中国 AI 大模型商业化报告(2026年上)

CHINA LARGE LANGUAGE MODEL COMMERCIALIZATION REPORT



出品：北京第四波科技智库
中关村天成创新研究中心

发布时间：2026年7月

中国 AI 大模型商业化报告（2026 年上）

出品： 北京第四波科技智库

中关村天成创新研究中心

策划： 章园园 符永康

编审： 张向东 程絮森 高玮

编制： 孙晓一博士 陈尚容博士 冯皓钦研究员 王晨研究员 卢杨研究员

发布时间： 2026 年 7 月 3 日

联系邮箱： fourthwave@vip.163.com

第四波科技智库

目 录

第一章 行业格局与商业化前景.....1

1.1 总体格局.....1

1.2 市场形势.....3

1.3 问题与挑战.....5

第二章 中国大模型市场规模与表现.....6

2.1 整体市场规模与构成.....6

2.2 主要参与者及市场表现.....9

第三章 商业模式选择与变现路径.....10

3.1 主要商业模式.....10

3.2 技术路线——商业模式的底层逻辑.....14

3.3 算力成本——盈利目标的现实考虑.....16

3.4 定价博弈——市场与价值的综合考量.....19

3.5 垂直场景——决定未来的生存空间.....21

第四章 典型案例分析.....24

4.1 百度：以“芯云模体”为基石，构建 AI 商业化全栈生态.....24

4.2 智谱 AI：以全自研 GLM 大模型为根基的独立商业化路径.....27

4.3 昆仑万维：“4+3”AGI 战略引领的多产品矩阵商业生态.....30

4.4 科大讯飞：“自主可控底座+行业深度应用”战略.....33

第五章 趋势与展望——商业化未来的 10 个判断.....35

第四报科技智库

2026 年上半年，中国 AI 大模型产业竞争明显激化，在延续技术竞赛、流量争夺、生态抢滩的同时，全面打响大规模商业化“落地战”，广阔应用前景得以验证，行业泡沫开始出清。行业从追逐参数规模的时代进入一个由效率、成本与场景深度定义的新阶段。同样，我们看到了市场的剧烈分化：进入二季度豆包开始计费，智谱 AI 逆势涨价，云厂商纷纷提价，而 DeepSeek 却实行了大幅降价，商业策略的复杂化，恰是行业走向理性的明确信号。

本报告旨在剖析中国 AI 大模型商业化的现状、路径与未来，解构从 MaaS 到私有化部署、从 C 端订阅到行业解决方案的多元化商业模式，选取代表性企业进行案例分析，揭示它们在“Token 经济”驱动下的商业化策略及深层逻辑。

第一章 行业格局与商业化前景

1.1 总体格局

从全球 AI 大模型市场格局看，以 OpenAI、Google、Anthropic 和 Meta 为领先梯队，这些企业不仅拥有世界领先的基础模型能力，还掌握了庞大的数据资源、强大的算力体系和成熟的开发生态。其中，OpenAI 凭借 ChatGPT 建立了最广泛的用户入口和 Agent 生态；Google 拥有搜索、YouTube、Android、TPU 以及云平台等完整资源体系；Anthropic 则在企业级市场和 AI 安全领域建立了明显优势；Meta 依靠 Llama 系列成为全球开源生态的重要推动者，这四家企业构成了当前全球大模型产业的头部力量。

根据 6 月最新数据，OpenAI 旗下 ChatGPT 月活跃用户超 10 亿，

第四波科技智库

采取“全民平台”路线在语言大模型中占到全球第一的市场份额，二季度 ARPU（每用户平均收入）约 2.2 美元。Anthropic 在 2026 年第一季度凭借企业级安全 AI 定位，ARPU 达到了 16.2 美元，实现从对话工具向生产力引擎的关键转型。Anthropic 在 4 月 80%营收来自企业客户，5 月进一步攀升至 85%，并在企业端 API 支出中占据约 30 至 40%的市场份额，成为 B 端变现效率最高的 AI 厂商之一。Google 依托 Gemini 系列及搜索引擎生态，深耕多模态场景，月活用户由 2 月的 7.5 亿攀升至 5 月突破 9 亿；Meta 则以开源的 Llama 4 系列构建用户壁垒，主打极低 ARPU 的普及化策略，并逐步向“分层付费与广告双轮驱动”转变。此外，xAI 的 Grok 4 系列宣称具备高达 200 万令牌的上下文窗口，并具有实时数据接入能力，这可能会显著提升其在长文本理解和最新信息处理方面的竞争力。

欧盟虽然在超大模型竞争中相对滞后，但在 AI 监管、隐私保护、伦理治理和标准制定方面拥有较强影响力，正在成为全球 AI 规则的重要制定者。与此同时，中东等能源资本力量也开始进入 AI 核心竞争体系。随着 AI 对电力、数据中心和能源需求急剧增加，AI 产业越来越呈现“能源化”趋势，数据中心、电网和算力基础设施的重要性迅速提升。

同期，中国大模型在 AI 应用落地、场景规模、工程优化和商业化速度等方面展现出强大竞争性。市场呈现出“大厂平台化竞争+创业公司创新突破+垂直应用深化”的格局。阿里、字节跳动、腾讯、百度和华为依托云计算、流量入口和产业资源，正在构建中国版的 AI

第四浪科技智库

超级平台；DeepSeek、月之暗面、智谱 AI、零一万物等独立创业公司则凭借技术创新和产品特色，在细分领域不断取得突破。在垂直行业大模型领域，科大讯飞、百川智能等依托庞大市场在教育、医疗等领域不断开发深度应用。

图一　AI 大模型产业链全景图



（公开资料整理，AI 辅助制图）

1.2 市场形势

AI 大模型的智能体和垂直应用场景开发得到了政策的重点支持。2025 年 10 月，《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出：全面实施“人工智能+”行动，以人工智能引领科研范式变革，加强人工智能同产业发展、文化建设、民生保障、社会治理相结合，抢占人工智能产业应用制高点，全方位赋能千行百业。2026 年政府工作报告首次提出“打造智能经济新形态”，并提出深化拓展“人工智能+”，促进新一代智能终端和智能体加快推广，推动

第四范科技智库

重点行业领域人工智能商业化规模化应用。智能体首次被写入政府工作报告，标志着 AI 从“工具”升级为“产业引擎”。2026 年 1 月，工信部、国家发展改革委等八部门印发《“人工智能+制造”专项行动实施意见》，进一步提出到 2027 年推出 1000 个高水平工业智能体等具体目标。

AI 大模型技术不断突破，形成**“高性能、低成本、可部署、可迭代”**的成熟技术体系。模型能力上，主流大模型支持百万级 Token 长文本，部分模型幻觉率降至 10%以下，可适配金融、法律等高精度严谨场景。依托 MoE 稀疏架构、低精度量化、知识蒸馏等组合优化技术，在同等模型性能对标传统稠密大模型的条件下，训练算力成本最高可下降 90%；短文本 C 端对话底层推理算力成本可控制到 0.001 元以内。Agent 智能体、RAG、端云协同技术成熟，AI 从“对话”升级为“做事”，大幅提升了企业商业价值与付费转化率。大模型技术在“实验室阶段”迈入产业应用阶段，开始注重规模化落地与盈利，中小厂商可快速落地 AI 应用。

根据创新扩散理论的一般规律，突破 40%普及率产业临界点，产业会遵循技术扩散 S 型曲线，进入指数级增长与变现周期。2025 年底生成式 AI 全民普及率达 42.8%[1]，正式跨越行业公认商业化临界点，渗透速度远超当年智能手机；企业开始转向高性价比轻量化模型，以及在法律、医疗、编程等垂直领域打造“专家模型”。同时，智能体的落地比拼成为核心，AI 加速从“能对话”转向“会办事”。用户、

[1]第五十七次《中国互联网络发展状况统计报告》，中国互联网络信息中心（CNNIC）发布

第四浪科技智库

数据、模型、体验形成正向循环，产业增长进入自我强化阶段。

商业化落地方面，AI 大模型实现 C 端、B 端、G 端、端侧硬件全域覆盖，从试点走向全流程深度嵌入，商业化需求真实且持续。C 端 AI 对话、内容创作、智能修图等成为日常工具，高频使用用户占比超 7 成；手机、PC 渗透率双双突破 50%；B 端办公、金融、工业、医疗、传媒等千行百业落地成熟，非互联网行业 AI 需求增速达 65%，企业付费意愿明确；G 端常态化应用：超 80%政务平台接入 AI 能力，公文写作、智能审批、政务咨询规模化落地。

1.3 问题与挑战

中国 AI 大模型商业化处于成长期而非成熟期，在技术、成本、模式等方面都面临挑战。短期的快速盈利和长期的生态构建，是商业化要解决的核心问题。

产业链上游的算力供给受限、成本高企、技术壁垒等严重影响 AI 大模型的商业化落地。当前 AI 产业高端芯片受限，高端算力供应受出口管制、生态成熟度和成本因素影响，国产算力仍处于适配和追赶阶段。算力成本仍然较高，训练千亿参数模型单次算力基础计算成本达到 30 万至 80 万美元；推理端成本随调用量激增，6 月豆包日均 token 调用量已达 120 万亿，GPU 集群压力巨大。通用数据的中文互联网高质量中文语料不足，金融、医疗、工业等领域高质量数据稀缺且垄断，金融高质量数据集估值高，中小企业难以负担。用户体验不稳定，幻觉、逻辑错误、知识过时、响应延迟等问题突出，C 端低价竞争激烈，MaaS 服务定价缺乏统一标准，客户付费意愿与价格敏感

第四浪科技智库

度矛盾突出。这些问题都会成为商业化落地过程中的不稳定因素。

当前的主要商业模式能够实现可持续盈利还有待验证。主流大模型依赖 API 调用、项目制、订阅，但 API 毛利率仅 10%–20%，项目制收入不可持续，订阅付费转化率低。场景落地方面，通用模型与行业需求错位，垂直适配成本高、周期长；企业要求效果、价格、安全三者平衡，本地化部署需求强但成本高。稳定可持续的商业模式还没有完全建立，头部企业年投入高，高度依赖融资，多数公司仍亏损，盈利周期普遍 3–5 年。训练数据版权、生成内容归属、侵权责任界定模糊，诉讼风险高。这些都会成为商业化的长期桎梏。

第二章 中国大模型市场规模与表现

2.1 整体市场规模与构成

根据多家机构的分析评估，中国 AI 大模型市场在过去六年间呈现规模化扩张，如艾媒咨询认为中国 AI 大模型市场从 2020 年的 16.23 亿元到 2025 年的 495 亿元，年复合增长率约 98%%；浙商证券报告认为从 2020 年的 15 亿元到 2025 年 318 亿元，年复合增长率约 83%。机构普遍认为 2026 年是商业化落地的关键期，尽管增速放缓，但未来五年仍将保持高速增长。艾媒咨询认为 2026 年中国 AI 大模型市场规模将达到 738 亿元，浙商证券预计市场规模达到 560 亿元，到 2028 年将达到 1179 亿元。而中商产业研究院预计，中国 AI 大模型市场规模 2026 年约为 680 亿元，赛迪顾问预计 2026 年中国大模型市场规模将突破 700 亿元。

综合上半年 AI 市场表现及多家机构对市场规模和增速的评估，

第四波科技智库

预计 2026 年全年中国大模型将形成 700 亿元左右的市场规模，乐观情景下市场规模可达到 750 亿元。

当前的市场竞争格局呈现出“总体集中、局部分化”的特征。在总体市场规模上，头部企业占据了绝大部分份额，但不同细分赛道的竞争态势存在显著差异。

在企业级服务市场，尤其是在 AI 云服务领域，市场集中度极高。阿里云、火山引擎、百度云、腾讯云、天翼云/华为云等的市场占比总和超过八成，头部效应尤为明显。在整体 AI 云与 AI IaaS 市场，阿里云保持着较强的市场份额；在 MaaS（模型即服务）与 Token 调用市场，阿里云与火山引擎表现出较强的竞争力；在 AI 大模型解决方案市场，百度智能云凭借全栈 AI 能力和自研昆仑芯，市场表现活跃。企业级客户对模型服务的稳定性、安全性和与现有云生态集成的强需求，拥有强大云基础设施的巨头具备天然优势。

在消费级市场，呈现出多家头部企业集中竞争的格局。百度文心助手凭借先发优势和广泛的生态集成，一季度月活用户达到 3.6 亿。字节跳动的豆包 APP 则凭借流量优势，一季度日活用户约 1.4 亿，月活用户约 3.45 亿。阿里巴巴的通义千问 APP 上线后增长迅猛，月活亦快速突破 3 亿。腾讯的混元模型虽在 C 端独立应用上布局较晚，但其通过微信、QQ 等超级入口的渗透潜力不容小觑。

从技术路线视角看，闭源与开源模型的竞争也构成了市场格局的重要维度。在需要高稳定性、强技术支持的政企核心业务场景，闭源模型仍占据主导。在成本敏感、数据可控性要求高或需要深度定制的

第四浪科技智库

场景，开源模型正快速渗透。以 DeepSeek、阿里的通义千问开源系列为代表的中国开源模型，已在全球范围内建立了庞大的开发者生态，对市场格局产生了深远影响。

图二 AI 大模型分类与产品示例（截至 2026 年 6 月）



（公开资料整理，AI 辅助制图）

第四浪科技智库

2.2 主要参与者及市场表现

过去一年中国大模型产业逐渐从单纯的“数量竞赛”迈向“合规落地与工程化应用”，截至 2026 年 5 月，累计有 868 款生成式人工智能服务在国家网信办完成备案，530 款在地方网信办完成登记；现存大模型相关企业达 1.2 万家，专注于大模型研发的核心企业数量在 100 多家左右。

这些企业是行业的主力军，持续进行技术迭代和商业化探索。参与主体呈现出显著的多元化特征，包括互联网巨头、独立 AI 公司、垂直行业公司及新型研发机构，每一类参与者都依托其独特的资源禀赋，选择了差异化的技术路线和市场定位，共同塑造了当前的市场生态。

表一 AI 大模型代表企业及产品

互联网巨头（自研通用大模型）		独立 AI 公司（专注大模型研发）		垂直行业公司（行业大模型）		新型研发机构	
企业	产品	企业	产品	企业	产品	企业	产品
百度	文心	智谱 AI	GLM 系列（GLM-4/5）	科大讯飞	讯飞星火	智源研究院	悟道
阿里巴巴	通义千问	商汤科技	日日新大模型	三六零	360 智脑		
腾讯	混元大模型	稀宇科技	MiniMax 系列	昆仑万维	天工大模型		
字节跳动	豆包大模型	DeepSeek	DeepSeek 系列	拓尔思	拓天大模型		
华为	盘古大模型	零一万物	Yi 系列通用大模型	云从科技	从容大模型		
小米	MiMo	月之暗面	Kimi 系列	百川智能	Baichuan-M4 医疗		
		面壁智能	BitCPM-CANN	快手	可灵、可依		

第四范式科技智库

				零犀科技	因果大模型、CaRhino		
				卓世科技	“璇玑玉衡”模型		
				佳都科技	“佳都知行交通大模型”		
				新石器无人车	无人车即时配送 RaaS		
				希迪智驾	运载智能体		
				企威科技	ColdStore AI		

在市场表现方面，互联网巨头依托庞大生态实现全栈自研与多渠道变现。在企业级大模型日均调用量市场中，阿里通义千问、字节豆包和 DeepSeek 等同时拥有强大闭源服务、开源生态及丰富应用场景的厂商，以及凭借极致性价比和开源策略快速崛起的厂商，正在获得市场的广泛认可。智谱 AI 在 GLM 系列代码能力与逻辑推理方面表现突出，港股上市后市值一度突破 5000 亿港元，6 月市值突破万亿港元；月之暗面主打“超长上下文”，估值突破 200 亿美元；DeepSeek 凭借推理性能与低成本优势，成为现象级独角兽，6 月完成 500 亿元融资，投后估值达 3500 亿元。MiniMax 以视频生成能力和 Agent 能力创下 AI 企业从创立至 IPO 的最快纪录，总市值一度逼近 4000 亿港元，6 月市值回落至约 1600 亿港元。

第四浪科技智库

第三章 商业模式选择与变现路径

3.1 主要商业模式

AI 大模型的商业模式日趋多元，当前已存在 10 种商业模式。其中，MaaS（API 服务）、本地化部署、行业解决方案、C 端订阅是当前最主流的商业模式，这四种模式构成了 AI 大模型商业化最基础的

矩阵。

表二 AI 大模型的 10 种商业模式

商业模式	核心逻辑	典型收入方式	应用场景
MaaS(API 服务)	通过 API 按调用量提供模型能力	按 Token/调用次数计费	智能客服、内容生成、代码辅助、翻译、数据分析、聊天机器人
本地化部署	模型部署在企业自有服务器，数据不出域	一次性项目费 + 年度维护费	金融风控与投研、政务审批与政策解读、医疗病历分析、企业知识库
行业解决方案	深度定制满足行业特定场景的 AI 应用	项目制 + 持续服务费	金融智能投研/智能客服、汽车故障诊断、广告创意生成、旅游行程规划
C 端订阅	个人用户按月/年付费获取 AI 产品使用权	月费/年费（如$20/月）	个人写作辅助、翻译、编程助手、学习辅导、日常信息查询
按效果/成果付费	按 AI 带来的实际业务效果（如转化率提升）付费	按效果分成或固定+浮动	营销广告效果提升、客服解决率提升、自动化流程效率优化
模型市场/交易平台	平台让第三方上架模型或应用，抽取交易佣金	交易分成（如 15%-30%）	垂直行业模型交易（如法律/医疗专用模型）、AI 插件扩展、RPA 技能市场
品牌/联名与 IP 授权	将模型能力与品牌 IP 结合，打造专属 AI 体验	授权费 + 销售分成	乳制品营养顾问、影视 IP 互动、游戏角色 AI 对话、奢侈品虚拟导购
广告变现	免费提供基础服务，通过展示广告获利	CPM/CPC 广告收入	搜索广告联动（AI 回答中嵌入广告）、内容推荐广告、AI 助手推荐体系
生态捆绑	将模型与云平台、算力、应用软件深度绑定	一体化解决方案套餐	云平台 AI 服务（大算力+模型一体）、企业办公 AI 助手、IoT 智能设备 AI 赋能
数据飞轮（间接支撑）	用户使用产生数据 → 优化模型 → 吸引更多用户	间接变现（提升其他模式收入）	通用对话助手、搜索推荐系统、内容社区生成

API 按量计费模式（MaaS）将大模型能力封装为标准化的接口服务，客户根据实际调用的 Token 数量付费。这种“用多少，付多少”

第四范科技智库

的灵活模式，已成为当前商业化成熟度最高、收入确定性最强的路径，其核心逻辑在于将前沿技术转化为可度量的标准化商品。MaaS 凭借低门槛、灵活计费的特点，成为中小企业和开发者快速接入 AI 的入口。

私有化部署（或称本地化部署）主要服务于对数据隐私、系统安全、业务合规及服务稳定性有极致要求的大型政企客户，通常采用“一次性软件授权/部署费 + 年度维护服务费”的模式。这是国产大模型厂商，尤其是独立 AI 公司，至关重要的收入基本盘和利润来源。行业解决方案则凭借数据安全和深度定制的优势，渗透到金融、政务、医疗等合规要求极高的行业。

行业解决方案是大模型商业化向“垂直场景深耕”的典型模式。AI 在这一模式中，深度嵌入到教务系统、投研平台、律所管理系统等核心界面中，成为专业工作流的一部分。其盈利模式正逐渐从单一的软件授权费，转向“基础订阅费 + 按席位/场景计费 + 效果分成”的复合结构。这种模式将数据回流、模型升级和场景优化纳入持续服务关系，可以为厂商提供更稳定的长期现金流。

面向个人用户及小微团队的订阅制模式，其核心挑战在于突破低付费率与产品同质化的瓶颈，目前仍处于“探路”阶段。随着模型能力向复杂、专业的任务场景渗透，该模式正探索一条可持续的路径。对于 C 端用户，订阅制形成了最简单直接的变现方式，但受限于较低的 ARPU，单纯依赖订阅难以支撑高昂的研发成本，故多数头部企业同时经营多条业务线。

贵阳市科技智库

表三 代表大模型厂商 token 套餐（截至 2026 年 6 月）

厂商	套餐档位	月费/首月特惠	额度/请求量	支持模型
火山引擎（字节·方舟）	Lite	40 元/月（首月 8.91 元）	1.8 万次/月	豆包 Seed-2.0、DeepSeek-V4、GLM-5.1、MiniMax-M3、Kimi-K2.6 等
	Pro	¥200/月（首月 ¥44.91）	9 万次/月	同上
阿里云（百炼）	Lite	¥40/月 （首月¥7.9）	1.8 万次/月	Qwen3.5-Plus、GLM-5、MiniMax-M2.5、Kimi-K2.5
	Pro	¥200/月（首月 ¥39.9）	9 万次/月	同上
百度智能云（千帆）	基础版	¥40/月 （首月¥40）	1.8 万次/月	GLM-5、Kimi-K2.5、DeepSeek-V4、ERNIE 5.1
	Pro	¥200/月	9 万次/月	同上
智谱 AI	Lite	¥49/月	400 次/周	GLM-5、GLM-4.7 等
	Pro	¥149/月	12 万次/月	GLM-5.2、GLM-5.1、GLM-5-Turbo 等
	Max	¥469/月	48 万次/月	同上
MiniMax	Starter	¥29/月	40 次/5 小时	M2.5 系列
	Max 极速版	¥199/月	300 次/5 小时	M2.5-highspeed
Kimi（月之暗面）	Moderato	¥99/月	未公开固定数值	Kimi K2.7 Code、K2.6、K2.5 等
腾讯云	Lite	¥40/月 （首月¥7.9）	1.8 万次/月	混元 2.0、GLM-5、Kimi-K2.5、MiniMax-M2.5 等

在其他模式中，按效果/成果付费按 AI 带来的实际业务效果（如客户转化率提升、成本优化比例）进行结算。此模式降低了客户决策门槛，但效果归因和度量体系是其落地的关键挑战。

模型市场/交易平台抽成构建类似应用商店的模型生态平台（如 Hugging Face、GPT Store），第三方开发者上架微调模型或 AI 应用，

第四波科技智库

平台收取交易佣金，这种模式轻资产且可规模化，但对冷启动和生态运营能力要求较高。

品牌/联名合作与 IP 授权将大模型能力与其品牌 IP 深度结合，实现差异化变现，该方式能提升品牌溢价，但规模化复制难度较大。

广告变现面向 C 端用户免费提供基础服务，通过嵌入广告或联动搜索推荐间接创收。目前部分头部大模型正在探索此路径，主要依赖庞大的用户流量。

生态捆绑将大模型与云平台、算力、应用软件深度捆绑，形成一站式解决方案，此模式客户粘性极强，有助于云厂商提升综合收入。

数据飞轮通过用户使用产生数据→模型迭代→吸引更多用户→产生更多数据的闭环，构建深度的竞争壁垒。该模式本身不直接产生收入，但通过强化模型能力间接提升其他商业模式的 ROI（投资回报率），形成持久护城河。

3.2 技术路线——商业模式的底层逻辑

从技术架构与性能看，闭源模型以追求极致的通用性能为目标，通常采用全参数密集型训练与多模态原生融合技术，在复杂推理、创意生成等需要高度泛化能力的任务上保持领先。其核心优势在于服务稳定性高、生态集成成熟，为企业提供“开箱即用”的可靠 AI 能力。相比之下，开源模型的技术创新更聚焦于效率与成本，部分开源或开放权重模型采用 MoE、量化、蒸馏等方式降低推理成本。在代码生成、数学推理等特定领域，顶尖开源模型的性能已实现“闭源替代”，部分指标甚至实现反超。

第四波科技智库

图三 AI 大模型技术路线对比

维度	闭源模型路线（追求极致通用性能）	开源模型路线（追求效率与成本优化）
代表模型	百度 文心 4.0；阿里云 通义千问 Qwen 2.5；华为 盘古 Pangu 5.0	DeepSeek V3.2；阿里云 Qwen3-Max；智谱 AI GLM-4.5；Mistral AI Mixtral 8x22B；月之暗面 Kimi K2
核心架构	全参数密集型（Dense Transformer）+ 多模态原生融合	MoE（混合专家）稀疏架构 + 注意力优化（MLA/DSA 等）
训练范式	超大规模全参数训练 + 强化学习对齐（RLHF/RLAIF）	高效训练范式 + 精准数据筛选 + 知识蒸馏 + 持续预训练
多模态能力	原生多模态（文本+图像+音频+视频+3D）深度融合	多模态支持（文本+图像为主，部分支持音频/视频）
性能表现（通用推理）	领先同期优秀开源模型 6-12 个月	特定领域已实现闭源替代，部分指标反超
成本效率（相对指标）	训练成本极高，推理成本高	训练成本低 10-100 倍，推理成本低 5-50 倍
优势领域	通用推理、复杂任务规划、创意生成、多模态理解、长文本处理	代码生成、数学推理、逻辑推理、专业领域、可定制化场景
劣势挑战	训练与推理成本高、模型不透明、定制化受限、数据隐私顾虑	通用泛化能力稍弱、服务稳定性依赖部署、生态集成度较低
生态与部署	服务稳定性高、API 生态成熟、企业级支持完善、开箱即用	开源透明、可定制化、本地部署灵活、社区生态活跃
适用场景	企业级应用、关键业务系统、需要高可靠性的场景、快速集成	成本敏感场景、垂直领域应用、科研教育、需要定制化的场景

（公开资料整理，AI 辅助制图）

第四波科技智库

商业化路径的差异直接源于技术特征。闭源模型依托其性能壁垒，构建了以 API 调用费与企业订阅为核心的高价值变现闭环。如 Anthropic 凭借 Claude Opus 系列大模型，商业模式依赖于锁定大型企业客户，提供稳定、合规且高性能的服务，但高昂的 API 成本也构成了使用门槛。智谱 AI 同样依托“API 调用+私有化部署+MaaS 平台”的商业模式，对标 Anthropic，深耕金融风控、政务公文、能源调度等场景。

开源模型践行“技术平权与生态共享”的逻辑，其商业化模式更

为多元和灵活。一是提供商业化 API 服务，以极低的定价吸引海量开发者与中小企业，如 DeepSeek 的 API 定价仅为国际头部闭源模型的数十分之一；二是私有化部署授权，满足金融、政务等对数据安全与可控性要求极高的客户需求；三是与云厂商合作分成，通过预装在云平台中获取算力收入；四是通过庞大的开源生态衍生增值服务。例如，阿里的 Qwen3 系列开源模型全球下载量已突破 10 亿次，衍生模型超 20 万个，构建了全球最大的开源模型生态之一，为其云业务和行业解决方案带来了巨大的引流和协同效应。

在下游 B 端商业化市场中，企业的应用选择呈现出清晰的场景分化逻辑。据 IDC、沙利文、阿里云研究院等多家机构的调研报告，高达 70-80%的企业有部署开源模型的意愿，其中 60%的企业认为开源模型综合成本更低，40%的企业选择开源的核心驱动力是数据可控性。因此，在实际部署中，企业普遍采用混合策略：在核心业务、快速产品原型验证及对稳定性要求极高的场景，倾向于采用闭源模型；而在非核心业务、涉及敏感数据处理、需要深度定制或成本敏感的场景，则优先选择开源模型。“闭源保核心、开源降成本控数据”的混合架构，正成为企业级 AI 应用的标配。

3.3 算力成本——盈利目标的现实考虑

AI 大模型的成本结构高度集中，算力成本构成了绝对主导的支出项，这是行业盈利困境的起点。综合行业数据，头部企业算力支出占总成本的比例高达 57%-70%，远超人才薪酬、研发及其他运营开支。这一成本可进一步拆解为训练、推理与能源三大刚性部分，每一环节

第四浪科技智库

都构成严峻的财务压力。

训练成本属于高额、周期性的资本性投入。单次训练前沿模型的成本动辄数千万至上亿美元,例如GPT-4o的训练成本超过1亿美元,而 DeepSeek-V3 使用了 2048 块英伟达 H800 GPU,官方披露成本约 557 万美元。尽管训练阶段的算力利用率(MFU)可优化至 50%以上,但其一次性投入巨大,且需持续进行以维持模型竞争力,构成了沉重的固定成本负担。

推理成本是持续运营中的核心变量成本,也是导致“规模不经济”现象的主因。大模型在推理解码阶段属于典型的显存带宽受限(Memory-Bound)架构。在低并发场景下,模型的实际算力利用率(MFU)极低,通常仅为 5%-10%。高达 80%以上的单 Token 成本,本质上并不是在为实际消耗的算力买单,而是为了维持大容量、高带宽显存(如 HBM)不间断读取模型参数与 KV Cache 所支付的成本。

能源成本随着算力规模指数级增长而日益凸显。运行先进 AI 数据中心需要巨大的电力支撑,英伟达 Blackwell Ultra/GB300 高密度液冷训练机柜,商业部署标准单机柜峰值功率 240kW,多家中国厂商已完成工程样机与试点部署,主流量产高密度训练机柜集中在 100–120kW 区间,如字节跳动单日整体算力总成本 2–5 亿元人民币,折算单日电费支出数千万元人民币。

第四浪科技智库

图四 代表通用大模型 Token 价格对比（截至 2026 年 5 月）



（公开资料整理，AI 辅助制图）

第四浪科技智库

当前头部 AI 企业的总支出规模约为其营收的 2-3 倍，行业整体仍处于“投入远超回报”的大规模投资阶段。任何商业模式的盈利性设计，都必须首先直面如何消化或转移这占比 57%-70%的算力成本挑战。

3.4 定价策略——市场与价值的综合博弈

2026 年，AI 大模型行业的定价策略告别了早期以近乎免费和断崖式降费为标志的粗放式价格战，进入一个由技术能力、场景价值与成本效率共同定义的价值战新阶段。

智能体（Agent）应用爆发带来的 Token 消耗指数级增长、英伟达 Blackwell 等先进芯片带来的算力成本激增，以及资本市场对盈利兑现的迫切要求，推动行业从追逐“模型稀缺性叙事”全面转向考核“单位 Token 经济价值”与“场景适配深度”。

表四 部分大模型厂商的商业模式与定价策略（截至 2026 年 6 月）

代表厂商	核心商业模式	核心定价变化与策略
智谱 AI	**开源引流+MaaS 与端侧授权**：通过极致开源策略降低门槛，吸引海量开发者；核心变现依靠云端 API 调用、高毛利的企业私有化部署，以及向终端硬件（如三星手机）植入 AI 能力的授权收费。	2026 年累计提价近 83%，发布 GLM-5.1 时再次上调 API 价格 10%。
腾讯/阿里/百度	**云基础设施捆绑（IaaS+PaaS）**：将大模型作为云计算的“超级插件”，通过 API 调用和算力消耗带动底层云基础设施（服务器、存储）的销售，实现 AI 与云业务的协同增长。	腾讯 AI 算力产品上调 5%，阿里云算力卡提价 5%-34%，百度智能云 AI 服务提价 5%-30%。
DeepSeek	**多元化 API 与私有化部署**：采用按调用量阶梯定价策略，以极低门槛（如 0.025 元起/百万 Token）抢占开发者与中小企业市场；同时针对	V4-Pro API 价格永久降至原定价的 1/4，降幅高达 75%。

智业投科技智库

	强监管行业提供高毛利的私有化部署与行业定制解决方案。	
字节跳动（豆包）	**C 端分层订阅与流量变现**：以免费基础功能维持庞大用户规模，通过高级功能订阅获取稳定现金流；并将 AI 能力嵌入短视频、电商等核心业务，提升原有广告与电商业务的转化效率。	2026 年 5 月正式落地商业化付费体系，采用三级会员制（标准版 68 元/月，加强版 200 元/月，专业版 500 元/月），基础聊天问答永久免费，针对 PPT 生成、数据分析等高算力复杂任务场景收费。
月之暗面（Kimi）	**高价值任务消耗型 MaaS**：聚焦长文本、复杂代码等高 Token 消耗场景，通过解决高难度痛点获取溢价；以 API 调用量为核心收入引擎，实现从“卖工具”到“卖生产力”的跨越。	Kimi K2.6 的输入价格（缓存未命中）从 4 元/百万 Token 上调至 6.5 元/百万 Token，涨幅达 58%，实现 API 价格的结构性提价，但同期调用量反而大幅增长。
小米(MiMo 大模型)	**硬件生态绑定与端侧赋能**：作为硬件巨头，大模型主要作为赋能终端设备（手机、AIoT）的底层组件，通过提升硬件产品的智能化体验来促进硬件销售，API 降价旨在快速繁荣开发者生态。	旗下 MiMo-V2.5 系列 API 实施永久降价，最高降幅达到 99%，Pro 版定价与 DeepSeek V4-Pro 对齐，并取消了输入长度分级，实行统一 Token 定价。

2026 年上半年，AI 大模型企业纷纷重构定价体系，定价策略调整的深层逻辑在于厂商判定市场已从“为功能付费”转向“为结果可靠性付费”。集体提价反映了企业级市场需求的爆发和客户对模型性能重视度的提升，价格本身不再是唯一决策因素。但从最新一轮的价格调整看，DeepSeek 的逆向降价并未引发新一轮价格战，反而通过展示革命性的成本控制能力，开辟了“普惠模型”的新赛道，促使行业竞争维度更加多元化。

DeepSeek V4-Pro 凭借底层技术架构的极致优化，将大模型推理成本降至极低水平，其永久降价是触发层级分化的关键催化剂。AI 大模型不同能力层级的模型表现出不同的定价逻辑。

基础层（免费/低价引流）：各大厂商保留免费基础服务或极低

第四浪科技智库

价版本，如智谱 GLM-4-Flash、豆包基础版。该层核心目标是维持用户规模、生态活跃度及获取数据反馈，服务于简单问答、内容生成等低算力消耗场景。

中间层（按量计费/订阅）：针对中端生产力场景，采用按 Token 计费或月度订阅制。例如豆包标准版（68 元/月）、智谱 GLM-5.1 的通用 API、阿里云百炼的按量付费套餐。该层在成本与性能间取得平衡，适合高频但非极端复杂的任务，是当前企业采购的主流选择。

旗舰层（高溢价/能力定价）：针对复杂长周期任务、高可靠企业级应用，提供高溢价服务。如豆包专业版（500 元/月）。该层模型凭借卓越性能、超长上下文、多模态深度理解等稀缺能力，拥有独立定价权，用户为确定性的结果支付显著溢价。

3.5 垂直场景——决定未来的生存空间

AI 大模型的价值最终要通过其在具体行业场景中的落地与赋能来体现。丰富的垂直应用场景，为大模型商业化提供了巨大的可能性，“通用底座+垂直深耕”的分工与协同模式，将为企业带来未来差异化竞争的宝贵机遇。当前中国 AI 大模型垂直商业化市场主要分布在金融、医疗、教育、智能制造及产业互联网、政务等核心应用领域，AI 大模型在这些领域的技术渗透深度和市场规模直接决定了商业化的可持续性与质量。金融业是数据密集型行业，对风险控制、效率提升和合规性有着极致追求，这使其成为 AI 大模型商业化落地最成熟、价值验证最清晰的领域。例如长亮科技推出的 DataMind 平台主要解决金融数据开发和经营分析难题。宇信科技自研的“星睿智调”AI

第四波科技智库

Agent 依托金融大模型、知识图谱等技术，打通了信息归集、风险穿透、智能研判、报告生成的全链条。预计到 2027 年，超过三分之一的金融机构将深度接入 AI 技术，推动金融科技细分市场规模突破 5000 亿元。

医疗健康是全球 AI 应用占比最高的领域，也是技术壁垒最高、社会价值最大的赛道。在应用层面，AI 已实现诊疗全周期覆盖；在药物研发领域，AI 可将传统研发周期缩短 18-24 个月，已有多款抗癌药物进入临床二期试验。例如医渡科技深耕医院核心流程多年，处理近 70 亿份医疗记录数据，开发了临床循证智能体“医渡智循”（Physical AI）。数坤科技建立了从单点诊断扩展到覆盖 “防-筛-诊-治-管”全链条的“数坤坤”大模型。

教育领域的 AI 应用以满足日益增长的个性化、公平化学习需求为核心。教育领域的增长逻辑在于将 AI 转化为可规模化的“数字教师”或“智能学伴”，其增长贡献度体现在通过技术手段释放家庭及社会在高质量教育上的消费潜力与预算。海亮科服打造了覆盖创思课堂、学生成长、教师提升等全场景的智能体产品矩阵，如“星乐读”数字人助教、“星学伴”精准教学平板等。好未来联合北京师范大学研发了紧扣新课标知识体系的基础教育大模型。

制造业的智能化升级是 AI 技术赋能产业的核心体现。AI 在制造业的渗透正从边缘环节向核心生产系统延伸。目前 AI 视觉质检系统的渗透率已达 38%，更深度的应用体现在柔性制造、智能排产与预测性维护等环节。蚂蚁工场研发的“蚂蚁智图大模型”用于工业零件加

崇阳发科技智库

工制造，实现秒级报价与批量工艺制定，工作效率等效 10 名工程师单日工作量。博清科技依托清华大学潘际銮院士团队研发的“际銮”焊接大模型，可将焊接工艺开发周期缩短 60%，综合生产成本降低 40%以上。

在产业互联网领域，大模型垂直应用正从“通用对话”向“产业决策与执行”深度演进，在处理复杂结构化数据，优化生产流程，提升管理效率等方面产生商业价值。如万联易达研发了覆盖国民经济全行业大类的全产业 AI 大模型万联摩尔，并正式上线主动型智能体管理平台 WanClaw，实现专家级技能封装，深挖垂类产业场景价值。浪潮集团依托浪潮云州开发的知业大模型，融合了平台上的工业机理与隐性经验，具备工业数据分析、知识推理和代码生成能力。

政务领域的 AI 应用具有鲜明的政策驱动特征，中央网信办与国家发展改革委于 2025 年 10 月联合印发的《政务领域人工智能大模型部署应用指引》明确了其“辅助型”定位并加强了经费保障，应用场景可全面覆盖政务服务、社会治理、机关办公及辅助决策。如拓尔思研发的拓天政务行业大模型，基于自有的 10 亿政务数据积累，已在智能政策助手、办事咨询助手和公文辅助写作等场景落地，并支持根据政府部门需求进行私有化部署。北京大学研发的城市治理大模型 UGGPT，定位于面向城市治理场景的智能体系统，可直接参与治理全流程。

第四范科技智库

第四章 典型案例分析

报告选取了百度、智谱 AI、昆仑万维及科大讯飞作为深度对标分析的典型案例，以分别分析互联网巨头、独立 AI 公司及垂直行业公司在 AI 大模型商业化中的不同路径与模式。

4.1 百度：以“芯云模体”为基石，构建 AI 商业化全栈生态

百度是中国最早布局人工智能领域的科技巨头之一，依托自主研发的文心大模型系列、百度百舸算力平台与千帆大模型平台，构建了一个覆盖基础设施、模型训练、应用开发与行业落地的全栈 AI 商业化生态。

（一）产品与服务

文心大模型系列，经过多次迭代，已进化至文心大模型 5.1 版本。该系列模型具备文本理解与生成、多模态感知、逻辑推理与知识问答能力，是百度对外提供 AI 服务的核心引擎。

基于文心大模型，百度构建了三层技术产品架构：

算力层：百度百舸平台。面向大规模深度学习的高性能云原生 AI 计算平台，具备异构算力统一管理、高效调度、训推加速等能力。在万卡集群上，模型有效训练时长占比超过 **98.8%**，线性加速比与带宽有效性均超过 95%，是业界领先的算力效能。

平台层：千帆大模型平台。提供从数据管理、数据标注、训练配置、模型纳管到服务发布的全流程工具链，支持企业进行大模型开发与微调，降低应用开发门槛。

应用层：行业解决方案与 AI 原生应用。基于强大的模型与平台

贵阳软科技智库

能力，百度面向多个行业输出定制化解决方案，并推出如“文心快码”“秒哒”等开箱即用的 AI 原生应用。

百度智能云依托文心大模型、千帆平台及百舸算力平台，将 AI 能力广泛嵌入各行业核心业务场景。关键落地领域与成效如下：

智能终端：赋能荣耀、三星、vivo、OPPO 等厂商，打造端云协同的智能助手，实现智能创作、多文档问答、会议纪要等功能。

具身智能：为北京人形机器人创新中心、千寻智能等提供高性能算力支撑，将集群有效训练时长提升至 98.8%，训练与推理效率分别提升 30%和 60%。

智能汽车：助力理想、长安、吉利等车企，构建智算中心与混合云平台，资源利用率提升超 40%；与蔚来共创端云多模态大模型 NOMI GPT。

AIGC：支持 Vidu 视频大模型训练，使其在全球评测中超越 Sora；助力 3D 大模型 Tripo 实现 8 秒快速纹理生成。

智慧科教：联合北京大学、厦门大学等高校及高途、好未来等企业，实现 K12 判题准确率达 95%，作业批改业务覆盖率提升 283%。

AI 游戏：为网易《逆水寒》提供超 1 亿玩家的实时 AI 竞技场景，并帮助三七互娱将客服 SLA 达成率从 61%提升至 99.3%。

智慧金融：服务邮储、招行、泰康等近 600 家客户，助邮储银行将模型迭代时间从 43 天缩短至 10 小时，保险文档检索准确率超 85%。

智慧工业：赋能国家电网、中国中车、华润等央国企，助中车空气动力学仿真从 24 小时缩短至 10 秒；为深圳燃气提升视频巡查效率

第四浪科技智库

15 倍以上。

政务服务：为全总、海淀区、深圳水务等提供“接诉即办”与行业智能体服务，将 3 天数据统计缩短至 1 分钟，服务覆盖超 32 万用户。

（二） 商业模式与定价策略

百度智能云的核心商业模式——云基础设施捆绑（IaaS+PaaS），即：将大模型作为云计算的“超级插件”，通过 API 调用和算力消耗带动底层云基础设施的销售，实现 AI 与云业务的协同增长。同时，百度智能云亦采取 API 按量计费、私有化部署授权、行业解决方案项目制等多种模式并行。

在定价方面，根据报告中的“代表厂商定价策略对比”表格显示：百度智能云 AI 服务整体提价幅度为 5%-30%。其文心一言 4.0 的 API 定价区间为 0.008-0.03 元/千 Token，与阿里、字节等头部厂商保持竞争力对齐。这种适度的提价策略，既体现了大模型技术价值的持续提升，也反映了市场需求依然旺盛。

（三）商业化分析

百度展示了互联网巨头在大模型商业化中的“全栈式”策略：通过底层算力、中台能力、上层应用的全面打通，构建强大的生态壁垒。百度将大模型商业化与其云基础设施、百度搜索、智能交通、自动驾驶等业务深度耦合，形成了极强的协同效应。

技术方面拥有从芯片（昆仑芯）、框架（飞桨）到模型（文心）、平台（千帆）、应用的全栈自主技术体系；行业落地案例覆盖智能终端、具身智能、智能汽车、AIGC、智慧科教、AI 游戏、智慧金融、智

第四波科技智库

慧工业、政务服务等十余个行业、数百个头部客户；百度百舸平台在算力调度、训推加速、国产芯片适配等方面有效支撑了各类大规模 AI 任务。这种“芯云模体”的商业模式，使其不仅在大模型技术能力上保持领先，更能在广泛的行业场景中获得持续、稳定且高价值的商业化收入。

4.2 智谱 AI：以全自研 GLM 大模型为根基的独立商业化路径

智谱 AI 成立于 2019 年，源自清华大学技术成果，是中国 AI 大模型领域的独立派领军企业。公司致力于打造新一代认知智能大模型，专注于做大模型的中国创新，从 GLM 预训练架构研发起步，历经 GLM-130B 开源、ChatGLM 系列对话模型、GLM-4 新一代基座大模型等里程碑，如今已成为国产大模型的代表性力量之一。

（一）产品与服务

智谱 AI **全自研的 GLM 系列大模型**，在 2021 年完成百亿参数模型 GLM-10B 并开源，2022 年发布高精度千亿模型 GLM-130B，2023 年推出千亿基座对话模型 ChatGLM 并两次升级，2024 年推出新一代基座大模型 GLM-4，整体性能大幅提升，2025 年推出 GLM-4.7，主打推理能力，2026 年 6 月，推出 GLM-5.2，支持 1M 稳定上下文窗口。

基于全自研基座大模型的强大能力，智谱 AI 构建了极具竞争力的 AIGC 模型产品矩阵，包括：AI 提效助手智谱清言、高效率代码模型 CodeGeeX、多模态理解模型 CogVLM 以及文生图模型 CogView 等。其中，CodeGeeX 已能每天帮助程序员编写 2000 万行代码，成为业内知名的代码生成工具。

第四波科技智库

在技术定位上，智谱 AI 选择了一条“开源与闭源并行”的差异化路径。它既开源了 ChatGLM-6B 等小参数模型，让大模型开发者的本地微调和部署成为可能，在开源社区受到广泛欢迎；同时又以闭源方式提供高性能的 GLM-4/GLM-5 商用 API 服务，在性能与生态之间找到了平衡点。

（二） 商业模式与定价策略

智谱 AI 践行 **“Model as a Service”（MaaS）** 的市场理念，其商业化路径清晰且多元，形成了“开源引流+MaaS 云端服务+企业私有化部署”的三层结构。

经过近六年的商业化实践，智谱 AI 已在**金融、传媒、政务、汽车、办公、招聘、旅游、能源**等千行百业落地了商业化案例，助力客户实现降本增效与智能化升级。典型合作伙伴包括：

德勤中国：私有部署 GLM 系列大模型，赋能报告生成智能助手，在中文环境下表现超过同类模型，整体员工工作效率大幅提升。

分众传媒：助力开发“众智 AI”营销行业大模型，进行广告语生成、文案创作等，对比 GPT-4，微调后的 GLM 系列模型生成的广告更加简练、准确。

华信永道：在政务公积金场景落地 24 小时智能客服，显著提升了政策解答的准确率和效率。

马蜂窝：打造 AI 旅行伙伴“AI 小蚂”，全链路测评评分 80 分以上，为用户提供个性化旅游服务。

蒙牛：联手打造 MENGNIU.GPT 与“AI 蒙蒙”营养健康助手，提升

第四浪科技智库

用户在私域和公域的消费意愿超 10%。

上汽集团：赋能汽车维修 AI 师傅，单轮对话抽取准确率达 80%，多轮对话修正后可达 90%。

金山办公：推出 WPS AI 智能办公助手，用户满意度高达 95%+。

智己汽车：构建 700+座舱意图支持，识别准确率超 95%，响应速度提升超 30%。

智联招聘：实现简历筛选时间节省超 80%、求职者写简历效率提升 100%、招聘面试效率提升 50%。

在营收表现上，智谱 AI 的 MaaS API 平台在 2025 年 ARR 达到约 17 亿元，同比增长 60 倍，验证了其商业化模式的增长潜力。2026 年，公司实现了累计提价 83%（如 GLM-5.1 发布时再度上调 API 价格 10%），但调用量反而暴增，证明了高端市场对其技术能力和服务价值的认可。

在定价体系上，智谱 AI 遵循“分层定价+差异化服务”的原则。其 API 定价区间为 6-30 元/百万 Token，与国内外主流大模型保持竞争力对齐。公司可同时提供**公有模型服务（MaaS）**，有定制化需求的**专有模型服务**以及**私有化部署**。

在成本控制方面，智谱 AI 通过多项技术手段优化算力效率：采用 MoE（混合专家）架构降低推理成本，GLM-5 模型采用 744B MoE 架构，在激活参数较少的情况下保持高性能；通过量化、蒸馏等技术降低训练成本；持续优化 AI Infra，提升算力利用率。

（三）商业化分析

面对 DeepSeek 掀起的行业降价潮，智谱 AI 并未跟风，反而逆势

第四浪科技智库

提价。这一决策的背后是其对“高端模型稳守溢价”发展战略的坚持。提价后调用量暴增，表明客户对其模型在实际场景中“结果可靠性”的认可，使得智谱 AI 在高阶 K 型市场中站稳了脚跟。

其竞争优势突出表现在全自研 GLM 系列不依赖海外大模型，满足国产化、信创要求，对于政务、金融等对数据安全要求极高的行业具有天然优势。在金融、政务、汽车、办公等千行百业均有成熟案例，形成了可复用的行业解决方案。开源 Chat GLM 系列模型在全球范围内广受欢迎，下载量超 1300 万，构建了庞大的开发者社区，为商业化引流提供了坚实基础。

智谱 AI 在港股上市后市值一度突破 5000 亿港元，体现了资本市场对其技术和商业模式的认可。从未来发展看，智谱 AI 将继续沿着**“开源引流+MaaS 与端侧授权”**的核心商业模式演进。充分展示了独立 AI 公司在大模型商业化中的生存与发展之道——通过极致的开源策略构建开发者生态和品牌影响力，以技术自主可控和中文场景的深度优化建立差异化壁垒，再通过从云端 API 到私有化部署、从软件应用到硬件授权的多元商业化路径实现价值变现。这条路径帮助智谱 AI 在巨头的夹缝中找到了属于自己的增长空间，也为其他独立 AI 公司提供了可参考的商业化范式。

4.3 昆仑万维：“4+3”AGI 战略引领的多产品矩阵商业生态

昆仑万维是中国 AI 大模型商业化领域的一家重要垂直行业公司，其核心优势在于将强大的技术基础与多元化的应用场景紧密结合，并以此构建了面向全球市场的商业化生态。

第四浪科技智库

（一）产品与服务

昆仑万维以通用大模型和垂直行业大模型为技术底座，构建了**“4+3”AGI 战略**。其技术体系覆盖视频模型、音乐音频模型、世界模型、基座文本与多模态模型四大 SOTA 级人工智能模型，并以此支撑 AI 短剧、AI 音乐、AI 游戏三大 AI Native 平台经济体。在产品端，公司已形成短剧和 AI 短剧平台、AI 音乐、AI 游戏、AI SuperAgent、AI 社交五大核心产品矩阵。

在商业化落地与营收结构方面，昆仑万维已取得显著成效。其核心营收来源当前以**C 端订阅及短剧平台流水**为主，**B 端 API 和定制化部署服务**正在快速增长。据公司 2025 年年报显示，全年实现营收 81.98 亿元，同比增长 44.78%；其中，天工 AI 板块收入 23.69 亿元，同比增长 251.74%。短剧及 AI 短剧平台业务作为核心增长引擎，收入达 16.17 亿元，同比增长 864.92%。海外业务贡献突出，占总收入比重由 2025 年的 94.2%进一步提升至 2026 年一季度的 96.8%，全球化布局持续深化。

（二） 商业模式与定价策略

昆仑万维综合考量多项核心指标来评判大模型商业化的成熟度，包括：**营收增速与 ARR**（短剧及 AI 短剧平台 ARR 超 5.7 亿美元）、**C 端付费率与转化率**、**B 端客户数量与客单价增长**（2025 年 API 业务收入同比增长 210%）、**毛利率水平与盈亏平衡节点**，以及**项目交付周期与实施效率**。

目前昆仑万维整体仍处于战略性投入期。2025 年全年归母净利

第四浪科技智库

润为-15.93 亿元。但部分 AI 应用已开始实现正向毛利：2025 年第三季度，公司实现净利润 1.9 亿元，单季度业绩扭亏为盈；同年 11 月，AI 音乐产品 Mureka 在扣减营销费用后，毛利首次转正。2026 年一季度，公司经营活动产生的现金流量净额也实现由负转正，达到 8932.41 万元，标志着公司内生“造血”能力正逐步增强，正从高强度投入期向收获期过渡。

当前，昆仑万维最大的成本集中在**销售费用与市场推广成本**（2025 年为 41.82 亿元，同比增长 81.53%）和**研发投入**（2025 年为 16.76 亿元）两个方面。为应对成本压力并提升盈利能力，公司计划通过**模型结构优化**（蒸馏、量化、剪枝）、**规模效应和供应链议价**、以及**生态开放与协同**等方式进一步压缩算力成本。

在定价模式上，昆仑万维构建了多元化的体系，包括：**API 按 Token 量计费**（如 SkyClaw-v1.0 旗舰版输入 0.5 元/百万 Token、输出 4 元/百万 Token）、**年费/月费会员订阅**、**项目打包定价**以及**阶梯定价**。面对行业降价潮，公司坚持在保证服务水平与研发投入的前提下适时调整，新发布的 SkyClaw-v1.0 定价已具备竞争力。

（三）商业化分析

昆仑万维认为当前客户落地的最大卡点在于**落地适配和集成难度**以及**效果稳定性和可控性问题**。基于对**未来 2-3 年是垂直领域大模型规模商业化落地的关键窗口期的判断**，其对 AI 短剧、AI 音乐、AI 智能体等赛道市场空间持乐观态度，认为头部企业将呈现出多产品协同放量、海外高增长的发展格局。凭借“4+3”战略和全球化布局，昆

第四浪科技智库

仑万维有望在这场商业化浪潮中占据有利位置。

昆仑万维的案例展示了垂直行业公司如何通过**技术驱动、多产品矩阵和全球化市场策略**构建差异化的商业化路径。它不局限于单一的API服务，而是通过将大模型能力与内容创作（短剧、音乐）、社交、游戏等具体场景深度结合，并依托强大的海外运营能力，实现了收入的快速增长和"造血"能力的初步验证，为行业发展提供了有价值的参考。

4.4 科大讯飞："自主可控底座+行业深度应用"战略

科大讯飞是中国人工智能领域的代表性企业，尤其在智能语音与自然语言处理方向拥有深厚的技术积累。在大模型浪潮下，科大讯飞并未盲目追求通用大模型的极致参数规模，而是选择了一条更加务实的商业化路径——**聚焦通用大模型、垂直行业大模型与专项能力模型的差异化布局**，并以此构建商业闭环。

（一）产品与服务

科大讯飞的核心产品为**星火认知大模型**，具备文本生成、语言理解、知识问答、逻辑推理、数学能力、代码能力、多模态等七大核心能力，采用"1+N"的整体布局。其中，"1"是指通用底座大模型（星火大模型）；"N"是指大模型在教育、医疗、金融、汽车、运营商、工业等各个垂直领域的落地应用。

在商业化落地方面，星火大模型已全面渗透至C端软硬件与B端业务。**C端**产品包括讯飞AI学习机、智能办公本、翻译机、AI录音笔、AI会议耳机等，其搭载星火大模型后，在2025年"双11"期间

第四浪科技智库

荣获五大品类销售额冠军，AI 硬件销售额同比增长 30%。截至 2026 年 3 月，讯飞星火 APP 在 Android 端累计下载量已超 4.2 亿，C 端用户基础庞大且活跃度高。**B 端**业务则覆盖教育、医疗、汽车、金融、工业、能源等多个行业，通过提供行业解决方案实现商业化。在营收结构上，基于讯飞开放平台的 AI 平台及授权服务收入达到 12.52 亿元，其中大模型 API 及 MaaS 平台服务收入为 3.85 亿元，同比增长高达 263%。

（二）商业模式与定价策略

目前，科大讯飞大模型业务处于**规模化商业落地的加速阶段**，商业化收入保持增长趋势。公司坚持长期投入与商业回报并重，并采用多元化的考核指标，如商业化收入规模及增速、客户续费率及复购率、单位算力成本下降幅度、Agent 渗透率等来评判商业化成熟度。

在算力成本优化方面，科大讯飞采取多维策略：通过模型架构优化、推理效率优化、国产算力适配、场景化模型部署及智能体协同机制等方式进一步压缩成本。这一策略有效地支持了其稳健的商业化扩张。

在定价体系上，科大讯飞采取多元化的定价模式，包括 API 按量计费、企业年费订阅、私有化部署授权费及项目制解决方案收费。面对行业大模型降价潮，公司认为市场竞争正从“价格竞争”转向“价值竞争”，未来，客户将更关注模型的实际效果、数据安全与合规、行业知识深度及长期服务能力，因此预计客户接受度与留存率将持续改善。

第四浪科技智库

（三）商业化分析

科大讯飞认为当前客户落地转化的主要卡点在于 ROI 评估周期长、业务流程重构难度大、企业数据治理水平参差不齐等。随着行业最佳实践的逐渐成熟，这些问题正在改善。展望未来，科大讯飞判断**未来 2-3 年是垂直领域大模型规模化落地的重要窗口期**。其重点深耕的教育数字化与个性化学习、医疗辅助诊疗与健康管理、智能汽车座舱与车载助手等领域被视为高潜力市场。公司将持续坚持“自主可控底座能力+行业深度应用”的发展路径，推动人工智能向千行百业深度赋能。

科大讯飞的案例展示了垂直行业公司在大模型商业化中的差异化路径：不依赖单一的通用 API 调用模式，而是通过深入理解教育与医疗等传统行业的真实痛点，将大模型能力深度嵌入核心业务流程。这种“行业深度应用”模式，虽然在短期规模化扩张上不及互联网巨头，但其在特定领域建立的行业知识壁垒和客户粘性，为长期可持续的商业化奠定了坚实基础。

第四波科技智库

第五章 趋势与展望——商业化未来的 10 个判断

中国 AI 大模型产业正从“技术验证期”跑步进入“产业落地期”。基于对市场、技术、资本和政策的综合研判，我们对未来商业化的发展趋势提出以下 10 个核心判断：

一、综合竞争取代单点突破。未来的商业化成功，本质上已不是单一的模型能力竞争，而是围绕算力、芯片、数据、能源、资本与生态体系展开的“系统级”综合竞争。单点技术的领先将越来越难以转

化为持久的商业优势。

二、算力成本与供应链风险是最大挑战。大模型的训练与推理高度依赖高性能 AI 芯片。全球芯片供应链的紧张、价格波动以及地缘政治风险带来的断供可能，是悬在所有企业头上的达摩克利斯之剑，将直接影响运营成本、技术迭代和应用推广。

三、市场将根据定价逻辑分层分化。市场竞争演变为“价值战”，将清晰地分化为三层：高端模型市场以智谱 GLM-5.1 为代表，凭借技术壁垒和“结果可靠性”维持高毛利。通用模型市场既不具备高端模型的绝对性能护城河，又在性价比上被普惠模型超越，若不寻求差异化，生存空间将急剧收窄。普惠模型市场以 DeepSeek 为代表，通过技术革新将大模型从“高价服务”推向“可规模化交付的基础设施”，通过极致成本推动技术普及，将开辟巨大的增量市场。

四、B 端垂直行业应用成为初创企业商业化主流。大模型商业化的主战场将坚定不移地转向 B 端企业级应用，尤其是在金融、医疗、制造、政务等垂直行业。深度嵌入核心业务流程、解决行业痛点的定制化解决方案，将贡献市场收入的最大份额。初创公司凭借深厚的行业知识、合规准入能力以及与关键行业客户的深度绑定，避开与通用大模型巨头的正面竞争，转而深耕金融、医疗、法律等特定行业，也可获得可观的市场空间。

五、C 端市场将转向“公共服务化”与“准 B 端化”。通用大模型的 C 端市场将出现明显分化。服务于大众日常简单需求的通用大模型，将成为类似水电的“公共服务”，由国家 AI 基础设施支撑。而服

崇阳设科技智库

务于 C 端专业需求（如编程、高级设计）的通用大模型，其商业模式将更多转向 B 端，通过服务小微企业实现价值变现。

六、投资逻辑将转向商业可行性。AI 大模型代表着这个时代最核心的机遇。对于投资者而言，可以接受某一次具体的投资失败，但绝不能因为犹豫而错过站上这一时代浪潮的机会。但资本市场的逻辑也会随着市场的成熟而转变，虽然 2026 年人工智能领域的投资事件和金额依然庞大，但资本的逻辑最终要从“赢者躺赢”转向商业可行性。缺乏自我造血能力和场景落地能力的初创企业将面临生死抉择。商业化闭环的验证、单位经济效益的健康度、垂直行业壁垒的深度以及应对上述风险的能力，将超越技术参数的短期波动，成为评估标的价值的核心依据。

七、技术迭代与颠覆性风险仍将持续存在。AI 技术本身仍处于快速演进期。若出现非 Transformer 的颠覆性架构，或在某些关键能力上出现突破性进展，可能使现有基于 Transformer 架构的大模型及相关投入面临技术过时风险。此外，当前火热的 Agent、具身智能等技术路径，若其商业化落地进度长期滞后于市场预期，可能导致相关产业发展周期延缓，在短期对产业形成冲击。

八、AI Agent（智能体）开始新的产业“蓝海”。初创企业在该领域的创新集中于让大模型“对话与生成”，工具进化为能够理解复杂指令、规划任务步骤、调用外部工具并完成执行的“自主智能体”。这类技术的商业化天然与企业服务场景深度绑定，因为 Agent 需要与企业的内部系统（如 CRM、ERP、数据库）进行集成以执行实际任务。

第四浪科技智库

该赛道的初创公司普遍采用 SaaS 订阅制或项目制定制化部署的商业模式，通过为企业提供自动化工作流解决方案来获取稳定收入。预计 2033 年全球 AI Agent 市场规模到超过 1300 亿美元。

九、部分产业泡沫将加速破裂，商业化不成功的产品面临快速消亡。AI 大模型进入商业化落地阶段后，市场规律将强力修正行业泡沫，部分商业化不成功的企业不可避免面临淘汰。2026 年上半年已有十几款 AI 应用下架，大厂主动放弃部分产品研发，其中不乏明星产品，这一趋势随着资本市场的兑现压力还会继续强化。AI 大模型行业可能形成大厂主导、少数独立公司并存、垂直小公司细分深耕的格局。

十、商业化未来引爆点有望落在物理 AI。具身智能与人形机器人代表了 AI 与物理世界交互的终极前沿。当期对 AI 大模型应用于物理空间的技术突破集中在高精度运动控制、多模态环境感知与实时决策等方面。该赛道的商业模式尚在早期探索阶段，主要以技术授权、为特定工业或服务场景提供定制化机器人解决方案为主。其高昂的估值反映了资本对技术突破可能催生全新产业的预期，当前商业化大规模落地仍需跨越成本、可靠性、安全性等多重障碍，但未来从数据空间向物理空间应用是商业化落地的必然选择。

第四浪科技智库

第四波
The Fourth Wave

TOX
天成创新

联系电话：13520443026
邮　　箱：fourthwave@vip.163.com