人工智能系列（一）：大模型从技术竞赛到商业闭环，AI 生态收敛与竞争格局前瞻

评级：强于大市（维持）

主要观点：

1.全球大模型使用需求保持高增长，行业集中度同步提升。

OpenRouter 数据显示，平台周度 Token 调用量由 2025 年 9 月 5.26T 增至 2026 年 9 月约 127T，最新数据中 DeepSeek、Google、OpenAI 三家合计份额已超过 60%，中国模型在 TOP10 中的合计份额超过 40%。海外市场中 OpenAI 与 Anthropic 在用户规模、企业客户和商业化进展方面领先，Google 依托云、搜索、安卓及自研算力形成较完整的产业布局。国内市场逐步形成大厂系模型、独立模型厂商、本地化部署厂商等不同发展路径。模型能力仍处于快速迭代阶段，单一模型的领先时间有限，用户入口、企业客户、工作流嵌入和数据积累对长期竞争力的重要性持续提高。

2.大模型能力仍有较大的提升空间，算力、数据和工程效率共同决定后续迭代速度。

Scaling Law 仍是当前头部模型迭代的重要基础。推理模型和 Agent 显著增加单次任务的计算量，长上下文进一步扩大代码库、长文档及复杂任务的处理范围。长上下文的情况下，单次任务能够容纳更多历史信息、任务状态和工具调用结果，同时明显提升 Token 消耗。国内模型在参数效率、推理成本和开放部署方面进展较快，DeepSeek、GLM、Kimi 等厂商通过稀疏架构、低精度训练、缓存复用和通信优化降低单位 Token 成本，并通过开放权重、本地部署和接口兼容扩大开发者及企业客户覆盖。随着国内外模型能力差距缩小，成本、部署方式和行业适配能力的重要性将进一步提高。

3.Coding 已经成为大模型商业化较为明确的应用场景之一，Agent 正在扩大模型在企业工作流程中的使用范围。

软件开发任务具有较清晰的输入、输出和验证标准，代码提交、测试通过率、缺陷修复时间等指标可以直接衡量AI带来的效率提升，因此 Coding 较早形成付费需求。当前 Coding 产品已经覆盖代码库理解、终端操作、测试执行和多文件编辑等环节，产品能力逐步覆盖更完整的软件工程流程。Agent 进一步进入办公、研究、销售、财务和客户服务等知识工作场景，并带动模型调用次数、上下文长度和任务

报告日期：2026-9-30

执业信息
王炤杰 电子行业分析师
执业编号：S1750526070001
电话：18925230381
邮件：wangzj@jypscnet.com

免责声明
报告观点基于逻辑分析、数据规律推演，并不能预知不同市场主体在复杂约束下的多样选择结果。

相关报告
20260803 电子行业半年度策略：AI 硬件主升浪后的分化与再配置

执行时间增加。随着任务复杂度提高，Token 消耗将更多反映实际使用深度，API、订阅、Credits 以及企业部署等多种收费方式预计将长期并存。

4.商业化进展开始成为大模型厂商估值的重要依据，国内厂商之间的差异更多体现在客户、收入和成本端。

2025 年以来，Coding 付费、Agent 企业应用以及模型调用量增长，使市场对大模型公司的观察指标更加具体。除模型能力外，Token 调用量、API 价格、ARR、企业客户数量、续费情况、推理成本和毛利率的重要性持续上升。国内大厂具备流量入口、云基础设施和企业客户基础，能够依托原有业务体系承接 AI 需求。独立模型厂商对融资能力、模型迭代速度和收入增长要求更高。本地化及垂直厂商主要依靠专有数据、部署能力和行业交付能力获取订单。国内 C 端调用量与 B 端 ARR 排名并不一致，用户规模并不能单独反映模型公司的商业价值。

5.投资建议：大模型行业已经进入商业化验证阶段。模型排名和参数规模仍是观察技术能力的重要指标，但 Token 需求、API 定价、ARR、企业客户和毛利率等经营指标的重要性持续提高。建议重点关注三类厂商：一是具备用户入口、云基础设施和企业客户资源的大型平台公司，其优势主要体现在客户获取和商业化稳定性。二是模型能力、推理成本和收入增速领先的独立模型厂商，其业绩弹性较高，同时对融资和算力投入较为敏感。三是在政企、金融、制造等领域具备本地部署、专有数据和系统集成能力的厂商，其订单获取更多取决于行业需求和项目落地情况。

6.风险提示：模型技术迭代不及预期；开源模型竞争加剧导致 API 价格和行业利润率承压；高端算力供给受出口管制及地缘因素影响；行业竞争加剧；算力资本开支较高、商业化及盈利进展不及预期；基础模型能力提升对现有软件及 AI 应用产品形成替代。

更多免费行业报告 并购家 www.ipoipo.net

目录

一、百模大战落幕，全球 AI 大模型进入商业格局定型的下半场 5
二、技术路线分岔后再收敛，差异化空间转向上下游 10
三、商业闭环：Coding 率先兑现，Agent 打开第二增长曲线 16
四、国产大模型竞争由技术投入转向商业兑现 22
五、投资建议 25
六、风险提示 27

图表目录

图表 1：OpenRouter 平台主要大模型厂商 Token 使用份额 5
图表 2：OpenAI 与 Anthropic 构筑前沿 AI 第一梯队竞争格局 6
图表 3：Epoch AI 追踪登顶领先中位数 7 周 6
图表 4：海外六强格局定位 7
图表 5：国内多元竞争下的梯队分化 8
图表 6：厂商优势定位 9
图表 7：监督微调、奖励模型与强化学习 10
图表 8：2026 年下半年国内外主流基座模型参数规模预期 12
图表 9：国内外主流厂商 Agent 产品发布 13
图表 10：主流大模型上下文窗口对比 13
图表 11：智能指数与加权平均输入价格对比 14
图表 12：架构效率与算力基础设施和能源协同优化 15
图表 13：OpenRouter 月度大模型 Token 使用量 16
图表 14：AI 任务复杂度提升与 Token 需求变化 17
图表 15：主流 Coding Agent 产品竞争力对比 17
图表 16：2025E-2030E AI 编程市场规模测算（TAM/SAM） 18
图表 17：2023-2028 年中国 Al 代码生成市场规模及增速 18
图表 18：主流 Agent 平台竞争力对比 19
图表 19：主流 AI Agent 产品能力对比 20
图表 20：大模型商业闭环模式 21
图表 21：大模型商业化的三阶段演进 22
图表 22：大模型关注重点的迁移 22

图表 23：国产大模型厂商竞争格局 23
图表 24：OpenAI 与 Anthropic 最新公开融资及估值情况 23
图表 25：大模型厂商投融资活跃，头部企业估值跃升 23
图表 26：生态型厂商 25
图表 27：独立模型厂商 25
图表 28：垂直与本地化模型 26

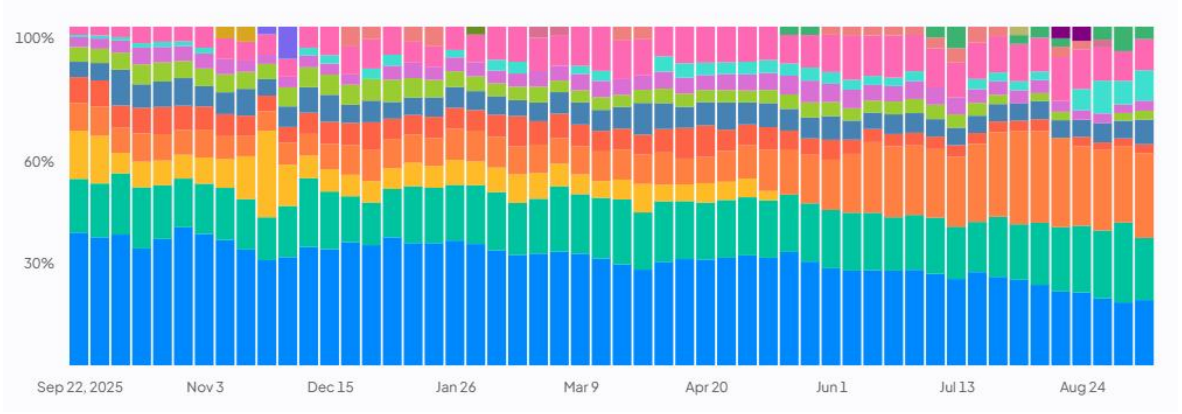
更多免费行业报告 并购家 www.ipoipo.net

一、百模大战落幕，全球 AI 大模型进入商业格局定型的下半场

1.1 全球大模型生态加速演进，Token 需求增长与竞争格局重塑同步推进

全球大模型竞争正进入以真实 Token 调用为核心的加速重构期，需求扩张与份额洗牌同步上演。全球最大的中立大模型 API 路由市场 OpenRouter 数据显示，各模型 Token 调用规模整体持续攀升，周调用总量从 2025 年 9 月的 5.26T 一路增至 2026 年 9 月初的 127T，增长 24 倍。Token 流量加速向头部厂商聚拢，OpenAI 以 23.6%的市占率暂居榜首，DeepSeek 与 Google 分别以 22.6%与 18.6%紧随其后分列二、三位，TOP3 厂商合计占比已突破 60%。格局重构之下，DeepSeek、腾讯混元、智谱 GLM 及阿里千问四家国产厂商跻身 TOP10，合计占比超 40%，成为本轮市场格局重构中最为活跃的中坚力量。

图表 1：OpenRouter 平台主要大模型厂商 Token 使用份额



20260907 平台大模型厂商 Token 使用份额

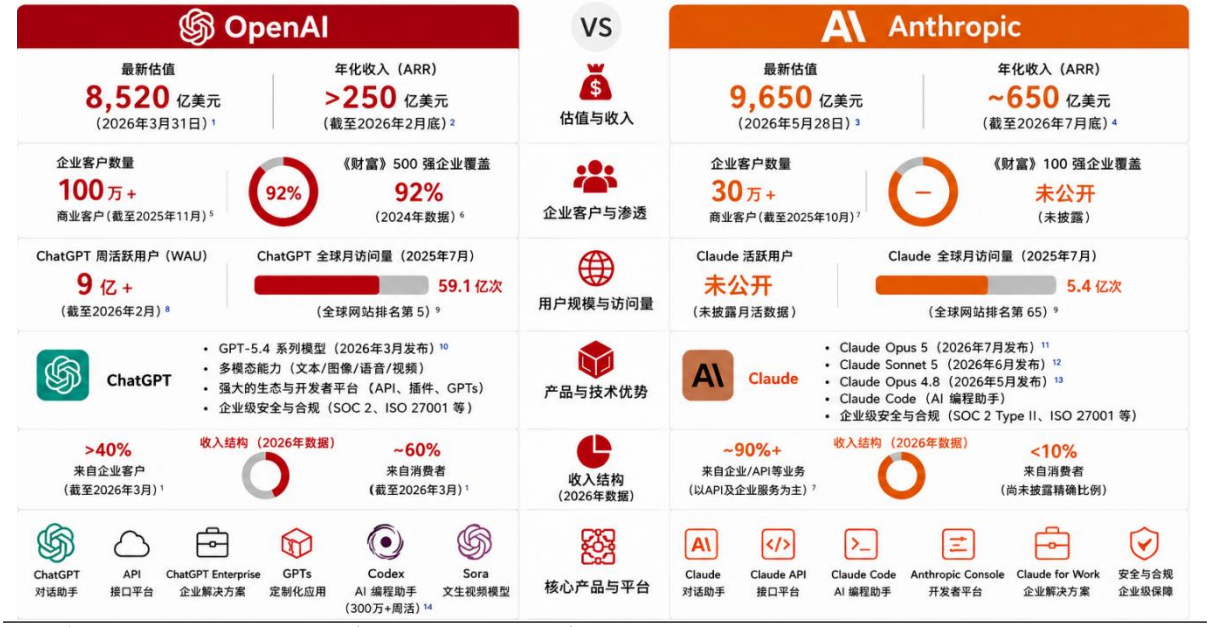
公司名称	图例	使用份额	公司名称	色块	使用份额
DeepSeek		25.1%	腾讯		3.8%
Google		19.2%	Anthropic		3.0%
OpenAI		18.3%	Mistral AI		2.6%
智谱		9.0%	Meta-llama		2.6%
千问		7.3%	others		9.2%

数据来源：OpenRouter-Market Share、金圆统一证券研究所

1.2 海外：双寡头领跑，多元模式竞逐

海外大模型市场已形成以 OpenAI 与 Anthropic 为核心的第一梯队格局，二者凭借模型能力、商业化规模与生态优势构筑竞争壁垒。OpenAI 依托 ChatGPT 形成全球最大的消费者入口，截至 2026 年 2 月已拥有超 9 亿周活用户，并通过 API、企业服务等渠道持续扩大商业化规模。Anthropic 则凭借 Claude 系列模型和 Claude Code 快速切入企业级场景，收入结构更加偏向 B 端，公司管理层预计 2026 年底 ARR 将达到 1,000 亿-1,200 亿美元，并计划在 2026 年秋季 IPO，目标估值冲击 2 万亿美元。

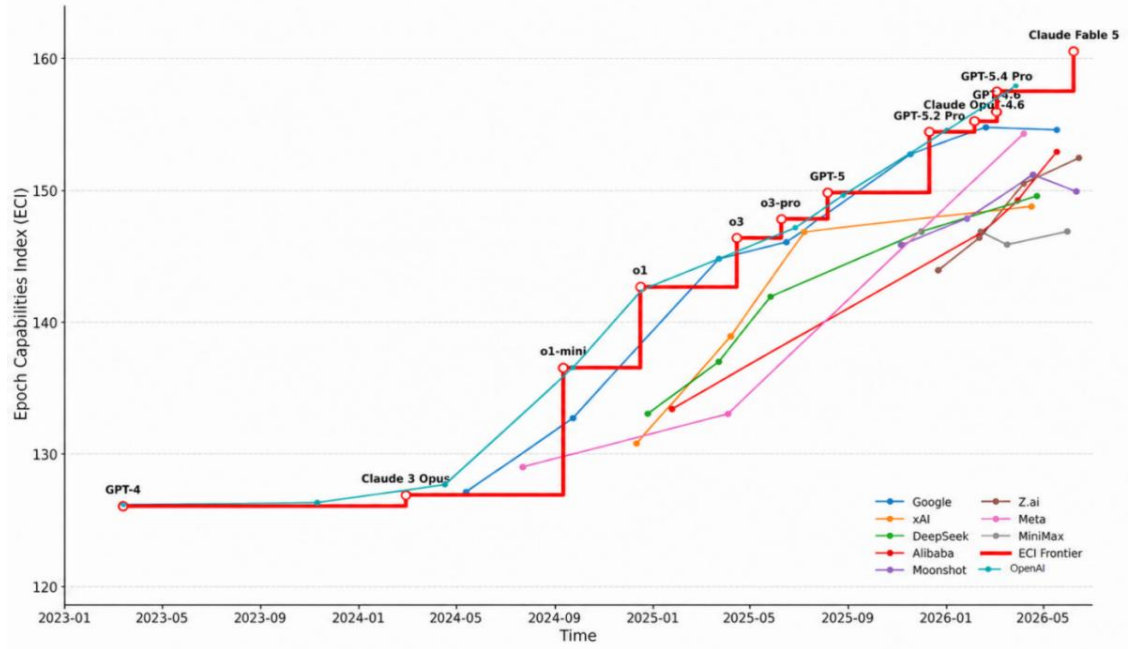
图表 2：OpenAI 与 Anthropic 构筑前沿 AI 第一梯队竞争格局



数据来源：OpenAI、Anthropic、证券时报网、金圆统一证券研究所

商业格局的头部集中并未形成技术能力的长期垄断，全球模型能力排名仍处于快速迭代之中。Epoch AI 统计显示，Anthropic 发布 Opus 3 后共有 17 个模型轮流登顶，平均领先约 7 周，技术领先优势高度动态。尽管收入与份额向头部集中，但技术领先窗口极短，任何厂商都难以长期垄断榜单。海外的收敛是商业意义上的收敛而非技术意义上的终局，真正的护城河沉淀在切换成本而非能力排名之中。

图表 3：Epoch AI 追踪登顶领先中位数 7 周



数据来源：Epoch AI、金圆统一证券研究所

海外六强发展路径各异，但透过能力来源、变现逻辑与护城河的三维透视，其商业模式最终殊途同归，收敛于三大核心形态：

模式一：闭源前沿双寡头（OpenAI、Anthropic）。以最强闭源基座模型为核心资产，通过

更多免费行业报告 并购家 www.ipoipo.net

API 与订阅直接变现，护城河来自先发优势、全栈产品矩阵与企业级切换成本。OpenAI、Anthropic2026 年底前预计均将触及 1000 亿美元 ARR，但抵达路径分化明显：

1、OpenAI 依托 ChatGPT 的消费级订阅规模与开发者生态广度，收入结构更均衡；

2、Anthropic 高度依赖 Coding 与 Agent 场景的企业级高客单价客户，并借 Claude Code 沉淀的真实使用轨迹形成数据飞轮。

模式二：全栈纵向整合型（Google）。Google 是唯一实现芯片、云业务、模型、应用四层全线贯通的厂商，其纵向一体化的成本结构构筑了竞争对手难以逾越的护城河。公司 VO3 技术路线坚持多模态原生架构，从底层将图文视频能力融为一体，而非行业常见的后期拼接。配合搜索安卓及 YouTube 构成的超级分发网络，Google 形成了生态纵深叠加原生多模态的独特壁垒。该模式变现逻辑迥异于单纯的 API 收费，核心在于通过 AI 升级既有用户体验来实现价值回收。这种内生增长逻辑不仅赋予了其在价格战中的极强耐受力，更使其成为冲击现有双寡头格局的最大扰动变量。

模式三：开源与差异化数据型（Meta、xAI、Mistral）。Meta、xAI、Mistral 均弱化模型层直接收费，转而依托主业生态兑现 AI 价值。Meta 通过开源策略全面赋能社交主业，同时持续压制闭源厂商的定价空间。xAI 凭借 X 平台独有的实时高噪语料构筑数据新鲜度壁垒，并依托超大规模 GPU 集群全力追赶技术代际。Mistral 则深耕 GDPR 及 AI 法案框架下的合规优势，牢牢占据欧洲主权 AI 及数据本地化这一排他性市场。这类厂商的共性在于变现路径相对间接，其成长高度依赖主业景气度、算力扩张节奏以及区域政策红利的释放。

图表 4：海外六强格局定位

厂商	代表模型	定位	核心优势	商业化进展
OpenAI	GPT-6	闭源基座 消费级入口	先发品牌与开发者心智 训练—推理—应用全栈矩阵 持续融资支撑算力储备	2026 年底前 ARR 预计达 1000 亿美元量级
Anthropic	Claude Claude Code	闭源第二极 企业级主导	Coding 与 Agent 长链路任务统治 Claude Code 真实轨迹数据飞轮 高客单价客户黏性	ARR 由 2025 年底 90 亿到 2026 年 5 月 470 亿，年内首超 OpenAI 并实现单季经营盈利，百万美元级客户 500→超 1000 家
Google	Gemini V3	原生多模态	搜索＋安卓＋Cloud 生态纵深 V3 原生多模态架构 自研芯片—云—模型—应用纵向整合	依托既有超级入口分发，从原生多模态切入重构交互范式
Meta	Llama 系列	开源阵营旗手	开源生态与社区标准 社交业务巨量多语言多模态数据 自有算力内部循环	不以模型层直接变现，通过开源压制闭源定价并反哺广告主业
xAI	Grok 系列	实时数据 算力规模路线	深度绑定 X 平台的实时社媒语料 协同大规模 GPU 集群的算力扩张速度	产品化与企业生态成熟度落后于双寡头，变现仍处于早期
Mistral	Mistral 开源	欧洲主权 AI	GDPR 与《AI 法案》下的合规适配 开源权重维系欧洲开发者社区 数据本地化首选供给	开源引流＋商业 API 服务企业客户的双轨变现

资料来源：Geek Salad、金圆统一证券研究所

1.3 国内：百模大战落幕，头部梯队领跑

C 端调用量排序与 B 端 ARR 排序并不一致。豆包、千问凭借大厂流量入口在 C 端调用量上遥遥领先，但其收入贡献更多体现为对母体生态的战略价值而非独立模型收入。智谱 GLM 的 C 端调用量位居第二梯队，却在 B 端 ARR 上占据首位，智谱以本地化部署为主的交付模式更贴合政企客户的合规与数据主权要求，单客户合同金额显著高于公有云 API。

图表 5：国内多元竞争下的梯队分化

定位	公司		
生态	腾讯、阿里、字节、百度		
API	DeepSeek、智谱、Kimi、MiniMax		
开发者	百川、阶跃		
梯队	第一梯队	第二梯队	第三梯队
ToC	豆包、千问 （日调用量：5 万亿+）	DeepSeek、智谱 GLM （日调用量：2-2.7 万亿）	阶跃、百川 （日调用量：1.5-1.8 万亿）
ToB	智谱 GLM （以本地化部署为主）	Kimi、MiniMax （B/C 端并重）	阶跃、DeepSeek

资料来源：OpenRouter、QuestMobile、金圆统一证券研究所

国内厂商的竞争差异，核心在于商业逻辑的分化，而非单纯的实力排位。以营收来源与护城河壁垒为标尺，可将其梳理为四类典型阵营：

模式一：生态入口型大厂模型（腾讯混元、阿里通义、字节豆包、百度文心）。共同特征是模型能力嵌入母体超级 App 与云业务，边际获客成本趋近于零，收入更多体现为对母体生态的战略价值而非独立模型收入。腾讯以微信、QQ、游戏、广告与腾讯云构成的矩阵占据入口，其云 API 输入价格一度飙升 463%，侧面印证生态入口型厂商的供给端定价能力。阿里以阿里云算力、钉钉企业入口与电商语料形成闭环。字节以推荐分发与周级迭代速度将豆包做成国民级应用，2026 年 6 月单家日均 Token 调用量达 180 万亿量级。百度则以搜索沉淀的中文语料与智能云 toB 能力构建纵深优势。该类厂商的存续取决于母体战略意愿，天然处在收敛后的存活名单之内。

模式二：API 平台与开源性价比型（DeepSeek、Kimi、MiniMax）。以开源权重或高性价比 API 扩散生态，再通过定价能力或多模态产品变现。DeepSeek 以极致性价比重构成本—能力曲线，直接压低全行业通用任务的 Token 价格中枢，其 V4 Flash 与小米 MiMo V2.5 包揽 2026 年 7 月全球 Token 调用量前两席。Kimi K3 以全球首个开源 2.8 万亿参数模型跻身全球前五，同时将未命中缓存输入价由约 6.5 元 / 百万 Token 提至 20 元，出价由约 27 元提至 100 元，需求不减，证明开源与溢价并不互斥。MiniMax 则以语音、音乐、视频的端到端多模态布局与 Talkie 等海外 C 端产品开辟第二战场。该类厂商必须在窗口期内完成商业闭环，是竞争最激烈，也最具业绩弹性的一类。

模式三：本地化部署与政企交付型（智谱 GLM）。以私有化部署契合数据安全与信创诉求，换取高客单价与高续费黏性，构成海外厂商无法进入的结构性市场。智谱 ARR 已达 10 亿美元、2026 年以来增长约 15 倍，API 价格提升 83%后调用量仍增长约 400%，是 B 端变现跑通最直接的注脚。技术侧 GLM 5.2 实现百万级上下文、整体对齐 GPT-4.6 并在部分场景达到 GPT-5.5 级别，环比增长 306%。该模式的显著特征是 C 端调用量与 B 端收入排序背离，流量不占优，变现效率却最高。

模式四：软硬一体国产化型（华为昇腾＋盘古）。以芯片—框架—模型的全栈自主构成闭环，在出口管制与地缘约束下具备不可替代性。WAIC 2026 上昇腾 950 超节点实现真机落地，国芯＋国模正循环初步成形。其意义超出单家厂商的份额之争，它补齐了国产算力与国产模型协同的关键一环，使国产模型对国产算力的适配从能用走向大规模必用。

更多免费行业报告 并购家 www.ipoipo.net

四类之外，百川智能与阶跃星辰代表垂直与开发者生态型路线：百川智能深耕医疗等行业，形成强合规壁垒的差异化数据。阶跃星辰以多模态体系化积累与端侧部署探索卡位。二者日调用量均处 1.5－1.8 万亿的第三梯队。

图表 6：厂商优势定位

厂商	代表模型	核心优势	海外样板	对标
腾讯	混元	微信/ QQ /游戏/广告/腾讯云构成的超级 App 矩阵带来明确入口与流量优势；社交内容生态真实交互数据；"云+模型"一体化交付；企业微信、腾讯会议等 B 端触点；腾讯云 API 输入提价 463% 印证供给端定价能力	Meta	都是超级平台用AI反哺存量生态，不靠模型SOTA单挑； AI嵌入社交、内容、广告、企业服务兑现价值
阿里	通义千问	阿里云国内领先的算力与云基础设施；钉钉企业级入口与现成 B 端工作流；电商交易与客服语料适配垂直场景；开源侧持续投入积累开发者生态	Google（弱版全栈）	云+钉钉企业入口+电商语料形成"云+模型"闭环，试图复制Google的四层贯通
字节跳动	豆包	推荐算法驱动的精准分发；抖音/头条超级 App 流量池；周级产品迭代速度；被共识视为收敛后留存的国民级 AI 应用	MetaxGoogle 分发	借推荐算法和分发优势做国民级AI应用， 类似 Meta AI 助手覆盖30亿用户的路径
百度	文心	搜索沉淀的海量中文语料与用户意图数据；自动驾驶/文库/网盘多元落地场景；百度智能云 toB 交付能力	Google+Microsoft混合	搜索语料+智能云toB能力 +文心一言智能体平台
智谱	GLM 5.2	百万级上下文，对齐 GPT-4.6 、部分场景达 GPT-5.5 ，环比+306%；面向政企的本地化部署契合信创与数据主权；高客单价与续费黏性； ARR 达 10 亿美元、年内约 15 倍增长；下一代全新架构参数远超 GLM-5	AnthropicxIBM /政企交付	同样押注Coding+Agent作为核心突破口， 同样靠 API/Token为企业客户创造价值
DeepSeek	V4 Flash	2026 年 7 月全球 Token 调用量前十中与小米 MiMo V2.5 同居前两席；开源社区广泛适配与生态扩散；极致性价比重构成本 — 能力曲线；有限算力下的训练效率优势	Meta开源xOpenAI	权重完全开源(像Meta/Llama)， 但同时自营低价API变现(像OpenAI)
Kimi （月之暗面）	K3	2026 年 7 月 17 日发布，全球首个开源 2.8 万亿参数模型，全球前五、开源第一；长上下文能力持续强化；输入价 6.5→20 元、输出价 27→100 元（超 3 倍）仍需求不减，具备强定价权	OpenAI（ToC订阅方向）	追求前沿模型能力+C端订阅 +开发者生态
MiniMax	MiniMax 系列	语音、音乐、视频等多模态端到端布局； Talkie 等产品的海外 C 端较早布局与跨区域用户基础；共识中保持核心竞争力的四家之一	OpenAIxMeta 多模态	端到端多模态布局(语音/音乐/视频)+Talkie海外C端， 类似OpenAI多模态+Meta社交分发的混合
百川智能	Baichuan 系列	医疗等垂直领域深耕形成的差异化数据与行业 know-how ；开源路线维护开发者社区与下游适配	Anthropic 垂直版	深耕医疗等行业的合规壁垒 与差异化数据
阶跃星辰	Step 系列	多模态能力体系化积累；端侧部署探索使能力向手机、 IoT 下沉，契合 2028 年端侧与物理 AI 落地趋势	Meta 多模态探索者	多模态体系化积累 +端侧部署卡位
华为	盘古＋昇腾 950	WAIC 2026 昇腾 950 超节点真机落地，"国芯＋国模"正循环；国产算力自主在出口管制下不可替代；芯片 — 框架 — 模型全栈软硬协同；政企信创市场深度覆盖	Google（全栈纵向整合）	唯一实现"芯片一框架一模型"全栈自主闭环， 最接近Google的四层贯通

资料来源：36Kr、OpenRouter、金圆统一证券研究所

二、技术路线分岔后再收敛，差异化空间转向上下游

2.1 大模型能力演进的关键环节

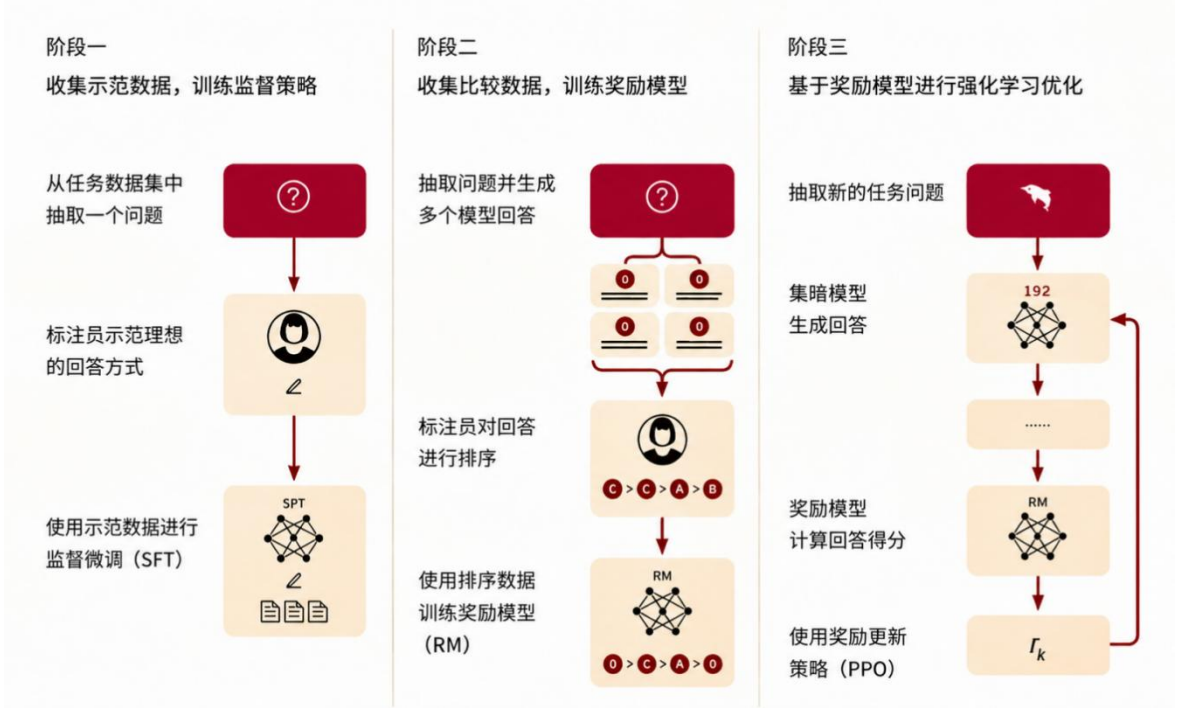
大模型从知识学习到任务执行的能力塑造，需要经历预训练、监督微调、强化学习三个阶段：

预训练：大模型能力构建的基石，其核心在于通过海量无标注文本数据进行自监督学习，让模型在万亿级别的 token 中广泛吸收语言的语法规则、逻辑结构、常识信息以及跨领域的世界知识。这一阶段相当于为模型建立了一个庞大的知识库，使其具备了丰富的认知储备，此时的模型更像是一个博览群书的学者，它知道很多却尚未学会如何根据人类的具体指令去组织这些知识，以恰当的方式回应问题，因此需要后续的引导来激活其任务执行能力。

监督微调（SFT）：连接知识储备与任务执行的关键桥梁，它通过高质量的人工标注数据对预训练模型进行有监督训练，教会模型准确理解人类指令的意图，并将已学到的知识转化为具体的任务输出形式，如问答、摘要、翻译、代码生成等。这一阶段模型被真正唤醒了指令遵循能力，它不再只是被动地预测下一个词，而是能够根据写一封邮件、解释这个概念、帮我改这段代码等明确指令，主动调用内部知识完成对应任务，实现了从懂知识到会做事的本质跨越。

强化学习（RLHF/DPO）：对模型任务执行质量的精细打磨与价值对齐，它先通过收集人类对模型不同回答的偏好数据训练出一个奖励模型，再利用强化学习算法或直接偏好优化算法来迭代优化模型策略，使其输出在有用性、逻辑性、安全性和事实准确性等维度上持续向人类期望靠拢。模型在复杂推理、长文本生成、多轮对话等任务上的表现更加可靠自然，同时大幅降低了有害输出和幻觉现象，使模型的行为方式与人类价值观和社会规范实现深度对齐。

图表 7：监督微调、奖励模型与强化学习



资料来源：OpenAI、金圆统一证券研究所

更多免费行业报告 并购家 www.ipoipo.net

2.2 AI 竞争迈向算力+数据双轮驱动时代

2026 年全球大模型技术路线共同收敛于大总参数量、稀疏化激活、长链推理与强工具调用的统一范式。海外阵营延续扩张稠密参数规模的路径，持续探索模型能力的天花板，国产阵营则普遍转向混合专家（MoE）架构，在总参数与激活参数之间寻求精妙平衡，以优化推理部署的经济性。当模型架构日渐同质化，厂商的竞争壁垒将更多取决于数据质量、工程效率与工具生态的厚薄，单纯依靠架构创新所能撬动的估值溢价正加速消退。技术突破仅为价值创造提供了供给侧的根基，而一项能力是否真正被需要、被使用、被付费，终究只能由应用场景中的真实数据作答。综合海外与国内厂商的禀赋，决定长期竞争力的核心维度可归纳为五项：

1. **算力获取。**在高端芯片受约束的环境下，谁掌握稳定、规模化的算力供给，谁就掌握商业闭环的底座；
2. **数据飞轮。**真实长轨迹数据的积累速度，正超越架构本身成为能力分水岭；
3. **生态入口。**超级 App 与云、电商、社交的流量分发，决定模型触达用户的成本与速度；
4. **性价比。**以更低成本达成可比性能，是生产场景规模化与开源扩散的前提；
5. **本地化部署能力。**在政企与信创场景中，数据安全与自主可控构成硬约束。

这五个维度是投资主线的底层逻辑，分别对应到具体厂商，即华为之于算力自主、腾讯之于生态入口、Anthropic 之于数据飞轮、DeepSeek 之于极致性价比、智谱之于本地化部署，五维能力的不均衡发展，正是当前梯队分化的根源。五个维度中，算力获取与数据飞轮的权重之所以持续上升，根源在于能力边界本身仍在延展，天花板未至，意味着继续投入仍能换来能力增益，而能力增益只能通过算力与数据兑现。

2.3 技术演进主线一：模型能力天花板未至，规模定律（Scaling Law）延续

业内对现有 Transformer 架构与训练范式继续 scale 的进步空间普遍乐观，甚至认为可直达 AGI。2026 年下半年，海外头部基础模型的参数规模预计进一步向 10 万亿级迈进，GPT-6 预计达到约 8 万亿参数，部分下一代模型的参数规模可能更高。**参数规模迈向十万亿量级，对产业链的直接影响体现在训练与推理两端：**

1. **训练侧：**十万亿参数模型的单次完整训练需数万张高端加速卡持续运行数月，对集群互联带宽、故障恢复能力与能耗管理提出前所未有的要求，这也正是超节点技术在 2026 年集中落地的直接动因，规模化互联技术已成为决定大模型训练效率的关键瓶颈之一。
2. **推理侧：**参数规模扩张对显存容量与访存带宽的要求依然刚性提高，这构成了高端 Token 成本刚性的技术根源。

图表：2026 年下半年国内外主流基座模型参数规模预期

地区	基座模型	总参数（万亿）	能力定位
国外	GPT-5.6 Sol	~5.0	通用旗舰、复杂推理、Coding 与 Agent
	GPT-6	~8.0	下一代通用旗舰、深度推理与 Agent
	Gemini 3.1 Pro	~4.6	原生多模态、复杂推理与长上下文
	Gemini 3.5 Pro	~6.0	下一代多模态旗舰、Coding 与 Agent
	Claude Opus 5	~5.0	Coding、复杂知识工作与长程 Agent
	Grok 4.6	~3.0	实时信息、复杂推理、Coding 与 Agent
	Grok 5	~6.0	下一代旗舰、多模态、Coding 与 Agent
国内	Kimi K3	2.8	长上下文、多模态、Coding 与 Agent
	Qwen3.8-Max	2.4	Coding、办公、多模态与长程 Agent
	DeepSeek V4-Pro	1.6	高性价比推理、Coding 与 Agent
	Doubao-Seed-2.1-Pro	1.0	通用推理、多模态与企业应用
	MiMo-V2.5-Pro	1.0	多模态、Coding 与端云协同
	GLM-5.3	0.7	长程 Agent、Coding 与企业部署
	MiniMax M3	0.4	多模态、长上下文与 Agent
	Tencent Hy3-preview	0.3	推理、Coding、Agent 与企业应用

资料来源：Artificial Analysis、LifeArchitect、金圆统一证券研究所

Kimi K3 以 2.8 万亿参数成为全球首个该量级开源模型，战略意义超出技术本身：既为全球开发者提供不受地缘约束的高性能底座，也对闭源厂商定价形成结构性约束。但该量级模型的自主部署门槛极高，能独立承载其推理服务的机构有限，所以开源更多体现为技术透明与生态号召而非直接的成本替代，这也解释了为何 Kimi 在开源同时仍能大幅提价。

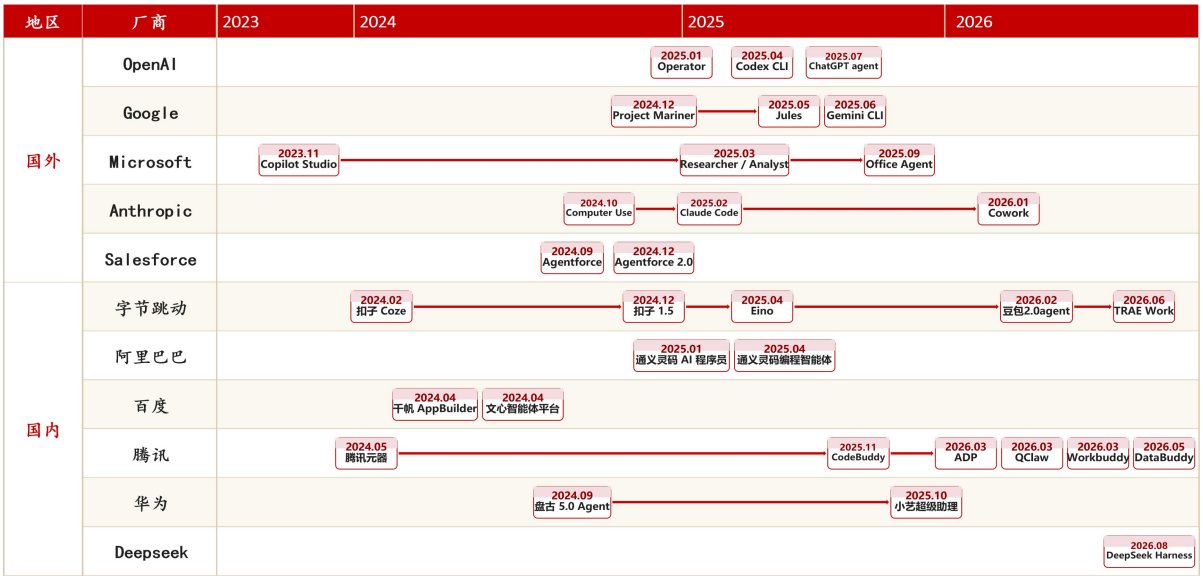
2.4 技术演进主线二：从被动应答到主动执行，推理 Agent 化提速

2026 年大模型产业进入以推理模型和 Agent 能力为主线的阶段。推理能力的提升，使模型从被动应答转向主动规划与执行，这是大模型从工具走向生产力的关键一跃，模型已具备处理长链条、多步骤复杂任务的能力。系统对话的输入输出比例相对均衡，单次交互的输出长度有限，而推理模型在给出最终答案前会生成大量中间思考 Token，Agent 任务更需要在工具调用、结果观察、计划修正之间反复循环。这导致单位任务的计算量呈数量级放大，同时也使得每 Token 价值出现分化，用于最终交付的 Token 与用于中间推理的 Token，在客户感知价值上并不相同，这正是分层定价得以成立的技术基础。

全球 Agent 产业正由对话辅助工具加速向具备自主规划、工具调用和任务执行能力的智能体演进，国内外市场已形成多厂商、多场景并行发展的格局。海外厂商起步较早，OpenAI、Google、Microsoft 和 Anthropic 等率先布局办公协同、软件开发、浏览器操作及企业服务，并逐步向通用任务执行与跨应用协作延伸。国内厂商虽然整体启动稍晚，但自 2024 年起明显提速，字节跳动、阿里巴巴、百度、腾讯、华为和 DeepSeek 等围绕代码开发、办公协同、数据分析和企业应用密集推出 Agent 产品。国内外产品的集中发布表明，Agent 竞争已由单一模型能力比拼转向模型、工具链、应用入口与产业生态的综合较量。

更多免费行业报告 并购家 www.ipoipo.net

图表 9：国内外主流厂商 Agent 产品发布



资料来源：金圆统一证券研究所

2.5 技术演进主线三：百万级上下文与架构创新，可用性边界大幅拓宽

百万级上下文正在从单纯的模型参数指标，转向实际的产品能力。海外方面，OpenAI GPT-4o、Google Gemini 1.5 Pro、xAI Grok 2 等模型的上下文窗口均已达到约 1M token。国内跟进速度同样迅猛，月之暗面 Kimi、DeepSeek-V3 将上下文扩展至 1M，阿里 Qwen2.5 支持 262K，智谱 GLM-4 达到 128K。相较 32K 时代，百万级上下文使单次调用可承载的信息量提升约 30 倍，完整代码库、数百页技术文档、多轮任务轨迹及复杂 Agent 执行过程，得以在更少的切分与压缩操作下被统一处理。DeepSeek-V3 甚至将 1M 上下文作为官方服务的标准配置，并通过 token 压缩与稀疏注意力机制降低长上下文带来的计算和显存开销。

图表 10：主流大模型上下文窗口对比

市场	厂商	代表模型	上下文窗口	32K 倍数	典型长上下文场景
海外	OpenAI	GPT-5.6 Sol	1.1M	32.8×	长代码库、复杂专业任务、Agent
	Google	Gemini 3.1 Pro	1.0M	32.8×	长文档、代码库、多模态资料分析
	xAI	Grok 4.20	1.0M	31.3×	长文本分析、Agent、工具调用
	Anthropic	Claude Sonnet 4	1.0M	31.3×	长文档、代码库、复杂工作流
国内	月之暗面	Kimi K3	1.0M	31.3×	长程编程、Agent、知识任务
	DeepSeek	DeepSeek-V4 Pro	1.0M	31.3×	Agent、代码、长文档处理
	阿里	Qwen3-Max	262K	8.2×	Agent、工具调用、复杂推理
	智谱	GLM-4.5	128K	4.0×	软件工程、Agent、网页浏览

资料来源：51Token、KIMI、金圆统一证券研究所

长上下文的商业价值，在于推动 AI 应用的任务组织方式从多轮拆分走向一次成型。在 32K 上下文时代，模型理解大型代码仓库往往需要按文件或模块切分任务，通过多轮调用并配合复杂的检索增强工程来维持信息连续性，不仅工程成本高，效果也容易受检索遗漏和上下文割裂影响。而当上下文扩展至 128K、256K 乃至 1M 后，代码仓库、历史交互、任务状态等信息可以被更完整地一次性注入，AI 应用由回答局部问题进一步跃升为理解并执行完整任务。值得注意的是，单次输入规模的成倍扩张往往会显著抬高整体 token 用量。这说明长上下文的核心价值是将原本需要多轮交互、复杂编排才能完成的高难度任务压缩进更短的交互链路，在提升产品可用性和任务完成质

量的同时，也显著提升单次调用的价值密度，并由此催生对高价值 token 的新增需求。以完整代码仓库理解为例，百万级上下文可以大幅降低任务切分、检索、状态管理等工程复杂度，而单次调用的输入 token 却可能较过去提升一个甚至两个数量级。对模型厂商而言，这是客单价和 token 消耗的自然提升。对客户而言，则是工程投入、开发周期和系统复杂度的显著下降。供需双方同时获益，正是高价值 token 商业逻辑得以成立的重要场景基础。

2.6 国产大模型的技术路线：从能力追赶到效率、生态与场景适配

国产模型进入前沿能力区间，稀疏架构与全链路工程优化共同提升单位算力的有效产出。以 DeepSeek V4Pro、GLM-5.2、Kimi K3 等为代表的国产模型，已进入 Artificial Analysis Intelligence Index（AI 智能指数）的前沿比较区间。国产模型的竞争重点在于每单位参数产生更多有效能力，DeepSeek V4Pro 采用约 1.6T 总参数、49B 激活参数的稀疏架构，GLM-5.2 则采用约 744B 总参数、40B 激活参数的设计，推理时只调用与当前任务相关的专家模块。本质类似于把大型工厂拆分成多个专业车间，订单到来时只启动所需车间，从而减少无效计算。与此同时，低精度量化、KVCache、多级缓存、通信调度优化以及稀疏注意力等技术，进一步压缩显存占用和重复计算，使架构优势能够转化为真实的推理效率。

图表 11：智能指数与加权平均输入价格对比



模型名称	图例	公司	智能指数得分	加权平均输入价格（美元/每百万 Token）
Claude Fable 5.1		Anthropic	53.4	1.5
通义千问 3.8 Max		通义千问	53.4	0.54
GPT-6 Astra		OpenAI	52.8	3.62
Claude Opus 5		Anthropic	50.7	1.25
Claude Fable 5		Anthropic	49.7	3.49
GPT-5.6 Sol		OpenAI	47.1	0.69
通义千问 3.8 Max(0902)		通义千问	45.4	0.51
GLM5.3		智谱	44.9	0.37
Grok 4.6(high)		x-ai	44.4	0.91
KimiK3(max)		月之暗面	43.8	0.64

灰底模型位于帕累托曲线

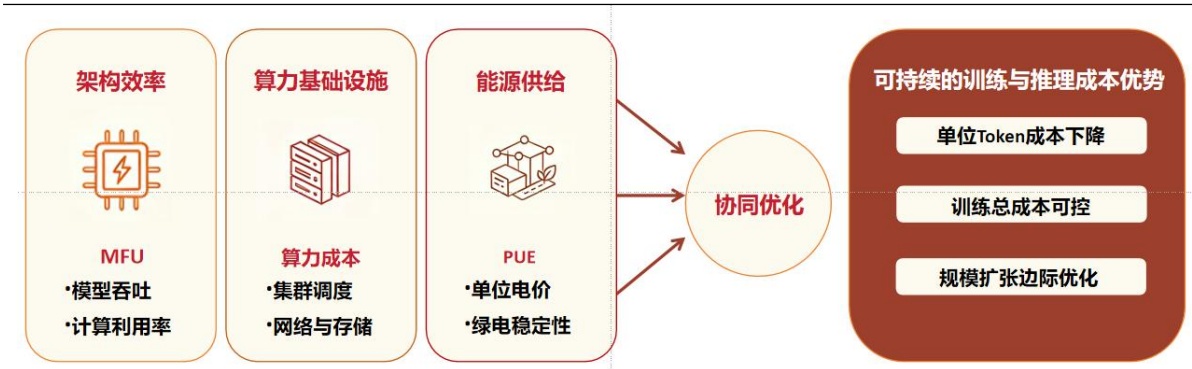
数据来源：ArtificialAnalysis，OpenRouter-Benchmarks，金圆统一证券研究所

架构效率只有与算力基础设施和能源供给结合，才能转化为可持续的训练与推理成本优势。大模型成本并不只由参数规模决定，还取决于训练集群利用率、显存访问效率、通信开销、推理吞吐和单位电力成本等因素。国产团队通过原生低精度训练、算子融合、通信流水线优化以及缓存复用，降低了每个有效 Token 所需的计算资源。国内算力中心扩张和低成本电力供给，也为模型的大规模训练与长期部署提供了基础条件。由此形成的成本优势是从芯片利用率、模型结构到数

更多免费行业报告 并购家 www.ipoipo.net

据中心运营效率的系统性改善。DeepSeek V4Pro 官方 API 的输出价格为每百万 Token 0.87 美元，GLM-5.2 也被 Artificial Analysis 评价为同时处于智能水平与任务成本的较优区间，说明国产模型正在把低成本转化为可量化的产品能力。

图表 12：架构效率与算力基础设施和能源协同优化



数据来源：Artificial Analysis，金圆统一证券研究所

国产大模型正凭借着开放权重与自主部署，加速扩大生态版图。GLM-5.2 以 MIT 许可、不设地域限制的方式开源。DeepSeekV4 系列坚持权重开放。Kimi 从 K2 到 K3 持续释出完整权重。Qwen 在 2026 年首次把 Max 级旗舰以开放权重形式放出。MiniMax 的 M2.1、H3 同样走开放路线。当权重可下载、许可允许商用，试用、迁移、微调、蒸馏的成本被压到接近于零，生态扩散不再依赖厂商的销售队伍，而由全球开发者自发完成。竞争随之从 API 订阅延伸到部署、二次开发与生态分发：

1. **闭源模型。**以 API 调用和企业订阅锁定客户，优势在于模型迭代快、服务质量可控、商业边界清晰，适合对稳定性和持续技术支持要求高且数据可以外送的场景。代价是客户被锁定在厂商的服务体系内，无法自主掌控模型与数据流向。

2. **开源模型。**通过权重开放、接口兼容与宽松授权，把试用和迁移成本压到最低。对开发者和中小企业，这意味着可以围绕具体行业做微调、蒸馏和工具链改造，长出更强的任务适配能力。对制造、能源、金融、政务等数据敏感行业，本地化部署则让核心数据不必离开企业边界，满足合规与数据主权要求。

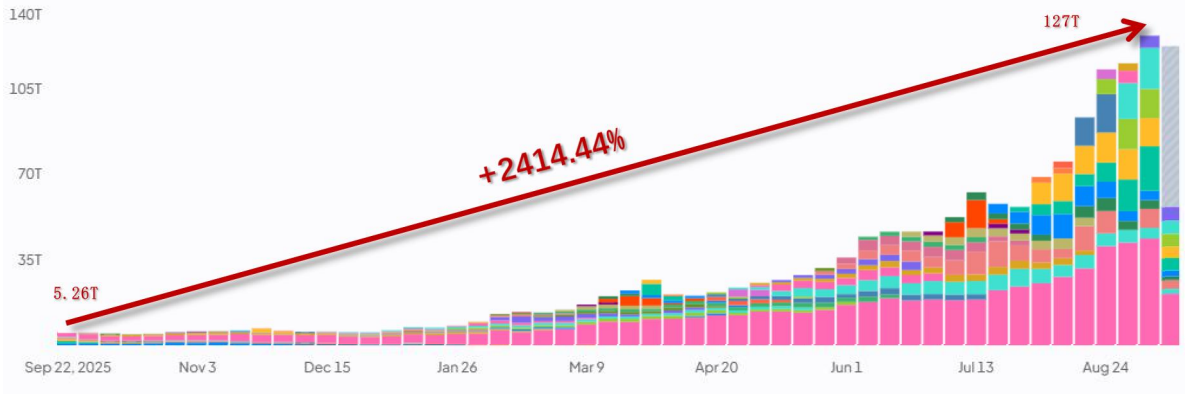
国产大模型的快速追赶，本质是换了一个更容易规模化复制的做事方式。海外闭源模型在复杂推理、前沿科研和高端企业服务上仍有优势，但国产模型依靠效率、成本与开放性的组合，已经在对部署形态和综合成本敏感的大规模产业场景中建立起替代能力。

三、商业闭环：Coding 率先兑现，Agent 打开第二增长曲线

3.1Token 消耗从对话次数转向任务时长与执行深度

Token 已成为观察 AI 应用渗透程度的重要高频指标。与用户数量、模型调用次数相比，Token 能进一步反映单次交互的信息处理规模，因此更适合衡量推理型 AI 应用的使用强度。全球主流大模型聚合网关 OpenRouter 数据显示，平台周度 Token 调用量已从去年 9 月的约 5T 快速增长至今年 9 月的超 127T，年增长约 24 倍，表明 AI 应用的 Token 消耗已进入较高量级。其中，3 月出现明显跃升，与月初 OpenClaw 等龙虾 Agent 应用集中出圈有关，自此 Token 消耗持续加速。从调用结构看，OpenAI 稳居模型使用份额第一，腾讯、智谱、DeepSeek、小米等多家中国厂商亦跻身榜单，国产模型在海外分发渠道的渗透力持续提升，构成总量高增、国产崛起的鲜明特征。

图表 13：OpenRouter 月度大模型 Token 使用量



20260909 平台大模型厂商 Token 使用量

公司名称	模型名称	Tokens	WoW	公司名称	模型名称	Tokens	WoW
OpenAI	GPT-5.6 Luna	17.3T	+34%	小米	MiMo-V2.5	7.83T	+87%
腾讯	Hy4 preview	12.0T	-39%	腾讯	Hy3	4.47T	+25%
智谱	GLM5.3 Flash	11.3T	-9%	DeepSeek	DeepSeek V4 Flash0423	4.24T	-12%
DeepSeek	DeepSeek V4 Flash0731	11.2T	-9%	Nvidia	Nemotron 3 Ultra(free)	3.49T	-4%
DeepSeek	DeepSeek V4.1 Flash	10.3T	-	智谱	GLM5.3	2.46T	-25%

数据来源：OpenRouter-Top Models、金圆统一证券研究所

Token 增长正在经历结构性转变。早期生成式 AI 主要用于短文本问答、内容生成和信息检索，单次任务的上下文规模和执行步骤相对有限。随着推理模型、Coding Agent 和通用 Agent 快速发展，AI 使用逐渐转向包含任务理解、规划、检索、工具调用、结果验证和多轮修正的复杂工作流，单个任务所需的模型调用次数、上下文长度和输出规模随之增加，Token 消耗也由简单问答转向长上下文、多轮交互和持续执行。

海外市场的使用深度变化已初见端倪。OpenRouter 平台的实测数据显示，大模型单次交互的复杂程度显著加深，同时期平均序列长度从不足 2,000tokens 攀升至 5,000tokens 以上，增幅超过 2.5 倍。其中，编程与 Agent 类任务作为复杂度提升的核心驱动力，其序列长度的增长尤为迅猛。AI 商业化的增长动力正在从简单的调用次数扩展至更高频、更复杂、更具实际产出的任务执行，Token 消耗也因此逐渐成为衡量 AI 应用使用深度和商业价值的重要指标。

更多免费行业报告 并购家 www.ipoipo.net

图表 14：AI 任务复杂度提升与 Token 需求变化

使用阶段	典型任务	执行方式	Token 需求特征	代表应用
基础问答	信息检索、简单问答	单轮/少轮交互	单次消耗较低	Chatbot
内容生成	文案、摘要、翻译	多轮生成与修改	上下文长度增加	AIGC
推理任务	数学、科研、复杂分析	长链推理	推理 Token 增加	Reasoning Model
Coding Agent	编码、测试、Debug 部署	多步骤工具调用	单任务 Token 显著增加	Claude Code，Codex
通用 Agent	研究、办公、销售	跨系统连续执行	长上下文+多工具调用	企业 Agent

资料来源：51Token、金圆统一证券研究所

3.2 Coding：从代码补全升级为软件工程执行层

Coding 是当前最先形成完整商业闭环的 AI 应用场景之一。软件开发具有明确的任务边界和可验证的结果，编译、测试、代码审查和持续集成等环节能够持续反馈 AI 生成结果，使 AI 的工作质量和效率具有较强的可观测性。同时，软件开发本身具有较高的时间价值，代码编写、缺陷修复、系统迁移等任务可以直接对应开发者的人力成本。因此，AI 调用产生的 Token 消耗能够较直接地嵌入具体的软件工程任务，并通过代码提交、测试结果和部署状态转化为可衡量的生产效率。

Coding 工具的演进可以划分为三个阶段：

1.代码补全阶段：以 GitHub Copilot 初代、Codex 为代表，核心能力是单行/片段代码补全、函数生成，仅作为 IDE 插件，被动响应开发者输入，不具备自主任务规划、调用终端与执行测试能力。模型仅做文本生成，几乎没有多轮工具调用，单次会话 Token 消耗很低。

2.代码理解与局部修改阶段：工具开始读取完整代码库，支持跨文件检索、重构、简单 bug 修复。可以基于整个项目上下文给出修改方案，但任务仍由人主导拆分，AI 不自主循环验证。

3.Coding Agent 阶段：以 Devin、Cursor、Claude Code、字节 TRAE 为标志性产品，正式进入 Agent 时代。Coding Agent 接收高层开发需求，自主完成任务拆解、读取全量代码库、调用终端运行代码、执行测试、定位报错、多轮迭代修复，最终输出可直接提交的 PR（代码合并请求）。Cursor 官方数据显示，Coding Agent 单次会话平均工具调用次数持续上行，对应 Token 消耗从一次性文本生成转向长上下文、多轮循环执行。

图表 15：主流 Coding Agent 产品竞争力对比

公司	模型	软件工程	终端基准	前沿软件工程	程序	代码	长周期软件工程
OpenAI	GPT-5.6 Sol	73	88.8	71.3	77.6	64.8	39
OpenAI	GPT-5.5	67	83.4	64.9	70.8	69	14
Anthropic	Fable 5	70	84.6	86.6	76.8	76.9	35
Anthropic	Opus-4.8	59	84.6	66.7	71.9	71.7	40
智谱	GLM-5.2	46.2	82.7	67.3	63.7	64.2	13
月之暗面	Kimi K3	67.5	88.3	81.2	77.8	72.9	42

资料来源：KIMI、财联社、金圆统一证券研究所

AI 编程的市场空间需从存量替代与增量创造两个维度综合评估。

存量市场：构成赛道价值的基石，其核心逻辑是对现有专业开发者工作流的替代与升级。以全球约 2870 万专业开发者为基数，假设中性 ARPU 为 400 美元/年，理论总潜在市场（TAM）约为 115 亿美元。参考当前 GitHub Copilot 仅约 5%的渗透率，若中期渗透率提升至 40%，即可对应约 46 亿美元的可服务市场（SAM）。这是一个具备高付费意愿、高黏性的稳固存量基本盘。

增量市场：赛道想象力的核心来源，其本质是通过“代码平权”赋能数亿知识工作者。我们将具备软件创造需求但缺乏专业技能的群体定义为“泛开发者”，其潜在规模至少是专业开发者

的 5-10 倍，即 1.5 亿至 3 亿人。假设该群体 ARPU 为 100-200 美元，对应的增量工具市场潜在规模高达 150 亿美元。

更重要的是，这仅仅是工具层的显性收入。由数亿新晋创造者涌入所催生的应用与服务生态，其衍生的经济价值将呈几何级数增长，这才是 AI 编程赛道真正的价值蓝海。

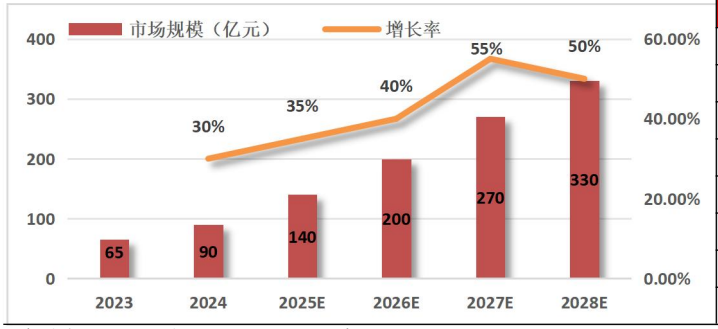
图表：2025E-2030E AI 编程市场规模测算（TAM/SAM）

市场类型	2025E	2027E	2030E	核心逻辑
专业开发者市场	18 亿美元	51 亿美元	100 亿美元	渗透率与 ARPU 双升
泛开发者市场	8 亿美元	45 亿美元	150 亿美元	用户基数与付费习惯养成
合计市场规模	26 亿美元	96 亿美元	250 亿美元	存量替代+增量创造

资料来源：慧博智能投研，金圆统一证券研究所

全球 AI 编程市场规模预计在 2031 年突破 244.6 亿美元，年复合增长率高达 24.3%。中国市场的增长势头更为迅猛，AI 代码生成市场规模在 2023 年为 65 亿元人民币，预计至 2028 年将飙升至 330 亿元，年复合增长率达 38.4%。从行业渗透来看，互联网与游戏行业是 AI 编程工具落地最快的领域，渗透率分别高达 90%和 80%。这主要得益于这两个行业密集、规范且高度项目化的开发特性，与 AI 代码生成的技术逻辑高度契合。此外，人工智能行业渗透率同样达到 80%，而金融渗透率为 60%，能源为 55%，医疗设备为 50%，电信为 30%，政务为 15%，整体呈现出阶梯式的渗透格局。

图表 17：2023—2028 年中国 AI 代码生成市场规模及增速



行业	渗透率	驱动因素
互联网	90%	开发密集、规范、项目制
游戏	80%	代码量大、迭代快
人工智能	80%	技术接受度高
金融	60%	合规与安全要求高
能源	55%	数字化转型需求
医疗设备	50%	嵌入式开发需求
电信	30%	基础设施复杂
政务	15%	安全与合规优先

资料来源：亿欧智库、金圆统一证券研究所

Coding 产品正在从 IDE 内的代码辅助插件，向自主执行的软件工程 Agent 演进。以字节跳动 TRAE 为典型案例，其产品定位从最初的 IDE 内 AI 编程助手逐步升级为能够理解完整代码仓库、自主规划开发任务、调用终端执行编译测试的 Coding Agent，并在此基础上推出 TRAE SOLO 与 TRAEWork，将 Agent 能力从软件开发延伸至产品设计、数据分析及市场研究等通用专业工作领域。

Coding Agent 的核心竞争力，已经脱离单点代码生成能力，转向理解、执行、验证的完整软件工程链路。具体而言：模型推理能力决定任务复杂度的处理上限。代码库长上下文窗口决定可理解的项目规模。终端、沙箱、浏览器等工具定义执行边界。单元测试、代码评审及版本控制系统构成结果的自验证闭环。在企业环境中，权限隔离、审计日志、私有代码库安全管控及私有化部署能力进一步成为企业采购的关键约束条件。

3.3 Agent：从工具调用走向可治理的数字劳动力

Coding Agent 是通用 Agent 能力的先行试验场，其成功验证了 Agent 在复杂工程场景中的自主执行能力。而通用企业 Agent 则代表着 Agent 能力从研发场景向全公司业务流程规模化延伸的战略

更多免费行业报告 并购家 www.ipoipo.net

方向。**Agent 产品正在形成不同的商业化路径，但共同方向是从工具调用走向长期任务执行和企业工作流嵌入。**

1. OpenAI：从通用 Agent 走向企业工作流执行。

OpenAI 通过 Workspace Agents 让企业用户创建能够调用不同工具、执行多步骤任务并共享给团队的 Agent，并进一步加入管理员权限和使用监控等企业治理能力。其后续计费也开始引入 Credits，将 Agent 任务与实际 Token 消耗联系起来，推动 Agent 从个人工具向企业工作流基础设施演进。

2. Google：从模型能力延伸至 Agent 基础设施。

Google 的重点在于提供 Agent 的构建、部署、运行和治理环境，将模型、企业数据、安全和运行资源纳入统一体系。其商业模式也逐步覆盖模型 Token 以及计算、存储、工具调用等运行资源，体现出从模型订阅向 Agent 基础设施收费延伸的趋势。

3. Microsoft：依托办公生态切入企业 Agent。

Microsoft 将 Agent 直接嵌入 Microsoft 365，使其能够调用 Teams、Outlook、SharePoint 等企业数据，并在 Word、Excel、PowerPoint 等应用中执行多步骤任务。依托成熟的企业软件入口和数据体系，Microsoft 的优势在于将 Agent 直接嵌入现有办公流程，并通过 Copilot Credits 进一步实现用量计费。

4. Anthropic：聚焦知识工作与长期任务执行。

Anthropic 通过 Claude Cowork、Projects、Skills 和 Plugins 等产品，将 Agent 能力从代码开发延伸至文件处理、研究分析和其他知识工作，并允许用户将工作流程和专业知识固化到 Agent 中。其产品路径更强调长任务执行和专业工作流自动化。

国内厂商也正在沿着企业工作流嵌入的方向推进。腾讯推出的 WorkBuddy 将办公 Agent 与 Coding Agent 纳入统一的 Buddy AI 体系，支持文档处理、数据分析、深度研究、邮件处理和多任务执行，并进一步通过腾讯云向企业提供 Agent 开发与部署能力。阿里则依托 Qwen 模型、钉钉和阿里云推进企业 Agent 落地，QwenWork 进一步向桌面、云端和企业协作场景延伸。字节跳动则从 TRAE Coding Agent 向 TRAE Work 扩展，将 Agent 的任务范围拓展至产品设计、数据分析、研究和市场等专业工作。由此看，国内 Agent 产品也正在从单一问答工具向能够连接企业数据、调用业务工具并持续执行任务的工作平台演进。

图表 18：主流 Agent 平台竞争力对比

公司	模型	GDPval- AA	AA- Briefcase	自动化任务	职业场景	电子表格	网页交互
Anthropic	Fable 5	1760	1583	29.1	57.4	34.7	88
OpenAI	GPT-5.6 Sol	1748	1495	29.7	46.5	32.4	90.4
Anthropic	Opus-4.8	1600	1354	27.2	48.4	31.6	84.3
OpenAI	GPT-5.5	1494	1158	22.7	38.3	29.1	84.4
智谱	GLM-5.2	1514	1260	12.9	43.4	28.1	-
月之暗面	Kimi K3	1668	1548	30.8	52.9	34.8	91.2

资料来源：财联社、金圆统一证券研究所

Agent 商业化的关键变化是 AI 开始从单项工具进入能够持续运行的企业工作流程。与单轮问答相比，Agent 能够围绕一个目标连续完成规划、信息获取、工具调用、结果检查和任务修正，因

此其价值不再局限于一次输出，而是体现为对完整工作流程的持续接管。随着权限管理、数据连接、状态管理和审计能力逐步成熟，Agent 能够进入研究、办公、销售、财务和客户服务等更复杂的知识工作场景，AI 的可服务市场由单点任务进一步扩展至连续业务流程。

3.4 入口决定获客，工作流决定留存，数据决定迭代

AI 应用层的竞争优势正在从模型能力向入口、工作流和数据延伸。模型能力决定 Agent 能够处理的任务范围，但用户是否持续使用，更取决于 Agent 能否进入已有的软件环境、调用必要的数据并完成实际工作。相比独立 AI 工具，嵌入 ChatGPT、Microsoft 365、Google Workspace、钉钉等既有平台的 Agent 能够直接承接用户和工作流，因此更容易形成稳定的使用入口。

入口、工作流与数据正在共同构成应用层壁垒。入口决定用户触达，工作流决定使用深度，企业代码库、知识库和业务数据则决定 Agent 能否提供持续的个性化服务。Coding Agent 已经在软件开发中验证了长任务执行的价值，通用 Agent 进一步将这一模式扩展至办公、研究、数据分析和企业运营等场景。未来 AI 应用的竞争重点将从单纯的模型能力转向谁能更深入地嵌入用户入口、工作流程和数据体系。

图表 19：主流 AI Agent 产品能力对比



资料来源：QuestMobile、金圆统一证券研究所

数据正在成为 Agent 持续迭代和形成竞争壁垒的关键。与传统聊天记录相比，Agent 能够沉淀从任务目标、执行路径、工具调用到最终结果的完整任务轨迹，企业可以据此识别失败环节、优化工作流并沉淀组织知识。国内外厂商因此都在强化 Agent 的权限管理、知识库、任务记录和使用分析能力，并将这些能力嵌入办公软件、企业 SaaS 和云平台。虽然海外厂商更多依托办公软件与企业 SaaS 生态，国内厂商更强调超级应用、协作平台与云基础设施联动，但核心目标都是让 Agent 进入真实工作场景并持续获取任务与数据。

AI 应用层真正形成壁垒的关键是入口、工作流和数据能够形成相互强化的使用闭环。入口决定任务能否持续进入系统，工作流决定 AI 能否从单点功能深入核心流程，执行数据则决定系统能

更多免费行业报告 并购家 www.ipoipo.net

否持续优化任务完成效果。当 Agent 不断获得真实任务并积累执行轨迹后，厂商能够据此优化提示、工具调用、模型路由和流程设计，进一步提高任务成功率和用户黏性。

3.5 当前主流商业闭环：从卖席位到卖可持续完成任务的能力

当前 AI 应用的商业闭环正在从单一席位销售向模型、软件、生态和企业服务的多元价值承接模式演进。将底层模型能力嵌入不同的价值承接体系，从而形成不同的收入来源：

第一类是模型/API 直供型，通过 API、订阅和用量服务直接变现模型能力；

第二类是应用路由型，通过模型聚合、上下文管理和工作流整合将底层模型能力封装为软件产品；

第三类是生态内生型，依托云、办公软件和超级应用，将 AI 嵌入既有用户体系并通过席位、云服务及其他业务交叉变现；

第四类是企业交付型，通过部署、实施、授权和持续服务满足金融、制造、政务等行业对数据安全和系统集成的要求。

AI 商业模式的差异最终取决于应用承担的任务类型和客户价值形成方式。低复杂度、高频任务更适合订阅与标准化用量模式，中高复杂度任务则更适合 Credits、按任务或按工作流等方式进行价值回收，而企业级场景还需要叠加部署、服务和治理收入。随着 AI 从单轮问答进入持续任务执行，商业模式的核心也从单纯出售模型访问权转向持续提供可交付的任务能力。

图表 20：大模型商业闭环模式

模型类型	典型代表	收入结构	成本/风险承担	核心护城河
模型/API直接型	OpenAI, Anthropic, DeepSeek、阿里云、百度智能云	API订阅+Token容量/服务等级，按量付费为主	厂商承担推理成本与价格竞争压力，需持续优化算力效率	极致的模型能力、庞大的开发者生态、海量数据与算力壁垒
应用/模型路由型	Cursor、各类Coding Agent、垂直场景智能应用	订阅制底座+按量资源包，或增值功能付费模式	应用方承担上游模型成本，通过路由优化降低成本	高频工作流入口、智能路由算法、极致的用户体验
生态内生型	Google、Microsoft、阿里、腾讯等生态巨头	席位订阅+云资源消耗，叠加企业软件与生态收入	利用主业现金流交叉补贴，降低AI业务的盈利压力	超级流量入口、海量企业客户基础、数据闭环生态
私有化/行业交付型	政企、金融、制造领域的AI解决方案服务商	项目制交付+软件授权费，叠加运维与算力服务	客户承担定制实施成本，部署与运维成本共担	合规资质、本地化部署能力、行业Know-how与服务
任务/结果计量型	Codex、WorkBuddy及各类任务型Agent产品	基于任务完成量的Credits付费，或按结果计价	按实际任务消耗转移成本，风险共担，结果导向	精准的任务执行能力、结果可验证性、工作流深度嵌入

资料来源：QuestMobile、沙利文、金圆统一证券研究所

AI 商业闭环最终需要同时满足需求增长、价值验证和收入回收三个条件。Coding 率先具备明确的任务边界和结果验证机制，因此能够较早证明 AI 带来的效率提升。Agent 进一步将这一模式扩展至办公、研究、销售、财务和客户服务等知识工作，并通过持续执行提高 AI 在企业流程中的渗透率。当应用能够稳定完成任务、客户能够持续感知价值、厂商能够覆盖计算成本并获得合理利润时，模型能力、Token 消耗、应用使用和收入增长才能形成真正意义上的商业闭环。

四、国产大模型竞争由技术投入转向商业兑现

4.1 商业化进程推动估值体系持续调整

2025 年以来，大模型行业已完成从能力竞赛向商业化验证的阶段性切换。2025 年上半年，Coding 率先形成清晰的付费场景，原因在于软件开发任务的边界、结果和成本相对容易量化，模型能力可以直接映射至开发效率和人力成本。2025 年下半年，模型调用代码和工具的能力进一步提升，传统 SaaS 和技术工具中的标准化功能开始面临重新定价，资本市场同步重估软件公司的产品壁垒。进入 2026 年，Agent 由单轮问答向持续任务执行演进，企业内部的数据权限、信息安全、系统集成和 ROI 约束逐步成为落地条件。到 2026 年下半年，头部厂商进一步向 Harness、工具生态和开发者入口延伸，行业竞争开始由单一模型能力扩展至完整的 Agent 技术栈。

图表 21：大模型商业化的三阶段演进

阶段		变化
2025H2	软件资产进入重估	标准化功能承压，协作、数据、行业 Know-how 和业务流程的重要性上升。
2026H1	Agent 进入企业工作流	企业级落地开始关注权限、安全、集成和 ROI，本地化部署及实施能力得到重视。
2026H2	竞争向执行层延伸	Harness、工具生态和真实任务轨迹的重要性提升，开发者入口成为新增竞争要素。

资料来源：腾讯云、金圆统一证券研究所

资本市场对大模型的估值锚正由技术能力向真实使用和商业兑现迁移。早期市场主要围绕模型参数、Benchmark 和融资能力定价，Coding 商业化后，调用量、付费用户和单位任务 ROI 开始进入估值框架。软件资产重估阶段进一步验证了场景和数据的重要性，同一模型能力在不同产品中的变现效率出现明显分化。Agent 进入企业工作流后，商业化关注点继续向企业客户、续费、部署方式和 Token 成本迁移。专家交流中也反复提到，通用 Agent 进入企业后需要面对成本与信息安全约束，海外 FDE 和国内成熟的软件实施体系因此成为重要的落地方式。市场对大模型的判断正在由技术可行性逐步转向收入质量和商业模式的可持续性。

图表 22：大模型关注重点的迁移

市场关注	核心指标与要素
模型能力	参数规模，Benchmark，推理与 Coding 能力，训练与推理成本
真实使用	Token 调用量，活跃用户，开发者覆盖，任务复杂度
商业兑现	ARR，API 定价，企业客户，毛利率与续费
平台壁垒	生态入口，Harness 与工具链，工作流渗透，真实任务数据

资料来源：腾讯云、金圆统一证券研究所

4.2 国产大模型厂商资源禀赋决定差异化路径

国内大模型竞争已由百模并立进入头部集中与定位分化并行阶段。大厂系模型拥有超级应用、云业务和企业客户基础，模型业务短期盈利并非其唯一目标，核心价值更多体现为对原有生态的效率提升和交叉变现。DeepSeek、智谱、Kimi、MiniMax 等独立模型厂商则需要在模型迭代、用户增长和收入扩张之间形成更快的正循环，对融资能力和商业化效率的要求更高。部分中部厂商开始向医疗、端侧、多模态等细分方向集中资源。

模型能力仍是竞争基石，而生态入口、成本效率与企业交付正决定长期分化，长期竞争力可归纳为算力获取、数据飞轮、生态入口、性价比与本地化部署五大方面。大厂在生态入口与客户触达上占优，独立厂商更依赖单位算力产出、API 定价能力与开发者生态，本地化部署则直接影响

更多免费行业报告 并购家 www.ipoipo.net

政企场景变现效率。值得注意的是，国内 C 端调用量与 B 端 ARR 排序并不完全一致，表明流量规模无法单独定义模型公司商业价值，随着行业进入商业化深水区，企业客户质量、定价能力与交付效率的重要性将持续提升。

图表 23：国产大模型厂商竞争格局

厂商类型	代表厂商	主要资源禀赋	商业化路径	核心观察指标
大厂生态型	腾讯、阿里字节、百度	超级入口、云业务、企业客户、数据生态	模型嵌入原有业务体系，通过云办公、电商、广告等方式交叉变现	AI 渗透率、云消费、用户时长、企业客户使用深度
独立通用型	DeepSeek	成本效率、开源生态、开发者影响力	API、开源生态及 Agent 执行层延伸	Token 调用、开发者覆盖、API 定价、生态扩张
	智谱	企业交付、本地化部署、政企客户	API、私有化部署、企业级服务	ARR、企业客户、续费、本地化部署收入
	Kimi	长上下文、开源模型、产品用户基础	C 端产品、API 及开发者生态	Token、付费转化、API 价格、开发者活跃度
多模态型	MiniMax	语音、音乐、视频等多模态能力	消费产品、海外市场及 API	海外收入、付费率、多模态调用量
垂直及端侧型	百川、阶跃	行业数据、端侧能力、特定场景适配	医疗、终端、产业合作等垂直方向	行业订单、项目落地、场景渗透率

资料来源：沙利文、金圆统一证券研究所

4.3 全球前沿模型融资潮下的估值逻辑迁移

全球前沿模型融资规模继续快速抬升，OpenAI 和 Anthropic 已将一级市场估值推升至接近万亿美元量级。OpenAI 于 2026 年 2 月宣布新一轮 1100 亿美元投资，投前估值 7300 亿美元。3 月底最终完成 1220 亿美元融资，投后估值达到 8520 亿美元，其中 Amazon、NVIDIA 和 SoftBank 分别承诺 500 亿、300 亿和 300 亿美元。Anthropic 在 5 月完成 650 亿美元 Series H 融资，投后估值达到 9650 亿美元，较 2 月约 3800 亿美元估值继续大幅提升。两家公司融资规模与估值水平已经明显脱离传统软件公司的定价框架，背后对应的是前沿模型训练、推理算力和全球企业市场扩张所需的持续资本投入。

图表 24：OpenAI 与 Anthropic 最新公开融资及估值情况

公司	时间	融资进展	估值（投后）	ARR
OpenAI	2026 年 3 月	完成 1220 亿美元融资	约 8520 亿美元	250 亿美元
Anthropic	2026 年 5 月	完成 650 亿美元 Series H 融资	约 9650 亿美元	650 亿美元

资料来源：AI Wiki、金圆统一证券研究所

国内融资正在从百模竞赛时期的广泛撒网转向头部集中，DeepSeek、Kimi、智谱等公司已经形成数十亿美元级融资能力。DeepSeek 于 2026 年 6 月完成约 74 亿美元融资，投后估值超过 500 亿美元，后续新一轮融资对应估值可能达到约 750 亿美元，并已启动科创板 IPO 筹备。Kimi 的估值提升速度更快，2025 年底估值约 43 亿美元，2026 年 5 月完成约 20 亿美元融资后估值超过 200 亿美元，7 月底再完成约 35 亿美元融资并将估值推升至 350 亿美元。智谱和 MiniMax 则率先进入公开市场，智谱 1 月港股 IPO 募资约 43.5 亿港元，7 月又完成约 40 亿美元增发，9 月进一步启动约 50 亿美元股权及可转债融资。MiniMax 1 月 IPO 募资约 6.2 亿美元。阶跃星辰 1 月完成超过 50 亿元人民币 B+轮融资，并曾以约 100 亿美元估值推进上市筹备。融资能力已成为模型迭代、算力采购和人才竞争的重要壁垒。

图表 25：大模型厂商投融资活跃，头部企业估值跃升

公司	时间	融资金额	投后估值
DeepSeek	2026 年 4 月	首轮外部融资 500 亿元人民币，创始人梁文锋领衔，腾讯、宁德时代、京东、网易、IDG 等参投	超 500 亿美元

	2026 年 8 月	第二轮融资计划募资 500 亿元人民币，投前估值约 5000 亿元，较首轮提升约 43%	约 775 亿美元
月之暗面 Kimi	2025 年 12 月	C 轮融资 5 亿美元，IDG 资本领投 1.5 亿美元，阿里、腾讯及美团联合创始人王慧文等老股东超额认购	43 亿美元
	2026 年 2 月	C+轮融资超 7 亿美元，九安医疗、五源资本领投，阿里巴巴、腾讯等参投	100 亿美元
	2026 年 5 月	D 轮融资 20 亿美元，美团龙珠领投超 2 亿美元，中国移动、CPE 等参投	超 200 亿美元
	2026 年 7 月	F 轮融资 35 亿美元，投后估值 350 亿美元，K3 发布后融资份额抢手	350 亿美元
智谱	2026 年 1 月	港股 IPO 募资总额约 43.48 亿港元，公开发行价 116.2 港元/股	约 65.5 亿美元
	2026 年 7 月	上市后配售：H 股净募资约 313.75 亿港元，按 1588 港元/股向 6 名投资者配售	约 950 亿美元
	2026 年 9 月	上市后配售+可转债：净募资约 392.74 亿港元（约 20 亿美元配售+约 30 亿美元零息可转债），配售价 714 港元/股	约 431 亿美元
MiniMax	2025 年 2 月	Pre-B 轮融资 6.54 亿美元，阿里巴巴、经纬等参投	26 亿美元
	2025 年 6 月	Pre-B+轮融资 1.23 亿美元	31 亿美元
	2025 年 8 月	Pre-B++轮融资 3.9 亿美元，Aspex、Janchor 等参投，投后估值 42.4 亿美元	42 亿美元
	2026 年 1 月	港股 IPO 募资总额约 55.4 亿港元，基石投资者含 Aspex、Eastspring、Mirae Asset 等 14 家	约 64 亿美元
	2026 年 7 月	上市后配售+可转债：募资约 160 亿港元，配售价 268 港元/股，另发行 65 亿港元可转债	约 116 亿美元
阶跃星辰	2026 年 1 月	B+轮融资超 50 亿元人民币，上海国投先导基金、国寿股权领衔，浦东创投、徐汇资本等国资及华勤技术等产业资本参与，腾讯、启明、五源等老股东跟投	约 28 亿美元
	2026 年 5 月	Pre-IPO 轮融资近 25 亿美元，中兴通讯、华勤技术、腾讯投资等联合注资，投前估值约 40 亿—60 亿美元	约 120 亿美元
百川智能	2024 年 7 月	A 轮融资 50 亿元人民币，阿里巴巴、小米、腾讯、亚投资本、中金及北京/上海人工智能产业投资基金、深创投等参投	约 28 亿美元

资料来源：证券时报网、36Kr、新华财经、金圆统一证券研究所

从估值分布看，国内模型厂商已经明显分层，头部私人公司集中在 350 亿—750 亿美元区间，上市公司则开始接受收入、亏损和融资摊薄的持续定价。DeepSeek 和 Kimi 仍主要依赖一级市场定价，估值受模型能力、融资供给和上市预期共同影响。智谱和 MiniMax 进入港股后，模型发布、收入增长、亏损、融资计划均会直接反映在股价中。以 9 月 11 日数据计算，智谱市值约 3692 亿港元，MiniMax 约 943 亿港元，二者市值差距已经超过三倍。该分化说明资本市场正在对企业级收入、模型能力、算力投入效率和后续融资需求进行差异化定价。对于尚未上市的模型公司，未来估值能否延续快速上行，关键仍在于收入增长能否跟上融资和估值扩张速度。

Kimi 是本轮国产大模型估值重估最典型的案例，估值在不到一年内由 43 亿美元提升至 350 亿美元以上，商业化数据同步快速增长。从 2025 年 12 月底 C 轮约 5 亿美元、投后 43 亿美元起步，月之暗面在 2026 年 1 至 2 月通过同一轮三段交割共计 19 亿美元，把估值从约 100 亿推到约 180 亿，5 月 D 轮约 20 亿美元、投后超 200 亿，6 月 E 轮投前约 315 亿，7 月 Kimi K3 发布后完成超 35 亿美元 F 轮、投后 350 亿，随后 Pre-IPO/G 轮目标投前约 500 亿。经营端同步加速，ARR 由 3 月破 1 亿、4 月超 2 亿、5 月约 2 亿、6 月中旬破 3 亿，到 K3 发布后称 8 月超 10 亿美元。 按 5 月约 200 亿估值对 4 月约 2 亿 ARR 粗略折算约 100 倍 ARR，按 7 月 F 轮 350 亿对 8 月约 10 亿 ARR 约 35 倍、按 500 亿目标估值对 8 月约 10 亿 ARR 约 50 倍。估值绝对数半年翻数倍，但收入从 2 亿到 10 亿的跃升更快。

更多免费行业报告 并购家 www.ipoipo.net

五、投资建议

大模型产业进入商业化兑现阶段，建议围绕生态入口、独立模型和国产化场景三条主线布局。模型能力仍是竞争基础，但单次榜单排名和参数规模对估值的解释力预计下降，市场将更关注 Token 需求、API 定价、ARR、企业客户和毛利率。结合不同类型厂商的资源禀赋，我们建议从三条主线把握投资机会：

1.生态型厂商——核心配置

生态型厂商的优势在于以存量业务现金流对冲大模型投入，AI 变现主要通过云推理、广告提效、企业 Agent 和订阅增值释放，短期业绩弹性未必最高，但客户入口、算力供给与商业化确定性最强。

图表 26：生态型厂商

公司	简介
腾讯控股 00700.HK	依托混元大模型，深度结合 WorkBuddy、企业微信等办公协同工具，将 AI 能力广泛应用于其强势的广告投放、游戏业务及企业级服务中。
阿里巴巴-W 09988.HK	依托通义千问大模型，通过阿里云“百炼”平台对外提供 MaaS 能力，赋能电商、钉钉等自有业务推动公有云算力与 AI 收入增长。
百度集团-SW 09888.HK	拥有国内起步较早的文心大模型，主要通过百度智能云及千帆大模型平台对外提供 API 和定制化服务。
小米集团-W 01810.HK	聚焦端侧 AI 落地，将 AI 能力深度融入其庞大的 AIoT 设备群与小米汽车生态中。
金山办公 688111.SH	依托 WPS AI 及 WPS 365，积极探索订阅+AI 用量的复合计费模式，2026 年上半年订阅收入即达 4.97 亿元。
金蝶国际 00268.HK	深耕企业管理云服务，针对复杂的财税、ERP 等核心业务场景，推出财税领域的 AI Agent。

资料来源：、金圆统一证券研究所
注：仅为统计列示，并非个股推荐

2.独立模型厂商——弹性配置

独立模型厂商处于高投入、高增长阶段，估值对模型排名、Token 成本、融资环境和盈利能力敏感，上行空间大但波动更高。

图表 27：独立模型厂商

分类	公司	简介
通用基座与 MaaS	科大讯飞 002230.SZ	依托星火大模型，在教育、医疗、政务等垂直场景深度落地，是国内 AI 应用落地较早、场景覆盖最广的通用基座厂商之一。
	商汤-W 00020.HK	以日日新大模型为核心，聚焦多模态和视觉技术，广泛服务于智慧城市、工业质检、自动驾驶等领域，具备从算法到算力的全栈能力。
	智谱 02513.HK	依托 GLM 大模型，主攻政企私有化部署与 API 服务，面向 B 端客户输出安全可控的大模型解决方案。
	MiniMax-W 00100.HK	主打多模态模型能力，覆盖视频生成、语音合成等前沿方向，并积极拓展海外 API 业务，在创意内容领域具有差异化优势。
应用驱动型自研	昆仑万维 300418.SZ	基于天工大模型，拥有 SkyReels、Mureka、Opera 浏览器等产品矩阵，海外业务与多模态内容生成弹性较大，是全球布局较深的应用型 AI 企业。
	三六零 601360.SH	以智脑大模型为核心，结合自身网络安全主业与 AI 搜索能力，侧重安全私有化、面向 C 端的 AI 搜索双轮驱动。
行业 NLP/政务私有化	拓尔思 300229.SZ	依托拓天大模型，专注于政务、媒体、金融等领域的大语言模型文本处理与私有化部署，在政务数据治理方面积累深厚。
	云从科技 688343.SH	定位为行业 NLP 与政务私有化解决方案提供商，面向政府及大型企业输出定制化的 AI 落地方案。
	第四范式 06682.HK	企业级决策 AI，提供从数据到决策的端到端解决方案，服务于金融、能源、零售等高价值行业客户。

资料来源：上交所、港交所、深交所、金圆统一证券研究所
注：仅为统计列示，并非个股推荐

3.垂直与本地化模型——结构配置

本地化及垂直模型的市场分散但客户黏性高，项目壁垒来自专有数据、合规资质与系统集成，适合以行业景气度与私有化订单验证为筛选条件。

图表 28：垂直与本地化模型

分类	公司	简介
工业	中控技术 688777.SH	自研流程工业大模型 TPT，已覆盖化工、石化等 13 个细分领域并在中石化、万华化学等头部客户批量落地，商业模式从软件买断转向工业 AI 订阅服务。
	宝信软件 600845.SH	依托宝武体系打造宝联登钢铁行业大模型，落地 170 个应用场景（炼铁、炼钢、轧钢全流程），获 2025 世界 AI 大会 SAIL 奖，同步建设宝之云智算中心向算力运营转型。
	赛意信息 300687.SZ	自研企业级 AI 中台善谋 GPT 并入选 IDC 工业大模型领导者阵营，与华为共建盘古制造大模型，PCB 行业大模型已规模商用。
	索辰科技 688507.SH	国产 CAE 龙头战略变轨至物理 AI，发布“天工·开物”平台，聚焦特种行业、商业航天、具身智能，从软件工具商转型 AI 模型服务商。
金融	恒生电子 600570.SH	自研 LightGPT 金融大模型，构建大模型+光子中间件体系，投研、投顾、风控等智能体已服务超 100 家金融机构，迈向 SaaS 订阅计费。
	同花顺 300033.SZ	自研问财 HithinkGPT 金融垂类大模型，问财升级为原生 AI 对话产品月活超 3800 万，C 端 AI 已开启阶梯收费。
	东方财富 300059.SZ	自研“妙想”金融大模型以万亿级金融语料训练，通过 Skills+MCP 向华为小艺、腾讯 WorkBuddy 等外部 Agent 开放，Skills 下载量份额达 86%，以免费 AI 技能为漏斗反哺证券开户、交易与两融变现。
	长亮科技 300348.SZ	全面推进 AINative 战略，以“N+M”大模型协同打造 AISAgent 平台及 Luban 研发工作台，AI 智能体已覆盖合同审查、反洗钱等六大场景并进入交付阶段。
医疗	卫宁健康 300253.SZ	自研医疗大模型 WiNGPT 以 Copilot 形态内置 WiNEX，落地近 150 家医疗机构、100 余个临床场景，成立 AI 医疗事业部推进订阅收费。
	创业慧康 300451.SZ	基于 BsoftGPT 打造 MedCopilot 智能医生助手，覆盖病历生成、质控等 50+场景，与浙大共建联合实验室并适配昇腾国产算力。
	东华软件 002065.SZ	旗下东华医为推出 iMEDWAY 医疗大模型平台，服务超 1200 家医疗机构。
	科大讯飞 002230.SZ	星火医疗大模型驱动的智医助理覆盖全国 828 个区县、7.7 万家基层机构，推动医疗业务从项目制向 ARR 订阅转型。
政务/企业软件	用友网络 600588.SH	YonGPT+LOM 本体大模型构建企业超级智能体 YonWork，AI 已融入 2500+业务场景，签约超 400 家头部客户，转向订阅制。
	金蝶国际 00268.HK	推出超级入口“小柯”及近 20 款 AI 智能体，苍穹+星瀚订阅收入快速增长，预计 2026 年 AI 套件收入超 10 亿元。
	金山办公 688111.SH	WPSAI 个人端驱动付费用户增至 4825 万，企业端发布 WPSComate 落地 26 个场景，探索订阅+AI 用量商业模式。
	税友股份 603171.SH	发布 AIBM 财税 AI 战略，自研“犀友”垂域大模型推出财税 Agentic 工场。
	石基信息 002153.SZ	推出酒旅 AI 智能体 eGenius，以六大垂直 AI 引擎组成营销决策大脑，从信息化工具商向旅游大消费 AI 服务平台演进。
	深信服 300454.SZ	自研安全 GPT 为国内首个通过网信办双备案的安全垂域大模型，构建全域安全智能中枢，国家攻防演练中 0day 检出率超 95%。
	用友网络 600588.SH	YonGPT+LOM 本体大模型构建企业超级智能体 YonWork，AI 已融入 2500+业务场景，签约超 400 家头部客户，转向订阅制。

资料来源：、金圆统一证券研究所
注：仅为统计列示，并非个股推荐

更多免费行业报告 并购家 www.ipoipo.net

六、风险提示

技术迭代不及预期。模型能力瓶颈可能早于预期到来。尽管业内对 Scaling Law 延续普遍乐观，但架构与数据的边际收益存在不确定性，若能力瓶颈早于预期出现，将影响全产业链的估值逻辑。高质量训练语料的可得性、长轨迹数据的采集合规性，均可能构成硬约束；一旦能力提升停滞，客户为高价 Token 付费的意愿将迅速衰减，提价仍能带动调用量增长的正弹性现象亦将消失。

开源生态冲击。开源模型快速追赶闭源，可能压缩利润空间。开源模型的性价比优势，对闭源厂商的定价权构成持续挑战，尤其在商品化 Token 领域。若开源前沿在 12 至 18 个月内实质性追平闭源能力，占 Token 消耗 27.6%的通用任务将首当其冲，模型厂商向 18%长期 EBIT 利润率收敛的路径也将被显著延长。

出口管制与地缘风险。高端芯片获取受限，国产算力训练侧成熟仍需 3－5 年。地缘因素仍是国产算力完全替代的最大外部约束。管制范围若进一步扩展至先进封装、高带宽存储或电子设计工具环节，将同时抬高国产算力的成本曲线并延后成熟时点，进而影响以国产算力为底仓的投资逻辑。

行业竞争加剧。API 价格战可能持续，部分厂商盈利承压。在收敛完成前，价格竞争可能侵蚀行业整体利润率。当前的涨价潮建立在供给紧张之上，若 2027 年产能集中释放而需求增速回落，价格逻辑可能快速逆转，中部厂商为维持份额被迫降价将拖累行业整体定价水平。

资本开支压力超预期。算力投入持续高企，盈利拐点存在不确定性。若需求增速低于预期，重资产投入将放大盈利波动。分成模式虽提升了算力方的收益弹性，但同时也削弱了其收入的确定性，一旦所绑定模型的调用量不及预期，保底条款之外的超额分成将迅速归零，而折旧与资金成本仍需刚性支付。

传统软件替代风险。通用大模型能力持续下沉，传统软件标准化功能面临被替代风险。这一风险虽指向传统软件厂商，但也提示大模型自身的护城河需持续加固。同样的逻辑亦适用于 AI 应用层：今日依靠工程封装建立的产品优势，可能在下一代基础模型发布后被原生能力覆盖，唯有沉淀了独占数据与深度工作流的应用方能长期存续。

投资评级声明：

分类	评级说明		
股票评级	自报告发布之日起 6 个月内，相对所属板块指数涨幅	买入	股价相对强于基准 20％以上；
		增持	股价相对强于基准 10%—20%以上；
		持有	股价相对基准波动±10%以内；
		卖出	股价相对弱于基准 10%以上；
行业评级	自报告发布之日起 6 个月内，相对沪深 300 指数涨幅	强于大市	相对强于基准 10％以上；
		中性	相对基准波动±10%以内；
		弱于大市	相对弱于基准 10%以上；

分析师声明：

负责本报告全部或参与部分内容的分析师在此声明，本人具有证券投资咨询执业资格，并在中国证券业协会注册登记为证券分析师，以勤勉的职业态度，独立、客观地出具本报告；本报告基于市场公开的合法合规信息进行撰写，本报告所表述的所有观点准确反映了分析师本人的研究观点；本人薪酬的任何组成部分不曾与，不与，也将不会与本报告中的具体分析意见或观点直接或间接相关。

特别声明：

法律许可的情况下，金圆统一证券可能会持有或交易本报告中提及的标的证券，也可能为标的证券提供股债融资、财务顾问、资产证券化和理财投顾服务。因此，投资者应当考虑到金圆统一证券及/或相关人员可能存在影响本报告观点客观性的潜在利益冲突，投资者请勿将本报告视为决策的唯一依据。

免责声明：

金圆统一证券有限公司（以下简称“金圆统一证券”或“本公司”）具有中国证监会批复的证券投资咨询业务资格。本报告由金圆统一证券制作并发布。 本公司及其客户使用。
本公司不会因收到本报告而视其为本公司的当然客户。

本报告是基于金圆统一证券认为可靠的已公开信息编制，但金圆统一证券不保证所载信息的准确性和完整性。本报告所载的意见、评估及预测仅为本报告最初出具日的观点和判断，本报告所指的证券或投资标的的价格、价值及投资收入可能会出现不同程度的波动，涉及证券或投资标的的历史表现不应作为日后表现的保证。在不同时期，或因使用不同假设和标准，采用不同观点和分析方法，致使金圆统一证券发出与本报告所载意见、评估及预测不一致的研究报告，对此金圆统一证券可不发出特别通知。

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，也没有考虑到客户特殊的投资目标、财务状况或需求。客户应考虑本报告中的任何意见或建议是否符合其特定状况，若有必要应寻求专家意见。本报告所载的资料、工具、意见及推测仅供参考，并非作为或被视为出售或购买证券或其他投资标的的邀请或向他人做出邀请。在法律允许的情况下，金圆统一证券或其关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能会为这些公司正在提供或争取提供投资银行业务服务。

更多免费行业报告 并购家 www.ipoipo.net

本报告版权仅为金圆统一证券所有。未经金圆统一证券书面同意，任何机构和个人不得以任何形式翻版、复制、发布、转发或引用本报告的任何部分。若金圆统一证券以外的机构向其客户发放本报告，则由该机构独自为此发送行为负责，金圆统一证券对此等行为不承担任何责任。本报告同时不构成金圆统一证券向发送本报告的机构之客户提供的投资建议。如未经金圆统一证券授权，私自转载或者转发本报告，所引起的一切后果及法律责任由私自转载或转发者承担。金圆统一证券将保留随时追究其法律责任的权利。

金圆统一证券营业网点

厦门地址：厦门市思明区展鸿路 82 号国际金融中心大厦 10-11 楼

深圳地址：深圳市南山区粤海街道高新区社区科苑南路 3099 号中国储能大厦 4801

上海地址：上海市中国（上海）自由贸易试验区陆家嘴环路 1233 号汇亚大厦 2301A、2302A 室

北京地址：北京市西城区武定侯街 2 号、4 号 10 层 F2-1（B）1001-05A