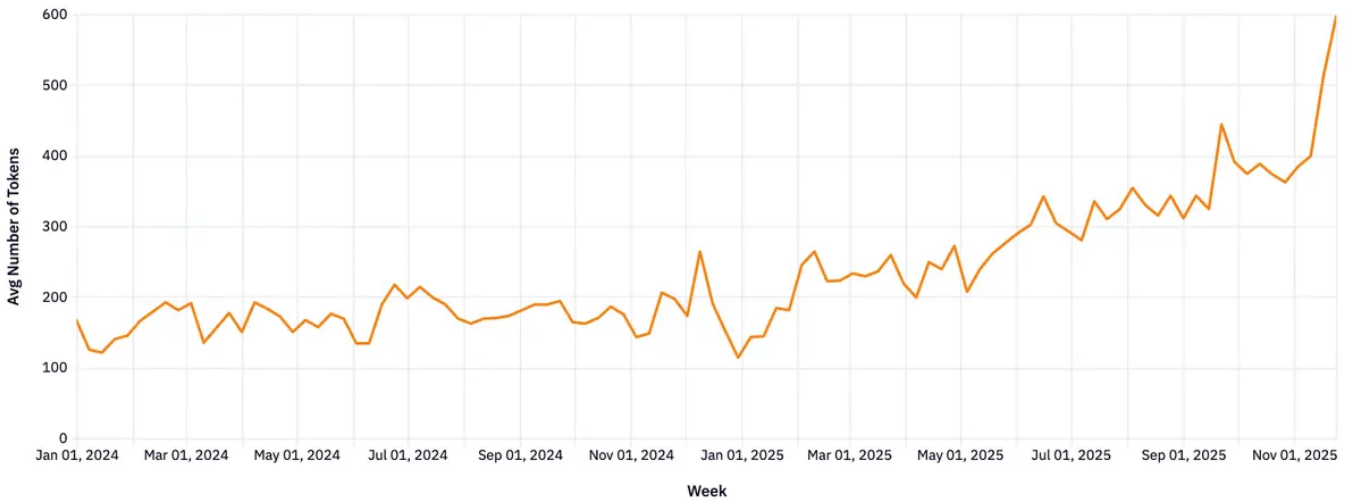
- ② 更长的输入和输出。根据 OpenRouter 年终总结报告, 用户平均每个 Prompt 的长度增长了 3 倍, AI 的平均回复长度增长了 2 倍。

图64: 从 2024 年 1 月到 2025 年 11 月, OpenRouter 上用户的 Prompt 平均长度增长了 3 倍



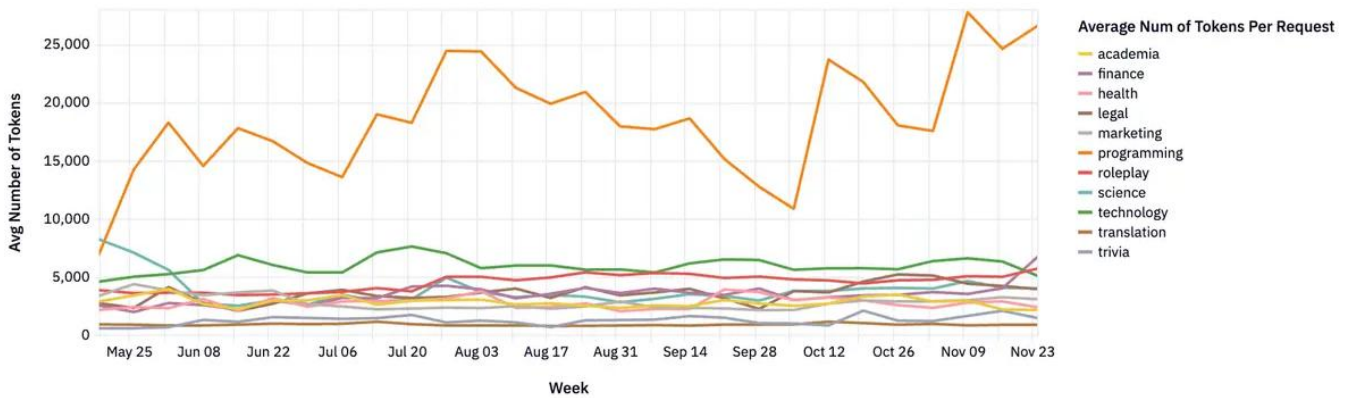
数据来源: OpenRouter, 东吴证券研究所

图65: 从2024年1月到2025年11月, OpenRouter上AI的回复平均长度增长了2倍



数据来源: OpenRouter, 东吴证券研究所

图66: 编程场景的每条输入输出的平均 tokens 长度远高于其他场景



数据来源: OpenRouter, 东吴证券研究所

- ③ **更长的思维链。** 推理模型在给出答案前会展开更长的推理链条, 一个原本 100 个 Token 能解决的问题, 可能变成 3000 到 5000 个 Token 的消耗。这并不等于浪费, 恰恰意味着 AI 开始把更复杂的任务纳入可交付范围, 而复杂任务本身就是更高价值的需求。
- ④ **Agent 模式:** 从人类消耗 Tokens 到 Agent 消耗 Tokens。过去 Tokens 主要来自人类对话的输入输出, 现在多 Agent 并行、Agent 协作会让单任务 Token 成倍增长: 一个负责需求, 一个负责架构, 一个写代码, 一个做测试, 彼此沟通本身就产生 Token。我们认为, 真正决定 Token 指标长期走势的不是用户数, 而是任务范式从聊天转向执行后的系统性调用, 这也是为什么 2026 年被普遍

视作 Agent 规模化的关键窗口。

5.3. toB: 谁在构建不可替代性

AI 应用场景包括 toC、toB 和科技巨头内部赋能。

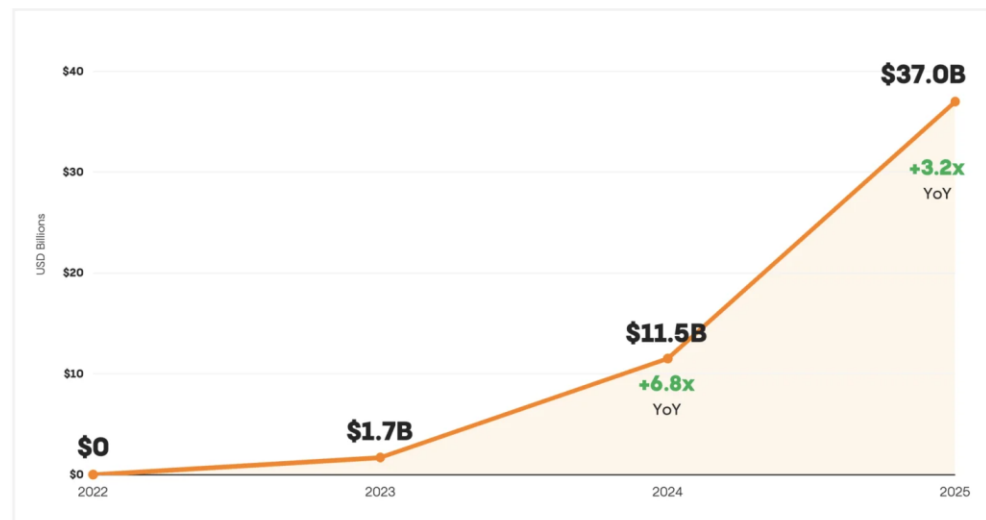
ToB 与 ToC 本质上都是服务人类，并无优劣之分。

在 toB 端，Menlo Ventures 报告显示，2025 年企业生成式 AI 总支出已达 370 亿美元，同比增长超过 2 倍；其中应用层 190 亿美元，基础设施层 180 亿美元。AI 正在从边缘工具演变为企业核心生产力基础设施。相比传统 SaaS 动辄需要十年才能达到类似渗透率的历程，生成式 AI 仅用 3 年就占据了全球 SaaS 市场 6% 的份额。

图67: 全球 toB AI 市场规模快速增长

Enterprise AI Is the **Fastest-Scaling Software Category in History**

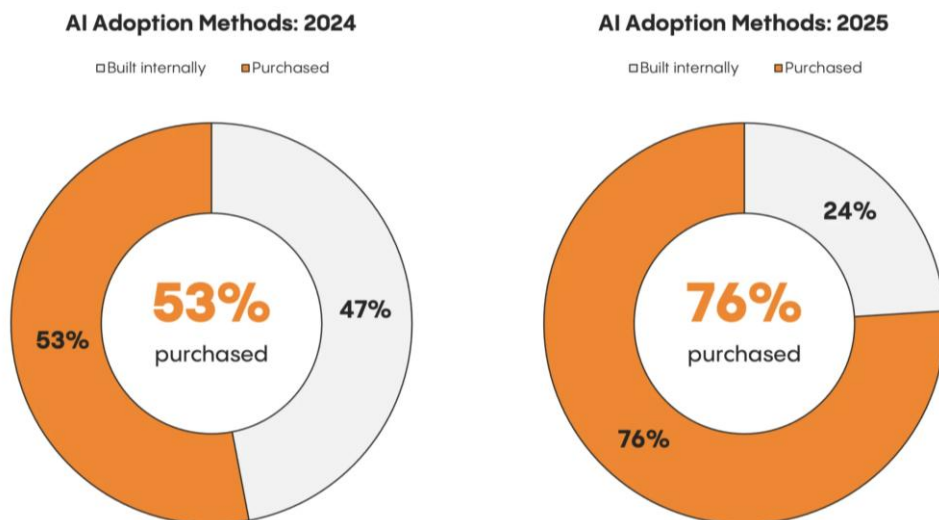
Now capturing **6%** of the ~\$300B global SaaS market*



数据来源：Menlo Ventures, Gartner, 东吴证券研究所

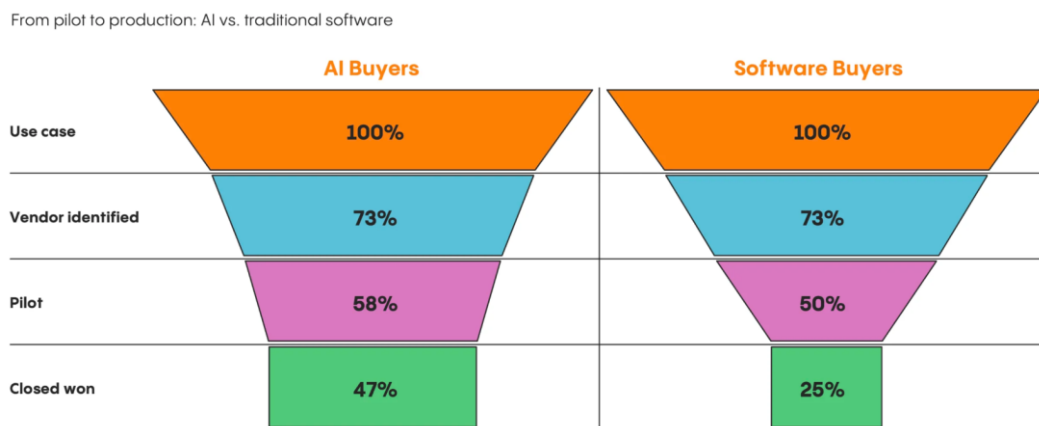
更关键的转变在于采购路径：企业从“自己建”转向“直接买”。2024 年时，企业内部自研和外部采购的比例还大致是 47%对 53%，但到 2025 年，这一比例已经逆转为 24%对 76%。这意味着客户越来越愿意为稳定部署、持续升级、专业实施、数据安全合规支付溢价，而不愿自己承担模型选型、微调、运维的复杂风险。一旦核心业务流程深度依赖某家厂商的服务，切换成本就会急剧上升。AI 从企业评估到最终进入生产的转化率高达 47%，几乎是传统 SaaS 25% 转化率的两倍。这反映出企业采购 AI 的意愿强烈——他们不是在“试试看”，而是带着明确的 ROI 预期直接下单。

图68: 企业从“自己建”转向“直接买”



数据来源: Menlo Ventures, 东吴证券研究所

图69: 相较于传统 SaaS, toB AI 的“试用→采购”转化率更高



数据来源: Menlo Ventures, 东吴证券研究所

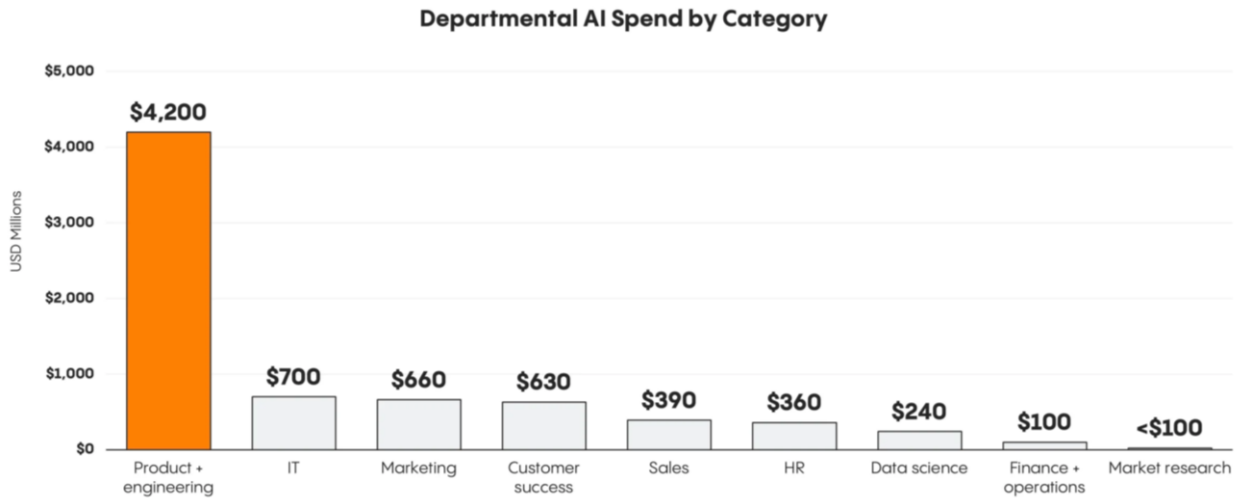
注: 图中的数据指的是各个环节的转化率

这种高意图采购的背后, 是 AI 在企业内部已经证明了其价值。编程领域是最典型的突破口。2025 年, 企业用于 AI 编程工具的支出达到了 42 亿美元, 占部门级 AI 支出的 55%, 成为整个应用层最大的单一品类。开发者每天使用 AI 编程工具的比例已达 50%, 从代码补全到多文件编辑、diff 审批、自然语言指令, 再到完整的代码代理和应用构建平台, AI 已经覆盖了软件开发生命周期的几乎所有环节。直接转化为更快的功能迭代、更低的工程师人力成本, 以及更强的市场竞争力。更重要的是, 编程场景的成功为其他部门树立了标杆: 只要模型能力达到经济意义上的可用阈值, 生产力红利就会迅速释放。

编程之外, 其他部门级 AI 应用也在快速跟进。IT 运维工具达到 7 亿美元, 营销内

容生成与优化 6.6 亿美元，客户自动化 6.3 亿美元。这些领域有一个共同特征：都存在大量重复性、结构化程度不高但又极耗人力的 workflow。

图 70：2025 年部门 AI 支出规模



数据来源：Menlo Ventures，东吴证券研究所

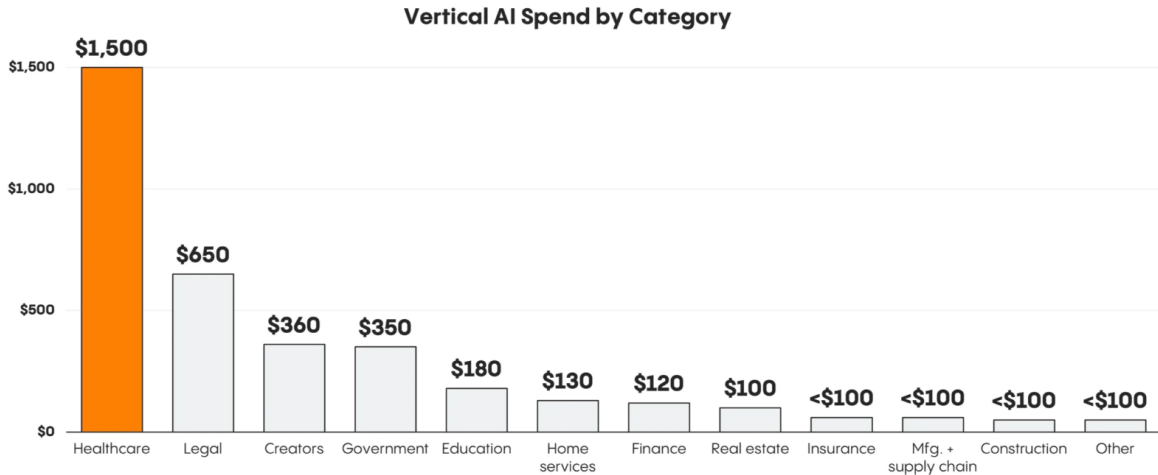
在 B 端市场，真正能够形成长期护城河的玩家，往往是那些愿意沉下心来去做“脏活累活”的垂直深耕者。他们花大量时间深入医院、律所、投行、工厂，一家家谈判、接触，采集那些通用大模型通过公开互联网数据根本无法获取的私有案例、内部文书、结构化报告、质检记录和行业流程规范，再围绕这些专有数据进行深度定制和流程重构。这种“本地脏数据+行业 know-how”的组合，正是大模型原厂最不愿意投入精力的领域——它们更倾向于专注通用能力，把垂直场景留给生态伙伴或第三方玩家，从而为这些垂直公司留出了宝贵的生存与突围空间。

2025 年，整个垂直行业 AI 支出已达到 35 亿美元，同比接近 3 倍增长，其中医疗健康领域独占 15 亿美元，远高于其他行业。医疗行业长期饱受行政负担、人员短缺和利润率压力的困扰，而 AI 环境听写（ambient scribe）类产品直接将医生每 5 小时诊疗对应的 1 小时文档时间削减 50%以上，大幅解放临床人力并提升诊疗质量。Abridge、Ambience 等新兴玩家因此迅速跻身独角兽行列，与 Nuance 的 DAX Copilot 共同主导了这一规模已达 6 亿美元的细分市场。医疗之外，法律 AI 支出达到 6.5 亿美元，创作者工具 3.6 亿美元，政府领域 3.5 亿美元。这些传统上被视为“软件渗透率低”、高度依赖人工判断和非结构化信息的行业，反而因为工作流的复杂性和数据私有性，成为生成式 AI 最肥沃的落地土壤。

海外的 Harvey（法律 AI）、Glean（企业知识管理）、openEvidence（医疗 AI）正是典型代表，而国内在金融合规审查、医疗报告结构化、工业质检等方向也正在涌现类似玩家。它们的核心优势并不在于模型本身有多领先，而在于愿意深入行业一线，积累那些无法通过爬取公开数据获得的专有知识，并将其转化为难以被通用模型直接替代的定

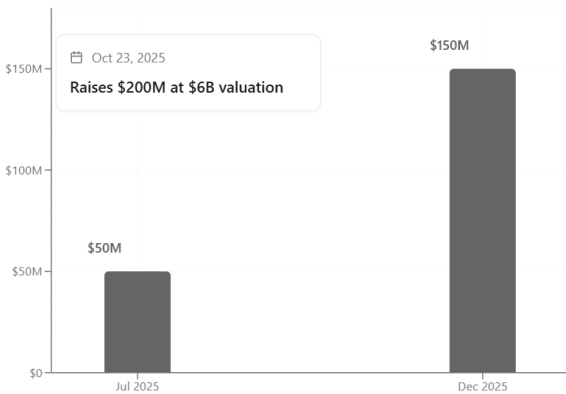
制化解决方案。

图71: 2025年垂直行业 AI 支出规模 (百万美元)



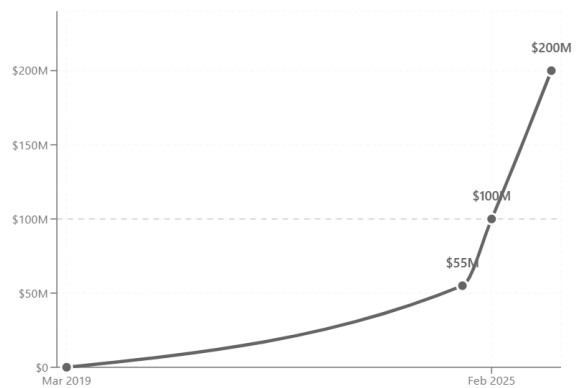
数据来源: Menlo Ventures, 东吴证券研究所

图72: OpenEvidence ARR (百万美元)



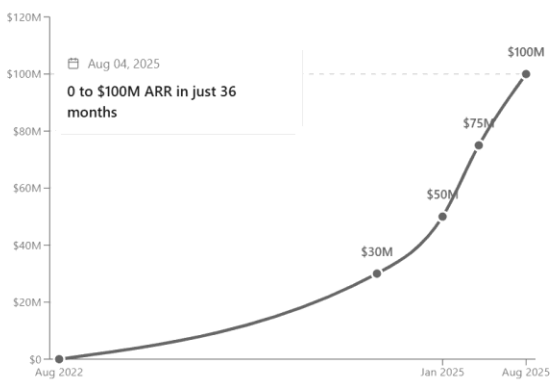
数据来源: ARR Club, 东吴证券研究所

图73: Glean ARR (百万美元)



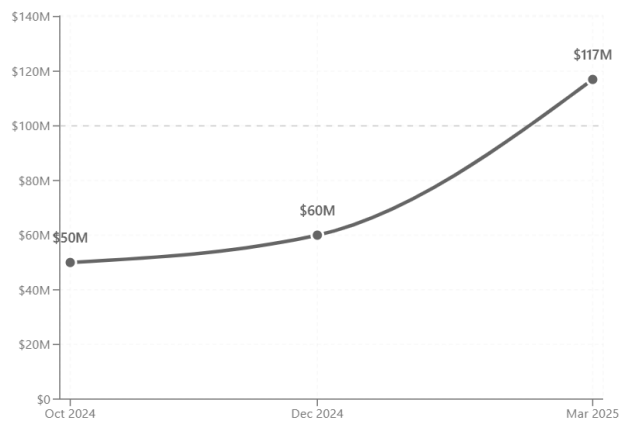
数据来源: ARR Club, 东吴证券研究所

图74: Harvey AI ARR (百万美元)



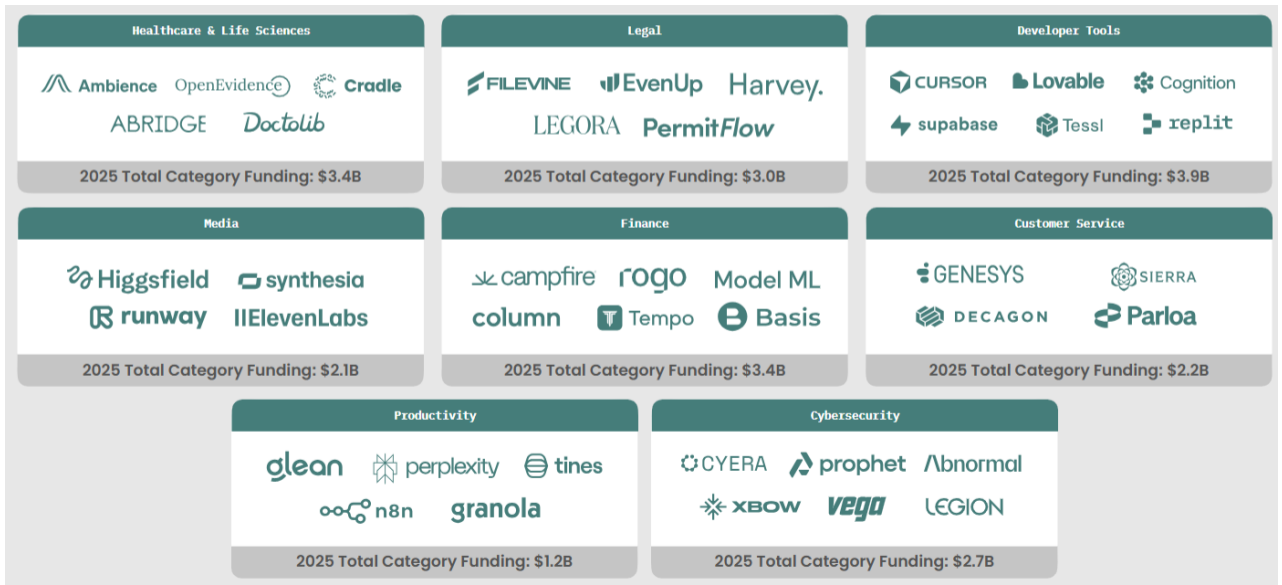
数据来源: ARR Club, 东吴证券研究所

图75: Abridge ARR (百万美元)



数据来源: ARR Club, 东吴证券研究所

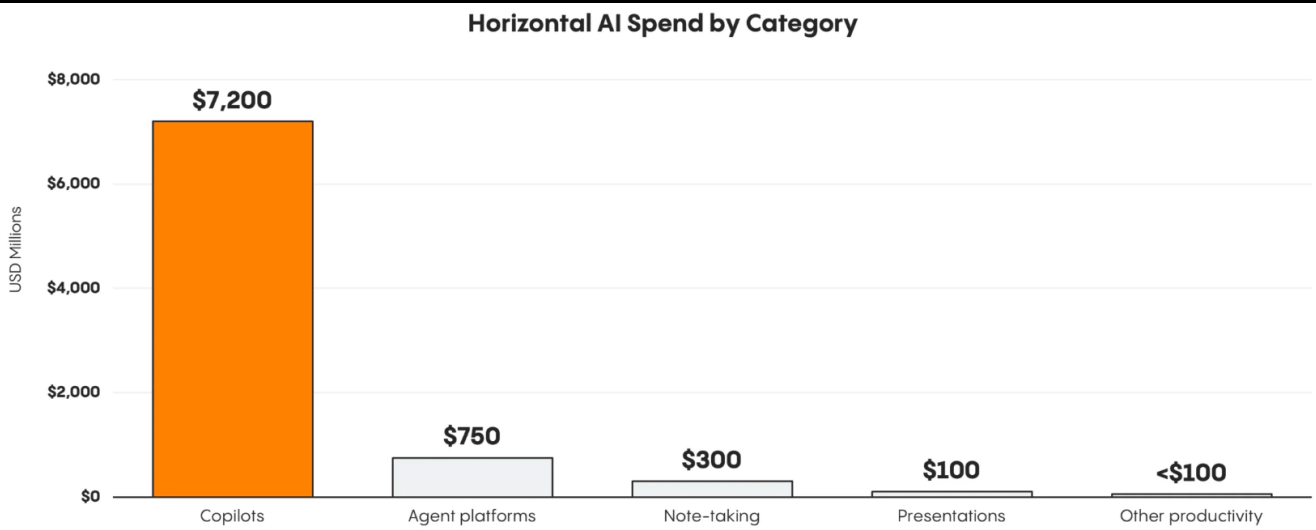
图76: AI native 应用大量涌现



数据来源: McKinsey 《The state of AI in 2025》, 东吴证券研究所

横向 AI (通用助手) 仍是体量最大的子市场, 2025 年达到 84 亿美元, 其中通用 copilot 类产品占据 86%、约 72 亿美元。ChatGPT Enterprise、Claude for Work、Microsoft Copilot 等产品成为大多数企业的第一入口。相比之下, 真正具备规划-执行-观察-迭代闭环的 agent 平台只占 10%, 约 7.5 亿美元。这反映出企业更倾向于“辅助型”AI 而非“全自动替代型”AI, 但随着模型在复杂任务上的可靠性持续提升, agent 的占比或在未来几年快速爬升。

图77: 2025 年横向 AI (通用助手) 支出规模

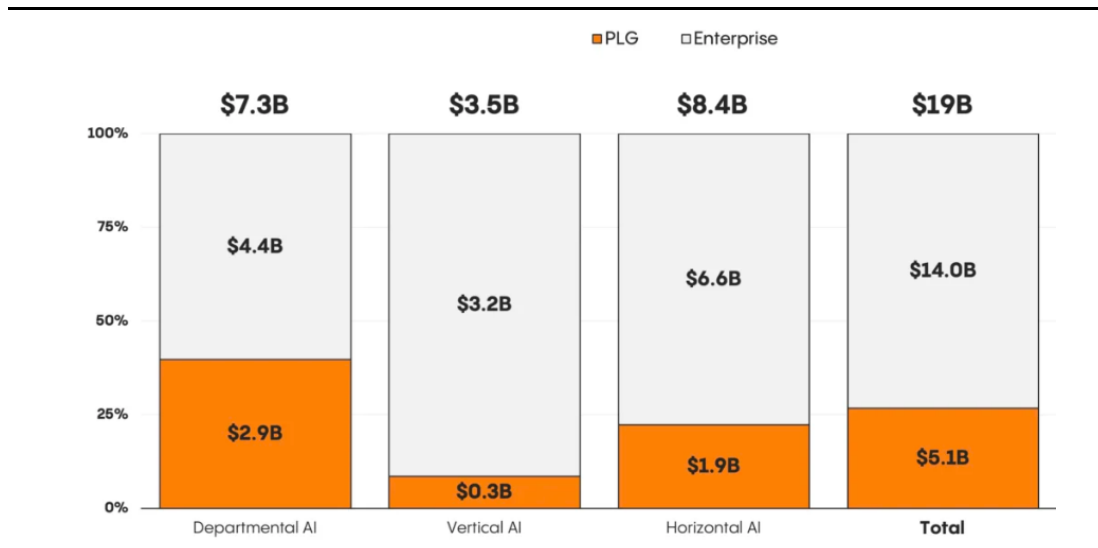


数据来源: Menlo Ventures, 东吴证券研究所

在采购路径上, ToB AI 与传统企业软件截然不同。产品驱动增长 (PLG) 已经成为

主流落地方式，27%的应用层支出直接来自 PLG 渠道，约为传统 SaaS 的 4 倍。如果再算上员工用个人信用卡支付 ChatGPT Plus 等“影子 AI”支出（其中约 27%用于工作），PLG 实际占比可能接近 40%。Cursor、n8n、ElevenLabs、Gamma 等公司，都是先靠开发者或个人用户自发传播，再逐步转化为企业合同的典型案例。这种自下而上的渗透路径，大大缩短了销售周期，也让 AI 工具能够在正式采购流程启动前就在企业内部形成标准。

图78：2025 年 toB AI 支出中 PLG 驱动的占比可观（单位：十亿美元）



数据来源：Menlo Ventures，东吴证券研究所

创业公司与传统巨头的竞争格局也因此分化。在应用层，2025 年创业公司已拿到 63% 的收入份额，尤其在产品工程、销售、财务运营这些敏捷部门，凭借更快迭代、更纯粹的 AI 原生体验，正在持续蚕食传统软件厂商。Cursor 击败 GitHub Copilot、Clay 绕开 Salesforce、Rillet 抓住 Intuit 行动迟缓的窗口，都是典型案例。基础设施层则相反，2025 年传统巨头仍占 56% 的收入份额，企业更倾向于继续使用 Databricks、Snowflake 等既有平台支撑 AI 负载。

我们预计 2026 年 toB AI 市场将延续两极分化的趋势：一端是平台巨头通过“模型即应用”的内部赋能，把 AI 变成基础设施，追求规模效应与存量效率跃升；另一端是少数垂直深耕者，通过独特数据壁垒、行业实施壁垒、情感或生产力壁垒，构建高单价、高粘性的护城河；中间大量缺乏场景根基、数据积累或分发能力的纯 AI 原生工具，则不可避免地面临被收编、替代或自然死亡的命运。

5.4. toC：模型即应用的降维打击

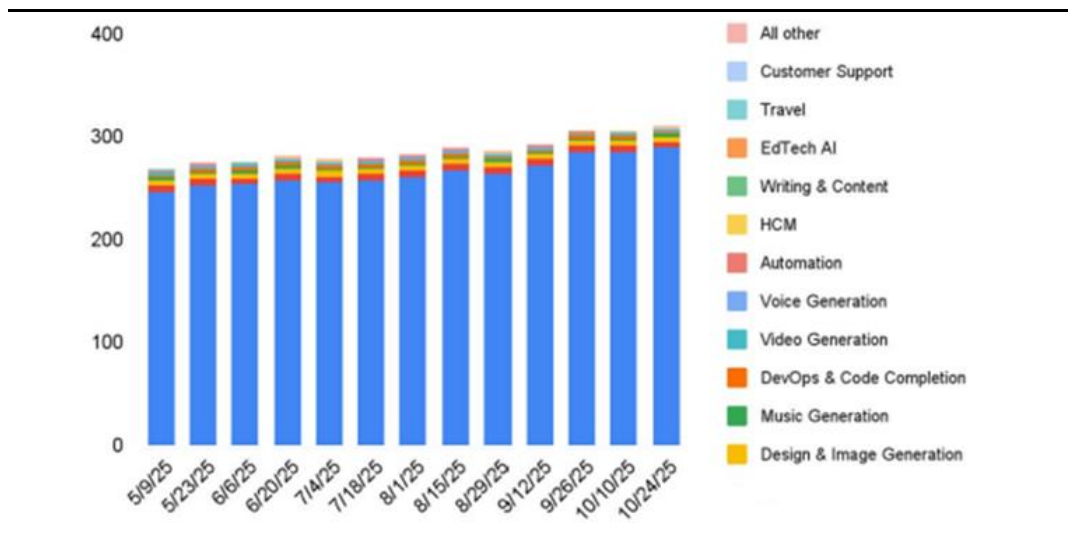
在 C 端 AI 产品市场，通用 AI chatbot 作为超级入口主导流量和用户时长，而垂直 AI 应用则面临“模型吞噬软件”的挑战，许多“薄壳”产品迅速消亡。与 B 端垂直 AI 不同，C 端缺乏私有数据壁垒和深度行业 know-how，导致通用大模型通过功能整合（如自定义助手、Gems 或应用商店）迅速蚕食独立垂直应用的生存空间。

然而，并非所有垂直 AI 都注定被吞噬——那些构建独特数据闭环、情感连接或高付费意愿的场景（如编程助手和角色扮演陪伴），正展现出顽强生命力。

通用 AI Chatbot 成为超级入口

2025 年，通用 AI chatbot 已占据 C 端 AI 的主要份额。根据 SimilarWeb 和 Sensor Tower，通用型 chatbot（如 ChatGPT、Gemini、Claude）占据了 80% 以上的 AI 网页访问量。“模型即应用”的趋势越发显著：用户无需下载多个独立 APP，一个统一界面即可完成写作、搜索、图像生成、简单编程等任务。2025 年，OpenAI 的 ChatGPT Atlas 浏览器和 Gemini 的 Deep Research 功能，进一步强化了数据闭环——用户行为直接反馈训练模型，形成“产品更好用→用户更多→模型更强”的正循环。国内市场类似，腾讯元宝接入 DeepSeek 后迅速登顶 App Store 免费榜，字节豆包日消耗 token 达数百亿，证明大厂生态+强模型的优势。

图79：按照类别划分的全球 AI 网页日均访问量（百万次/天）



数据来源：similarweb，东吴证券研究所

图80：2025 年按照类别划分的 AI 网页端访问量增速

12wk Change	5/23	6/6	6/20	7/4	7/18	8/1	8/15	8/29	9/12	9/26	10/10	10/24
General	23%	22%	21%	6%	5%	6%	6%	4%	5%	12%	11%	11%
Character & Chat	-1%	-5%	-4%	-5%	-9%	-8%	-5%	-1%	2%	4%	7%	4%
Design & Image Generation	-11%	-8%	-10%	-22%	-4%	-2%	0%	2%	-2%	-8%	-9%	-8%
Writing & Content	-16%	-19%	-20%	-18%	-14%	-12%	-11%	-10%	-8%	-8%	-15%	-14%
EdTech AI	-3%	-10%	-13%	-2%	-2%	7%	-1%	5%	4%	4%	-1%	-10%
Customer Support & Experience	-11%	-15%	-12%	-1%	8%	7%	11%	23%	16%	-5%	-20%	-22%
All other	-15%	-19%	-12%	-2%	15%	13%	10%	13%	3%	4%	-1%	1%

数据来源：Similarweb，东吴证券研究所

注：12wk change 指的是当期两周访问量相较于三个月前的两周访问量增速

然而，在各个通用 AI chatbot 中，用户的跨平台切换成本较低——流量增长更多取决于模型性能迭代。GPT-5 系列不及预期导致 OpenAI 网页增速放缓，而 Claude 凭借文本/coding 领先、Gemini 靠多模态和便宜 Flash 版逆势增长。这反映出 C 端用户行为的核心逻辑：**实用性和新鲜感驱动迁移**。谁的模型更“聪明”、更便宜、更拟人，用户就流向谁。

我们认为，未来通用 chatbot 将进一步向“超级 App”演化。2025 年底，Agent 模式普及（如 ChatGPT 的自主任务执行）标志着从“被动问答”向“主动代理”转型。2026 年，通用平台将整合更多外部工具（浏览器、购物、社交），渗透率进一步提升。

AI Graveyard 统计了过去 2023-2025 年消失的所有 AI 应用。在平台收录的 5000 多个 AI 应用中，已有 1457 款应用退市，仅 2025 年就接近 400 个（截至 2025/12/31）。

垂直 AI：大量死亡，但分化明显

与 B 端垂直 AI（如医疗的 Abridge、法律的 Harvey）通过私有数据构建壁垒不同，C 端垂直 AI 面临“壳薄”问题：离大模型能力越近的产品，越易被吞噬。

AI Graveyard 统计显示，2023-2025 年 5000+ AI 应用中 1457 款退市，2025 年近 400 个。其中，死亡率最高的是 AI 图像生成、AI 设计、AI 创意写作类 APP，原因很简单——它们离大模型太近。几乎所有 AI 应用都可以称之为“套壳产品”，问题不在于“是否套壳”，而在于“壳有多厚”。离模型能力越近、壳越薄的产品，越容易被大模型的发展所吞噬——即“**模型吞噬软件**”。

当 ChatGPT 推出应用商店、Gemini 提供 Gems 自定义助手功能后，用户可以在一个统一的界面中获得各种垂直场景的服务，而不需要为每个场景单独下载一个 APP。对于用户来说，为什么要在手机上安装十几个不同功能的 AI 应用，当一个 ChatGPT 就能完成所有任务？这种用户体验上的便利性优势，是独立 C 端应用很难逾越的障碍。

死亡率最高的是图像生成、设计、创意写作类——这些直接被通用模型的内置功能（如 ChatGPT 的 GPT-4o 图像、Gemini 的 Nano Banana）取代。

图81：AI Graveyard 上已经消失的 AI 应用（截至 2025/12/31）



数据来源：AI Graveyard，东吴证券研究所

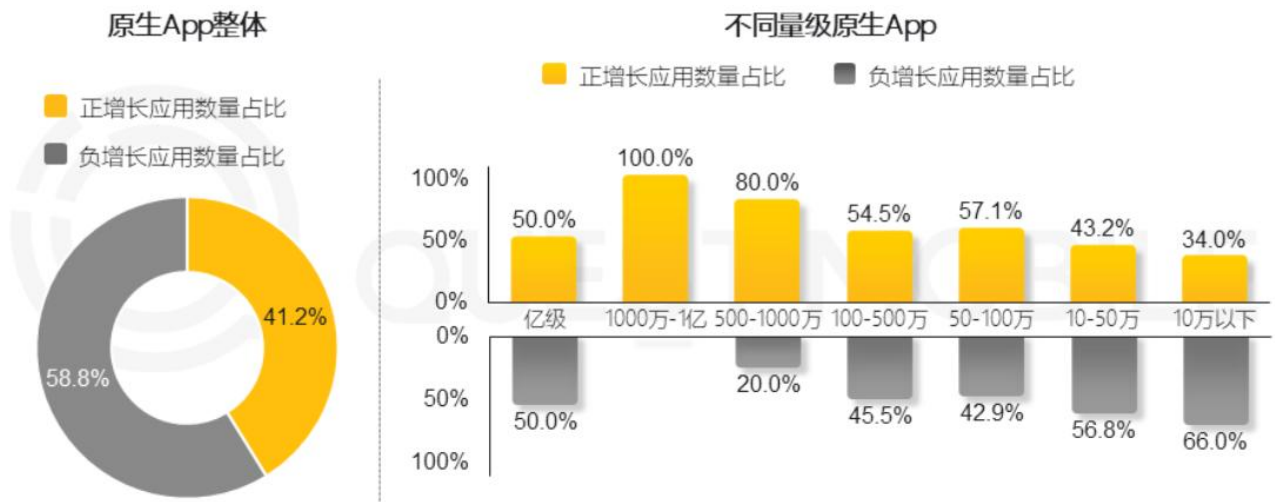
图82: AI Graveyard 上死亡率最高的产品类型 (截至 2025/12/31)

POPULAR CATEGORIES	POPULAR INTEGRATIONS
AI Image Generators from Text Tools	AI Telegram Bots
AI Design Tools	AI Microsoft Word Tools & Generators
AI Art Generator & Creator Tools	AI Wordpress Plugins & Tools
AI Writing Tools and Generators	Twitter AI Bots & Tools
AI Video Maker Generators & Tools	AI Google Docs Tools
AI Chatting Tools	Shopify AI Apps & Tools
Business AI Tools	Zapier AI Tools & Zaps
AI Assistant Tools	Facebook AI Tools & Generators
AI Chatbot Tools	Slack AI Bots & Summarizers
Audio AI Tools	Whatsapp AI Chatbots & Tools

数据来源: AI Graveyard, 东吴证券研究所

国内 APP 面临同样的困境。QuestMobile 统计显示, 25Q3 国内约 59% 的 AI native 应用用户数同比下降, 只有 41% 在增长。而且, 用户体量越小的长尾应用, 下滑越严重。体量在亿级的头部应用能维持五五开, 而中长尾应用的下滑比例更高。

图83: QuestMobile 统计的 25Q3 国内原生 AI APP 用户增长表现



注: 1、复合增长率, 统计2025年7月至2025年9月AI应用月活跃用户规模增长表现, 其值= $[(9月规模/7月规模)^{(1/2)}]-1$; 2、正(负)增长应用数量占比=正(负)增长的应用数量/应用合计数量; 3、部分应用以实际收录作为起始时间计算。

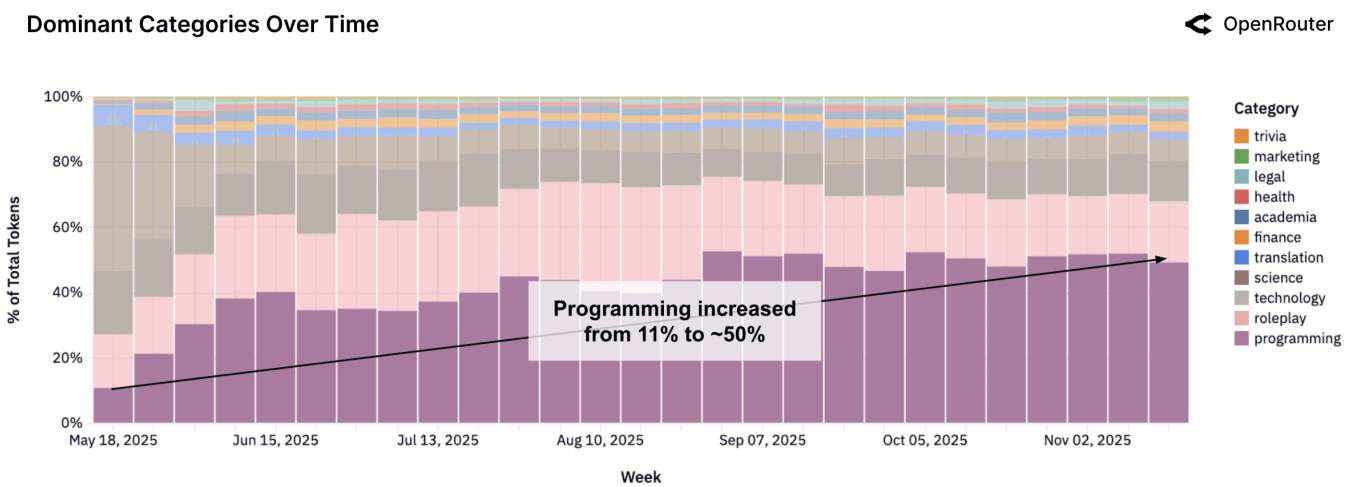
数据来源: QuestMobile, 东吴证券研究所

并非所有垂直 AI 都会消亡。生存者往往拥有独特数据壁垒或高情感/专业粘性:

编程工具类: OpenRouter 数据显示, 2025 年 6-12 月, 编程和角色扮演占平台总 Tokens 消耗量 60% 以上。

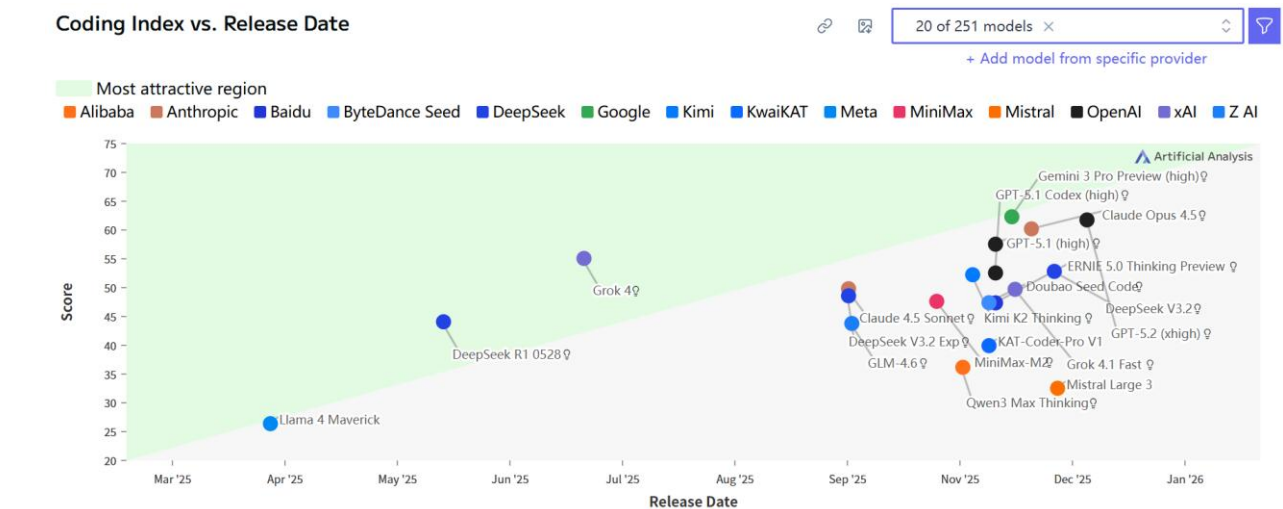
编程场景之所以率先爆发，因为：**刚需、高价值、可验证**。程序员每天都在写代码，AI 的帮助可以立刻体现在开发速度、缺陷率、重构效率、文档生成等环节；代码的语法规则明确，错误可以快速被编译器与测试发现，形成即时反馈闭环，产品迭代速度自然更快。更重要的是付费意愿高且 ROI 清晰，尤其企业端按席位购买时，一点点效率提升就可能对应巨大的人力成本节省。举例来说，一个 100 人工程团队每人每月 30 美元的工具成本，一年也就 3.6 万美元，但如果效率提升 20% 的量级成立，节省的人力价值远超成本。我们认为编程不仅是 AI 最先实现商业化的场景，也会成为 Agent 化软件开发的起点。

图84: OpenRouter 平台上 tokens 消耗最多的场景是编程和角色扮演



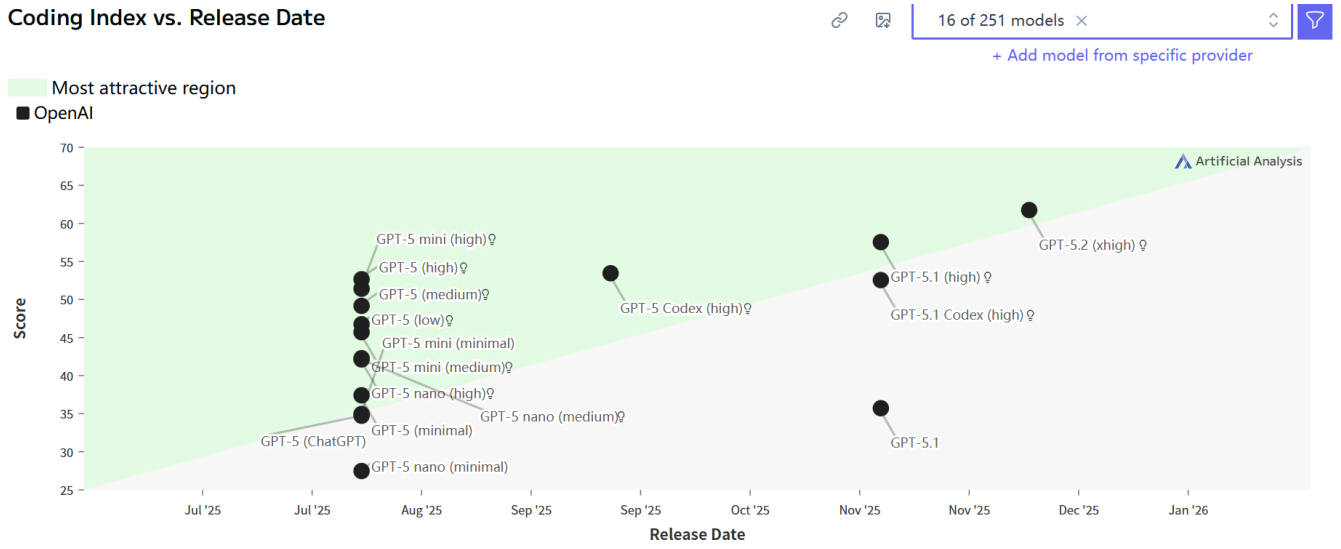
数据来源: OpenRouter, 东吴证券研究所

图85: 新发布的模型大多定向优化了 coding 能力



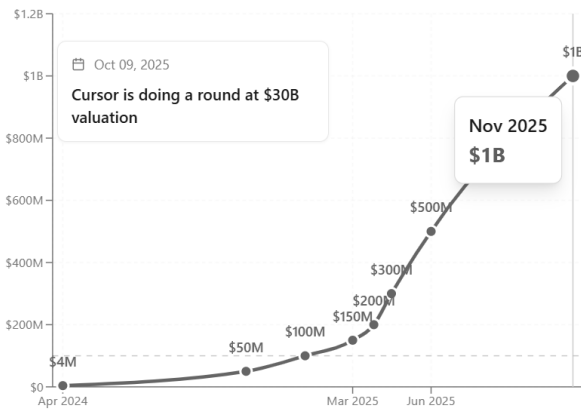
数据来源: Artificial Analysis, 东吴证券研究所

图86: OpenAI 此前的版本分为 high/medium/low, 而后专门推出 codex 版本



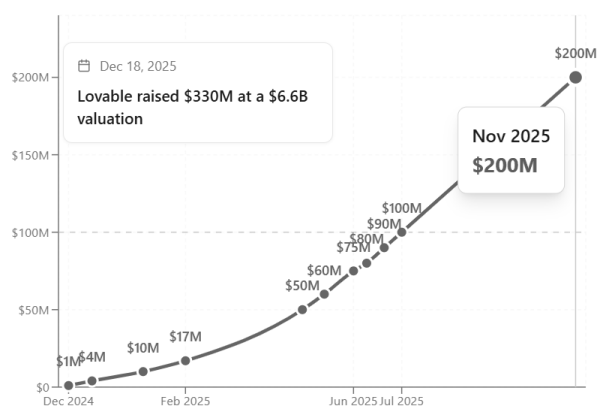
数据来源: Artificial Analysis, 东吴证券研究所

图87: Cursor ARR (十亿美元)



数据来源: ARR Club, 东吴证券研究所

图88: Lovable ARR (百万美元)



数据来源: ARR Club, 东吴证券研究所

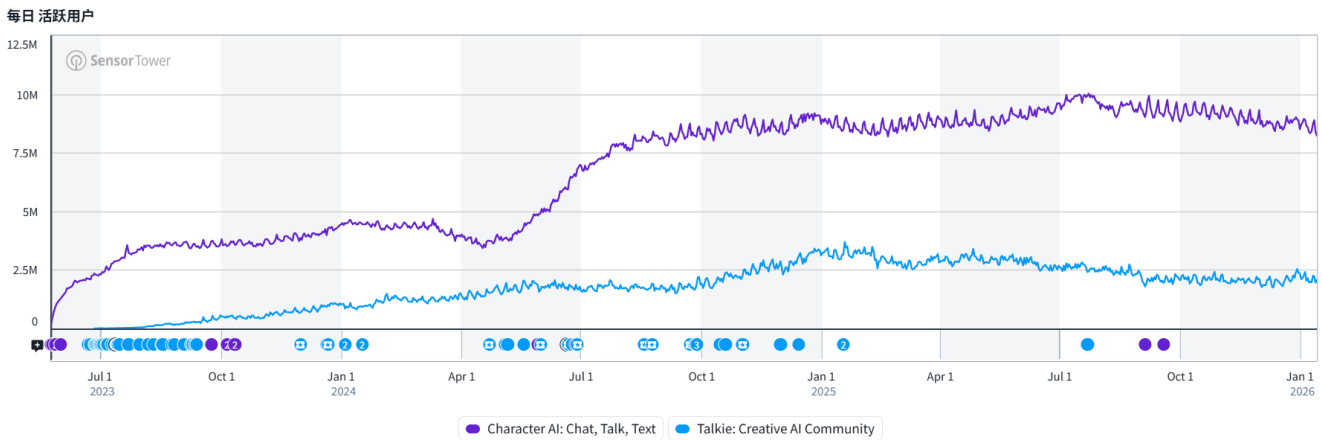
角色陪伴类: 角色陪伴 AI 不像编程助手那样能快速兑现高 ROI, 也不像通用 chatbot 那样靠模型性能碾压一切, 而是靠长时段、情感化的交互来驱动使用量。根据 OpenRouter 等平台的数据, 角色扮演和陪伴场景占据了 token 消耗的较大比例——用户一次对话可能持续几十甚至上百轮, 远超普通问答的几百字输入输出。

然而, Character AI 和 talkie 的 DAU 都在 2024 年快速增长, 而在 2025 年增长放缓。但我们认为, 原因并非需求不存在——open router 上角色扮演类 tokens 消耗快速增长的趋势证明, 需求是真实的。

最核心的挑战是**记忆不足与成本高企**。2025 年多数产品仍受上下文窗口和推理成

本限制，难以维持超长记忆；公司为了盈利往往主动压缩上下文，进一步恶化体验。

图89: character AI 和 talkie 的 DAU (百万人)



数据来源: sensortower, 东吴证券研究所

尽管如此，我们看好 AI 陪伴赛道的长期前景。一方面，目前面临的挑战有望解决：模型能力快速迭代，记忆机制（如更长的上下文、更便宜的 token、更好的长期记忆微调）正在逐步解决；tokens 成本也在下降（见于我们在前一章分析的 tokens 定价逐年下降趋势），允许平台放开上下文限制。

另一方面，AI 陪伴产品的最大优势在于情感数据闭环。用户一旦与某个角色建立“情感联结”，迁移成本较高。这种“感情绑定”正是核心壁垒：对话历史、个性记忆、专属剧情都沉淀在平台内，换到其他产品等于一切归零。用户投入的时间和情感越长，积累的专有对话数据就越丰富，这些数据不仅强化了角色记忆和个性一致性，还成为平台难以被通用模型直接复制的护城河。

通用 chatbot 短期内难以完全取代这类产品。通用 chatbot 如 ChatGPT 或 Gemini 虽能模拟角色，但很难在单一平台内沉淀出“专属情感关系”。你可以在 Gemini 里让它扮演任何人，但它不会记住你上个月的生日惊喜，也不会有其他用户共同构建的社区剧情。陪伴的核心价值恰恰在于这种私有化、连续性的情感投射，而这需要平台级的生态而非单次对话。我们认为，随着记忆与成本瓶颈突破，角色扮演与陪伴可能会成为少数几个不被通用模型完全吞噬的垂直领域之一。

5.5. 科技巨头内部赋能

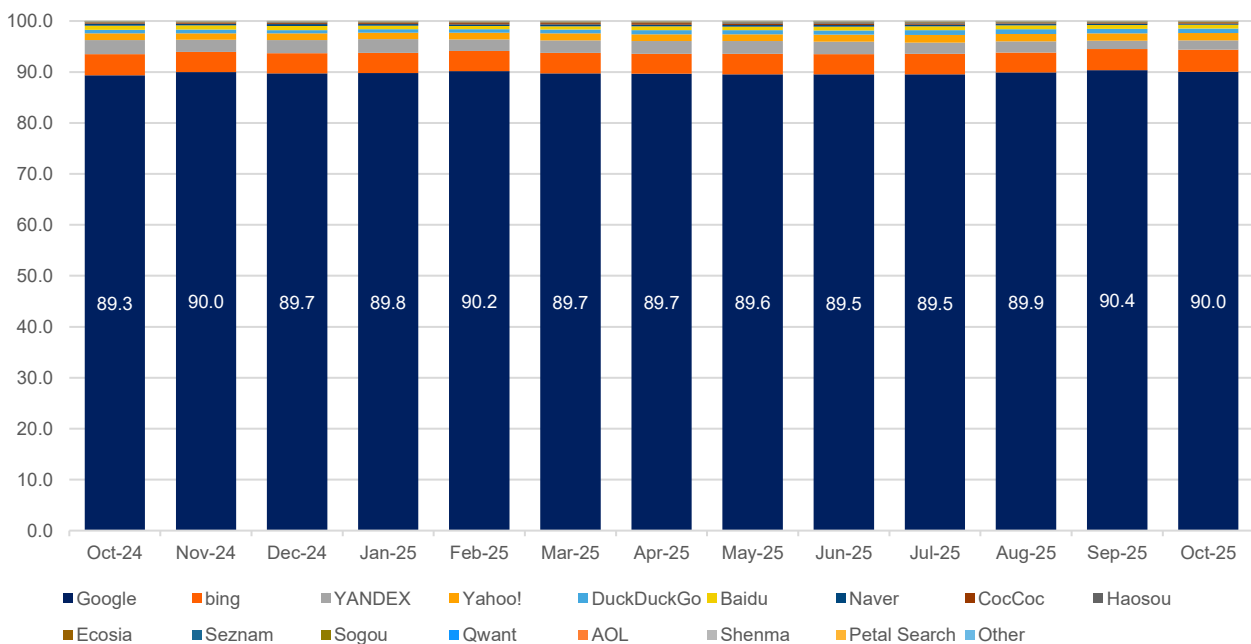
最后一条路径，也是我们认为最被低估的路径，是科技巨头的内部赋能。很多人盯着独立 AI 应用的订阅收入，却忽视巨头把 AI 深度嵌入自身业务带来的看不见的价值创造。

- ◆ 谷歌用 AI Overviews 拉动查询量并保持货币化率，打消 AI 蚕食广告担忧；YouTube 用 AI 推荐提升 Shorts 货币化效率；研发侧代码生成占比提升，直接提高工程师生产力。

- ◆ Meta 的 AI 广告工具提升效率，但其增量含金量需要更严格拆解，同时也要警惕自动化黑盒可能削弱对广告主的议价能力。
- ◆ 亚马逊通过 AI 优化物流效率、机器人部署与广告业务外延扩张，把 AI 的价值直接落在成本与效率上。
- ◆ 苹果的策略则更像把 AI 当作硬件体验的免费增配，只要能提高生态粘性与换机意愿，就不需要 AI 功能单独赚钱。
- ◆ 百度、阿里、腾讯将 AI 嵌入原有的生态。在 ToC 领域，用户对模型智能的感知已逐渐钝化——ChatGPT 从 2025 年到现在体验提升，对大多数普通人而言已不再震撼，大家更多将其当作加强版搜索引擎。因此，ToC 产品的核心瓶颈不再是模型本身更强，而是更丰富的上下文与环境输入。对于许多生活相关的问题，无论模型多强，仅依赖模型的历史知识来回答，都很难令人满意，只有接入真实生活数据（如天气、位置、社交记录）才能产生真正价值。这也解释了为何腾讯等 ToC 基因公司更关注模型与产品的垂直整合：模型能力必须与用户环境强耦合、快速迭代，才能持续带来惊喜。

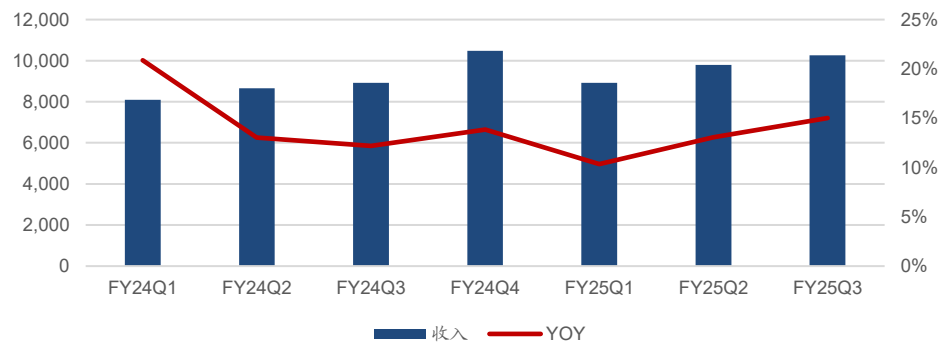
我们认为，内部赋能的共同特征是价值稳健、兑现路径清晰，但难以在财报上被单独标注；它更像把 AI 变成企业的全要素生产率提升，而不是一条新业务线。

图90: 谷歌的搜索引擎份额 (%) 稳定在 90%左右



数据来源: statcounter, 东吴证券研究所

图91: Alphabet 谷歌搜索收入增速企稳回升 (百万美元, %)



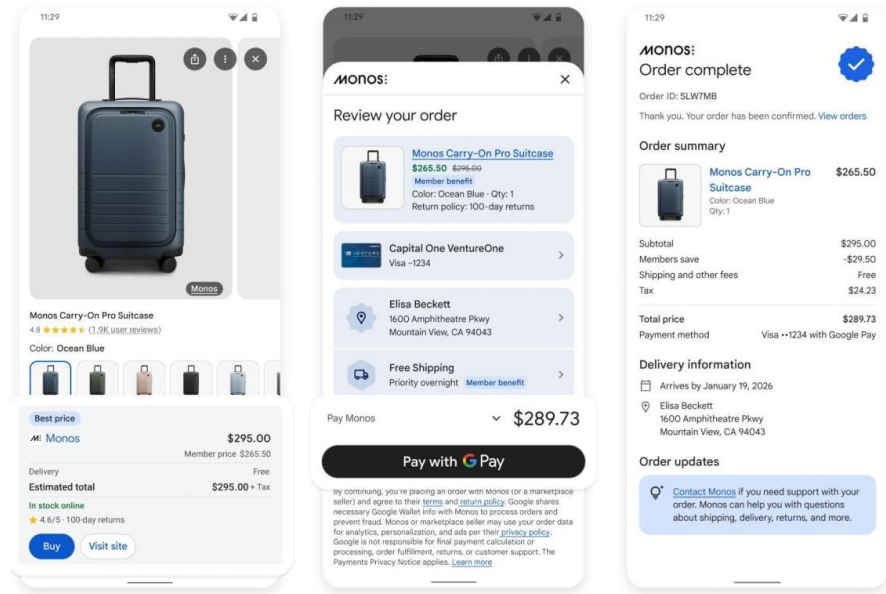
数据来源: wind, 东吴证券研究所

2026年1月, Google 正式推出 Universal Commerce Protocol (UCP), 这是一个开放的代理式商务标准, 允许用户通过 Gemini app 和 AIMode 直接完成产品发现、购买和售后支持。该协议由 Google 与 Shopify、Etsy、Wayfair、Target 等共同开发, 并获得 Adyen、American Express、Best Buy、Flipkart 等 20 多家支付和零售巨头的支持。核心目的是建立统一的语言, 让各种 AI 代理能在消费者界面、商家系统和支付平台之间无缝协作, 而无需为每个代理单独建立连接。

UCP 的推出意味着 Gemini 将从信息工具升级为交易工具。用户在对话中提到想要买东西, Gemini 不仅能推荐, 还能直接引导完成下单、支付和跟踪订单。Google Search 也将引入“Business Agent”功能, 进一步把搜索结果变成可执行的购物行为。这种代理式购物 (agentic shopping) 把原本分散在多个 App 和网页的操作, 集中到一个对话界面里完成。

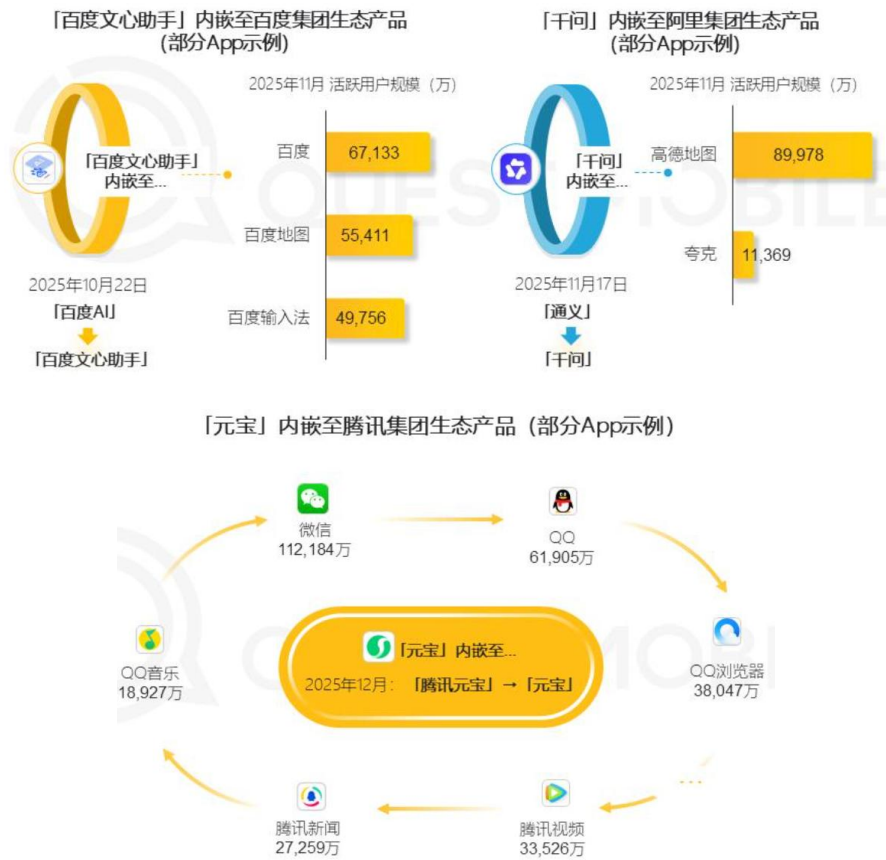
UCP 通过开放标准降低商家接入门槛, 避免了每个 AI 平台都要单独对接的碎片化问题, 这比封闭生态更有扩展性。Shopify 和 Etsy 这类平台型商家尤其受益, 它们本来就依赖第三方流量, 接入 UCP 后能直接触达 Gemini 的亿级用户, 而不用额外开发专属技能。支付方如 Adyen 和 American Express 也能借此扩大交易量, 整体生态形成正循环。

图92: Google 推出 UCP 协议, 支持 Gemini app 和 AI Mode 直接结账



数据来源: 9to5google, 东吴证券研究所

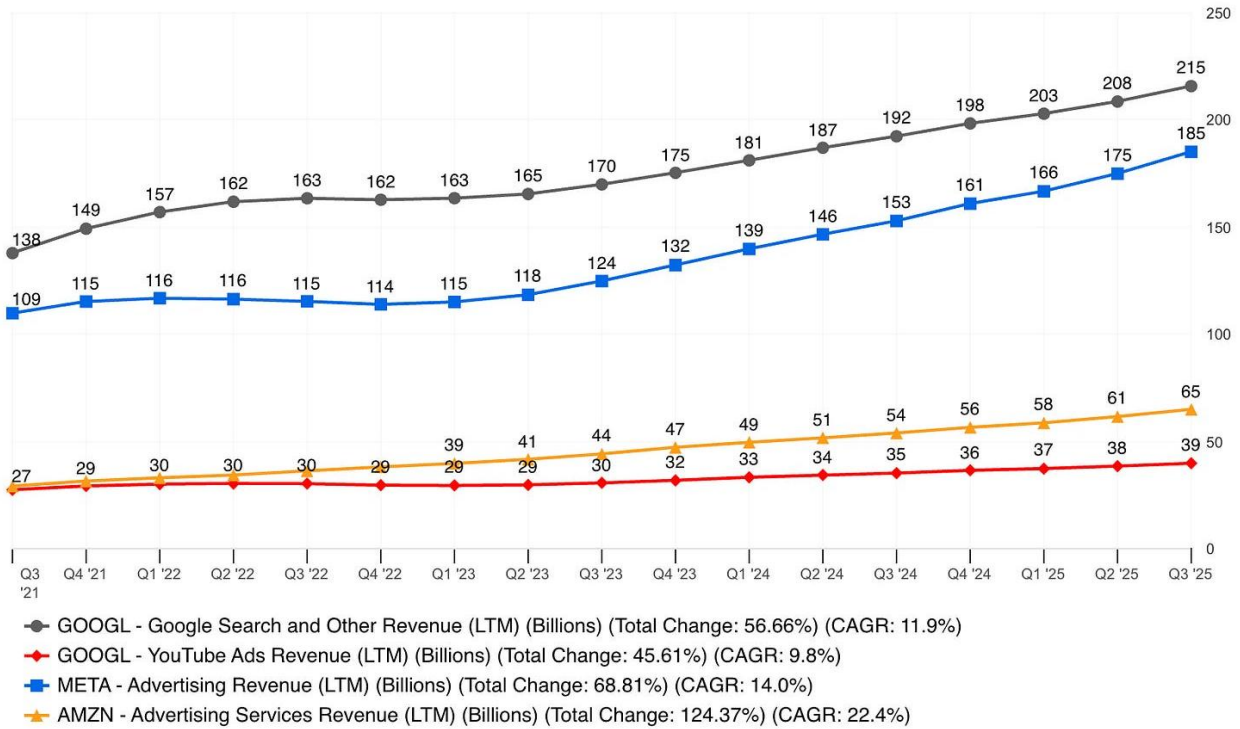
图93: 百度、阿里、腾讯将 AI 能力嵌入原有的生态产品



数据来源: QuestMobile, 东吴证券研究所

注: 下图中的数字表示 2025 年 11 月的月度活跃用户数

图94: 海外大厂广告收入-TTM (十亿美元)



数据来源: app economy insight, 东吴证券研究所

我们认为应用层商业化可以总结为：规模与盈利的平衡。 Tokens 消耗告诉我们需求在快速增长，甚至可能在 Agent 普及后指数级增长；ARR 告诉我们商业化健康度仍需验证，尤其在补贴、降价与成本下降同步发生时，收入质量与毛利结构需要进一步分析。编程场景给出了最清晰的健康答案，陪伴场景仍需等待模型能力进一步提升和 token 降价；企业级市场可能是最大的增量；巨头内部赋能最稳健，却最难被市场准确计价。

6. 投资建议

我们认为 2026 年 AI 将进入需求兑现与效率竞争并重的全新阶段。狂欢虽暂告段落，但新周期已然开启，投资机会不再是泛 AI 概念的普涨，而是聚焦于具备真实商业闭环、成本控制优势与长期护城河的核心玩家。局部泡沫虽存，但产业链上游硬件与主流云厂商基本面稳健，模型与应用层的分化则孕育结构性机会。我们看好 AI 产业链，核心思路是“上游强 alpha+云巨头稳健 beta+应用层精选赛道”。

在上游算力硬件环节，我们看好英伟达。尽管推理时代为其传统护城河带来一定挑战，但 Blackwell 系列供不应求、毛利率高位稳定，以及通过收购 Groq 快速补齐低延迟推理短板，显示其生态防御能力依然强劲。英伟达不仅是 AI 资本开支的最大受益者之一，更通过 CUDA 兼容性与旧卡向推理迁移，实质性拓宽了长期 ROI 空间。

建议关注谷歌、亚马逊。谷歌凭借 TPU 全栈优势、TCO 优势以及 Gemini 与谷歌生

态的深度融合，已展现出强利润韧性与增长潜力，AI Overviews 的成功落地进一步打消了市场对广告业务的担忧。亚马逊则通过 Trainium 自研集群与 AWS 规模效应，在成本与客户黏性上占据主动，同时物流、电商等多元化业务为其 AI 投入提供充裕现金流支撑。相较之下，微软虽有 Azure 订单支撑，但与 OpenAI 深度绑定的潜在风险使其价值略逊。

模型与应用层则需精选具备独特壁垒的赛道与个股。我们建议关注国内头部大模型公司 MiniMax-WP 和智谱。MiniMax-WP 在多模态与角色陪伴场景（Talkie 海外版）已实现较强商业化变现，海外营收占比高、用户粘性强，伴随记忆机制与成本瓶颈突破，长期增长潜力显著。智谱则凭借极致性价比的编程模型（Coding Plan）与开源生态影响力，快速抢占 B 端与开发者市场份额。

港股互联网公司中，我们推荐配置具备 AI 深度嵌入与生态优势的标的。阿里巴巴-W 背靠通义千问与阿里云，模型能力快速提升的同时，已在电商、娱乐、金融等多场景实现落地，性价比路线助力其在国内外开源市场份额持续扩张。腾讯控股通过元宝与 DeepSeek 合作快速起量，并在游戏、社交、内容生态中系统性赋能 AI，变现路径清晰。快手-W 在短视频与直播场景的多模态能力突出，可灵视频模型已展现较强竞争力。美图公司则在影像与美颜垂直领域形成 AI 原生优势，用户付费意愿强、商业化效率高。这些公司拥有庞大的用户基础与数据飞轮，AI 将成为其存量业务效率跃升与新增长曲线的核心催化剂。

7. 风险提示

全球企业与消费者对 AI 推理服务的实际需求增长不及市场预期：例如大模型在 B 端应用的付费转化率偏低，或 C 端用户对付费功能的接受度未能快速提升，可能导致云厂商与模型公司的收入兑现节奏放缓，进而传导至上游硬件采购的资本开支收缩，相关标的短期面临估值回调压力。

模型商业化路径出现重大变化：如开源模型的快速迭代大幅压缩闭源模型的差异化优势，或监管政策对数据隐私、内容生成合规性提出更严格要求，可能削弱部分玩家的变现能力与竞争壁垒，引发市场对增长可持续性的重新评估。

地缘政治因素：若中美科技摩擦进一步升级，导致先进芯片与算力设备的出口管制持续收紧，或全球供应链出现新的断点，将直接加剧算力资源约束，推高训练与推理成本，影响整个产业链的利润空间与扩张节奏。

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。

本报告的版权归本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制和发布。经授权刊载、转发本报告或者摘要的，应当注明出处为东吴证券研究所，并注明本报告发布人和发布日期，提示使用本报告的风险，且不得对本报告进行有悖原意的引用、删节和修改。未经授权或未按要求刊载、转发本报告的，应当承担相应的法律责任。本公司将保留向其追究法律责任的权利。

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>