东吴证券 SOOCHOW SECURITIES

计算机行业深度报告

解析大模型行业：从发展历程到投资视角

2026 年 06 月 18 日

增持（维持）

证券分析师 王紫敬
执业证书：S0600521080005
021-60199781
wangzj@dwzq.com.cn

投资要点

- **当前大模型产业正处于基础设施高强度扩张、模型能力快速扩散、Agent 初步产品化、企业逐渐实现 ROI 验证的关键发展阶段。**从技术演进看，大模型已完成从 Transformer 奠基（2017-2019）、预训练规模化（2020-2021）、指令对齐与产品化（2022-2023）、多模态开源与企业化（2023-2024）的四轮跃迁，2025-2026 年正式进入以推理模型与 Agent 能力为核心的第五阶段，竞争焦点从 "回答问题" 转向 "完成任务"，评价标准升级为复杂目标规划、工具调用、动作执行与结果验证能力。

- **利润池沿产业链持续迁移，投资逻辑从追逐模型能力转向更低成本、更高可靠性完成真实任务。**早期利润池集中于训练 GPU，ChatGPT 问世后扩展至 API、订阅与 AI 原生应用，多模态与开源阶段私有化部署、RAG、向量数据库、数据治理成为增量，当前推理模型与 Agent 阶段进一步外溢至推理 GPU/ASIC、HBM、高速网络、电力、液冷、Agent 平台与工作流软件。**大模型产业链价值优先流向三类环节：一是 GPU/ASIC、HBM、先进封装、光模块等稀缺资源；二是 CUDA 生态、云平台、企业数据等高切换成本能力；三是代码助手、客服、办公等可量化 ROI 应用入口。**当前利润池集中在基础设施，GPU 与 ASIC 显性受益，HBM 与先进封装是供给瓶颈，光模块与交换机等是集群扩张后的第二弹性；云厂商是企业 AI 主要入口但资本消耗较快；应用层长期空间最大但短期分化显著，代码、客服、办公等高频场景落地最快。开源模型压缩通用能力超额利润，推动多模型路由成为常态，竞争焦点转向入口、数据、成本控制与工作流嵌入。**之于大模型本身，当前商业模式仍处在探索阶段，尽管已经有相对稳定的付费模式，但用户心智、使用场景、工作流嵌入实际等均未稳定。考虑到大模型是未来智能时代的智能底座，大模型本身同样具有投资价值。**

- **关键大模型标的：智谱、Minimax、科大讯飞。（1）智谱：**从项目型 AI 公司向 MaaS 平台转型，Coding 是其从模型能力转向收入规模的最短路径；**（2）MiniMax：**通过架构优化实现成本与能力平衡，走全球化与全模态平台化路线，Token Plan 构建多模态智能额度池的商业模式；**（3）科大讯飞：**核心价值在于国产算力全流程训推能力，是国内唯一实现全国产算力上全栈模型训练的厂商，在政企自主可控采购中具备稀缺性。公司 2025 年大模型相关项目中标金额 23.16 亿元，超过第二名至第六名总和，印证我们的观点：公司卡位的稀缺性具备极强的市场竞争力。

- **大模型的发展路径尚未明确，我们认为未来大模型行业的发展可能存在三种情况。（1）闭源模型持续领先并放大代际差距：**利润将持续集中在 AI 硬件以及头部模型厂；**（2）开源模型持续追赶并缩小代际差距：**开源模型持续追赶推动模型能力商品化，形成多模型分层路由格局，模型层 API 毛利承压但应用层与基础设施受益；**（3）出现新的模型架构突破：**可能催生一批全新的企业。

- **投资建议：建议关注大模型与大模型相关产业链。1、大模型：**建议关注智谱（港股）、Minimax（港股）、科大讯飞、阿里巴巴（港股）；**2、其他大模型相关板块：建议关注：（1）AI 芯片**（国产算力、半导体设备）；**（2）数据中心**（AI 服务器、光模块、PCB/交换设备、液冷/电源/电网电力设备）；**（3）AI 应用；（4）物理/端侧 AI。****详见正文 P45“6.4 市场投资框架”。**

- **风险提示：**技术迭代不及预期、行业竞争加剧、算力供应链风险、监管政策不确定性、商业化落地慢于预期。



行业走势



相关研究

《金刚石散热：新一代 AI 散热方案》

2026-06-11

《SpaceX 商业概览：从火箭、通信到 AI》

2026-05-26

仅供内部参考，请勿外传！

内容目录

1. 大模型技术演进史 5
1.1. 从 Transformer 到 Agent：五个阶段 5
1.2. 每个阶段解决的问题和引入的新瓶颈 6
2. 大模型发展阶段映射 9
2.1. 大模型发展到了什么阶段？ 9
2.2. 与历史技术周期的比较 12
2.2.1. 互联网 1995—2000 年：相似的叙事与不同的成本结构 13
2.2.2. 云计算 2008—2015：最接近 AI 的历史模板 15
2.2.3. 移动互联网 2009—2014：AI 会重构入口，但入口更分散 15
3. 大模型产业链拆分与价值分配 17
3.1. 产业链价值分配的核心逻辑 17
3.2. 产业链利润为什么首先流向基础设施 17
3.3. 产业链价值分配图谱：当前利润池主要在哪里？ 18
3.4. 开源模型会压缩哪些利润，放大哪些需求？ 20
4. 竞争格局的本质 21
4.1. 竞争元素趋于多元，各大模型厂暂时在不同领域站稳脚跟 21
4.2. 重点厂商展开 21
4.2.1. OpenAI：从 GPT scaling 到工作型智能体平台 22
4.2.2. Anthropic：企业安全、长程任务 23
4.2.3. Google DeepMind / Gemini：全栈平台型模型厂 25
4.2.4. xAI / Grok：模型厂商+大规模 AI 基础设施运营商 25
4.2.5. DeepSeek：高性价比、开源冲击与百万上下文推理效率 26
4.2.6. Qwen：从开源生态到 Agent 全栈云平台 28
4.2.7. 字节：产品化、推荐数据和多模态生产部署 30
4.2.8. MiniMax：长上下文、Hybrid MoE 与低成本 test-time compute 31
4.2.9. Kimi：从长上下文聊天到长程编码与 Agent Swarm 32
5. 关键标的：智谱、Minimax、科大讯飞 33
5.1. 智谱：从私有化大模型公司走向 Coding + Agent MaaS 平台 33
5.1.1. 智谱的重估逻辑正在从项目型 AI 公司切向模型能力驱动的 MaaS 平台 33
5.1.2. 从 GLM-5 到 GLM-5.2，押注 Agentic Engineering+长程任务 34
5.1.3. 从本地部署到 MaaS，智谱正在寻找中国版 Anthropic 路径 36
5.1.4. 企业客户和国产芯片协同 37
5.2. Minimax：从 AI 原生应用走向全模态 AI 平台 37
5.2.1. MiniMax 的独特性在于全模态+全球 C 端+开放平台” 38
5.2.2. 模型路线：从 M2 到 M3 的 1M 上下文+原生多模态+Agentic Coding 39
5.3. 科大讯飞：唯一国产算力全流程训推的大模型底座标的 40
6. 未来大模型格局推演与投资建议 43
6.1. 情景 A：前沿模型继续 Scaling，闭源巨头优势扩大 43
6.2. 情景 B：开源模型持续追赶，模型能力商品化 44
6.3. 情景 C：架构或新范式突破，Transformer 优势被削弱 44
6.4. 市场投资框架 45
7. 风险提示 45

请务必阅读正文之后的免责声明部分

图表目录

图 1： 大模型演进示意图 6
图 2： 大厂 CapEx 呈加速上升趋势 9
图 3： 英伟达数据中心收入持续高增 10
图 4： 微软 AI 业务收入快速增长 10
图 5： 开源模型虽然与头部闭源模型能力依然有差距，但仍在持续追赶 10
图 6： 以 GitHub Copilot Business 为例，2026 年价格为 228 美元/人/年，对比 BLS 发布的 2025 年 5 月软件行业平均时薪测算的回本周期 12
图 7： GitHub Copilot 与人力成本对比 12
图 8： 调研中认为 AI 在该领域带来了改进的比例 12
图 9： 美国互联网公司融资情况，1995-2000 14
图 10： AI 用户渗透率尚处早期 14
图 11： AI 时代与互联网时代前期的纳指走势高度相似（横坐标 n 为起始日后第 n 年，我们此处定义互联网时代起始于 1994 年 1 月 1 日，AI 时代起始于 2023 年 1 月 1 日；纵坐标为纳指上涨倍数，以起始日期为基准） 14
图 12： 云产业链的价值链传导方向与 AI 类似 15
图 13： 2009-2014，移动操作系统市场最终形成安卓+iOS 双寡头格局（纵坐标为市占率） 16
图 14： 截至 2026 年 5 月，ChatGPT 全球 AI 助手市场的份额降至 46.4%，大模型或将进一步进入多平台并存阶段 17
图 15： DeepSeek-V4 保留 Transformer 架构和 MTP 模块，同时引入 mHC、CSA+HCA 28
图 16： 智谱收入边际增长转向云端 API、Coding、Agent 和 MaaS 33
图 17： GLM5 系列表现与同期大模型相比能力亮眼 35
图 18： GLM5.1 Coding 能力较前代显著提升，在 SWE-Bench Pro、NL2Repo、Terminal-Bench 2.0 等任务上具备更强表现 36
图 19： Minimax 业绩基本情况 38
图 20： 科大讯飞业绩基本情况 41

表 1： 大模型演进时间表以及投资逻辑迁移一览 8
表 2： 当前部分主流模型公司的 Agent 产品 11
表 3： AI 全产业链价值分配图谱，当前利润池集中在硬件与底层资源 18
表 4： 全球主要大模型厂商汇总表 22
表 5： GPT 的几个发展阶段 23
表 6： Claude 发展历程一览 24
表 7： Grok 系列产品一览 26
表 8： 千问系列产品 29
表 9： 豆包大模型发展梳理 31
表 10： Minimax 迭代历程 32
表 11： Kimi 系列产品 33
表 12： 如果闭源模型持续扩大对开源模型的领先优势，将持续利好 AI 硬件等受益环节 44

仅供内部参考，请勿外传！

表 13： 开源模型持续追赶的情形下，AI 应用等板块优先受益，AI 硬件板块可能受到冲击 44
表 14： 建议关注标的列表 45

请务必阅读正文之后的免责声明部分

1. 大模型技术演进史

1.1. 从 Transformer 到 Agent：五个阶段

过去九年，大模型能力栈大致经历了五轮重构：从底层架构突破，到预训练规模化，再到指令对齐、产品化、多模态与企业化，最终进入以推理能力和 Agent 执行为核心的新阶段。**每一次重构都不仅是模型能力的提升，更对应着训练范式、工程体系、商业模式和应用边界的系统性变化。**

第一阶段是 Transformer 奠定基础，2017—2019。Transformer 的关键突破在于以自注意力机制（Self-Attention）替代 RNN 的递归结构和 CNN 的局部卷积，使模型能够更高效地并行训练，并具备捕捉长距离依赖关系的能力。这一变化表面上是自然语言处理架构的升级，本质上则为后续大规模训练奠定了工程基础。由于 Transformer 的主体计算以矩阵乘法为核心，天然适配 GPU、TPU 等现代加速器，因此其意义并不局限于算法层面，而是同时打开了模型规模化训练的硬件效率空间。Transformer 原论文提出了完全基于 Attention 的网络架构，并在机器翻译任务中验证了其效果和训练效率，由此成为后续大模型发展的基础范式。

第二阶段是预训练规模化，2020—2021。GPT-3 的出现证明，大规模自回归语言模型可以在少样本学习（Few-shot Learning）条件下完成多类型任务，而不再需要针对每一个具体任务单独训练模型。这一阶段的产业意义在于，基础模型（Foundation Model）开始成为行业共识，模型 API 商业化也由此起步。此前，AI 项目更多围绕单一场景、单一任务进行定制化建模；GPT-3 之后，产业范式逐步转向“通用基础模型+Prompt+API+微调/RAG”的平台化模式。**模型不再只是某个任务的工具，而开始成为承载多类应用能力的底座。**

第三阶段是指令对齐与产品化，2022—2023。InstructGPT 和 ChatGPT 的核心贡献，在于解决了大模型能够续写文本但未必能够理解和执行用户指令的问题。通过监督微调（SFT）、人类反馈强化学习（RLHF）以及安全策略约束，大模型从研究系统转化为普通用户可直接使用的交互式产品。这一阶段的关键变量已经不再只是参数规模，而是后训练能力、交互体验、安全边界和低延迟服务能力的综合结果。换言之，模型真正进入产品化阶段，靠的不只是“更大”，而是“更可控、更好用、更稳定”。

第四阶段是多模态、开源、长上下文与企业化并行推进，2023—2024。GPT-4 证明，大模型能力可以从文本扩展到图像理解以及更复杂的专业任务；与此同时，Llama、Qwen、DeepSeek、Mistral、Claude、Gemini 等模型推动闭源与开源路线并行演进。企业在这一阶段逐渐形成更务实的认知：并非所有业务场景都需要调用最强闭源模型。对于客服、文档处理、代码生成、知识库问答和行业流程自动化等任务，开源模型叠加 RAG、私有化部署和业务系统集成，往往已经能够满足可用性和成本要求。**因此，模型竞争开始从单点能力比拼，转向“性能、成本、部署方式、数据安全和行业适配”的综合权衡。**

第五阶段是推理模型与 Agent 能力成为主线， 2025—2026。OpenAI o 系列、GPT-5.5、Claude Opus 4.7、Gemini 3.5 Flash、DeepSeek R1/V4、Qwen 3.7-Max、Kimi K2.6、MiniMax M 系列、腾讯 Hy3、百度 ERNIE 5.0 等模型的演进表明，前沿竞争正在从“回答问题”转向“完成任务”。模型能力的评价标准不再只是知识覆盖、语言生成或单轮问答质量，而是能否围绕复杂目标进行规划、调用工具、执行动作，并对结果进行验证和修正。OpenAI GPT-5.5 强调跨工具完成复杂任务，Qwen 3.7-Max 明确面向 Agent 时代，DeepSeek V4 则试图通过 1M 上下文和高效注意力机制改善长程 Agent 的成本结构。**由此看，大模型产业正在进入新的能力重构阶段：模型不只是信息生成器，而是逐步演化为能够嵌入工作流、调用外部系统并承担部分执行职能的智能操作层。**

图1：大模型演进示意图



数据来源：东吴证券研究所绘制

1.2. 每个阶段解决的问题和引入的新瓶颈

大模型能力栈的每一次跃迁，本质上都是在解决上一阶段的核心约束，同时又引入新的工程瓶颈和产业问题。

Transformer 解决了传统序列模型难以并行训练、长距离依赖捕捉能力不足的问题，使大规模预训练成为可能；**但随着上下文长度持续扩展，模型又面临长上下文计算成本上升、KV Cache 显存占用快速膨胀等新瓶颈。**

GPT-3 进一步解决了模型任务泛化能力不足的问题，证明大规模语言模型可以通过少样本学习完成多类任务；**但与此同时，训练成本、高质量语料供给以及输出可靠性问题开始成为行业关注重点。**

初代 ChatGPT 则解决了大模型与普通用户之间的交互门槛问题，使指令跟随和自然语言交互成为主流产品形态；**但其背后也带来了更高的对齐成本，以及幻觉、安全、推理成本等一系列落地难题。**

进入多模态与开源阶段后，大模型开始突破文本边界，向图像、音频、视频等更多模态扩展，同时开源模型降低了企业部署门槛；**但这一阶段也伴随模型价格战加剧、版权争议、跨模态幻觉以及企业合规压力上升。**

推理模型与 Agent 的出现，则进一步推动大模型从“对话工具”走向“任务执行系统”，提升了复杂问题拆解、工具调用和工作流执行能力；但相应地，**单任务 Token 消耗更高，工具调用安全、长链路错误累积、任务评测体系以及每成功任务成本，开始成为新的产业核心矛盾。**

从投资角度看，上述大模型技术演进的主线并不是单纯的模型能力提升，而是利润池沿着算力、模型、应用、数据、安全和工作流持续迁移。 早期利润池主要集中在训练 GPU；ChatGPT 问世后，商业化空间扩展至 API、订阅制产品和 AI 原生应用；多模态与开源阶段，企业私有化部署、RAG、向量数据库、数据治理和安全合规成为重要增量；而在推理模型与 Agent 阶段，利润池进一步外溢至推理 GPU/ASIC、HBM、高速网络、电力、液冷、模型路由、Agent 平台以及工作流软件等方向。

因此，大模型产业链的投资重点正在从追逐模型能力逐步转向谁能以更低成本、更高可靠性完成真实任务。 这意味着，未来产业竞争的关键不只在模型参数和榜单表现，更在推理效率、系统稳定性、企业部署能力以及单位任务成本的持续优化。

表1：大模型演进时间表以及投资逻辑迁移一览

年份	代表模型/事件	核心技术	能力跃迁	产业影响	投资影响
2017	Transformer	自注意力机制、并行训练、位置编码	模型能够在**长距离 token 之间建立依赖关系**，并比 RNN 更适合 GPU/TPU 并行训练	自然语言处理架构范式改变，为后续大规模预训练奠定基础	**GPU 训练需求开始被长期放大，**AI 计算开始从任务模型走向基础模型
~2019	BERT、GPT-1、GPT-2、T5	预训练+微调、自回归生成、编码器/解码器架构分化	语言理解和生成能力显著提升，迁移学习成为 NLP 主线	预训练模型成为通用语言底座，任务模型开始被基础模型替代	**云训练、数据标注、NLP 应用和模型服务**开始受益
2020	GPT-3	大规模自回归预训练、少样本学习、提示词接口	模型不再需要针对每个任务单独微调，可以通过 prompt 执行多类任务	基础模型成为产业共识，模型 API 商业化起步	**GPU、云算力、模型 API、开发者平台**开始被重估
2021	Codex、Switch Transformer	代码预训练、多专家混合	编程辅助能力出现，稀疏激活证明可以在不等比例增加计算量的情况下提升模型容量	AI 编程商业化开端，MoE 成为后续降本路线	**开发者工具、云 API、MoE 训练和推理系统**受到关注
2022	InstructGPT、ChatGPT	监督微调、人类反馈强化学习、对话产品化	模型从会续写文本变成能听指令、能对话、能拒答的大众产品	大模型从研究系统进入 C 端消费和企业试点，AI 订阅与 API 收入出现	**AI 应用、搜索、办公、客服、教育和开发者工具**开始重估
2023	GPT-4、Claude、Llama、Gemini	多模态模型、开放权重、长上下文、企业级对齐	模型进入图文理解、专业考试、代码和企业知识场景	闭源与开源并行，企业开始关注私有化、RAG 和合规	**云厂商、模型厂、数据库、向量库、安全治理**受益
2024	Llama 3、Qwen2.5、DeepSeek-V3、Claude 3.5、GPT-4o	多专家混合、低成本训练、实时多模态、模型价格战	模型性价比显著提升，开源模型接近闭源中高端能力	通用模型 API 溢价被压缩，AI 应用成本下降	**硬件仍受益，**但模型层毛利压力上升；**应用层** ROI 改善
2025	DeepSeek-R1、OpenAI o 系列、Claude 4、Gemini 2.5、Qwen3、Kimi、MiniMax-M1	可验证奖励强化学习、测试时计算扩展、工具调用、Agent 工作流	模型在**数学、代码、复杂推理和多步任务执行**上显著提升	产业焦点从聊天转向完成任务，推理 token 和工具调用需求上升	**推理 GPU/ASIC、调度、网络、电力和 AI 云**受益
~2026.5	GPT-5.5、Claude Opus 4.7、Gemini 3.5 Flash、DeepSeek V4、Qwen3.7-Max、	**智能体式编码、**计算机使用、百万上下文、混合注意力、多智能体编排	模型从回答问题进入**跨工具、跨文件、跨系统执行任务**的阶段	Agent 平台、企业 workflow、AI 云、私有化	投资重点从单纯训练侧 capex 扩展为**训练+推理+**

	Kimi K2.6、Seed2.0			部署和应用 ROI 成为主线	**电力+Agent 应用 +数据治理**

数据来源：东吴证券研究所整理

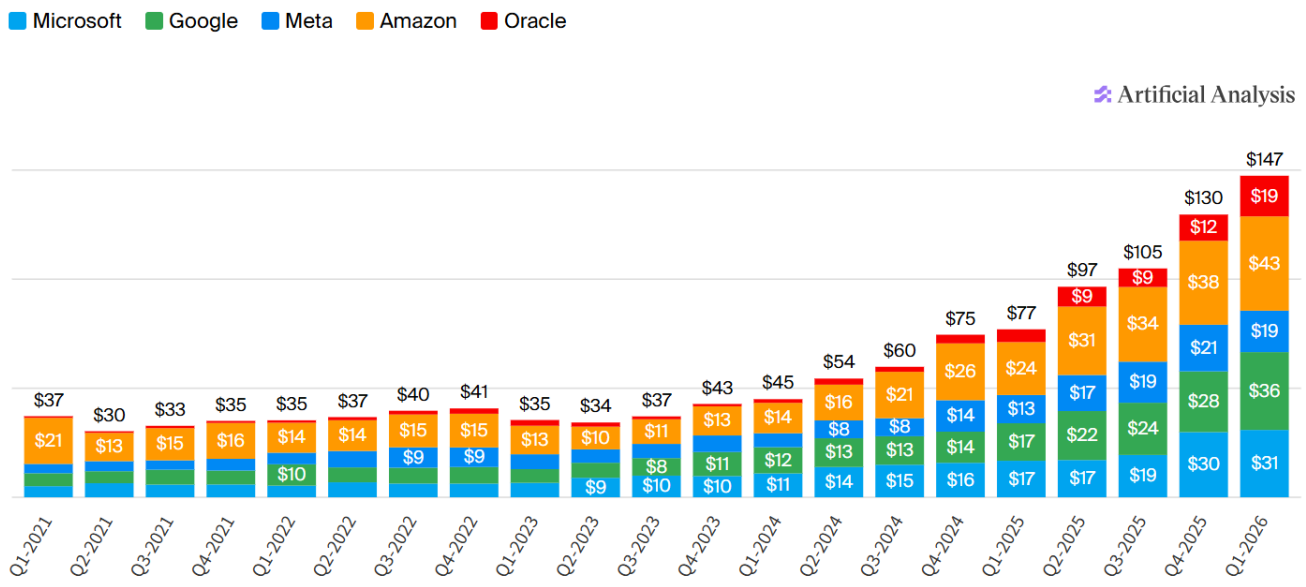
2. 大模型发展阶段映射

2.1. 大模型发展到了什么阶段？

当前 AI 发展正处在基础设施高强度扩张、模型能力快速扩散、Agent 初步产品化、企业逐渐实现 ROI 验证的时期。

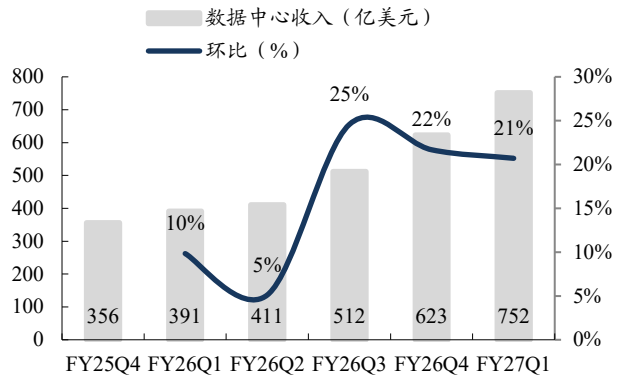
基础设施仍在快速扩张，GPU/ASIC、HBM、先进封装、网络、液冷、电力和数据中心仍处于强需求阶段。从 CapEx 节奏来看，自 2023 年以来大厂资本开支呈加速上升趋势；NVIDIA FY2027Q1 数据中心收入同比增长 92%，Microsoft AI business 年化收入超过 370 亿美元，Amazon AWS 同比增长 28%。这些数据说明，AI 已经从实验性预算，进入云和基础设施的真实收入兑现阶段，AI 产业链已经在数据中心收入、云计算增速和硬件订单中形成了可验证的产业趋势。

图2：大厂 CapEx 呈加速上升趋势



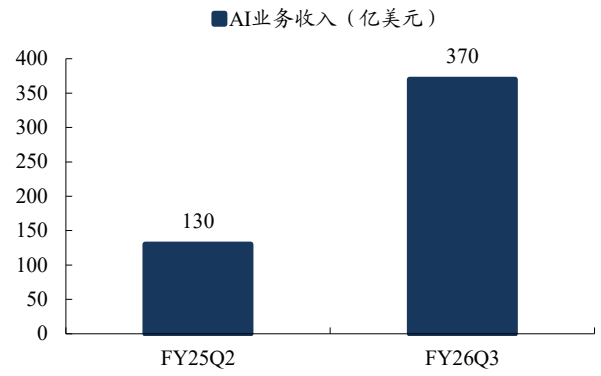
数据来源：Artificial Analysis，东吴证券研究所

图3：英伟达数据中心收入持续高增



数据来源：英伟达公告，东吴证券研究所

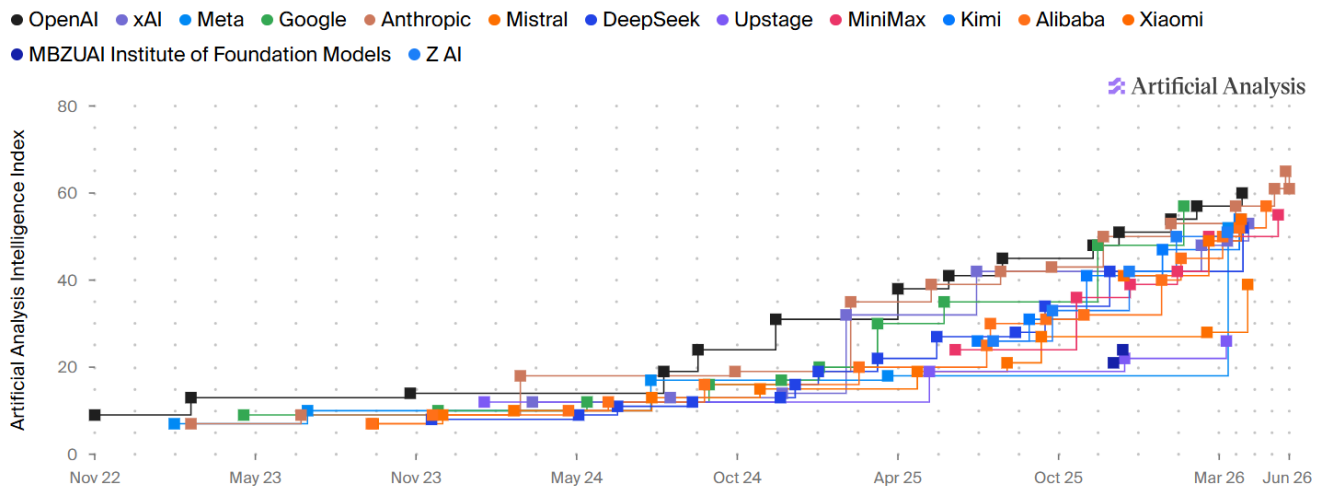
图4：微软 AI 业务收入快速增长



数据来源：微软公告，东吴证券研究所

模型能力快速扩散，闭源旗舰模型依然保持领先，但开源模型和高性价比模型的追赶速度明显加快。DeepSeek V4、Qwen3.7-Max、Kimi K2.6、MiniMax M2.7 等开源模型表明，长上下文、MoE、Agent、工具调用、代码和多模态能力，已经不再只属于少数闭源模型，开源模型正逐步具备解决实际工程问题的能力。

图5：开源模型虽然与头部闭源模型能力依然有差距，但仍在持续追赶



数据来源：Artificial Analysis，东吴证券研究所

Agent 产品化早期、企业逐渐实现 ROI 验证。AI Agent 已经从模型能力演示逐步进入真实产品形态，但仍处于从“可用”走向“稳定可规模化交付”的过渡阶段。当前 Agent 已经能够在编程、文档生成、信息搜索、数据分析、办公自动化等场景中产生明确价值，其核心变化在于大模型不再只是回答问题，而是开始具备任务拆解、工具调用、文件读写、代码执行、多步骤规划和结果校验等能力。因此，Agent 的产业定位正在升级为任务执行入口。

这一阶段的典型特征是：产品形态已经成立，但可靠性边界仍然清晰。一方面，Claude Code、Grok Build、Google Antigravity、Gemini API Managed Agents、Qwen 的 MCP 与多 Agent 能力等产品或能力，说明 Agent 正在进入开发者工具、企业工作流、移动端应用开发和办公自动化体系。尤其在编程场景中，Agent 可以读取代码库、修改文件、运行命令、执行测试、生成提交，已经形成端到端软件工程协作的能力。**另一方面，Agent 尚未达到完全自主、长时间稳定、低错误率、低成本、可无监督运行的成熟状态。**在复杂任务中，Agent 仍可能出现规划偏差、工具调用错误、上下文遗忘、权限误判、长链路错误累积和结果不可验证等问题，因此当前主要在人类监督下承担高频、标准化、可验证的工作流。

表2：当前部分主流模型公司的 Agent 产品

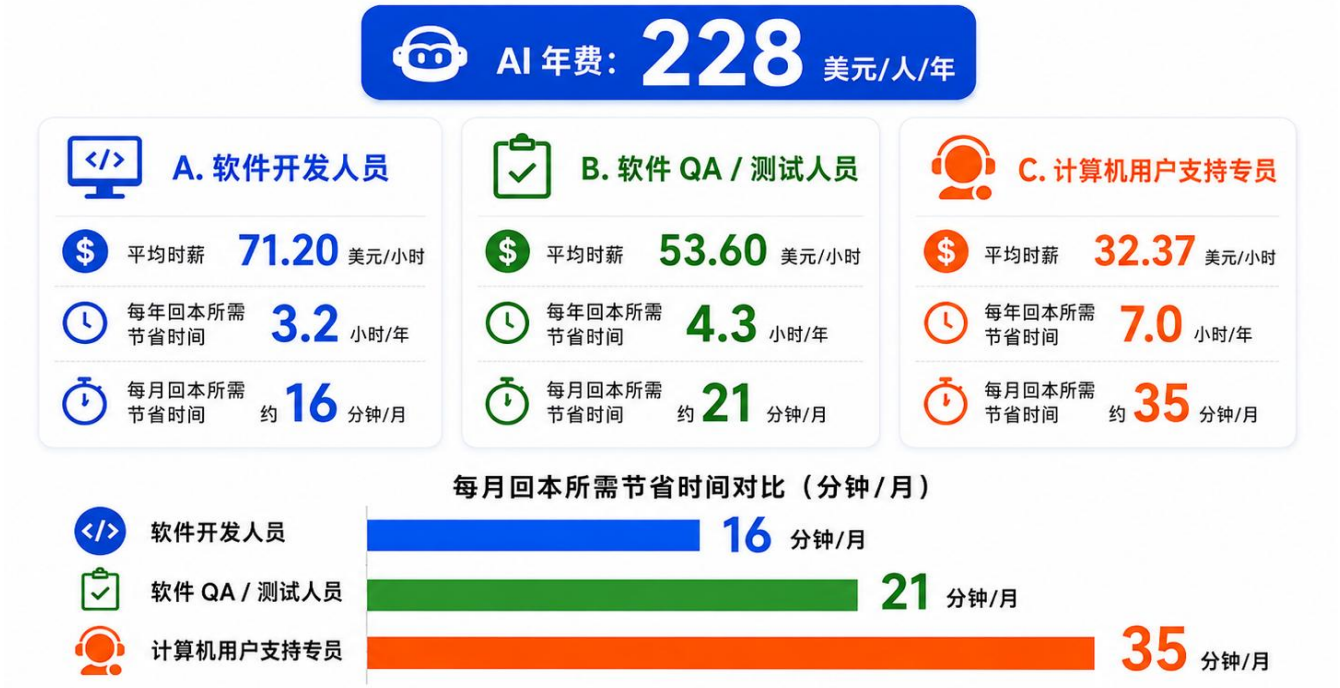
厂商/生态	Agent 产品	主要形态	核心能力	典型场景	产品化阶段判断
Google	Google Antigravity / Gemini API Managed Agents	云端 Agent 平台+开发者工具	支持在安全云沙箱中运行 Agent	应用开发、代码执行、工具调用、企业级 Agent 部署	从模型 API 进入“可部署 Agent 平台”阶段
Anthropic	Claude Code	终端/IDE 编程 Agent	可编辑文件、运行 Shell 命令、发起网络请求，并通过权限模式控制人工确认频率	代码库理解、代码修改、测试运行、软件工程协作	编程 Agent 率先本地产品化，但仍需权限控制
xAI	Grok Build	终端编码 Agent/Agent Client Protocol	可通过交互式 TUI、脚本、机器人或 ACP 协议使用	Web 开发、代码生成、调试、自动化开发流程	面向开发者的早期 Agent 产品
Qwen	Qwen Code / Qwen-Agent / Qwen3-Coder	开源编码 Agent+Agent 框架	支持工具使用、规划、记忆、代码解释器、浏览器助手、自定义助手等能力	开源代码开发、代码库理解、Agent 应用构建	开源生态开始补齐 Agent 框架与工具链
OpenAI	Codex / Codex CLI / Codex Web	云端软件工程 Agent+本地终端编码 Agent+IDE/桌面入口	可在云端沙箱中并行处理任务，支持编写功能、回答代码库问题、修复 Bug、提出 PR；CodexCLI 可在本地终端读取、修改并运行代码	代码库理解、功能开发、Bug 修复、PR 生成、代码审查、终端内软件工程协作	编程 Agent 产品化的核心代表，覆盖云端、终端和开发者工作流，在人类控制的安全边际内执行指定任务

数据来源：东吴证券研究所整理

BLS 在 2026 年 5 月发布的 May 2025 OEWS 新闻稿显示，软件开发人员平均年薪 148,100 美元、平均时薪 71.20 美元；软件 QA 分析师和测试人员平均年薪 111,490 美元、平均时薪 53.60 美元；计算机用户支持专员平均年薪 67,330 美元、平均时薪 32.37 美元，以 GitHub Copilot Business 为例，**若按美国软件开发人员平均时薪测算，每名开发者每月只需节省约 16 分钟即可覆盖订阅成本。**

根据 Deloitte 2026《State of AI in the Enterprise》报告中的调研成果，大型企业与 AI 的领先使用者，编程、客服、办公、营销、数据分析最先落地，因为任务高频、错误成本相对可控、ROI 可量化；医疗、金融、工业、法律和机器人长期空间更大，但需要更高可靠性和责任边界。

图6：以 GitHub Copilot Business 为例，2026 年价格为 228 美元/人/年，对比 BLS 发布的 2025 年 5 月软件行业平均时薪测算的回本周期



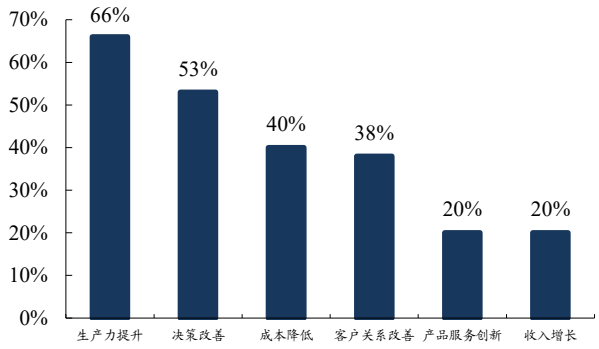
数据来源：GitHub，BLS，东吴证券研究所

图7：GitHub Copilot 与人力成本对比

项目	年成本/人
GitHub Copilot Business	228 美元
GitHub Copilot Enterprise	468 美元
Cursor Teams	480 美元
计算机用户支持专员中位年薪	60,340 美元
软件 QA/测试中位年薪	102,610 美元
软件开发人员中位年薪	133,080 美元

数据来源：Cursor，GitHub，BLS，东吴证券研究所

图8：调研中认为 AI 在该领域带来了改进的比例



数据来源：Deloitte 2026《State of AI in the Enterprise》，东吴证券研究所

2.2. 与历史技术周期的比较

大模型产业同时具备六类周期特征：

第一，它像 1995—2000 年互联网，因为用户增长快、创业公司大量出现、资本市场迅速重估新入口、新应用和新平台，且应用层想象空间极大。

第二，它像 2008—2015 年云计算，因为产业先由基础设施和平台层扩张，再逐步向企业应用迁移；规模经济、开发者生态、云端 API、企业采购和平台锁定非常重要。

第三，它像 2009—2014 年移动互联网，因为 AI 正在改变人机交互方式，并可能形成新的入口层，包括聊天框、搜索框、IDE、办公软件、浏览器、手机、PC、车机、眼镜和机器人。

第四，它像半导体先进制程周期，因为 AI 的供给侧高度依赖 GPU/ASIC、HBM、先进封装、光模块、交换机、液冷、电力和数据中心，具备高资本开支、高折旧、高供给瓶颈和强周期性。

第五，它像 SaaS 渗透周期，因为企业最终购买的不是“模型参数”，而是能否提升员工效率、降低成本、提高收入、改善留存和嵌入工作流。

第六，它像搜索引擎/操作系统平台化周期，因为未来的 AIAgent 可能成为新的任务入口、数据入口和工具调用入口，进而重塑软件、广告、搜索和企业系统。

AI 同时是基础设施周期、软件周期、平台周期、半导体周期、电力周期和应用周期的叠加。这也是为什么当前市场会同时出现 NVIDIA 这类硬件龙头的高速增长、Microsoft/AWS/Google Cloud 这类云厂商的 AI 收入兑现，以及应用层公司估值大幅分化。

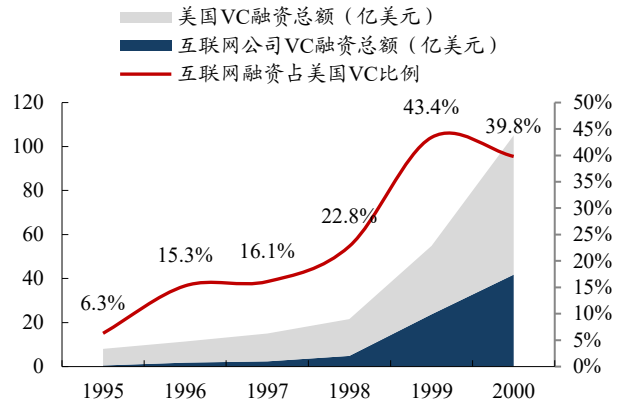
2.2.1. 互联网 1995—2000 年：相似的叙事与不同的成本结构

一、互联网 1995—2000 年的发展历程

1995—2000 年的互联网周期，本质上是新网络基础设施+新入口+新商业模式共振。早期用户通过浏览器接入互联网，门户网站、搜索引擎、电子商务、在线广告、邮件、社区和即时通讯快速出现。资本市场当时给予互联网公司极高估值，核心原因不是它们短期盈利强，而是市场相信互联网会重构信息分发、商业交易和用户入口。

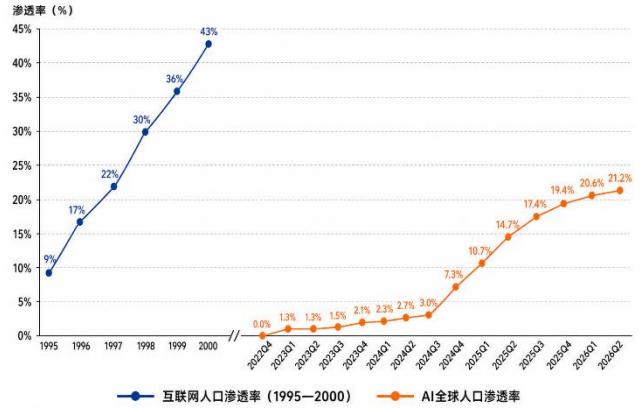
这一阶段的典型特征有四个：**第一，用户增长快；第二，创业公司密集出现；第三，商业模式尚未完全稳定；第四，资本市场从传统财务指标转向用户规模、流量、点击率、GMV 和网络效应。**而最终互联网泡沫破裂，系由于大量互联网公司诞生时行业处于探索期，商业模式与实际业绩无法消化前期高估值。

图9：美国互联网公司融资情况，1995-2000



数据来源：NVCA-PwC-National-Aggregate-MoneyTree-Q2，东吴证券研究所

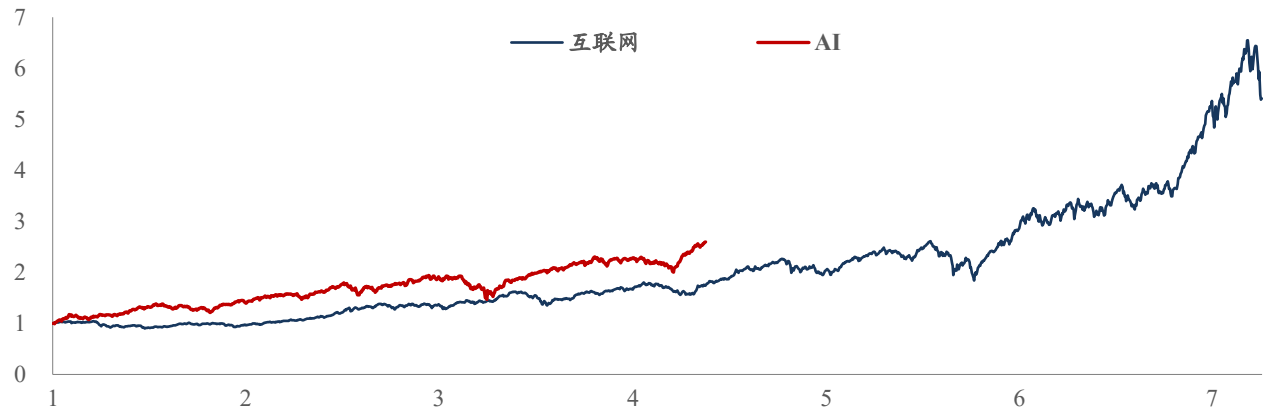
图10：AI 用户渗透率尚处早期



数据来源：国务院、Axios、Reuters、世界银行，东吴证券研究所测算

AI 大模型与这一阶段非常相似：2022 年 ChatGPT 发布后，用户教育极快完成，生成式 AI 应用大量出现，C 端聊天、AI 搜索、AI 写作、AI 编程、AI 视频、AI 设计、AI 教育和 AI 陪伴等方向迅速获得融资和估值重估。OpenAI、Anthropic、Perplexity、Cursor、Runway、Midjourney、Kimi、豆包、MiniMax **等公司和产品都在不同层面复制了互联网早期“新入口+新应用”的扩散速度。**

图11：AI 时代与互联网时代前期的纳指走势高度相似（横坐标 n 为起始日后第 n 年，我们此处定义互联网时代起始于 1994 年 1 月 1 日，AI 时代起始于 2023 年 1 月 1 日；纵坐标为纳指上涨倍数，以起始日期为基准）



数据来源：Stooq，东吴证券研究所

互联网泡沫破裂后，最终留下的是搜索、电商、广告、云和平台公司。同样地，纯概念的 AI 应用、无付费闭环工具和缺乏成本控制能力的产品风险更高；真正具备安全边际的是基础设施、平台能力和能带来明确 ROI 的高频应用。

2.2.2. 云计算 2008—2015：最接近 AI 的历史模板

总的来看，云计算是最接近 AI 的历史类比，因为两者都遵循基础设施先行、平台能力扩张、应用生态释放价值的路径；但 AI 的需求曲线和成本曲线比云计算更不稳定。

云计算的产业演进非常清晰：先是 IaaS 提供弹性算力和存储，再是 PaaS 提供数据库、中间件、开发工具和安全服务，最后 SaaS 在企业工作流中大规模渗透。AI 当前也在走类似路径：先建设 GPU 集群、数据中心和网络；再提供模型 API、推理服务、向量数据库、MLOps 和 Agent 框架；最后进入办公、代码、客服、营销、金融、医疗、工业等应用场景。

图12：云产业链的价值链传导方向与 AI 类似



数据来源：东吴证券研究所绘制

这也是为什么 AI 早期最先兑现业绩的是算力链和云厂商，而不是应用公司。历史上，AWS 在 2013—2015 年收入从 31.08 亿美元增长至 78.80 亿美元，2015 年同比增长 70%；Amazon 同时披露，AWS 增长主要来自客户使用量提升，但也受到持续降价影响。这个过程与今天 AI 云的逻辑高度类似：需求增长快，但平台也要通过规模化、降价和利用率提升来维持竞争力。

AI 与云计算的差异在于，**云计算需求相对稳定，**企业上云后工作负载持续存在；**AI 需求则同时受到模型能力、推理成本、开源模型、价格战、Agent 成功率和监管约束影响。**模型效率提升可能降低单位算力需求，Agent 普及又可能显著放大 token 消耗。因此，AI 云收入增长不必然等于利润改善，资本效率需要持续验证。

2.2.3. 移动互联网 2009—2014：AI 会重构入口，但入口更分散

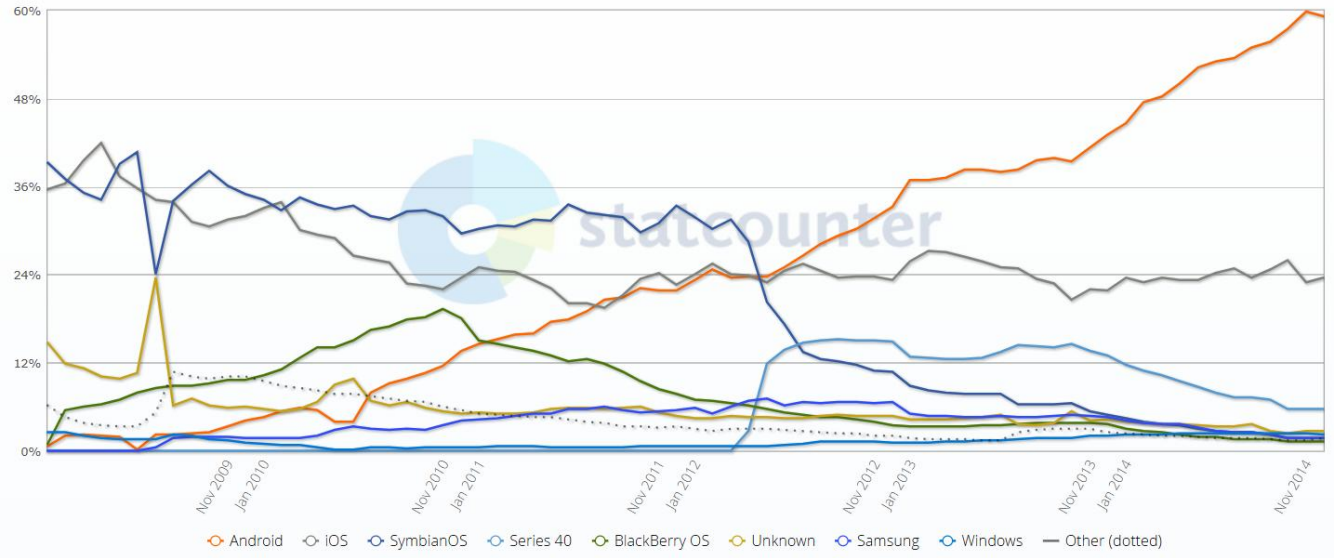
大模型与移动互联网的共同点，是都会带来新入口和新应用生态；但不同的是，移动互联网入口相对统一，而 AI 入口高度分散。

移动互联网的核心变化，是智能手机成为统一入口，AppStore 和 Android 应用生态重构软件分发。入口形成后，应用生态快速繁荣。Apple 在 2008 年披露，AppStore 上线首个周末下载量超过 1000 万次，上线初期已有超过 800 个原生应用。

AI 当前也在形成新入口。用户不再只通过搜索框、菜单和 App 完成任务，而是通过聊天框、语音、多模态输入和 Agent 直接表达需求。AI 入口可能分布在 ChatGPT、搜索引擎、IDE、Office、浏览器、手机、PC、车机、眼镜和机器人中。

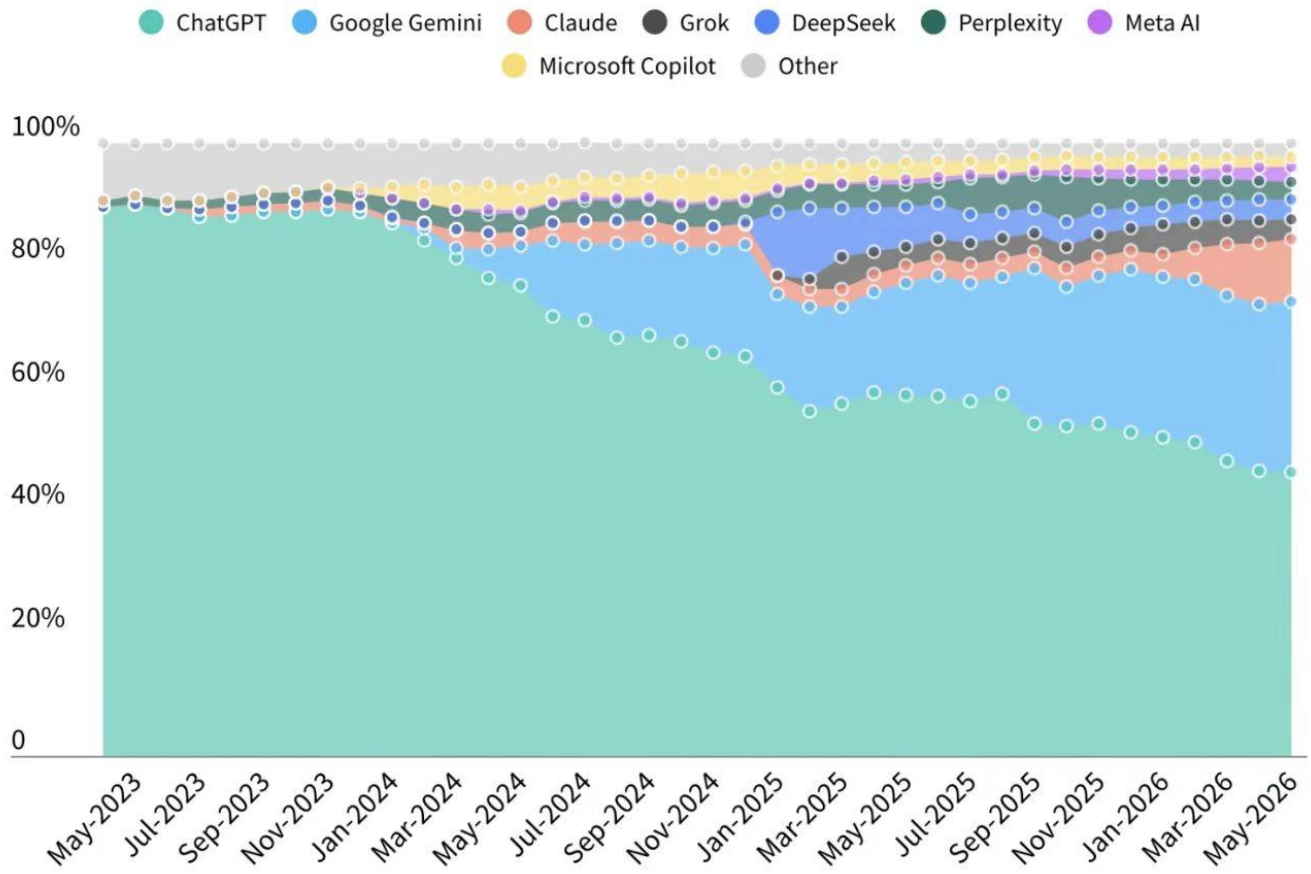
但 AI 和移动互联网的最大差异在于入口集中度。移动互联网时代，智能手机是相对统一的硬件入口，iOS 和 Android 形成强势双寡头。AI 入口则明显更分散，搜索、办公、代码、云、企业软件、终端和机器人都可能成为入口。因此，AI 平台垄断可能弱于移动互联网时代，最终更可能是多入口、多平台、多场景并存。

图13：2009-2014，移动操作系统市场最终形成安卓+iOS 双寡头格局（纵坐标为市占率）



数据来源：Statcounter，东吴证券研究所

图14：截至 2026 年 5 月，ChatGPT 全球 AI 助手市场的份额降至 46.4%，大模型或将进一步进入多平台并存阶段



数据来源：《State of AI 2026 Report》，东吴证券研究所

3. 大模型产业链拆分与价值分配

3.1. 产业链价值分配的核心逻辑

大模型产业链的价值流向三个板块：

第一，关键硬件资源。包括高端 GPU/ASIC、HBM、先进封装、800G/1.6T 光模块、交换芯片、低 PUE 数据中心、可用电力和低延迟网络。

第二，高切换成本能力。生态、云平台、企业数据、开发者工具、数据库、权限系统、安全治理和工作流。

第三，可量化 ROI 的应用入口。包括代码助手、客服、销售、办公、数据分析、搜索、广告、营销、设计和企业自动化流程。

3.2. 产业链利润为什么首先流向基础设施

大模型产业链的价值会优先流向三类环节。第一，稀缺资源，例如高端 GPU/ASIC、HBM、先进封装、高速网络、低 PUE 数据中心和可用电力；**第二，高切换成本能力，**例如 CUDA 生态、云平台、数据库、权限系统、安全治理和企业 workflow；**第三，可量化 ROI 的应用入口，**例如代码助手、客服、办公、搜索、广告和数据分析。

3.3. 产业链价值分配图谱：当前利润池主要在哪里？

表3：AI 全产业链价值分配图谱，当前利润池集中在硬件与底层资源

层级	关键环节	利润池判断	为什么有议价权	跟踪指标
硬件	GPU、ASIC、HBM、先进封装、光模块、交换机、服务器、电源。	当前最确定、最先兑现收入的利润池**仍在硬件与算力链。**	高端 GPU/ASIC、HBM、先进封装和高速网络处于供给瓶颈，客户愿意提前锁定产能。	GPU 交付周期、HBM 合约价、CoWoS 产能、800G/1.6T 光模块订单。
底层资源	电力、土地、水资源、数据中心、液冷。	当前利润池正在**从芯片扩散到电力和数据中心基础设施。**	高密度 AI 数据中心必须获得稳定**低成本电力、并网容量、散热能力和低 PUE**。	上电容量、PUE、电价、液冷渗透率、机柜功率密度。
系统软件	CUDA、编译器、推理框架、训练框架、调度系统、数据库、向量库。	战略价值很高，但商业捕获可能被云厂商或开源生态分散。	推理成本和训练效率高度依赖 kernel、编译器、调度、batching、KV cache 和低精度优化。	推理吞吐、延迟、GPU 利用率、缓存命中率、框架采用率。
云与算力平台	公有云、私有云、Neocloud、主权云。	收入高速增长，但利润率取决于利用率、折旧和电力成本。	企业需要云平台提供模型、算力、数据、权限、审计、安全和部署能力。	AI 云收入、云毛利率、capex/revenue、GPU 利用率、backlog。
基础模型厂	**闭源模型、开源模型、垂直模型。**	**收入增长确定，但通用 API 毛利不确定。**	**前沿能力、品牌、开发者心智、数据和工具生态构成壁垒。**	**API 价格、调用量、企业续费率、benchmark 领先期、每成功任务成本。**
应用层	办公、代码、搜索、广告、客服、设计、视频、游戏、教育、金融、医疗、工业。	长期空间最大，但短期高度分化。	真正的护城河来自用户入口、企业数据、流程嵌入、行业 know-how 和留存率。	客户留存率、ARPU、净收入留存率、付费转化率、任务完成率。
终端与具身智能	手机、PC、眼镜、汽车、机器人、IoT。	当前处于导入和产品验证阶段，重视长期投资价值。	端侧 AI 提供低延迟、隐私保护、本地感知和系统级入口。	AI PC/手机/眼镜出货量、NPU TOPS、本地模型调用、机器人交付量。

利润传导方向

数据来源：东吴证券研究所整理

第一利润池：GPU/ASIC、HBM、先进封装、网络

当前利润池首先集中在基础设施，原因是模型能力提升会先触发算力需求，而算力需求必须通过芯片、HBM、服务器、网络、数据中心和电力转化为供给。应用层长期空间更大，但收入兑现需要产品设计、客户教育、流程集成、权限治理和 ROI 验证，周期更长。

训练和推理都需要高算力、高内存带宽和高速互联。尤其是推理模型和 Agent 普及后，推理工作负载不再是简单单轮对话，而是长上下文 prefill、多轮 decode、工具调用、代码执行、搜索、验证和回滚。NVIDIA FY2027 Q1 数据中心收入同比增长 92%，数据中心 networking 收入同比增长 199%，这说明集群互联已经成为 AI 价值链中与 GPU 同等重要的环节。

投资上，GPU 和 ASIC 是显性受益，HBM 和先进封装是供给瓶颈，光模块和交换机是集群规模扩张后的第二弹性。过去市场常把 AI 硬件等同于 GPU，但 2026 年更准确的说法是 GPU/ASIC+HBM+封装+网络+电力+冷却的系统性短缺。

第二利润池：云与 AI 平台

云厂商是企业采用 AI 的主要入口。企业通常不愿自建完整训练和推理集群，也不愿自行维护模型服务、权限、监控、审计和安全。微软、Amazon、Google、阿里、腾讯、百度等云厂商的优势不只是“有算力”，而是有企业客户、数据库、开发者工具、身份权限、数据治理和软件生态。

但是云同样面临现金流压力。Microsoft 的 AI revenue run rate 增长很快，Alphabet Cloud backlog 和收入加速，但 Amazon 的自由现金流被 AI 相关投资显著压低，Alphabet 也明确披露技术基础设施投资会带来折旧和能源成本压力。换言之，AI 云是收入增长快但资本消耗也快的业务。

第三利润池：应用层

应用层长期空间最大，但短期分化最大。代码、客服、办公、营销、搜索、数据分析最快，因为这些场景有三个共同点：第一，任务高频；第二，错误成本可控；第三，ROI 可量化。医疗、金融、法律、工业和机器人虽然单任务价值更高，但错误成本、合规责任和系统集成复杂度更高，因此落地较慢。

应用层真正的护城河在于数据闭环、流程嵌入、权限系统、业务结果、用户习惯和分发入口。没有这些，只靠包装模型的中间层很容易被模型厂、云厂或 SaaS 厂商下沉替代。

基础大模型：一切 AI 产业链的核心，未来智能底座，利润模型仍在验证中。

当前商业模式仍处在从“产品付费验证”走向“利润模型验证”的中期阶段。不同于早期互联网或云计算的冷启动，大模型已经形成相对清晰的收入入口：C 端以订阅制为主，B 端以 API 调用、企业席位、私有化部署、Agent 工具和云平台绑定为主，代码、办公、

客服、营销、数据分析等场景也已出现较高频使用。**但这并不意味着商业模式已经完全成熟，因为当前用户对大模型的心智仍在变化，企业采购也仍在从“试用工具”向“重构流程”过渡。**换言之，市场已经证明用户愿意为大模型能力付费，但尚未完全证明大模型能够在所有场景中稳定转化为可量化 ROI 和高质量利润。

从需求侧看，大模型已经具备强产品牵引力，但用户心智尚未固化。以 ChatGPT 为代表的通用大模型产品，已经从早期问答工具演化为写作、编程、搜索、数据分析、图像生成、办公助理和 Agent 入口，用户规模和付费用户规模均快速提升。**这说明大模型不是一个单点功能，而更接近下一代人机交互入口。**但另一方面，用户到底将大模型视为搜索引擎、办公助手、代码工具、个人助理，还是未来的操作系统入口，目前仍在动态演化。产品形态、交互方式和杀手级场景尚未完全收敛，**因此当前大模型商业化仍处于“高增长、强探索、未定型”的阶段。**

从企业侧看，大模型的关键矛盾已经从“能不能用”转向“能否嵌入流程并形成可审计的 ROI”。企业已经不满足于简单试用，而是开始关注能否节省人工、提升转化率、缩短研发周期、降低外包成本、减少返工、改善客户体验或减少新增 HC。但现实中，AI 工具只有真正嵌入研发、销售、客服、财务、法务、运营等核心工作流，才能从“个体效率提升”转化为“组织效率提升”。因此，大模型公司的竞争重点正在从单纯模型能力比拼，转向模型能力、工具调用、权限体系、数据治理、工作流集成、安全合规和成本控制的综合竞争。

从供给侧看，大模型本身具备核心投资价值，核心原因在于它是未来智能时代的能力底座。大模型不仅是一款软件产品，更是承载自然语言交互、代码生成、多模态理解、Agent 执行和企业知识调用的基础设施。谁掌握更强的基础模型，谁就更有机会占据未来 AI 应用的能力源头、开发者生态、企业入口和数据反馈闭环。其商业价值不只体现在当前订阅费或 API 收入上，更体现在未来可能延展出的 AI 云、智能体平台、行业模型、AI 应用商店、商业交易入口和企业自动化基础设施。

综合看，当前大模型商业模式仍未完全成熟，但这并不削弱其投资价值，反而说明行业仍处于价值分配格局尚未定型的阶段。短期看，基建层、云、算力等更容易看到收入兑现；**中长期看，基础模型厂商若能持续保持能力领先、降低推理成本、沉淀开发者生态，并把通用模型升级为企业和个人的智能操作入口，则更有机会成为智能时代的核心平台型资产。**

3.4. 开源模型会压缩哪些利润，放大哪些需求？

开源模型压缩的是模型能力的超额利润。过去模型厂可以类比当年的互联网公司依靠垄断地位收取高价；但开源模型接近闭源能力后，客户会采用多模型分配任务：复杂任务用前沿闭源模型，中低复杂度任务用便宜模型，敏感数据用私有化模型，长尾任务

用开源模型微调。可以理解为：**由于企业与个人的趋利性，开源模型可以胜任的任务会全部被开源模型吞噬，因此闭源模型必须持续保持对开源模型的代际领先，否则赢者不一定通吃，但败者一定归零。**

DeepSeek 是典型的成本冲击。DeepSeek 的训练成本和架构效率让市场重新评估前沿模型是否必须依赖极高资本开支；DeepSeek-R1 则让 RL 推理能力扩散到开源社区。Qwen、MiniMax、Kimi 等模型继续推动中文、企业、代码、Agent 和长上下文能力扩散。开源模型的持续进步说明，**模型能力不再只由少数闭源厂商掌握；当基础能力扩散后，竞争焦点会转向谁拥有入口、数据、成本控制、工具生态和实际场景下的工作流。**而开源模型在这些问题上具有天然优势：模型开源→用户提升→更容易建立工具生态→建立实际场景下的工作流→获取真实数据，从而形成生态上的闭环。**这也是开源模型在模型能力落后于顶尖闭源模型的情况下能够从生态上实现换道超车的标准答案。**

4. 竞争格局的本质

4.1. 竞争元素趋于多元，各大模型厂暂时在不同领域站稳脚跟

全球大模型竞争已经从“单模型能力排名”进入“模型能力、算力供给、产品入口、企业客户、监管安全和资本效率”的综合竞争。我们认为，当前厂商分化主要沿三条主线展开：一是 OpenAI、Anthropic、Google、xAI 等前沿闭源模型继续提高长程任务、代码和科学能力；二是 DeepSeek、Qwen、Kimi、MiniMax、智谱等中国模型通过开源、高性价比和本土商业化扩大份额；三是平台型公司把模型能力嵌入搜索、办公、社交、短视频、云和终端，逐步形成入口壁垒。

整体格局来看，可以将大模型分为闭源、开源两大阵营。OpenAI、Anthropic、Google 和 xAI 代表前沿闭源模型路线；Meta、DeepSeek、Qwen、Kimi、MiniMax、ERNIE、Hunyuan、Seed 等代表开源、开放或高性价比路线。与此同时，Microsoft、Amazon、Tencent、Baidu 等云基础设施与平台型公司则试图用模型、算力、数据库、开发工具、权限系统和企业客户关系形成平台壁垒。

4.2. 重点厂商展开

我们总结了全球主要的大模型厂商，汇总如下。

表4：全球主要大模型厂商汇总表

厂商	最新重点模型/产品	路线特征	商业化重点
OpenAI	GPT-5.5、GPT-5.3/5.4-Codex、GPT-5.2 系列	旗舰推理、代码、企业 Agent、Microsoft 生态	ChatGPT、API、Copilot、企业知识工作
Anthropic	Opus 5.0 / Fable 5 / Mythos 5、Opus 4.8	安全对齐、长程 Agent、代码与科学任务	Claude Code、企业高 ARPU、Project Glasswing
Google DeepMind	Gemini 3.5 Flash、Gemini Spark、Omni、Veo/Gemini Live	模型+TPU+搜索+Android+Workspace 全栈	搜索、云、办公、终端和开发者工具
xAI	Grok 4.3、Grok Build	实时数据、X 入口、编码 Agent、Musk 生态	X、Grok 订阅、开发者和特斯拉/SpaceX 生态
Meta	Llama 4 Scout/Maverick、Meta AI、AI 眼镜	开放权重+社交入口+广告优化	广告、社交产品、AI 眼镜、创作者工具
DeepSeek	DeepSeek V4-Pro/V4-Flash	开源、高性价比、1M 上下文、国产算力适配	API、私有化、模型路由、开发者生态
智谱	GLM-5 系列、GLM-5.1、GLM-5.2	GLM 架构、开源、政企本地化、国产化适配	MaaS、私有化部署、政企工作流
阿里 Qwen	Qwen3.6、Qwen3-Max/Thinking、Omni	开源生态+阿里云+Agent 平台	百炼、阿里云、企业部署、MCP/多 Agent
字节	豆包、Seedance 2.0、Seed 系列	产品化、多模态内容、推荐数据	豆包、剪映/CapCut、广告、电商、创作者工具
MiniMax	MiniMax M3、Hailuo 2.3、Talkie/星野	1M 上下文、多模态、C 端产品、开放平台	AI 陪伴、视频、语音、Agent、API
Kimi	Kimi K2.6、Kimi Code、Agent Swarm	长上下文、代码、长程执行、多 Agent	长文档、投研/法律、代码和企业 Agent
腾讯混元	Hy3 preview、混元多模态、腾讯元宝	微信/QQ/文档/游戏/企业微信入口	办公、游戏、广告、腾讯云、企业微信
百度 ERNIE	ERNIE 5.1、ERNIE 5.0 统一多模态	搜索、知识增强、统一多模态与 Apollo	AI 搜索、文库、千帆、智能云、自动驾驶
小米 MiMo	MiMo-V2.5-Pro	推理优先、人车家生态、端云协同	手机、汽车、IoT、端侧 Agent

数据来源：各家公司官网，东吴证券研究所整理

4.2.1. OpenAI：从 GPT scaling 到工作型智能体平台

OpenAI 的技术路线可以分为四个阶段。第一阶段是 GPT-2/GPT-3 代表的预训练规模化阶段，证明 decoder-only Transformer 可以通过参数、数据和计算量扩大获得通用语言能力。第二阶段是 InstructGPT 和 ChatGPT 代表的指令对齐与产品化阶段，解决模型不听指令、不安全、不稳定的问题。第三阶段是 GPT-4/GPT-4o 代表的多模态与实时交互阶段，把文本扩展到图像、语音、实时对话和低延迟多模态交互。第四阶段是 o 系列、GPT-5+代表的推理模型与工作型智能体阶段，重点从“对话体验”转向“复杂任务

执行”。OpenAI 官方对 GPT-5.5 的描述明确强调，它能理解用户目标、承担更多工作、写代码和调试、在线研究、分析数据、创建文档和表格、操作软件并跨工具推进任务。

GPT-5.5 的商业定位也很清晰。GPT-5.5 标准输入价格为 5 美元/百万 token，缓存输入为 0.5 美元/百万 token，输出为 30 美元/百万 token；GPT-5.4 和 GPT-5.4mini 则形成更低价的专业工作与子代理模型层级。这说明 OpenAI 的策略不是所有任务都用一个旗舰模型，而是通过旗舰、大型、mini、缓存、Batch API、工具调用和企业预留容量构建价格分层。

表5：GPT 的几个发展阶段

阶段	代表模型	能力侧重	产业含义
GPT-3 阶段	GPT-3	少样本学习（Few-shot learning）、通用文本生成、提示词任务接口。	模型 API 成为可售卖的基础能力，云端大模型服务雏形形成。
ChatGPT 阶段	InstructGPT、ChatGPT	指令跟随、对话交互、安全拒答和人类偏好对齐。	大模型进入大众产品和企业试点，订阅与 API 商业化启动。
GPT-4/GPT-4o 阶段	GPT-4、GPT-4o	多模态理解、专业考试、图像输入、实时语音和低延迟交互。	搜索、办公、教育、客服和内容生成受到冲击。
GPT-5.5 阶段	GPT-5.5、GPT-5.5 Pro	智能体式编码、计算机使用、长程工具调用和复杂知识工作。	模型从回答者升级为工作执行者，推理 token 和工具调用需求上升。

数据来源：ChatGPT 官网，东吴证券研究所

4.2.2. Anthropic：企业安全、长程任务

Anthropic 的商业价值主要在高 ARPU 企业场景，用可靠性、安全和复杂任务成功率获取溢价，但自有 C 端推广弱于其他主流平台。

Anthropic 的路线一直围绕安全对齐、企业可信、长上下文和复杂任务可靠性展开，也是当前市场上公认落地项目完成度最高的模型。Claude 早期强调安全对话和长文本；Claude 3/3.5 强化多模态、代码和企业可用性；Claude 4 系列进入 coding agent 和 computer use；Claude Opus 4.7 则进一步强化长程智能体工作、知识工作、视觉、记忆和安全防护。Anthropic 的定位在于高价值任务不只需要更聪明，还需要更可靠、更可审计和更少幻觉。

当前，Anthropic 的路线已经从 Claude 4.7/4.8 进一步迈向 Opus 5.0、Mythos 5 和 Fable 5。Fable 5 是 Mythos 级能力的公开安全版本，定位为更强的长程 Agent、代码迁移、科学研究和复杂知识工作模型；Mythos 5 则主要面向受控场景，覆盖网络安全、生物、化学等高敏领域。Fable 5 发布时定价高于既有 Opus，10 美元/百万输入 token、50 美元/百万输出 token，显示其能力与风险均被放在最高档。

表6：Claude 发展历程一览

阶段	代表模型/产品	能力侧重	产业含义
Claude 早期	Claude 1/2	安全对话、长上下文和低伤害输出。	与 OpenAI 形成差异化，吸引对安全敏感的企业客户。
Claude 3/3.5	Claude 3 Opus/Sonnet/Haiku、Claude 3.5 Sonnet	代码、长文档、多模态和企业可用性。	在开发者工具、法律、金融、投研和知识工作中形成较强口碑。
Claude 4	Claude Opus/Sonnet 4.x	长程编码、工具调用、Agent 执行和复杂研究。	从聊天模型转向企业任务模型和 Agent 平台。
Claude Opus 4.7/4.8	Opus 4.7、Opus 4.8、Claude Code、Claude Managed Agents	长程编码、可治理 Agent、computer use、浏览器 Agent、复杂知识工作和更强诚实性。	Anthropic 强化“企业级可治理 Agent”定位，核心卖点是高价值复杂任务的稳定完成率和可信赖输出。
Claude Fable 5 / Mythos 5	Claude Fable 5、Claude Mythos 5、Project Glasswing	Mythos-class 模型，强调更强长程自主执行、软件工程、复杂知识工作、科学假设生成、视觉任务和多日级 Agent 工作流；Fable 5 为带更强安全护栏的广泛发布版，Mythos 5 为更高信任门槛版本。	Anthropic 进入“超前沿受控发布”阶段：Fable/Mythos 不只是 Opus 的线性升级，而是更接近高能力、高风险、高监管敏感度的前沿 Agent 模型。其产业意义在于，前沿模型竞争从企业可用性进一步进入国家安全、出口管制和高风险能力治理阶段。

数据来源：Anthropic 官网，东吴证券研究所

美国商务部近期对 Anthropic 发布的出口管制令，标志着 AI 产业竞争格局已发生深刻的结构性转变。从过去聚焦 NVIDIA 等先进 AI 芯片的硬件层面，全面延伸至模型实体与软件算法层面，这既反映了全球大模型能力竞争已正式进入国家安全博弈的深水区，也对我国自主可控的大模型底座建设提出了紧迫要求。面对这一地缘政治带来的技术封锁挑战，我们需要从国家安全战略、产业自主底座以及生态竞争格局三个维度，深入解析此次管制背后的深层逻辑与应对路径。

1、安全防线跃迁：模型实体已成大国博弈核心焦点

在大模型任务能力为王的背景下，大模型的能力边界直接关联到经济运行效率与关键基础设施安全。美国对特定商用模型实施紧急出口管制，实质上是将先进 AI 模型视为一种战略性国家资源。当模型本身具备了跨工具、跨系统执行任务的能力，其获取权便超越了常规的商业交易，成为了国家安全博弈的核心筹码。这种管制趋势标志着国际科技竞争已从单一的硬件封锁（如对先进 AI 芯片的限制）扩展至模型实体与软件算法的深度封锁，意味着大模型及其配套的算力底座已经正式纳入国家安全战略的宏观框架之中。

2、自主底座重构：加速全栈国产化训练进程

在当前国际地缘政治环境趋于复杂、高端算力供应链存在巨大不确定性的背景下，依赖海外模型底座存在被随时切断的重大业务风险。为确保国家产业的安全底线，实现

自主可控的大模型基座已刻不容缓，其核心挑战在于从“算力适配”向“全栈训练自主”的纵深突破。业界普遍关注模型适配国产算力推理，但真正的战略价值与核心壁垒集中在训练侧——即预训练、后训练、分布式强化学习、多卡通信与系统稳定性工程。以科大讯飞星火大模型为代表的实践证明，只有攻克全国产算力环境下的全栈模型训练难题，才能让大模型摆脱对海外 GPU 集群的路径依赖。如果未来国内主流模型仍完全依赖海外高性能算力进行训练，而仅在推理端迁移至国产芯片，那么中国 AI 底座并未实现本质上的自主可控。因此，高度重视并大力发展国产开源模型，是实现底座自主的必由之路。

4.2.3. Google DeepMind / Gemini：全栈平台型模型厂

Google 的核心定位是全栈 AI 平台型模型厂。通过模型研究、TPU、自有云、搜索、YouTube、Android、Chrome、Workspace、广告系统和开发者平台的组合实现独特的市场竞争力，这一点与国内的阿里颇为类似。Gemini1.0 强调原生多模态，Gemini1.5 强调长上下文，Gemini2.x 强化推理和工具，Gemini3+则把重点放在 Agent、开发者平台、多入口分发和与 Google 全生态的深度融合。

Google 的优势是全栈式产品闭环。第一，它拥有 TPU 和数据中心基础设施，可以在训练、推理和成本控制上形成自有算力优势。**第二，**Google 有搜索、Android、Chrome、Gmail、Docs、Sheets、YouTube 等高频入口，模型能力一旦成熟，可以直接嵌入用户日常工作和生活场景。**第三，**GoogleCloud 能将 Gemini、VertexAI、数据库、数据分析、安全和企业工作流打包销售，形成企业 AI 云平台。**第四，**Google 在多模态、视频理解、地图、搜索和知识检索方面有天然数据优势。

短板在于商业模式的内部矛盾，这一点与 Meta 类似。第一，AI answer 和 Agent 可能压缩传统搜索点击和广告链路，Google 必须在用户体验和广告变现之间重新平衡。第二，Google 内部产品线复杂，模型发布、命名和集成节奏曾经不够统一，影响开发者心智。第三，Gemini 在部分开发者群体中的“模型口碑”需要持续追赶 OpenAI 和 Anthropic，尤其是在高强度 coding agent 和复杂任务自主执行中。

Google 是大模型时代最不容忽视的长期玩家。**它未必需要在每个 benchmark 上第一，但只要 Gemini 深度嵌入搜索、Workspace、Android 和 GCP，将爆发巨大的商业弹性。**

4.2.4. xAI / Grok：模型厂商+大规模 AI 基础设施运营商

xAI 的路线起点是 Grok 与 X 平台实时数据结合。相较 OpenAI/Anthropic 的企业路线，xAI 更强调实时信息、社交语境、Musk 生态和快速迭代。随着 Grok 模型升级以及

Grok Build 等产品出现，xAI 正从实时问答模型转向编码 Agent、开发者工具和更广泛的工作流执行。同时，随着 Colossus 等超大规模 AI 数据中心建设推进，xAI 已开始具备明显的算力平台属性，其价值不仅来自模型能力，也来自训练和推理基础设施本身。

xAI 的优势主要来自数据入口、生态资源和基础设施能力。第一，X 平台提供实时社交数据和热点语境，这是 OpenAI、Anthropic、Google 在公共社交讨论中难以完全复制的入口。**第二，**Musk 生态包括 X、SpaceX、Tesla、xAI 算力集群以及潜在机器人和自动驾驶场景，长期可能形成从线上信息到物理世界的数据闭环。**第三，**xAI 在算力投入上极为激进，Colossus 等超级计算中心已经成为全球最大的 AI 集群之一，使其不仅能够训练 Grok，也具备向外部客户提供 AI 基础设施服务的能力，市场已开始将其视为兼具模型厂商与 AI 算力租赁双重属性的公司。**第四，**Grok 的传播属性强，产品更新容易形成市场关注。

表7：Grok 系列产品一览

阶段	代表模型/产品	能力侧重	产业含义
Grok 早期	Grok 1/2	实时信息、社交语境、与 X 平台结合。	形成与 ChatGPT/Claude 不同的数据入口。
Grok 3/4	Grok 3、Grok 4	推理、多模态、工具和更大训练集群。	xAI 尝试进入前沿模型竞赛。
Grok 4.3	Grok 4.3	更强工具访问、真实工作环境和 Agent 能力。	从回答实时问题转向完成工作任务。
Grok Build	Grok Build CLI	编码智能体、子智能体、MCP、插件和代码工作流。	进入开发者工具市场，与 Claude Code、Codex、Kimi Code、Qwen Code 竞争。

数据来源：xAI 官网，东吴证券研究所

从产业链竞争格局看，xAI 与 OpenAI、Anthropic 的核心差异在于竞争逻辑的双重性：不仅比拼模型能力，同时业务向算力租赁延伸，且能够提供不菲的收入。Anthropic 以约 12.5 亿美元/月的价格租用 Colossus/ColossusII 数据中心算力**至 2029 年 5 月。**xAI 也因此成为算力租赁运营商，一方面将固定资产变现，另一方面缓解 xAI 内部的现金流压力。

4.2.5. DeepSeek：高性价比、开源冲击与百万上下文推理效率

DeepSeek 的迭代可以分为三个阶段。第一阶段是 DeepSeek Coder 和 DeepSeek Math 代表的专项能力突破，通过代码、数学和可验证任务形成差异化。第二阶段是 DeepSeek-V2/V3 的 MoE 工程效率突破，V3 通过 MoE、MLA、低精度、通信优化和训练稳定性显著降低成本。第三阶段是 R1 与 V4 的推理和 Agent 化阶段，R1 证明 RLVR 可以激发推理行为，V4 则进一步把重点放在 1M 上下文、Agent 和低成本长上下文推理。DeepSeek V4 官方文档称其预览版已上线并开源，标志着“高性价比 1M 上下文时代”。

DeepSeek V4 首次由华为昇腾芯片参与训练。DeepSeek V4 Flash 是首个公开说明在后训练当中使用国产算力的通用大模型，通过三大核心设计初步试验去英伟达化的技术

布局。（1）引入 MXFP4 量化感知训练，对 MoE 专家权重与索引器 QK 路径实现 FP4 量化，降低了对 NVIDIA FP8 生态的绑定，可无缝适配华为昇腾、寒武纪等国产芯片；（2）采用 TileLang 领域专用语言开发底层算子，脱离 CUDA 生态强绑定，可跨硬件平台编译，降低向国产芯片的迁移成本；（3）自研 MegaMoE2 融合内核，实现专家并行的细粒度通信计算重叠，已在华为昇腾平台完成适配跑通，解决了国产硬件环境下 MoE 模型的通信瓶颈。

性能表现：整体跻身全球第一梯队，多项核心指标比肩甚至超越国际顶级闭源模型。

（1）知识储备：DeepSeek-V4-Pro-Max 在 SimpleQA-Verified 基准上取得 57.9 分，领先第二名开源模型 20 个百分点，中文 SimpleQA 得分达 84.4，大幅缩小与 Gemini-3.1-Pro 的差距，MMLU-Pro、GPQA Diamond 等教育知识基准均领跑开源赛道。

（2）推理与代码能力：Pro-Max 版本 Codeforces 评分达 3206，位列人类选手排行榜第 23 名，LiveCodeBench Pass@1 达 93.5，IMOAnswerBench 得分 89.8 仅略逊于 GPT-5.4；Flash 本 Codeforces 评分也达到 3052，推理性能追平 GPT-5.2 等闭源模型。

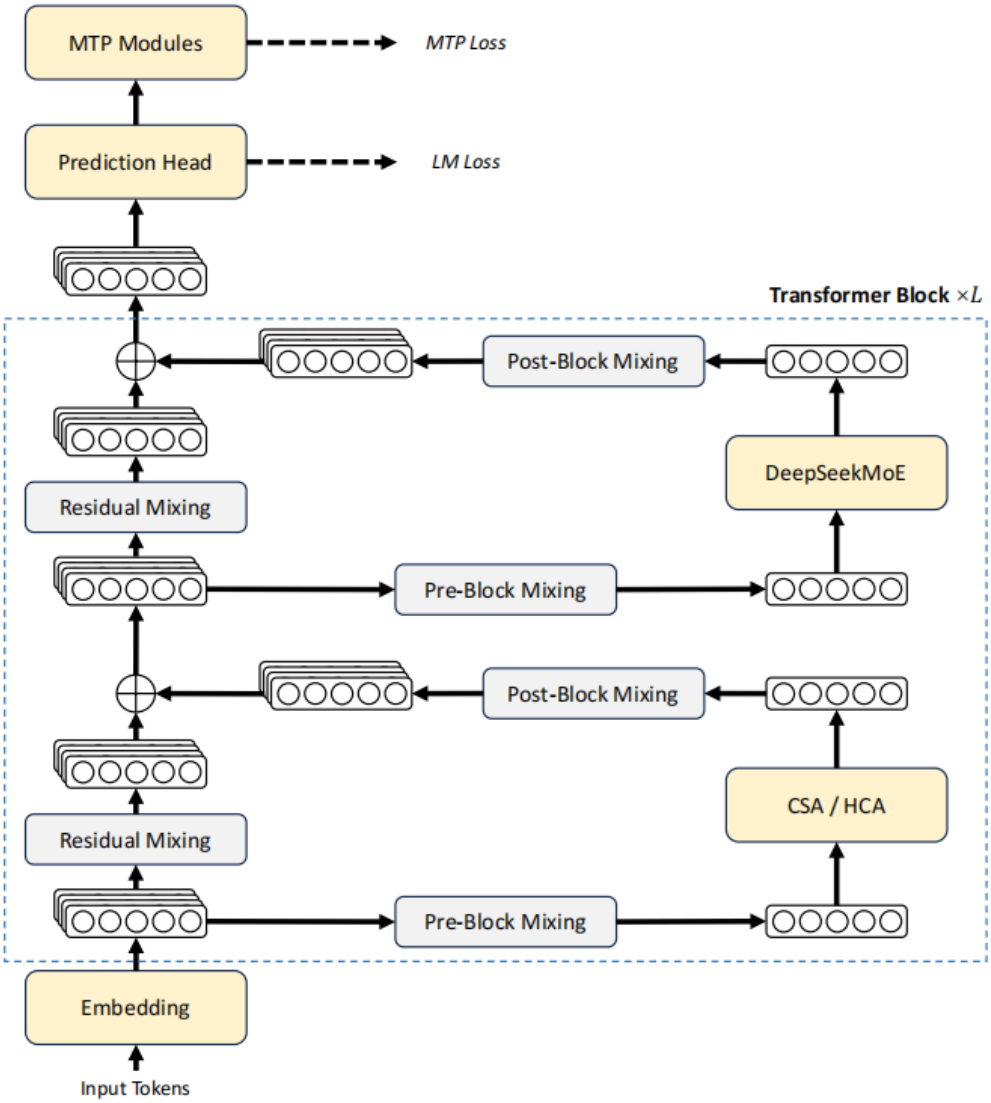
（3）Agent 能力：V4 Pro-Max 任务解决率达 80.6，与 Claude Opus 4.6 基本持平，Terminal Bench 2.0、MCPAtlas Public 等基准均处于开源模型第一梯队。

（4）长上下文能力：1M token 场景下，MRCR、CorpusQA 得分分别为 83.5、62.0，超越 Gemini-3.1-Pro，且 128K 上下文内检索能力保持高度稳定。

（5）中文创作：其功能性写作对 Gemini-3.1-Pro 胜率达 62.7%，创意写作质量胜率高达 77.5%，仅在高难度多轮约束场景略逊于 Claude Opus 4.5。

模型技术架构：CSA+HCA+mHC 进一步压缩推理成本。（1）首创 CSA+HCA 交替的混合注意力架构。通过分层 KV 缓存压缩与稀疏注意力结合，在 1M token 上下文场景下，Pro 版本单 token 推理 FLOPs 仅为 V3.2 的 27%，KV 缓存占用降至 10%，Flash 版本更是分别降至 10%与 7%，从底层解决了超长上下文的算力瓶颈；**（2）引入 mHC 流形约束超连接**升级传统残差结构，提升了深层模型的信号传播稳定性与表达能力，同时采用 Muon 优化器搭配预期性路由、SwiGLU 钳制技术，解决了万亿参数 MoE 模型训练的 Loss Spike 难题；**（3）采用领域专家独立训练+全词表在线蒸馏的后训练范式，**规避了多能力融合的性能退化问题。

图15：DeepSeek-V4 保留 Transformer 架构和 MTP 模块，同时引入 mHC、CSA+HCA



数据来源：DeepSeek V4 技术论文，东吴证券研究所

DeepSeek 的产业影响不只是模型能力，而是价格体系。V4-Pro 大幅降价会压缩通用模型 API 毛利，但对国产开源模型生态的建立、国产算力适配敞口的打开，均有重大意义。

4.2.6. Qwen：从开源生态到 Agent 全栈云平台

Qwen 的核心定位是“中国开源模型生态 + 阿里云企业 Agent 平台”。与 DeepSeek 更强调成本效率和开源冲击不同，Qwen 的战略更偏平台化：先用开源模型建立开发者生态，再与阿里云、百炼平台、数据库、企业客户、MCP、多 Agent 编排和自研 AI 芯片结合，形成企业 AI 云底座。

Qwen 的优势主要在生态和工程落地。第一，Qwen1/1.5 到 Qwen2/2.5 阶段已经在中文、多语言、代码、数学、多模态和开源生态中积累了较强开发者心智。第二，Qwen3 系列通过 thinking、MoE、推理和长上下文完成模型分层，既服务高性能任务，也服务低成本部署。第三，Qwen3.7-Max 明确面向 Agent 时代，强调编码、调试、自动化办公、多文件工程、MCP 集成和多智能体编排。第四，阿里云具备企业客户、算力、数据库、数据治理、安全和行业解决方案能力，能将模型能力转化为云收入和客户粘性。

Qwen 的短板主要在竞争强度和价格压力。第一，国内云厂商价格战激烈，模型 API 很难长期维持高毛利。第二，Qwen 的 C 端入口和用户心智弱于豆包、Kimi，海外市场又面临地缘和合规限制。第三，开源生态虽然利于扩散，但也可能削弱闭源旗舰模型的直接商业溢价。第四，企业 Agent 落地需要数据治理、权限、流程改造和 ROI 验证，销售周期可能较长。

Qwen 是中国最具平台化潜力的大模型底座之一。阿里云有望通过模型服务、Agent 平台、企业数据治理和私有化部署提升客户粘性。它最适合企业云、私有化、代码 Agent、办公自动化、多 Agent 编排和国产化替代场景。

表8：千问系列产品

模型/体系	定位	版本状态	主要能力	适用场景
千问总览	阿里云/通义千问团队研发的大模型体系	自 2023 年首次发布以来，已完成多轮重大版本迭代；2026 年进一步演进至 Qwen3.7 系列	覆盖通用语言、多模态/全模态、代码、智能体、开源/开放权重模型；Qwen3.7 系列进一步强化 Agent、编程、工具调用、长程自主执行与多模态交互能力	文本生成、问答、复杂推理、代码、工具调用、视觉理解、GUI 操作、音视频理解、企业级 Agent 应用
Qwen3.7-Max	千问当前旗舰闭源模型，面向 Agent 时代设计	2026 年 5 月发布；已在 Qwen 与阿里云百炼体系中上线	强化代码编写与调试、办公工作流自动化、长周期自主执行、多步骤工具调用与复杂 Agent 任务；强调可持续执行数百至上千步任务	复杂代码工程、长程 Agent、办公自动化、多工具调用、企业级智能体、复杂任务规划与执行
Qwen3.7-Plus	千问 3.7 系列高性能多模态 Agent 模型	2026 年 6 月发布；阿里云百炼支持 qwen3.7-plus、qwen3.7-plus-2026-05-26	在强文本能力基础上升级视觉-语言能力，同时保留编码、工具使用和生产力工作流能力；可感知真实世界场景、读取屏幕并操作 GUI、基于视觉参考生成代码、端到端导航移动应用	视觉 Agent、GUI 操作、移动应用导航、基于视觉的代码生成、多模态办公、企业多模态智能体
qwen3-max (Thinking)	前代旗舰闭源推理模型	2026 年 1 月发布；能力与 qwen3-max-2026-01-23 相近	支持思考与非思考模式；强化事实知识、复杂推理、指令跟随、人类偏好对齐和工具调用能力；上下文长度最高 26 万 tokens	企业级通用推理、复杂问答、长文本处理、知识分析、生产环境调用
Qwen3.6-Max Preview	Qwen3.7 之前的前沿旗舰预览版	2026 年 4 月上线，处于预览迭代阶段	相较 Qwen3.6-Plus，在世界知识、指令遵循、智能体编程能力上进一步提升	复杂代码任务、长程智能体任务、高阶知识推理、模型能力前沿测试
Qwen3.6-Plus	面向企业和开发者 API 的高性能托管模型	2026 年 4 月发布	默认支持 1M 上下文窗口；增强智能体编程、多模态感知与推理能力	代码开发、复杂任务规划、工具调用、视觉理解、企业级智能体应用

Qwen3.6-27B	开源/开放权重稠密模型	Qwen3.6 系列开源模型之一	270 亿参数稠密模型；在主要智能体编程基准上超过更大规模的 Qwen3.5-397B-A17B；兼顾文本、多模态与推理能力	本地部署、代码任务、智能体编程、边际成本敏感场景
Qwen3.6-35B-A3B	开源/开放权重 MoE 模型	Qwen3.6 系列开源模型之一	采用 MoE 路线，强调推理效率和部署灵活性；以较低激活参数实现较强推理与生成能力	本地化部署、高效推理、开发者和企业私有化场景
Qwen3.5-Omni	多模态/全模态模型	2026 年发布技术报告	支持文本、图像、音频、视频等多模态输入输出；面向实时语音、音视频理解和多模态交互	音频、视频、语音交互、音视频理解、实时多模态应用

数据来源：Qwen 官网、arXiv、Hugging Face，东吴证券研究所

Qwen 的商业价值不只是模型，而是与阿里云、百炼、数据库、企业客户、国产芯片和私有化部署结合。若 Qwen3.7-Max 能成为中国企业 Agent 的默认底座，阿里云可通过模型服务、Agent 平台、企业数据治理和私有化部署提升客户粘性。风险在于云价格战、海外市场地缘限制、API 毛利率和企业应用 ROI。

4.2.7. 字节：产品化、推荐数据和多模态生产部署

字节的核心定位是“C 端产品化+多模态内容生产+高并发推理部署”。与阿里、腾讯更偏企业云和场景化不同，字节的优势在产品迭代、流量入口、推荐系统、多模态内容数据和大规模用户服务。豆包提供 C 端入口，Seed 系列提供模型底座，抖音/TikTok、剪映、电商、广告和搜索提供落地场景。

字节的优势是产品闭环。第一，豆包具备强 C 端流量入口和使用频次，能够快速验证模型能力、收集反馈并推动产品迭代。第二，字节在推荐系统、内容理解、广告投放、短视频和创作者工具方面积累深厚，模型能力可以直接嵌入已有业务。第三，Seed2.0 提供 Pro、Lite、Mini 等不同规模通用 Agent 模型，说明字节不是只追旗舰能力，而是重视生产部署、成本分层和高并发服务。第四，Seedance、Seedream、Seed3D 等内容模型线可以直接服务剪映、抖音广告、电商素材和创作者经济。

字节的短板在企业云和海外监管。第一，火山引擎在企业云市场的心智和客户基础弱于阿里云、腾讯云、华为云，模型能力外部化仍需时间。第二，TikTok 面临海外地缘政治和数据合规压力，可能影响全球 AI 产品扩张。第三，内容生成模型虽然商业场景清晰，但也面临版权、深度伪造、内容安全和平台治理问题。第四，字节的大模型能力需要证明不仅能服务内部业务，也能在外部开发者和企业客户中形成独立平台价值。

字节是中国大模型产品化能力最强的公司之一。**它最适合 C 端助手、内容生成、广告创意、短视频生产、电商素材、搜索问答、客服和高并发低成本推理场景。**

表9：豆包大模型发展梳理

阶段	代表模型	能力侧重	产业含义
豆包-1.5	Doubao-1.5-pro	MoE、训练—推理一体化、高并发推理优化。	服务豆包、火山引擎和字节内部内容/广告场景。
Seed1.x	Seed1.6、Seed1.8、Seed1.5-VL	自适应思考、视觉语言、多模态、Agent。	从文本聊天转向多模态生产工具。
Seed2.0	Seed2.0 Pro/Lite/Mini	长程任务、复杂推理、GUI agent、全模态理解、生产部署分层。	字节模型更强调产品化落地和规模化服务。
内容模型	Seedream、Seedance、Seed3D	图像、视频、3D、音乐和内容生成。	直接服务抖音/TikTok、剪映、广告和电商内容生产。

数据来源：字节跳动官网，东吴证券研究所

字节的最大优势是流量和内容闭环。模型能力可以直接嵌入抖音/TikTok、剪映、电商、广告、搜索、客服和创作者工具。

4.2.8. MiniMax：长上下文、Hybrid MoE 与低成本 test-time compute

MiniMax 的核心定位是“长上下文 + 高性价比推理 + Agent 工作流”。它不是拥有最强入口的大厂，也不是单纯开源价格战选手，而是试图通过架构优化和低激活参数提高复杂任务执行效率。MiniMax-M1 强调 hybrid MoE、Lightning Attention、1M 上下文和测试时计算效率；M2/M2.5/M2.7 则进一步强化编码、工具调用、搜索、办公生产力和模型自我迭代。

MiniMax 的优势在于长上下文和效率创新。第一，长上下文是 MiniMax 的早期差异化标签，适合长文档、法律合同、投研材料、企业知识库和多文件工程。第二，Hybrid MoE 和低激活参数设计有助于在较低推理成本下获得较强能力。第三，M2 系列强调 agentic coding、deep search、office-task 和 reasoning，说明 MiniMax 正从模型参数竞争转向真实生产力任务。第四，若 M2.7 的模型自我迭代能力兑现，有望降低后训练和 Agent scaffold 开发成本。

MiniMax 的短板主要是入口和生态。第一，它缺少 OpenAI 的全球 C 端入口、阿里的云平台、字节的内容流量和腾讯的社交入口，获客成本更高。第二，资本强度和算力资源弱于全球巨头，持续追赶前沿模型需要大量投入。第三，开发者生态、企业销售体系和行业解决方案仍需扩展。第四，长上下文能力容易被 DeepSeek、Qwen、Gemini 等对手快速追赶，差异化需要持续迭代。

结论上，MiniMax 是技术差异化较强的第二梯队厂商，适合长文档、办公生产力、代码 Agent、搜索增强和多 Agent 工作流场景。它的关键不在于成为最大入口，而在于能否把长上下文和低成本 test-time compute 转化为高任务成功率和可持续商业化。投资

跟踪指标包括 M2/M2.7 API 调用量、企业客户数、长上下文任务成本、Agent 任务成功率、多模态产品收入、海外开发者采用情况和融资/算力投入。

表10：Minimax 迭代历程

阶段	代表模型	能力侧重	产业含义
Text-01 阶段	MiniMax-Text-01	超长上下文、文本理解和生成	在长文档和企业知识场景形成差异化
M1 阶段	MiniMax-M1	Hybrid MoE、Lightning Attention、1M 上下文、test-time compute	证明长上下文推理可以通过架构优化降本
M2/M2.5	MiniMax-M2、M2.5	编码、工具调用、搜索和办公生产力	从模型能力转向应用工作流
M2.7	MiniMax-M2.7	模型自我迭代、Agent Teams、动态工具搜索、办公与代码任务	尝试用模型自迭代降低开发和后训练成本

数据来源：Minimax 官网，东吴证券研究所

MiniMax 的机会在长上下文、低成本推理、多模态和 Agent 工作流。风险是入口、资本强度和生态规模弱于 OpenAI、Anthropic、Google、阿里、腾讯和字节。

4.2.9. Kimi：从长上下文聊天到长程编码与 Agent Swarm

Kimi 的核心定位是“中文长上下文工作流+长程编码+多 Agent 协作”。它最早形成的用户心智是长文档阅读，尤其在中文 PDF、网页、投研、法律、合同、会议纪要和材料分析中体验较好。**KimiK2 系列之后，Moonshot 将能力从“读长文档”扩展到“写代码、调用工具、长程执行、多 Agent 协作”。**

Kimi 的优势在于中文工作流和长上下文心智。第一，Kimi 在中国用户中形成了较强的“长文档助手”认知，这一点对投研、法律、咨询、学生、内容创作和办公人群非常重要。第二，KimiK2/K2.5/K2.6 逐步强化代码、搜索、工具调用和长程任务，正在从阅读助手向工作执行工具升级。第三，Agent Swarm 的定位体现出 Moonshot 对复杂任务拆解、多智能体并行和长步骤执行的投入，这与未来 Agent 竞争方向一致。第四，Kimi Code 能使其进入开发者工具市场，与 Claude Code、Codex、Qwen Code 等竞争。

Kimi 的短板是平台和商业化。第一，Kimi 的云平台、企业客户和行业解决方案弱于阿里、腾讯、百度等大厂。第二，C 端用户心智虽强，但订阅和付费转化需要持续验证。第三，长上下文不是永久壁垒，DeepSeek V4、Gemini、Qwen、MiniMax 都在强化 1M 上下文能力。第四，多 Agent 长程执行对稳定性要求极高，一旦出现上下文漂移、工具失败或错误累积，用户体验会明显下降。

结论上，Kimi 是中文长上下文和工作流 Agent 的代表性玩家。它最适合投研、法律、长文档阅读、网页分析、代码生成、前端开发、DevOps 和多资料综合研究场景；

表11：Kimi 系列产品

阶段	代表模型/产品	能力侧重	产业含义
Kimi 初期	Kimi Chat	长上下文、中文文档阅读、网页和 PDF 分析。	形成“长文档助手”的用户心智。
K2/K2.5	Kimi K2/K2.5	代码、搜索、工具调用、长程任务。	从阅读工具转向工作执行工具。
K2.6	Kimi K2.6、Kimi Code、Agent Swarm	长程编码、多智能体协作、前端/DevOps/性能优化。	进入 coding agent 和多 Agent 工作流市场。

数据来源：Kimi 官网，东吴证券研究所

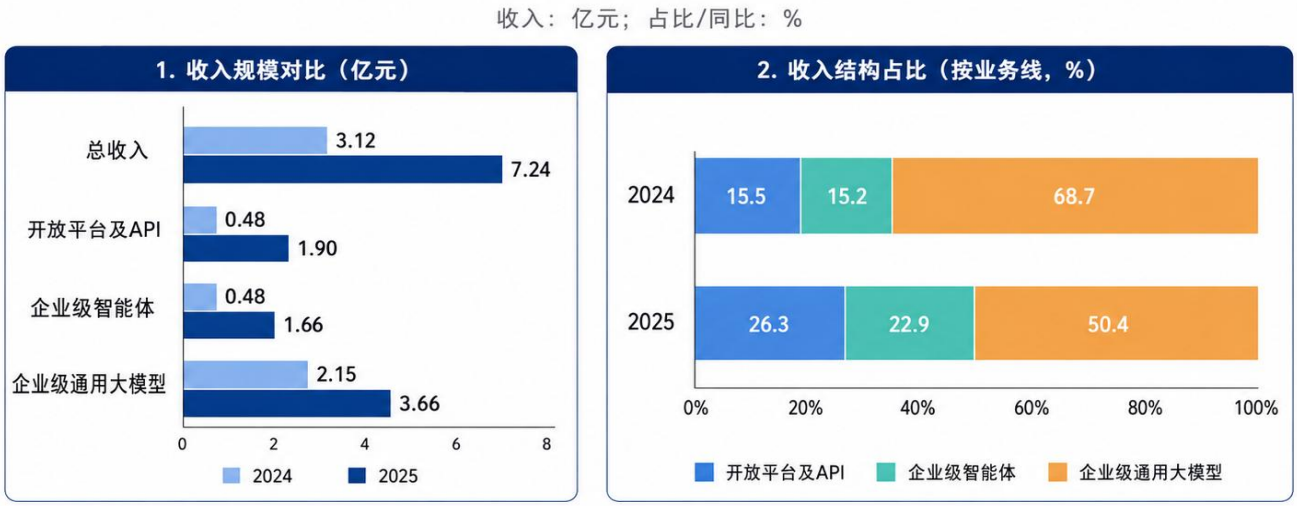
5. 关键标的：智谱、Minimax、科大讯飞

5.1. 智谱：从私有化大模型公司走向 Coding + Agent MaaS 平台

5.1.1. 智谱的重估逻辑正在从项目型 AI 公司切向模型能力驱动的 MaaS 平台

智谱当前最重要的变化，不只是收入同比高增长，而是收入结构、商业模式和技术路线正在同步切换。2025 年，公司收入达到 7.24 亿元，同比增长 131.9%；其中开放平台及 API 收入 1.90 亿元，占比从 2024 年的 15.5%提升至 26.3%；企业级智能体收入 1.66 亿元，占比提升至 22.9%；企业级通用大模型收入 3.66 亿元，占比则从 68.7%下降至 50.4%。本地化/私有化部署仍是公司业务基本盘，但边际增长已经明显转向云端 API、Coding、Agent 和 MaaS。过去市场对智谱的主要担忧，是其本地化部署占比较高，商业模式更接近项目制软件公司，收入确认和交付资源绑定较重，边际利润弹性有限。**而 2025 年 API 与 Agent 的合计收入占比已接近 50%，公司商业模型正在改善。**

图16：智谱收入边际增长转向云端 API、Coding、Agent 和 MaaS



数据来源：智谱年报、招股说明书，东吴证券研究所

5.1.2. 从 GLM-5 到 GLM-5.2，押注 Agentic Engineering+长程任务

1. GLM-5：从 Vibe Coding 走向 Agentic Engineering

智谱对 GLM-5 的核心定位，是从 Vibe Coding 走向 Agentic Engineering。官方 GLM-5 技术博客称，GLM-5 面向复杂系统工程和长程 Agent 任务；相比 GLM-4.5，模型参数规模从 355B 提升至 744B，激活参数从 32B 提升至 40B，预训练数据从 23T Token 增加至 28.5T Token。

这一路线和交流会纪要高度一致。管理层认为，早期 AI Coding 仍以人为主导，AI 主要负责补全、解释和复制粘贴；Vibe Coding 则允许用户用自然语言描述需求，让模型快速生成前端或应用原型。但 Vibe Coding 的问题也很明确：黑盒生成、技术债积累、代码审查困难、上下文漂移，以及安全与隐私隐患。

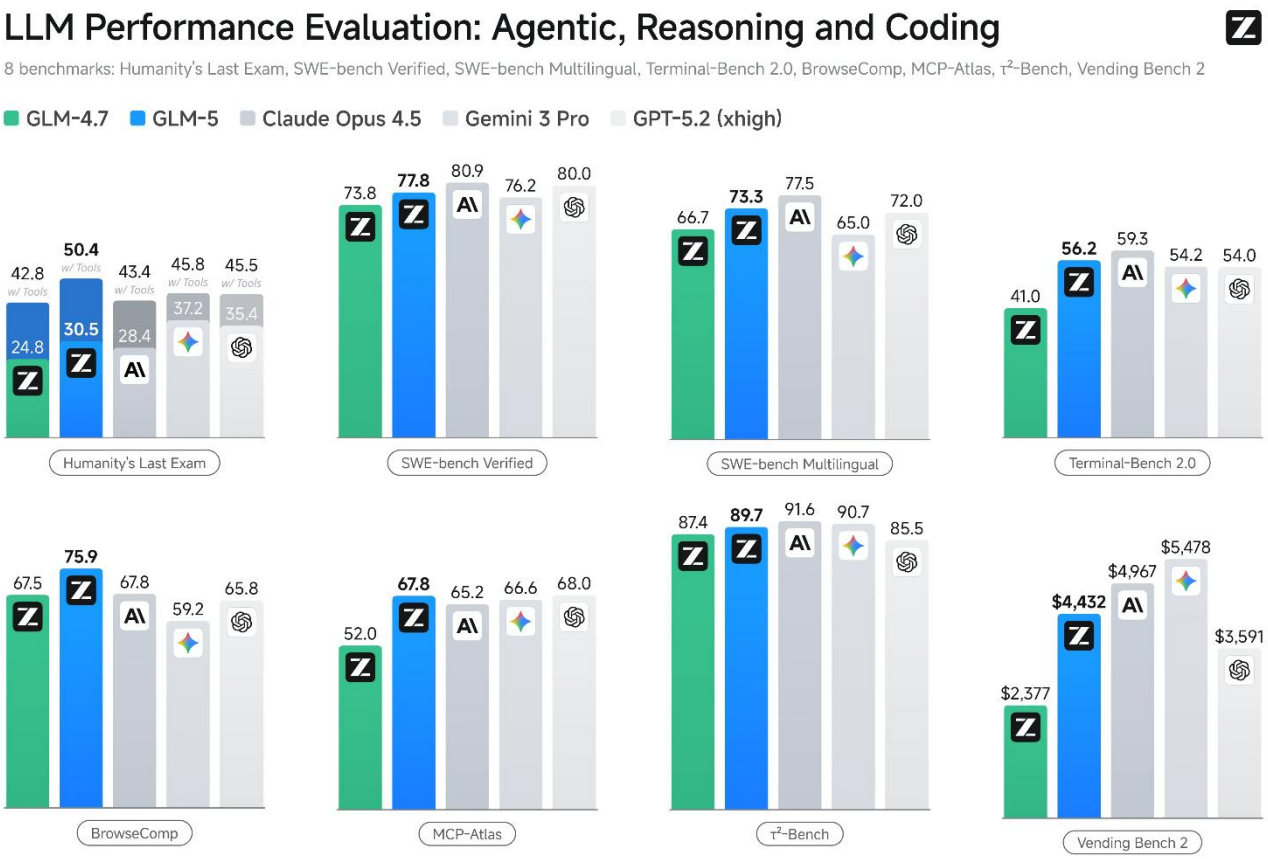
智谱所谓 Agentic Engineering，本质上是把“代码生成器”升级为“可规划、可执行、可审查的软件工程智能体”。其核心差异在于：模型负责理解意图、拆解任务和规划路径，Agent 在沙盒环境中执行，过程中允许人类介入、审查和修改。这有助于缓解 Vibe Coding 时代“一次生成很爽，后期维护困难”的问题。

2. GLM-5 的技术抓手：稀疏注意力、异步强化学习、芯模协同

GLM-5 的技术总结主要包括：**第一，采用改进 GQA 与稀疏注意力机制，**在 128K 长上下文中降低计算量；**第二，采用 MoE、MLA** 等机制提升解码效率；**第三，引入 SLIM 异步强化学习框架，**解决长时间任务中的 GPU 空转问题；**第四，加强国产芯片适配，**并进入模型公司与芯片公司的软硬协同设计阶段；**第五，建设面向多智能体系统的调度基础设施。**GLM-5 论文摘要显示，模型采用 DSA 以降低训练和推理成本，同时保持长上下文能力，并通过异步强化学习基础设施提升后训练效率，使模型能够从复杂长程交互中学习。

简单来说，智谱在尝试解决两个行业共性问题：一是长上下文和长程任务带来的推理成本爆炸；二是中国模型公司面临的算力受限和国产芯片适配问题。 如果智谱能够通过架构优化和芯模协同降低单位任务成本，其 API 毛利率和价格竞争力才有持续改善基础。

图17：GLM5 系列表现与同期大模型相比能力亮眼

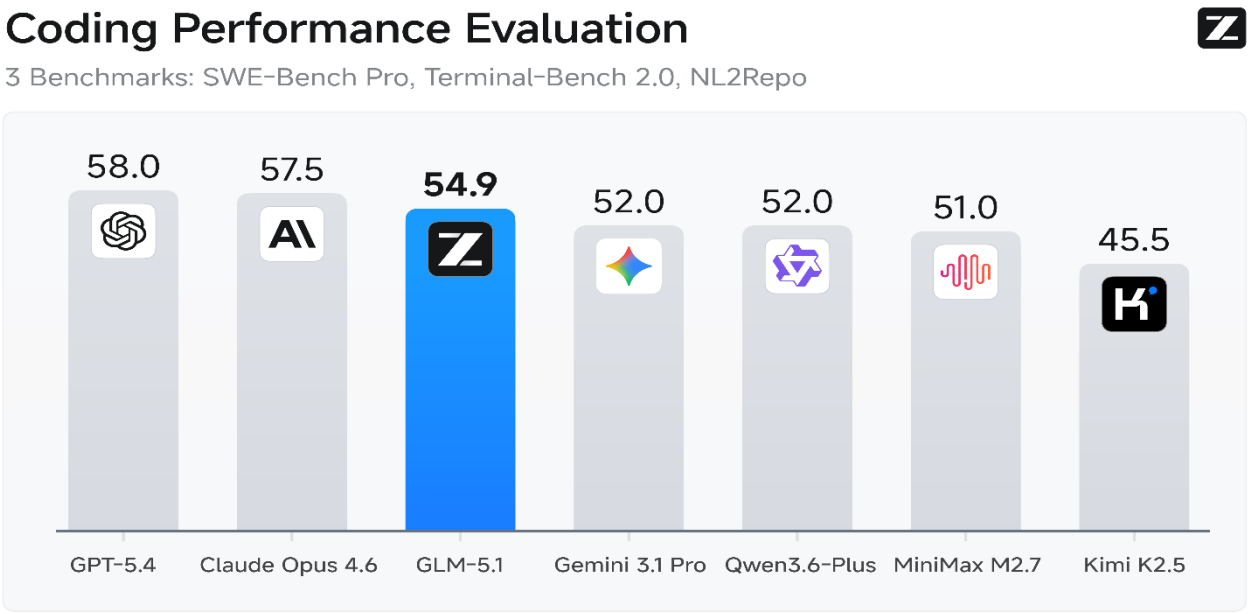


数据来源：智谱官网，东吴证券研究所

3. GLM-5.1 与 GLM-5.2：长程任务成为新主线

智谱官方对 GLM-5.1 的定位是面向 Agentic Engineering 的下一代旗舰模型，并强调其 Coding 能力较前代显著提升，在 SWE-Bench Pro、NL2Repo、Terminal-Bench 2.0 等任务上具备更强表现。

图18：GLM5.1 Coding 能力较前代显著提升，在 SWE-Bench Pro、NL2Repo、Terminal-Bench 2.0 等任务上具备更强表现



数据来源：智谱官网，东吴证券研究所

最新变化是 GLM-5.2。2026 年 6 月 13 日公开报道显示，智谱宣布 GLM-5.2 面向 GLM Coding Plan 全量用户开放，覆盖 Lite/Pro/Max/团队版；GLM-5.2 API 将于下周上线，模型也计划于下周正式开源，并遵循 MIT 协议。报道同时称，GLM-5.2 是智谱迄今能力最强的开源模型，支持真正可用的 1M 上下文，并在长程任务中继续保持领先。

GLM-5 是 Agentic Engineering 的起点，GLM-5.1 强化长程任务，GLM-5.2 则把 1M 上下文、长程任务和 Coding Plan 用户全量开放结合起来。对商业化而言，GLM-5.2 的关键不在于“又发布了一个模型”，而在于它能否提升 Coding Plan 留存、提升 API 调用深度，并把长上下文能力转化为更高客单价的任务型 Token 消耗。

5.1.3. 从本地部署到 MaaS，智谱正在寻找中国版 Anthropic 路径

MaaS 平台：ARR 印证未来确定性，市场“用脚投票”对公司 API 形成付费需求

截至 2026 年 3 月 31 日，开放平台及 API ARR 超过 2.5 亿美元，约合人民币 17 亿元；MaaS 平台日收入超过 460 万元；付费 Token 较 2025 年末提升约 4 倍；Coding Plan 推出约 6 个月，付费 Token 增长超过 15 倍。这组数据如果能够持续兑现，将推动智谱估值逻辑发生重大变化。传统软件项目型公司看订单、收入确认和毛利率；MaaS 平台型公司更看 ARR、净收入留存、API 调用量、单位 Token 毛利和开发者生态。智谱目前

还不能完全按海外头部 AI SaaS 或模型平台估值，**但 ARR 的爆发式增长说明，市场对其模型能力和 API 产品已经形成真实付费需求。**

Coding 是当前最清晰的高价值入口

智谱把 Coding 视为最优先的场景。原因在于：第一，需求高频且支付意愿强，开发者和企业软件团队能直接感知效率提升；第二，效果相对可评估，可以通过代码通过率、修复成功率、任务完成率等指标衡量；第三，Token 消耗大，天然利好 API 和订阅收入。公司在 2025 年围绕 Coding Plan 实现快速放量，并强调其在字节 Trae、腾讯 CodeBuddy 等 Coding 平台中作为第三方模型接入。同时，当前前十大互联网大厂中已有九家深度集成 GLM，其中不乏包括在 Coding 产品货架上接入 GLM。

Coding 或成智谱从模型能力转向收入规模的最短路径，探索中国式 Anthropic。如果 GLM-5.2 在 Coding Plan 用户中形成更强留存，并通过开源扩大开发者心智，智谱有机会复制 Anthropic 以 Claude Code/Coding 能力切入企业市场的路径。

5.1.4. 企业客户和国产芯片协同

智谱长期深耕企业级市场，虽然早期导致本地化部署占比较高，但也积累了大量 KA 客户、行业与交付经验。公司与互联网大厂和国产芯片公司形成两条重要商业化线索：一是前十大互联网大厂中九家深度集成 GLM；二是与寒武纪、沐曦、摩尔线程等国产芯片厂商进行深度适配，甚至进入软硬协同设计阶段。

这对中国模型公司非常关键。海外模型公司可以依赖 Nvidia 生态和全球云计算基础设施，而中国模型公司必须面对算力可得性、国产芯片适配、推理成本和交付稳定性问题。**智谱如果能在模型架构层面为国产芯片优化，将率先形成差异化成本优势。**

5.2. Minimax：从 AI 原生应用走向全模态 AI 平台

MiniMax 的重估逻辑不只是大模型，而是全模态模型+全球化应用+Token 平台化.

公司从成立早期就同时押注语言、视频、语音、音乐、角色互动、AI Agent 和开放平台。2025 年，公司收入 7903.8 万美元，同比增长 158.9%；其中 AI 原生产品收入 5307.5 万美元，占比 67.2%；开放平台及其他 AI 企业服务收入 2596.3 万美元，占比 32.8%；中国内地以外收入 5766.3 万美元，占比 73.0%。

MiniMax 当前收入仍以 C 端/开发者端 AI 原生产品为主，但开放平台收入增速更快，公司正在从“应用变现”扩展到“模型能力变现”。AI 原生产品收入主要来自用户参与度提升、付费意愿增强，以及 Hailuo AI 等产品商业化；开放平台及企业服务收入增长，则主要来自付费客户数量增加。

地域结构来看，2025 年，公司中国内地收入 2137.5 万美元，占比 27.0%；中国内地以外收入 5766.3 万美元，占比 73.0%。MiniMax 的收入具备明显全球化结构。

图19：Minimax 业绩基本情况



数据来源：Minimax 年报、招股说明书，东吴证券研究所

5.2.1. MiniMax 的独特性在于全模态+全球 C 端+开放平台”

1、AI 原生产品：Hailuo、Talkie、MiniMax Agent 构成全球化入口

MiniMax 与其他友商相比，更早把 AI 原生产品推向全球 C 端市场。Hailuo AI 对应视频生成，Talkie 对应角色互动/陪伴，Audio 和 Speech 对应语音合成，Music 对应音乐生成。这些产品的共同特点是：用户规模更大、国际化更快、付费场景更偏消费级和创作级，但也更依赖产品留存、内容生态和合规审查。

2、开放平台：从 API 到 Token Plan，再到 MiniMax Code

开放平台是 MiniMax 的第二增长曲线。2025 年，公司开放平台及其他 AI 企业服务收入 2596.3 万美元，同比增长 197.8%，增速高于 AI 原生产品。

模型平台和订阅产品在 2026 年初出现显著加速。2026 年电话会报道显示，MiniMax 2 月 ARR 已超过 1.5 亿美元，面向企业客户和个人开发者的开放平台新注册用户数达到 2025 年 12 月的 4 倍以上；M2 系列文本模型 2026 年 2 月日均 Token 消耗量较 2025 年 12 月增长 6 倍以上，其中 Coding Plan Token 消耗增长超过 10 倍。

3、Token Plan：把 Coding Plan 升级为全模态共享用量包

MiniMax 的 Token Plan 是理解其平台化战略的重要产品。官方文档称，Token Plan 是对原 Coding Plan 的扩展，使用量不再只覆盖语言模型，而是向文本、图像、语音、音乐等 MiniMax 资源共享额度扩展。

MiniMax 不想只卖某一个模型调用，而是希望把用户锁定在一个全模态额度池中。对开发者而言，一个订阅同时覆盖代码、文本、图像、语音、音乐和 Agent 使用，可以降低多模型、多供应商组合的复杂度；对公司而言，Token Plan 能够提升用户留存，增加多场景交叉使用，并把不同模态需求集中到统一付费账户中。

在 M3 发布时，MiniMax 官方披露，Token Plan 分为 Plus、Max、Ultra 三档，分别为 20 美元/月、50 美元/月、120 美元/月，并对应约 17 亿、51 亿、98 亿个 M3 使用 Token 配额；文本、图像、语音和音乐共享同一使用池。

5.2.2. 模型路线：从 M2 到 M3 的 1M 上下文+原生多模态+Agentic Coding

1、M2 系列：小激活 MoE 架构——成本与能力平衡的核心突破

MiniMaxM2 系列技术路线清晰，核心战略是通过稀疏激活 MoE 实现模型能力、推理效率与单位成本的最优平衡。旗舰 M2 模型总参数规模达 2299 亿，但每 Token 仅激活 98 亿参数，显著降低推理算力消耗。模型架构原生面向 Agentic 部署设计，涵盖三大核心模块：Agent 驱动数据管线、Forge Agent 原生强化学习系统，以及训练-推理-Agent 三层解耦架构。

产业意义：MiniMax 并未盲目追求总参数规模扩张，而是通过低激活 MoE 路线直击 AIAgent 商业化核心痛点。据产业链调研，M2.5 已实现复杂 Agent 的经济可扩展性：以每秒 100Token 输出速率连续运行一小时成本约 1 美元，据此测算 1 万美元预算可支撑 Agent 持续运行一年。

AI Agent 从演示走向真实生产力的核心前提是"聪明+便宜+快速+稳定"四要素兼备。MiniMaxM2 系列的技术叙事本质上是围绕单位任务成本展开，低激活 MoE 架构有望成为 Agent 规模化落地的关键技术路径。

2、M2.7：从 Agent 能力构建到"模型自我迭代"范式探索

2026 年 4 月 MiniMax 开源 M2.7，标志着技术路线从"构建 Agent 能力"向"模型参与自我迭代"延伸。官方披露，M2.7 可自主构建复杂 Agent Harness，完成高复杂度生产力任务；在研发流程中，模型可参与构建强化学习 Harness 中的复杂 Skills、更新 Memory，并优化强化学习过程与 Harness 本身。

能力指标验证：M2.7 在真实软件工程、端到端项目交付、日志分析、Bug 排查、代码安全与机器学习等场景表现突出：SWE-Pro 得分 56.22%，VIBE-Pro 得分 55.6%，TerminalBench2 得分 57.0%；在 40 个超过 2000Token 的复杂 Skills 场景中，Skills 遵循率保持 97%。

战略判断：M2.7 的核心价值不在于榜单刷分，而在于"模型参与构建自身训练闭环"的范式创新。该方向若持续验证，将推动模型迭代从"完全依赖人类工程师"向"模型辅助数据生成-模型辅助训练-模型辅助评估-模型辅助工具链优化"的半自动化方向演进。需强调的是，当前阶段仍处于技术探索早期，不等同于真正意义上的自主研发 AI。

3、M3：三大前沿能力统一——Coding/Agent+长上下文+原生多模态

2026 年 6 月 1 日 MiniMax 发布 M3 旗舰模型，战略定位实现关键跃升：从 M2 系列的"低成本 Agent"单一优势，升级为三大前沿能力集于一身的开放权重模型：Coding/Agent 能力、1MToken 长上下文、原生多模态（支持图片/视频输入+桌面操作）。

核心技术突破——MSA 稀疏注意力架构：M3 底层采用 MiniMax 自研 Sparse Attention（MSA）新型稀疏注意力架构，实现三大性能跃迁：

上下文能力：支持最高 1MToken 上下文窗口

计算效率：100 万上下文长度下，每 Token 计算量仅为上一代的 1/20

推理加速：prefill 阶段加速超 9 倍，decoding 阶段加速超 15 倍

MSA 论文进一步披露：在 109B 参数原生多模态模型上，MSA 在 1M 上下文下将每 Token 注意力计算降低 28.4 倍，H800 平台实现 14.2 倍 prefill 加速与 7.6 倍 decoding 加速。

商业化价值：长上下文能力的商业价值不在于"技术可实现"，而在于"成本可承受"。MSA 架构本质上是将"上下文长度"从模型能力指标转化为可规模化的商业能力，为代码仓库理解、长文档处理、复杂 Agent 执行、跨文件办公、金融研究、长视频理解等高价值场景打开商业化空间。

基准测试表现：M3 在 SWE-Bench Pro 得分 59.0%、Terminal-Bench2.1 得分 66.0%、MCP Atlas 得分 74.2%，官方强调其在真实研发流程中更注重长期协作、任务规划与人机协同效率。

5.3. 科大讯飞：唯一国产算力全流程训推的大模型底座标的

科大讯飞的估值逻辑正迎来历史性重构，应从传统的“教育+语音+智慧城市+消费

硬件”多元业务估值框架，**切换至国产 AI 算力生态中的大模型底座来看待。**

在中美科技博弈长期化、AI 算力供应链国产化加速、央国企模型采购自主可控要求持续提升的产业背景下，具备在国产算力平台上完成**"预训练-强化学习-推理部署-MaaS 服务-行业模型落地"全链条能力的厂商具备极强稀缺性。公司 2025 年大模型相关项目中标金额 23.16 亿元，超过第二名至第六名总和，这是极其具有说服力的市场答卷。**

图20：科大讯飞业绩基本情况



数据来源：科大讯飞年报、招股说明书，东吴证券研究所

核心主线：星火大模型是“国产算力全流程训推”的关键标的

1、核心壁垒：训练侧国产化才是真正的自主可控，训练与推理的难度有本质差异

当前行业普遍强调 "模型适配国产芯片" 或 "国产算力推理"，但**真正的技术壁垒与战略价值集中于训练环节**。推理本质是将已有模型部署到国产芯片运行，核心优化方向为低延迟、高吞吐与成本控制；而训练涉及预训练、后训练、强化学习、多卡通信、算子精度、分布式稳定性、训推一致性等复杂系统工程，技术难度呈指数级提升。

科大讯飞在投资者交流中明确指出：对于国产芯片而言，模型训练难度远高于推理，**国内目前仅讯飞一家真正实现全国产算力上的全栈模型训练**。

核心判断：科大讯飞的战略稀缺性在于星火能否在国产卡上持续完成下一轮模型训练与架构迭代。若未来主流大模型仍依赖海外 GPU 完成训练、仅推理端迁移至国产芯片，中国大模型底座并未实现真正意义上的自主可控。科大讯飞选择的是技术难度更高、但战略价值更为深远的全栈国产化路线。

2、产品迭代：从 "可运行" 到 "商业化可用" 的三级跨越

星火 X1：国产算力训练的里程碑突破

2025 年 1 月科大讯飞发布星火深度推理模型 X1，系**国内首个基于全国产算力平台**

训练、具备深度思考与推理能力的大模型，攻克了训练推理强交互、高吞吐推理优化、国产算子优化等核心技术难题，并率先在教育、医疗等刚需场景落地。

年报进一步披露，星火 X1 在模型参数较业界同类模型小一个数量级的情况下，实现了业界一流效果；2025 年 11 月发布的星火 X1.5，在全国产算力平台上攻克 MoE 模型全链路训练效率，支撑教育、医疗等行业大模型能力持续提升。

产业意义：X1/X1.5 将国产算力训练从"技术可运行"推进至"可用于深度推理模型"，标志着国产算力不再仅是低成本推理替代方案，正式进入深度推理大模型、长思维链、行业思维链、MoE 架构等前沿模型训练阶段。

星火 X2-Flash：中尺寸 MoE 的商业化突破

2026 年 4 月 29 日科大讯飞发布星火 X2-Flash，基于华为昇腾 910B 集群训练完成，采用 MoE 架构，总参数 30B，最大支持 256K 上下文并同步开放 API。该模型率先在国产算力上实现 DSA 稀疏注意力与 MTP 多 Token 预测结合的长文本高效训练，通过亲和国产芯片的算子与分布式训练策略优化，训练效率相较同规模 A800 集群从初期 20%提升至约 90%。

据投资者交流披露，X2-Flash 在 Skill 管理与调用、系统控制与执行等智能体任务上效果接近业界万亿级参数模型，同时 Token 消耗不到部分大尺寸旗舰模型的三分之一。

昇腾 950：2026 年核心技术催化

科大讯飞披露已与华为团队针对昇腾 950 芯片深度对接，联合攻坚更高效模型结构、混合 Attention 机制、智能体强化学习等关键技术；公司预计在 2026 年 1024 开发者节期间，基于昇腾 950 平台发布**中国首个对标业界最先进主流模型的旗舰大模型**。

这将是科大讯飞 2026 年最重要的技术催化。910B 时代公司已证明国产算力能够完成主流大模型训练；950 时代若显存、带宽、算力、低精度能力与多卡互联能力进一步提升，科大讯飞有望将国产算力训练从 "追赶可用" 推进至 "接近国际旗舰模型"。需关注核心不确定性：昇腾 950 的实际供给、集群稳定性、软件生态成熟度、训练效率与模型效果仍需正式发布后验证。

3、产业本质：为什么全流程训推国产化是胜负手？“海外训练 + 国产推理" 的路径局限”

当前行业主流路径为"海外 GPU 训练+国产芯片推理"，短期效率优势显著——海外 GPU 生态成熟、训练工具链完善、主流算法可快速复现；但核心短板在于底座模型的下一轮架构升级、预训练、强化学习与多模态扩展仍受制于海外 GPU 供应。

科大讯飞明确指出，国际主流算法在英伟达卡上可直接实现，但在国产卡上需额外解决算子库效率优化等问题，通常需要 3-6 个月适配周期。这 3-6 个月表面是时间成本，实质是国产算力生态的工程壁垒。科大讯飞与华为长期深度合作，持续解决底层 bug、算子、通信、训推一致性与分布式训练效率问题，这些工程能力正沉淀为公司相对其他模型厂商的先发优势。

当前，昇腾 910B 与 H200 存在显著硬件差距：

显存容量：910B 为 64GB vs H200 为 141GB

带宽：910B 为 1.6TB/s vs H200 为 4.8TB/s

智能体强化学习训练效率：910B 采样推理操作效率较 H200 差距可达 5 倍

这表明国产算力训练并不是简单将 PyTorch 代码迁移，而是需要重构模型结构、算子实现、并行策略、数据管线、采样推理与后训练框架。科大讯飞在 X2-Flash 上采用的 DSA+MTP、亲和国产芯片的算子与分布式训练策略，本质上是在硬件约束下重新设计模型工程路线。

6. 未来大模型格局推演与投资建议

根据当前大模型格局的演化，我们认为未来大模型发展存在三种情形：第一，**闭源模型持续领先开源模型并不断拉开差距；**第二，**开源模型持续追赶并实现工程化真实可用；**第三，**transformer 架构带来的模型边际提升彻底见顶，其他架构取代 transformer 架构成为主流。**

6.1. 情景 A：前沿模型继续 Scaling，闭源巨头优势扩大

核心假设：前沿模型在真实复杂任务上继续拉开差距，尤其是在代码库级工程、科学研究、金融分析、法律推理、多模态视频、计算机使用和长程 Agent 上，闭源模型领先开源模型 6—12 个月以上。若这种领先能显著提高任务完成率，企业会愿意为闭源模型支付高价。

这个情景下，OpenAI、Anthropic、Google、xAI 及其云伙伴最受益。GPU、HBM、先进封装、光模块、交换机、数据中心、电力和液冷继续是最确定的基础设施受益环节。

表12：如果闭源模型持续扩大对开源模型的领先优势，将持续利好 AI 硬件等受益环节

项目	判断
关键前提	闭源前沿模型在真实复杂任务中持续领先
领先公司	OpenAI、Anthropic、Google、xAI、Microsoft、AWS、GCP
受益板块	GPU/ASIC、HBM、先进封装、光模块、交换机、电力、液冷、云
受损板块	中小模型厂、无分发应用、低壁垒 API 包装商
触发信号	闭源模型在 SWE-bench、OS World、Terminal-Bench、HLE、企业 Agent 等指标持续领先
投资结论	继续配置算力—云—前沿模型入口
风险指标	API 降价、云毛利率下降、开源追赶周期缩短、capex 占收入比重过高

数据来源：东吴证券研究所整理

6.2. 情景 B：开源模型持续追赶，模型能力商品化

此为本报告的基准假设。如果开源模型能力持续追赶，就意味着大多数商业任务具备多种模型选择，也就回到了前文提到过的情况：复杂任务用旗舰闭源模型，中等任务用便宜闭源或开放模型，敏感任务用私有化模型，低价值高并发任务用小模型或端侧模型。

模型能力商品化会降低应用开发门槛，推动更多企业把 AI 接入真实流程。这个情景下，模型层 API 毛利受压，但应用层、推理基础设施、私有化部署、RAG、数据库、安全和端侧 AI 受益。

表13：开源模型持续追赶的情形下，AI 应用等板块优先受益，AI 硬件板块可能受到冲击

项目	判断
关键前提	开源/开放模型在 3—6 个月内追近闭源
领先公司	Meta、DeepSeek、Qwen、Kimi、MiniMax、Baidu、Tencent、云厂商、应用厂商
受益板块	AI 应用、推理云、私有化部署、RAG、数据库、安全、端侧 AI
受损板块	纯闭源 API、无入口模型厂、简单套壳应用
触发信号	企业采用多模型分工，开源模型真实任务完成率接近闭源
投资结论	从“买最强模型”转向“买最低成本完成任务的系统”
风险指标	开源安全事件、许可收紧、闭源模型再次拉开差距

数据来源：东吴证券研究所整理

6.3. 情景 C：架构或新范式突破，Transformer 优势被削弱

这一情景概率最低，但潜在影响最大。Mamba/SSM、linear attention、RetNet、Hyena、hybrid attention、memory-augmented models、world models、neuro-symbolic AI、brain-inspired architectures 和 sparse modular systems 都可能削弱纯 Transformer 的中心地位。

Mamba 论文显示，选择性状态空间模型可实现线性时间序列建模，并在推理吞吐和长序列扩展上具有优势；Flash Attention 则说明即使不替代 Transformer，也可以通过 IO-aware kernel 显著改善 attention 的实际效率。

但替代 Transformer 很难。新架构不仅要在小规模论文任务上表现好，还要在大规模预训练、多模态、后训练、工具调用、推理部署、硬件利用率、生态兼容和真实任务上全面胜出。因此更可能的路径是 Transformer + MoE + hybrid attention + memory + Agent system 的复合演进，而不是单一新架构完全替代。

6.4. 市场投资框架

综合来看，围绕大模型的投资框架依然是依据前文提到的 AI 利润池进行梳理：从大模型本身，到 AI 芯片（国产算力、半导体设备）、数据中心（AI 服务器、光模块、PCB/交换设备、液冷/电源/电网电力设备）、AI 应用、端侧 AI。

当前仍处在 AI 基建热潮当中，利润池集中在 AI 硬件，建议关注 AI 芯片、AI 服务器、光模块等 AI 硬件相关板块。

中远期来看，我们依然看好 AI 应用。尽管当前 AI 软件尚未迎来确定性大爆发，但随着开源大模型能力持续提升、物理 AI 需求倒逼行业发展，我们认为 AI 应用具备广阔的成长空间。

表14：建议关注标的列表

A 股板块	投资逻辑	国内代表公司
大模型	**AI 时代核心产品，智能底座**	**智谱（港股）、Minimax（港股）、科大讯飞、阿里巴巴（港股）**
AI 芯片	国产算力、推理加速、信创需求	寒武纪、海光信息、禾盛新材、景嘉微
AI 服务器	训练/推理服务器需求	工业富联、浪潮信息、中科曙光
光模块	800G/1.6T、高速互联	中际旭创、新易盛、天孚通信、光迅科技
PCB/交换设备	高速网络与服务器主板	沪电股份、胜宏科技、深南电路、紫光股份
半导体设备/材料	国产替代、先进封装	北方华创、中微公司、长电科技、通富微电
液冷/电源	高功率机柜和数据中心	英维克、科华数据、高澜股份、申菱环境
电网/电力设备	AI 数据中心并网和供电	国电南瑞、许继电气、思源电气、特变电工
AI 软件	办公、教育、语音、企业应用	金山办公、科大讯飞、用友网络、万兴科技、昆仑万维
数据安全	企业 AI 合规和安全	奇安信、启明星辰、安恒信息
机器人/汽车	端侧 AI 与具身智能	中控技术、埃斯顿、汇川技术、拓普集团

数据来源：东吴证券研究所整理

7. 风险提示

技术迭代不及预期：大模型架构与 Agent 能力仍处于快速演进阶段，若核心技术突破慢于预期可能影响商业化进程。

行业竞争加剧：开源模型与高性价比路线快速追赶，可能导致模型 API 价格战与利润率承压。

算力供应链风险：GPU、HBM 等核心硬件供给与地缘政治变化可能影响训练部署进度。

监管政策不确定性：AI 内容安全、数据合规与监管政策趋严可能增加企业合规成本。

商业化落地慢于预期：Agent 与企业级应用 ROI 验证周期可能长于市场预期。

仅供内部参考，请勿外传！

东吴证券 SOOCHOW SECURITIES

免责声明

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

本研究报告仅供东吴证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为客户。在任何情况下，本报告中的信息或所表述的意见并不构成对任何人的投资建议，本公司及作者不对任何人因使用本报告中的内容所导致的任何后果负任何责任。任何形式的分享证券投资收益或者分担证券投资损失的书面或口头承诺均为无效。

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

市场有风险，投资需谨慎。本报告是基于本公司分析师认为可靠且已公开的信息，本公司力求但不保证这些信息的准确性和完整性，也不保证文中观点或陈述不会发生任何变更，在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。

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