

中国大模型供给与厂商研究

厂商类型划分与差异化布局全解析

概览标签：人工智能、中国大模型

China Large Language Model Industry

中国の大規模言語モデル産業

报告提供的任何内容（包括但不限于数据、文字、图表、图像等）均系头豹研究院独有的高度机密性文件（在报告中另行标明出处者除外）。未经头豹研究院事先书面许可，任何人不得以任何方式擅自复制、再造、传播、出版、引用、改编、汇编本报告内容，若有违反上述约定的行为发生，头豹研究院保留采取法律措施、追究相关人员责任的权利。头豹研究院开展的所有商业活动均使用“头豹研究院”或“头豹”的商号、商标，头豹研究院无任何前述名称之外的其他分支机构，也未授权或聘用其他任何第三方代表头豹研究院开展商业活动。

研究目的与观点摘要

➤ 中国大模型产业正从模型发布密集期进入供给分层与商业化验证阶段，厂商竞争重点由参数规模、榜单表现和单点模型能力，转向推理成本、系统集成、数据治理、复杂场景交付和业务价值兑现。本报告从厂商类型、能力结构、开闭源路线和竞争格局演化四个维度展开研究，重点回答中国大模型供给侧由哪些类型厂商构成，厂商竞争能力如何变化，开源与闭源路线如何影响商业化路径，以及产业成熟后利润和定价权将向哪些节点转移。

1. 厂商能力：厂商能力的核心是否仍是模型参数和通用榜单？

不是，模型能力决定入场资格，工程化交付和场景价值决定商业化上限。

随着头部基础模型能力差距收窄，参数规模、通用榜单和单点交互质量只能决定厂商是否进入客户评估名单，已不足以解释商业化结果的分化。企业客户真正采购的不是单一模型，而是一套能够接入私有数据、嵌入业务系统、满足合规要求并产生可验证ROI的智能化系统。因此，厂商能力正在从“模型能力”转向“模型能力工程化变现能力”。这一能力主要体现在三方面：软硬协同决定单位推理成本、高并发服务效率和算力有效转化率；系统集成决定模型能否进入企业IT、数据和工作流体系；复杂场景交付决定厂商能否在医疗、金融、工业、科研等高价值场景中形成持续收入。

2. 供给格局：当前中国大模型厂商的供给格局如何划分？

中国大模型供给不是简单收缩，而是从厂商数量扩张转向供给结构分层。

中国大模型供给主体可归纳为三类：云厂商、大模型原生厂商和场景生态型模型厂商。这一划分的重点是不同主体掌握的资源禀赋和价值捕获位置不同。云厂商依托算力、云资源、企业客户和开发者生态，更接近模型服务的平台化入口；大模型原生厂商依托模型架构、算法迭代、Agent和多模态能力，更接近技术创新和开发者心智；场景生态型模型厂商则依托既有业务中的数据、场景、终端入口和客户关系，把模型能力嵌入原有业务流程。前沿基础模型受资金、算力、工程和数据门槛约束，将向云厂商、大型平台和少数AI原生头部集中；但开源模型、云托管、私有化部署、微调、蒸馏和端侧适配会降低模型服务供给门槛，推动行业模型、应用模型和场景模型持续扩散。

3. 路线博弈：开源与闭源路线的商业化差异是什么？

开源扩大生态覆盖，闭源保留高价值利润池，中国厂商大概率走向双轨制和混合变现。

开源与闭源的差异，本质上不是技术路线差异，而是价值优先在哪个环节被捕获。闭源路线强调将高价值模型能力保留在自有平台内；开源路线则强调先扩大开发者覆盖、模型装机量和生态影响力，再通过云托管、官方API、私有化部署、调优服务、模型蒸馏和行业交付回收价值。因此，中国头部厂商大概率不会走向纯开源或纯闭源，而是走向“分层开放、混合变现”的双轨制。基础模型、成熟能力或轻量模型开放给社区，以扩大生态覆盖和企业选型入口；旗舰模型、企业级能力、高成本推理服务和平台化能力则留在自有体系内，用于承接高价值调用、企业客户和持续服务收入。

4. 竞争演化：产业成熟后，厂商的竞争结构和利润分配将如何变化？

长期利润不取决于“是否有模型”，而取决于厂商能否占据算力、平台、入口和行业交付等强定价权节点。

产业成熟后，长期利润和定价权不会持续停留在单一通用模型层。随着模型供给增加、开源模型扩散、API价格下探和基础能力差距收窄，通用模型会逐渐从技术稀缺品转为基础能力，模型层单位利润将持续承压。真正能够沉淀长期利润的节点，将向更接近稀缺资源、客户入口和业务结果的位置迁移。算力基础设施掌握高端算力供给和推理成本控制能力；MaaS/云平台掌握模型调用、部署、治理、合规、安全和企业客户关系；应用入口掌握用户流量、交互数据和产品体验；行业解决方案掌握业务流程、行业数据和可验证ROI。

内容目录

◆ 厂商能力与产品体系	5
• 大模型厂商的供给方类型与能力分工	6
• 高并发推理需求下，软硬协同如何影响厂商能力？	8
• 系统集成将如何影响厂商能力？	9
• 厂商向复杂场景延伸的产品能力如何评估？	10
◆ 开源与闭源竞争路线	11
• 厂商开源与闭源的商业化差异是什么？	12
• 厂商开源/闭源战略切换的真实考量是什么？	13
• 开源和闭源模型厂商的竞合关系是怎样的？	14
◆ 供给演化与竞争结构	15
• 高资金和高技术门槛下，模型的供给将怎样演化？	16
• 边缘计算与端侧模型会创造哪些新生态位？	17
• 产业成熟后，利润和定价权将向哪些节点转移？	18
◆ 头豹业务合作介绍	19
◆ 方法论与法律声明	20

名词解释

- ◆ **前沿模型**：指在模型规模、推理能力、多模态能力、代码能力、Agent能力或综合评测表现上处于行业领先水平的模型。前沿模型通常需要较高训练成本、算力资源、工程团队和数据体系支撑。
- ◆ **MaaS (Model as a Service)**：模型即服务。指厂商通过云平台或模型平台向客户提供模型调用、部署、调优、评测、知识库、Agent编排和运维等服务。MaaS是大模型商业化的重要承接方式。
- ◆ **API调用**：指客户或开发者通过接口访问模型能力，并按照调用量、输入输出Token数量、并发规格或服务等级付费。API调用是闭源模型和平台型模型服务的典型收入来源之一。
- ◆ **Token**：大模型处理文本时的基本计量单位，可以理解为模型输入和输出内容被切分后的片段。模型API通常按照输入Token和输出Token数量计费，因此Token价格会直接影响企业使用成本。
- ◆ **MoE (Mixture of Experts)**：混合专家模型。指模型内部包含多个专家网络，在推理时根据任务动态激活部分专家，从而在保持较大总参数规模的同时降低单次推理计算量。MoE常用于提升模型性能与推理效率之间的平衡。
- ◆ **KV Cache (Key-Value Cache)**：注意力机制中的键值缓存。大模型在长文本生成和多轮对话中会缓存历史上下文信息，以减少重复计算。KV Cache管理能力会影响显存占用、长上下文处理能力、并发吞吐和推理成本。
- ◆ **高并发推理**：指大量用户或企业系统在同一时间向模型发起请求，要求模型在稳定延迟和可控成本下持续响应。高并发推理对算力调度、显存管理、芯片互联、推理引擎和平台稳定性提出较高要求。
- ◆ **推理引擎**：指负责模型推理任务执行、请求调度、批处理、显存管理和加速优化的软件系统。推理引擎效率会影响模型响应速度、并发吞吐、资源利用率和企业部署成本。
- ◆ **算力有效转化率**：指厂商将底层芯片、集群和云资源转化为实际模型服务能力的效率。算力有效转化率越高，单位推理成本越低，厂商越能在价格竞争中维持盈利能力。
- ◆ **SLA (Service Level Agreement)**：服务等级协议。指服务提供方对系统可用性、响应时间、稳定性、故障恢复等指标作出的承诺。在大模型企业级部署中，SLA反映模型服务能否满足生产环境要求。
- ◆ **TCO (Total Cost of Ownership)**：总拥有成本。指企业部署和使用大模型过程中产生的全部成本，包括模型调用、算力资源、数据治理、系统集成、运维、安全合规和人员投入等。
- ◆ **RAG (Retrieval-Augmented Generation)**：检索增强生成。指模型在回答问题前先从企业知识库、文档库或外部数据库中检索相关信息，再结合检索结果生成回答。RAG常用于企业知识问答、客服、文档分析和合规场景。
- ◆ **知识内化**：指通过预训练、微调、强化学习或领域数据训练，将专业知识、行业规则、诊疗指南、工艺流程等深度嵌入模型能力中，而不是仅依赖外部提示词或临时检索。

Chapter 1

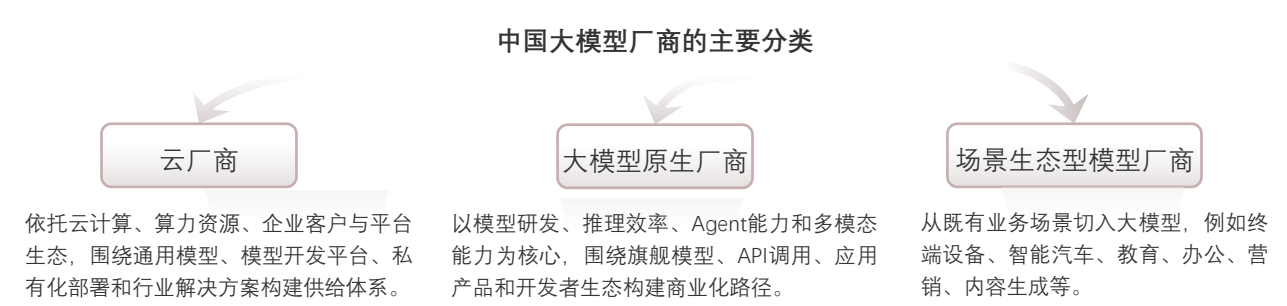
厂商能力与产品体系

- 大模型供给侧已从单一模型厂商竞争，演化为多类型主体共同参与的分层格局。云厂商依托算力、云资源、企业客户和开发者生态，把大模型作为云服务能力的延伸；AI原生厂商依托算法迭代、模型架构、Agent和多模态能力形成技术差异；硬件厂商通过芯片、终端和边缘设备控制模型入口；行业厂商则依托垂直数据和行业 Know-how，推动模型能力进入医疗、金融、工业、教育等场景。
- 厂商竞争的关键正在从“是否拥有模型”转向“能否把模型能力稳定、低成本地转化为业务价值”。软硬协同决定单位推理成本和算力有效转化率，系统集成决定企业落地效率和生产环境稳定性，复杂场景交付决定厂商能否进入高价值业务流程。

大模型厂商的供给方类型与能力分工

- 中国大模型行业正在形成三类厂商的供给格局：云厂商、大模型原生厂商和场景生态型模型厂商

大模型三种典型厂商



- 当前中国大模型供给方按资源禀赋可划分为三大类型。云厂商更接近平台化入口，大模型原生厂商更接近模型与Agent创新，场景生态型模型厂商则通过“模型+场景+数据+渠道”强化原有业务壁垒。

云厂商			大模型原生厂商		
系列	代表模型	特点	系列	代表模型	特点
 Qwen	qwen3.6-plus、qwen3.6-flash、qwen3.6-max-preview	“全栈均衡” 长上下文，代码、Agent、视觉、工具调用、产品线齐全	 智谱·AI	GLM-5.1、GLM-5V-Turbo、GLM-5-Turbo，以及AutoGLM体系	“长程任务+Agent” 关注长时间自主规划、执行、修复、交付
 腾讯混元	hunyuan-2.0-thinking、hunyuan-2.0-instruct	“内容生态+ToB 接口兼容” 把模型做成能直接进入企业工作流程和开发工具链的底座。	 KIMI	K2.6、kimi-k2.5、kimi-k2-thinking	“长上下文+Agent” 关注长上下文、研究分析、代码、Agent。近几代模型在往生产力 workflow 提升
 文心一言	文心5.0、文心4.5 Turbo、X1 Turbo	“原生全模态+企业应用” 文心5.0全模态统一建模，4.5Turbo/X1 Turbo则突出更低成本和更强推理与行业应用结合。	 MINIMAX	MiniMax M2.7、M2.5、Speech 2.8、Music 2.6、Hailuo视频模型	“Coding+多模态” 在语音、音乐、视频、多模态交互上全面铺开
 豆包大模型	豆包大模型1.8、doubao-seed-1.6/1.8等	“产品化+消费级体验” 模型能力关联内容生产、Agent、企业应用平台和大规模产品分发。	 阶跃星辰	step-3.5-flash、step-3、step-r1-v-mini、step-audio-r1.1、step-gui	“多模态+GUI Agent” 关联多模态和操作界面

- 云厂商把模型能力转化为云资源消耗、API调用、企业开发工具和应用生态。大模型的价值是放大云平台的客户黏性和算力消耗。
- AI原生厂商以模型迭代速度、推理成本控制和Agent能力争夺技术心智。它们短期更容易获得开发者和C端用户关注，但长期仍需要在订阅、API收入、Agent产品或垂直场景中证明商业闭环。

来源：企业官网、头豹研究院

大模型厂商的供给方类型与能力分工

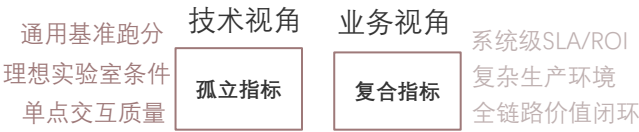
- 中国大模型行业正在形成三类厂商的供给格局：云厂商、大模型原生厂商和场景生态型模型厂商

大模型三种典型厂商（接上页）与业务视角切换

场景生态型模型厂商		
系列	代表模型	特点
	DeepSeek-V4-Pro/V4-Flash、V3、R1	“金融量化背景+通用模型研发” 积累金融量化中的算法与工程能力、高性价比、强工程优化
Xiaomi MiMo	MiMo-V2-Pro/V2-Flash/V2-Omni、MiMo-V2.5	“终端生态+推理模型+多模态” 从轻量推理模型切入，逐步扩展至视觉、音频等多模态能力，服务小米生态
	星火X1.5、星火X2等	“语音、教育” 大模型基于原有的数据积累和场景理解
	日日新V6.7	“视觉、数字人” 强调多模态长思维链、中长视频解析

- 场景生态型模型厂商布局大模型，并不只是参与通用基础模型的正面竞争，而是将既有行业中的数据、算法、算力、场景入口或客户资源转化为模型能力与商业化优势。

在高并发推理需求与业务落地压力下，对厂商能力的关注将从技术转向业务视角的复合指标



- 在“三类厂商”并存的供给格局下，厂商竞争不能只用模型参数、榜单成绩和单次推理成本衡量。云厂商的优势在算力、平台和企业交付；大模型原生厂商的优势在模型迭代、产品创新和开发者生态；场景生态型模型厂商的优势则来自既有业务中的数据、算法、终端入口和客户场景。
- 当Scaling Law边际效益递减，基础模型能力趋同化加速，技术视角的孤立指标将无法解释商业结果的分化。系统工程化交付能力，包括私有化部署周期、复杂生产环境的稳定性保障、全生命周期TCO，正成为B端采购的核心决策依据。这要求评估框架从“单点交互质量”升级为“系统级SLA达成率”，从“理想实验室条件”切换到“复杂生产环境适配度”，从“通用基准跑分”转向“业务场景ROI验证”。
- 同时，缺乏垂直场景工程化封装能力的厂商，在复杂场景交付中将面临成本与场景适配的双重压力。如Kimi、MiniMax的大模型原生厂商在“长上下文+Agent”与“多模态交互”环节已建立算法领先性，但其技术优势向跨学科、重资产场景转化的交付壁垒尚未完全兑现。场景生态型模型厂商（如科大讯飞、商汤）则依托星火X1.5、日日新V6等模型，将通用能力与“语音/教育”“视觉/数字人”等垂直Know-how深度封装，有效规避了基础模型层的同质化价格战。
- 这种业务视角的切换，意味着软硬协同、系统集成和复杂场景交付预计成为厂商能力的关键变量。

来源：企业官网、头豹研究院

高并发推理需求下，软硬协同如何影响厂商能力？

- 高并发推理把大模型竞争推向“模型-芯片-推理引擎-部署架构”的协同效率竞争，软硬协同能力将直接影响厂商的成本、性能和商业化空间

厂商能力——软硬协同

多重压力构成厂商推进软硬协同的动机

内存与带宽压力	能耗压力	互联压力	边缘侧压力	成本压力
---------	------	------	-------	------

- 大模型规模化商用后，算力消耗重心已由集中式训练向高并发推理服务转移。现有芯片架构下，仅靠堆叠算力或扩大参数规模，边际收益已显著递减，并直接推高企业部署的总拥有成本。例如，在同等并发条件，AMD MI300X在推理时的端到端延迟比英伟达H200高出37%至75%，差距主要来自CUDA/TensorRT-LLM等软件栈与推理框架成熟度。英伟达的优势在于CUDA、TensorRT-LLM、NCCL和开发者生态形成了较完整的软件闭环；AMD的挑战更多来自ROCm、vLLM/SGLang适配和工程优化不足。因此，软硬协同能力将影响厂商的算力有效转化率和单位推理成本。
- 厂商提升的路径在于从算法与硬件的独立迭代转向全栈协同优化。云厂商需依托IaaS层资源调度与编译栈优化，提升集群并发吞吐；硬件及终端厂商则通过自研指令集、算子融合与端侧部署，规避通用架构的带宽瓶颈。能否在KV Cache管理、动态批处理与底层驱动间实现技术闭环，将直接决定厂商的算力变现效率与企业级交付门槛。



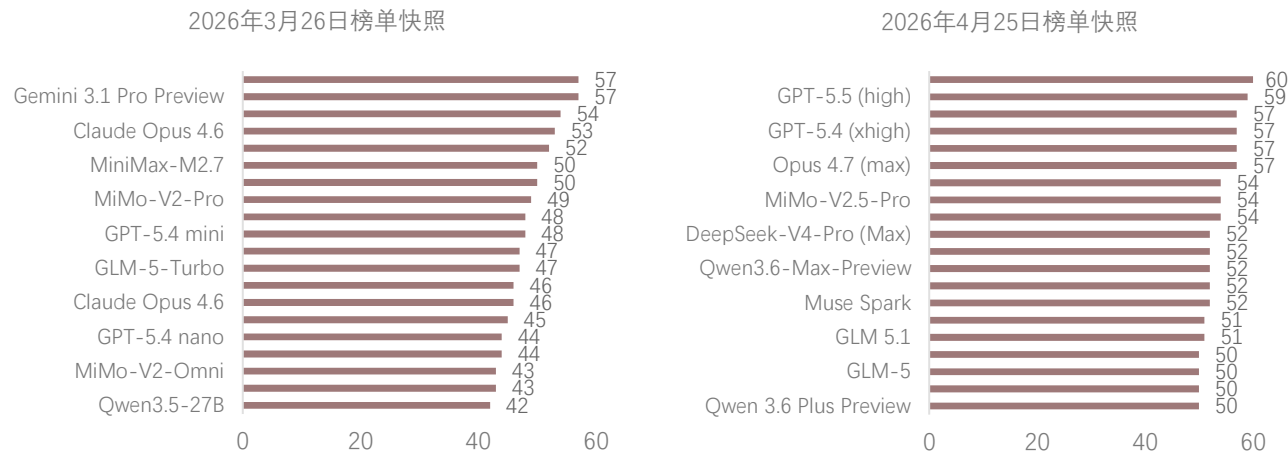
来源：企业官网、大模型技术报告、aimultiple、头豹研究院

系统集成将如何影响厂商能力？

- 模型能力领先的边际差距正在收窄，系统集成成为厂商能力的放大器：谁能更快、更稳、更低成本地把模型嵌入企业流程，谁就更容易形成商业化优势

厂商能力——系统集成

模型能力收敛后，系统集成交付的重要性凸显



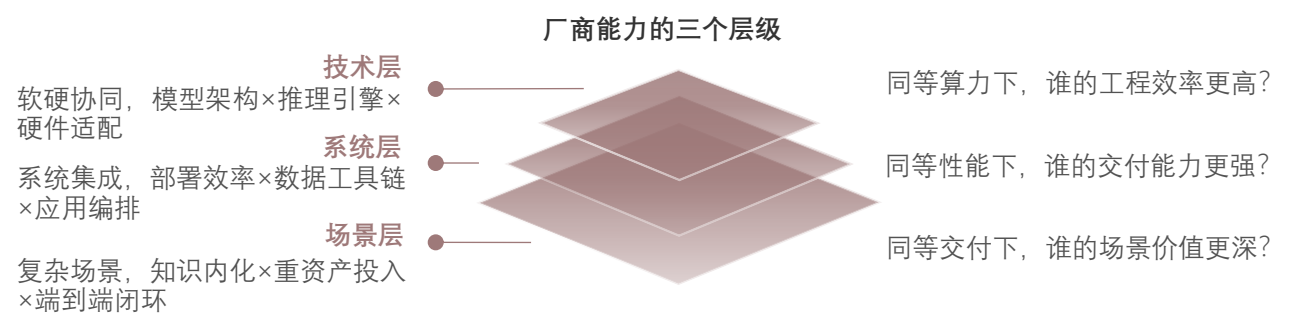
- 注：排行榜选取AA Intelligence Index数据，评估维度包含知识储备、逻辑推理、代码生成和智能体能力四个维度，该榜单反映模型通用能力。
- 从综合能力榜单看，头部模型之间仍存在分数差异，但差距正在缩小，已经从“能不能完成任务”转向“完成任务的稳定性、成本和体验差异”。对于企业客户而言，模型榜单分数只能证明通用能力上限，不能直接证明模型能否在企业内网、私有数据、既有ERP/CRM/OA系统和合规流程中稳定运行。因此，企业采购不再只看模型能力分数，而会进一步评估部署、数据、集成和运维能力。
 - B端客户采购的不是模型，而是一套可上线、可维护、可验收的智能化系统。**2026年Q1大模型落地中标项目数据表明，在全部大模型中标项目中，应用类大模型中标项目占59%，具体是智能审核与分析决策、智能客服与数字人、知识问答与知识平台三个最接近业务结果的类别。模型厂商如果不能承担数据治理、系统对接、权限控制、流程改造、效果评估和持续运维，就很难进入企业核心场景。
- | 厂商类型 | 代表厂商 | 部署效率 | 数据工程链条 | 应用编排能力 |
|-----------|--------------|------------------------|------------------|---------------------|
| 云厂商 | 阿里云/腾讯云/华为云 | 标准API与云上部署效率高 | 知识库+向量DB+内网VPC闭环 | 支持Agent/工作流/代码多模式编排 |
| AI原生厂商 | 智谱AI/MiniMax | PoC和API接入快，私有化部署依赖技术团队 | RAG构建+多模态解析单点突出 | Agent工作流编排能力强 |
| 场景生态型模型厂商 | 商汤/科大讯飞 | 半标准化+行业定制 | 垂直场景数据治理深度封装 | 行业Know-how与工作流程深度融合 |
- 系统集成能力正在成为模型厂商竞争的关键分水岭。模型能力分数决定厂商进入客户评估名单的资格，但部署效率、数据工程、系统兼容、流程嵌入和持续运维，决定模型能否真正进入企业核心业务。云厂商凭借云资源、数据工具链和企业IT入口，在标准化部署和平台化交付上更占优势；AI原生厂商在模型体验、Agent能力和任务创新上更强，但需要补齐交付网络；行业解决方案厂商则依托行业Know-how、客户关系和系统集成经验，在复杂场景中更容易形成长期客户黏性。

来源：企业官网、AA Intelligence Index、界面新闻、头豹研究院

■ 厂商向复杂场景延伸的产品能力如何评估？

- 系统集成解决模型进入业务系统的问题，复杂场景延伸则进一步检验模型能否在高专业、高流程、高责任的场景中稳定创造业务价值

厂商能力——复杂场景



- 软硬协同解决了大模型在物理约束下的推理效率问题，系统集成交付进一步打通了从模型到业务的落地路径。当厂商进入科研、医疗、工业等复杂场景时，复杂场景不是把通用模型接入某个行业，而是要求模型理解行业知识、接入多源数据、遵循业务流程，并在低容错环境下稳定输出。**能否完成这一过程，决定厂商是否具备真正的高价值交付能力。**
- 复杂场景存在三大门槛（1）**专业知识门槛**：场景涉及极深的专业领域知识，通用大模型存在幻觉高、术语生疏、逻辑断层的问题，无法直接作为决策依据。（2）**异构数据门槛**：需要处理PB级的非结构化数据，如天文观测图像、工业质检视频、复杂病历，涉及多模态融合与长序列处理。此外，大量科学和专业领域数据并不公开。（3）**极低容错率门槛**：在严肃场景中，错误的输出可能导致严重后果，如误诊、设备故障等。场景要求模型具备“可解释性”和“确定性”，而非概率性生成。



- 面对上述门槛，**厂商必须实现场景解决方案交付**：（1）知识内化，不再依赖提示工程，而是将行业知识库内化到模型中，构建垂直领域的“专家大脑”。（2）架构编排，引入RAG和Agent，将大模型从“执行者”升级为“调度者”。通过工具调用连接外部业务系统，实现数据流转与多步推理的自动化。（3）全栈交付，提供包含“数据+微调平台+推理引擎+硬件”的一体化方案。不仅解决模型效果问题，更解决私有化部署、低延迟响应和算力成本问题。

案例：百川智能×北京儿童医院——儿科AI医生 知识内化：整合300多位儿科专家经验、高质量病历数据、4万多份指南和专家共识、3800多万份科研文献，形成儿科专科知识体系。 架构编排：支持多轮病情询问、报告解读、诊疗建议生成，并在疑难罕见病多学科会诊中辅助医生决策。 全栈交付：采用“AI医生+真人医生”的双医生模式，在低容错医疗场景中作为辅助决策工具，而非直接替代医生。	案例：科大讯飞×中国石油——昆仑大模型 知识内化：面向勘探开发、炼油化工、装备制造、销售等油气全产业链，沉淀能源化工专业知识。 架构编排：围绕26条业务线、119个业务域构建场景体系，发布100个应用场景；在全波形反演场景中实现正反演全流程10倍以上效率提升。 全栈交付：与中国移动、华为、科大讯飞等合作，形成模型、算力、数据、场景应用和行业系统协同的交付体系。
---	---

来源：企业官网、新华网、中国青年报、人民日报、头豹研究院

Chapter 2

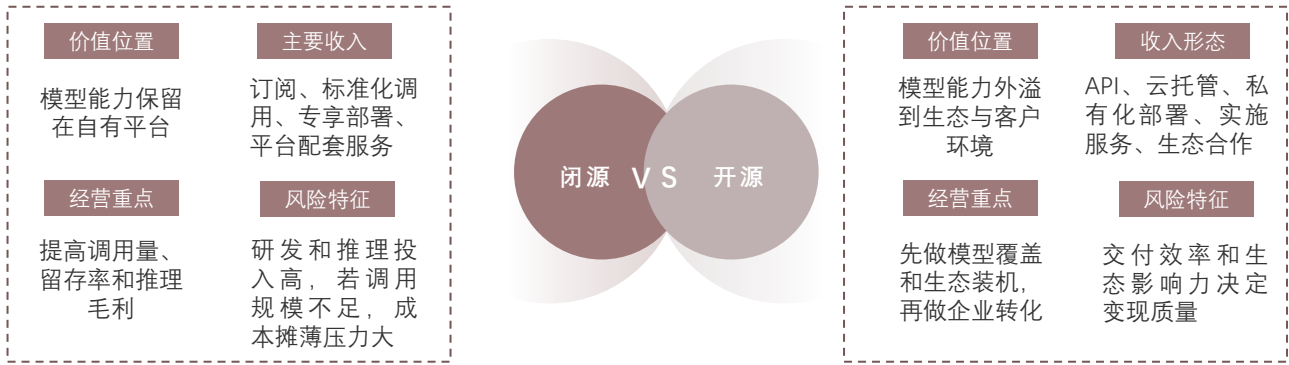
开源与闭源竞争路线

- 开源与闭源不是简单对立，而是模型能力在不同环节实现商业价值的两种路径。闭源路线强调将高价值模型能力保留在自有平台内，通过订阅、API调用、MaaS服务、专享实例、订阅和企业级平台能力持续变现，其核心是保留定价权、客户关系、推理收入和数据反馈。开源路线则强调先扩大开发者覆盖、模型装机量和生态影响力，再通过云托管、官方API、私有化部署、调优服务、模型蒸馏和行业交付回收价值。
- 中国头部厂商将从单一开源或单一闭源，走向“分层开放、混合变现”的双轨制。随着模型能力趋同和API价格下探，单纯依靠闭源通用模型维持高溢价的难度上升；但完全开源也会削弱厂商对高价值调用和企业客户的控制。因此，头部厂商更可能开放基础模型、成熟能力或轻量模型，以扩大生态覆盖；同时将旗舰模型、企业级能力、高成本推理服务和平台化能力保留在自有体系内。

■ 厂商开源与闭源的商业化差异是什么？

- 开源与闭源的商业化差异在于价值捕获位置不同：闭源重在将模型能力在平台内通过订阅或API等方式变现，开源重在生态扩散后通过服务、部署和企业项目变现

开源与闭源厂商的商业化差异



开源与闭源厂商的商业化差异在于模型能力优先在哪个环节实现价值

- 闭源厂商体现在将模型能力保留在自有平台。闭源模式的重点是，把高价值模型能力保留在自有平台和服务体系内，通过标准化调用、专享实例和平台配套能力持续变现。其收入更容易表现为订阅收费、API/MaaS收费、资源专享部署收费、模型调优和配套平台服务收费。
- 开源厂商体现在开放模型权重或降低使用门槛。开源模式先扩大模型使用范围和生态覆盖，再把使用量转化为托管推理、云上部署、私有化落地和企业服务收入。模型厂商本身未必只依赖模型授权收费，更常见的是通过API、云托管、模型蒸馏、企业服务和生态合作来实现价值。
- 从成本结构看，闭源模式的压力主要在模型研发投入、持续推理服务和平台运营，要求厂商把训练和推理成本摊薄到足够大的调用量上；开源模式的压力更多转向生态扩散后的托管、部署、适配和服务交付，要求厂商把模型影响力转成可计费的云资源和企业服务。从收入形态看，闭源模式更容易形成统一定价和持续复购，适合积累标准化API收入、专享部署收入和平台订阅收入；开源模式的收入来源更分散，既包括官方托管调用，也包括私有化部署、行业解决方案、实施服务和生态分成。从规模化看，闭源模式更依赖高频调用和大客户留存，以提升单位算力产出和推理毛利；开源模式更依赖模型采用规模、部署规模或生态适配量，再通过托管推理、专享资源和企业项目完成转化。
- 国内头部平台普遍形成了“标准化模型服务+部署与交付能力”并行的商业结构。以阿里云百炼为例，其提供模型API、调优和资源专享部署；百度千帆定位为企业级一站式平台，覆盖模型服务、应用开发和企业服务；DeepSeek在开源模型之外也提供官方API并按token计费；智谱也提供标准化模型价格体系，同时提供私有实例部署能力，以满足数据安全、性能隔离和资源可控需求。
- 因此，开源与闭源在中国市场对应的不是两套彼此排斥的收入模式，而是两种不同的商业化重心。闭源厂商更强调将高价值调用留在自有云和平台体系内，以标准化调用摊薄研发与推理成本；开源厂商更强调先扩大模型渗透率和生态覆盖，再通过托管推理、专享部署、私有化项目和企业服务回收价值。中国多数头部模型厂商实际采取的是开源和闭源的混合策略：既提供标准化模型服务，也提供专享部署、调优和企业交付。

来源：企业官网、头豹研究院

■ 厂商开源/闭源战略切换的真实考量是什么？

- 对于厂商而言，开源与闭源不是二选一，而是在商业回报、生态扩散和控制权之间的动态取舍；头部厂商走向“部分开源+平台服务+企业交付”的混合策略

厂商切换开源/闭源路径的考量

厂商模式选择与两大锚定点			
厂商模式	核心考量	战略逻辑	典型中国厂商
闭源	保护模型能力和商业控制权	通过订阅、API和MaaS变现，把高价值调用留在自有平台	豆包、Kimi早期模型、MiniMax早期模型
开源	扩大开发者采用和生态影响力	通过开放模型权重或降低使用门槛进入应用框架、开发者社区和企业选型，带动托管API、云资源、私有化部署、微调适配和服务收入	DeepSeek、面壁智能、零一万物、部分Qwen/GLM/ERNIE开源模型
开闭源并行的双轨制	在生态扩散与商业控制间寻找平衡	基础模型、部分旗舰模型或上一代模型开放，最新旗舰模型、企业级能力、高成本推理服务、Agent平台和专属部署能力留在平台内	Qwen、百度文心、智谱GLM、Kimi

财务回报 ROI

ROI决定厂商是否保留模型能力的商业控制权。当模型能力仍能形成调用溢价，且厂商具备平台流量、企业客户或云资源承接能力时，闭源或半闭源更有利。其核心目标不是单纯保护技术，而是保留定价权、客户关系、推理收入和数据反馈。随着模型能力趋同和推理价格下行，单纯依赖闭源API的回报率会下降，厂商会更倾向于开放部分模型，用生态扩散换取更大的转化空间。

生态影响力

生态影响决定厂商是否需要开放部分模型能力。开源模型可以降低开发者试用门槛，进入更多应用框架、Agent工具链、云市场和私有化部署场景。对后发厂商或需要争夺行业标准的厂商而言，开源能够更快抢占开发者心智和企业选型入口；对平台型厂商而言，开放部分模型也可以补强生态位置，但通常不会放弃API、专享部署和企业服务等商业化环节。

厂商选择开源或闭源是对财务回报、生态影响力和商业控制权的动态权衡



阶段 1：从闭源开始，平台化开放使用

2023年8月，文心一言面向全社会开放使用；随后通过C端应用、会员服务和百度智能云千帆平台积累用户反馈、企业调用和商业验证。

阶段 2：商业化压力与生态压力上升

2024-2025年，国内开源模型性能快速追赶。对百度而言，如果继续只依赖闭源平台化路径，虽然有利于控制商业化利润池，但会面临开发者生态弱化和价格竞争压力。

阶段 3：开源ERNIE 4.5系列，补强生态入口

2025年6月底，百度正式开源文心ERNIE 4.5系列模型，支持学术研究和产业应用。



阶段 1：从开源开始，扩大开发者覆盖

Qwen系列长期通过Hugging Face、GitHub等渠道发布开源模型，覆盖语言、多模态、代码等方向。这个阶段用低门槛模型扩大开发者采用和生态覆盖。

阶段 2：百炼API承接标准化和商业化服务

阿里云百炼上线模型API、模型部署和企业服务。其中包含Qwen3-Max、Qwen3.5-Plus、Qwen3.5-Flash等商业模型。

阶段 3：旗舰模型保留平台内，形成分层经营

基础模型和部分能力开放给社区，旗舰模型和企业级能力留在云平台内，用于承接高价值调用和企业客户。

- 因此，开闭源策略调整本质上是对模型能力开放边界和商业控制权的动态调整。厂商通常开放基础模型和部分成熟能力以扩大生态覆盖，同时保留旗舰模型、企业级能力和高成本推理服务，以维持商业化利润池。中国头部厂商走向的不是单一开源或单一闭源，而是分层开放、混合变现。

来源：企业官网、头豹研究院

 **头豹**  www.leadleo.com
LeadLeo 400-072-5588

13

■ 开源和闭源模型厂商的竞合关系是怎样的？

- 开源与闭源厂商的关系不是互相替代，而是分层竞合：在模型层竞争，平台层合作，企业交付层共生，应用生态层再竞争

开闭源模型厂商的竞合关系

层级	竞合关系	开源厂商角色	闭源厂商角色	典型表现
模型层	竞争	提供可替代的基础模型能力	维持旗舰模型差异化	价格下探、性能追赶、模型榜单竞争
平台层	合作	成为平台模型供给	提供分发、算力、计费和服务	百炼、千帆、腾讯云接入第三方模型
企业交付层	共生	降低试用和部署门槛	提供调优、评测、知识库、Agent和运维	模型托管、专享部署、私有化项目
应用生态层	竞争	争取成为任务路由中的低成本模型	争取成为高价值任务和默认平台入口	多模型调度、智能路由、成本优化

厂商在模型层竞争，平台层合作，企业交付层共生，应用生态层再竞争

- 在模型能力层，开源模型与闭源模型是直接竞争关系。随着DeepSeek、Qwen、GLM等开放模型能力提升，企业客户在模型选型时不再只能购买闭源API，也可以选择本地部署、云上托管或使用第三方平台调用开源模型。开源厂商主要以低门槛、可部署、可微调和成本优势争夺通用任务；闭源厂商则需要通过旗舰模型能力、低延迟服务、稳定SLA、企业级安全合规和平台工具链维持差异化。
- 在平台层，开源和闭源模型都是云平台的重要供给。以阿里云百炼为例，其既支持调用自研Qwen系列模型，也支持DeepSeek、Kimi、GLM等第三方模型；百度千帆也在编码、Agent和模型服务场景中接入多类主流模型。云厂商接入开源模型，不一定会削弱自身商业化，反而可能扩大客户覆盖，并带来推理算力、托管部署、统一计费、评测工具和企业服务收入。开源厂商提供模型供给，云平台提供分发、算力和企业级服务，双方共同扩大模型使用量。
- 在企业交付层，开源和闭源模型更多是互补关系。企业客户通常不会只比较模型榜单，而是关注能否稳定部署、能否接入内部数据、能否做权限管理、能否满足合规要求、能否控制成本。开源模型降低了企业试用、私有化部署和国产算力适配门槛，但企业真正落地时，仍需要云平台、系统集成商或模型厂商提供推理服务、调优、评测、知识库、Agent编排和运维支持，这些都需要闭源模型和平台的能力支持。
- 在应用层，开源和闭源模型会进一步被纳入多模型调度。企业不会把所有任务都交给同一个模型，而是根据任务复杂度、成本、响应速度和安全要求选择不同模型。简单任务可以交给成本更低的开源模型或轻量模型；中等复杂任务交给平台主力模型；高复杂度、强合规或高价值任务交给闭源旗舰模型、专享部署模型或私有化模型。这意味着，当企业同时使用开源和闭源模型时，模型竞争会从单模型替代转向多模型组合。
- 因此，开闭源模型厂商的关系不是单纯替代，而是分层竞合。这意味着中国大模型市场的竞争从单模型能力竞争，转向多模型供给、工程化交付和企业入口的综合竞争。

来源：企业官网、头豹研究院

Chapter 3

供给演化与竞争结构

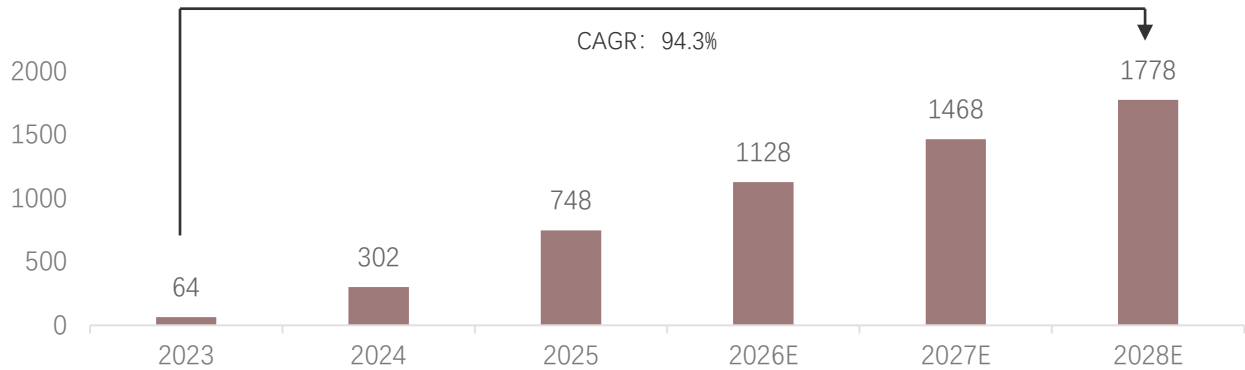
- 高资金和高技术门槛将推动基础模型供给向头部集中，但模型服务供给仍会持续扩散。前沿基础模型训练依赖长期资金投入、算力集群、工程团队和数据体系，重复训练通用基础模型的性价比正在下降，因此供给将向云厂商、大型互联网平台和少数AI原生头部厂商集中。但开源模型、云平台托管、行业微调、蒸馏量化、私有化部署和端侧适配降低了模型服务供给门槛，推动行业模型、应用型模型、边缘模型和端侧模型持续扩散。
- 产业成熟后，利润和定价权将从通用模型层转向算力、平台、入口和行业解决方案。随着模型供给增加、开源模型扩散和API价格持续下探，通用模型层的单位利润将被压缩。相比之下，算力基础设施掌握稀缺资源和推理成本控制能力，MaaS平台掌握模型调用、部署、治理和企业客户关系，应用入口掌握用户流量和交互数据，行业解决方案掌握业务流程和可验证ROI。未来大模型厂商的长期利润，将取决于其能否进入更接近资源、客户和业务结果的价值链节点。

高资金和高技术门槛下，模型的供给将怎样演化？

- 模型供给不会因高门槛而简单收缩，而会从厂商数量扩张转向供给结构分层：基础模型向头部集中，模型服务在云平台丰富，行业场景供给持续增加

大模型厂商的供给趋势：基座集中、服务丰富、场景分散

中国生成式人工智能模型/应用备案数量及预测（单位：款）■ 年末累计备案服务数



- 生成式人工智能备案服务数量继续扩张，供给端在扩容。从备案数据看，截至2025年底，中国生成式人工智能服务累计备案数量增至748款，2025年全年新增446款，预计生成式AI服务供给会继续增加。然而，前沿模型的数量占比非常低且增加有限。根据Stanford AI Index披露的中国前沿模型，2024年为15个，2025年为30个，这表明中国厂商能进入全球前沿模型统计的仍是少数头部厂商和研究机构。由此看出，中国大模型供给呈现“服务数量扩张、基座能力集中”的结构。

中国大模型市场将迈向基础模型供给集中，模型服务供给丰富，模型场景供给分散

基础模型头部厂商供给将集中：承担高成本预训练、多模态、复杂推理和旗舰模型迭代

- 前沿模型训练仍然需要长期资金、算力集群、工程团队和数据体系。从现有模型看，前沿模型的工程门槛持续提高。DeepSeek-V3采用MoE架构，公开技术报告披露其总参数量为671B、每token激活约37B参数；GLM-4.5也采用MoE架构，其总参数量为355B、激活参数约32B；Qwen3系列则覆盖Dense和MoE模型，形成从小参数模型到大参数旗舰模型的梯度供给。这一层需要持续训练投入、推理基础设施、顶尖团队和商业化承接能力，预计主要由云厂商、大型互联网公司和少数模型创业公司占据。

开源模型和云平台层供给将丰富：通过开放权重、API托管和模型市场扩大模型可获得性

- 高门槛会限制前沿训练厂商数量，但不会减少模型供给的多样性。原因是开源模型降低了其他厂商进入模型供给侧的门槛。企业不一定要从零训练基础模型，也可以基于开源模型做微调、蒸馏、量化和部署。同时，云平台会成为模型供给的重要分层：一方面保留自研模型，另一方面接入开源模型和第三方模型，为企业提供统一API、模型部署、调优、评测、知识库和Agent编排能力。这将强化云平台在模型供给中的入口地位。

行业和场景厂商供给将分散：基于通用模型做领域适配，把模型能力转成具体业务交付

- 在高资金和高技术门槛下，大量厂商重复训练通用基础模型是性价比低的路径。在金融、政务、制造、医疗、教育、汽车等领域，企业需要模型理解行业术语、业务流程、知识库、合规边界和系统接口。它们的竞争将会是行业数据、交付能力、成本控制和客户场景理解。不同厂商基于成熟基座做微调、知识增强和系统集成，将促进行业模型数量持续增加。

来源：网信办、斯坦福HAI、头豹研究院

■ 边缘计算与端侧模型会创造哪些新生态位？

- 边缘计算与端侧模型会创造新的生态位，但不是替代云端基座模型，而是把大模型能力下沉到手机、PC、汽车、工业现场和园区场景，形成“云端训练、边缘推理、端侧交互”的新分工

边缘计算与端侧模型的竞争生态



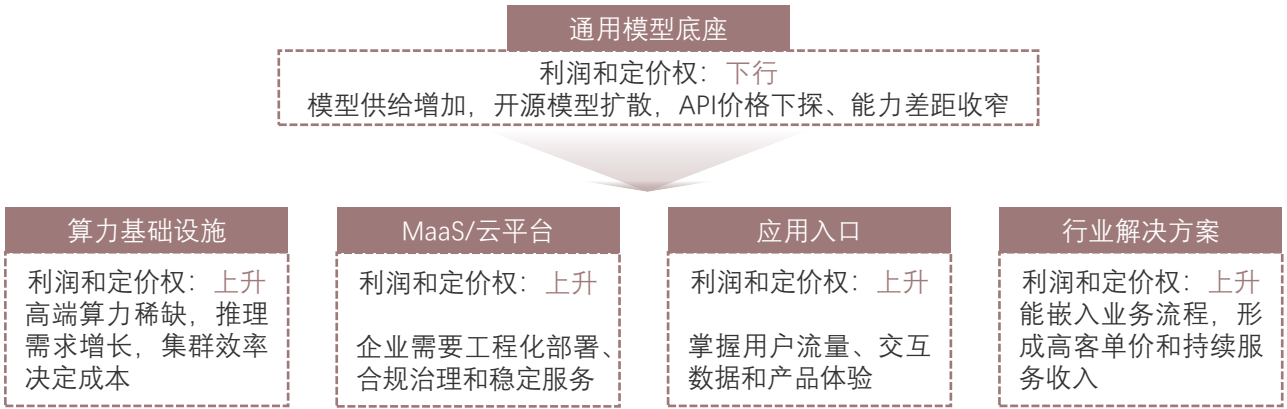
- 对于硬件厂商，现场运行效果将成为边缘与端侧模型的竞争主导因素。工业、交通、园区、零售、安防等场景中，模型需要部署在靠近数据源的位置，依赖边缘服务器、AI盒子、边缘网关和专用推理芯片完成本地推理。以华为Atlas 500 A2智能小站为例，该类边缘AI设备强调低功耗、宽温运行、视频分析、云边协同、模型远程更新和设备统一管理能力。对硬件厂商而言，竞争点不只是芯片算力，而是设备能否在复杂现场稳定运行、远程运维、批量部署，并与云端模型服务协同。
- 对于模型厂商而言，边缘与端侧模型让模型供给从单一旗舰模型转向矩阵。对基础模型厂商而言，端侧与边缘不是替代，而是要求其提供更多尺寸、更低成本和更易部署的模型产品。未来模型厂商不能只发布旗舰模型，还需要提供1B、3B、7B、14B等不同规模模型，以及量化版、蒸馏版、端侧版和行业版，适配不同算力环境。这也是Qwen、GLM、DeepSeek等开源或开放权重模型重要性上升的原因。它们不仅用于云端API，也被企业和开发者用于本地部署、私有化部署和边缘推理。
- 因此，边缘计算与端侧模型会创造新的生态位，但不会改变基础模型层头部集中的趋势。原因是端侧和边缘模型通常依赖云端基础模型蒸馏、压缩和版本更新，难以独立承担前沿能力迭代。它们真正影响的是“云端基座、边缘推理、端侧交互”之间的价值分配。

来源：企业官网、头豹研究院

产业成熟后，利润和定价权将向哪些节点转移？

- 产业成熟后，定价权将从单一模型API转向算力资源、MaaS平台、应用入口和行业交付。模型能力决定入场资格，客户入口和业务结果决定长期利润

产业利润和定价权转向基础设施、MaaS平台、应用入口和行业解决方案



- **单一通用基座模型API的单位利润将承压，模型层会进入成本效率竞争。**中国大模型供给已经快速扩张，同时模型API价格持续下降。阿里、百度、腾讯、字节、DeepSeek等厂商多轮降价，部分模型调用价格降幅达到80%以上。DeepSeek通过低成本模型和开源策略，进一步拉低了市场价格预期。通用模型正在从“技术稀缺品”变成“基础能力”。当模型效果差距缩小后，客户会更关注成本、稳定性和部署便利性，而不是单纯为模型能力支付高溢价
- **算力将从模型厂商的成本项，转变为产业链中的关键定价权来源。**模型降价并不代表算力价值下降。相反，模型使用门槛降低后，调用量和推理需求会进一步上升，算力资源的重要性反而增强。在高端GPU供应受限的背景下，英伟达B300服务器在中国价格达到约700万元人民币，接近美国价格的两倍；同时，租赁价格也达到每月约19万元人民币。能够稳定获得芯片、建设大规模集群、提升资源利用率并降低单位推理成本的厂商，将具备更强议价能力。
- **MaaS平台的定价权通常会强于标准化单一模型API，因为平台控制了模型调用、部署托管、统一计费、权限治理、评测调优和企业客户关系。**企业客户通常不会直接采购单一模型，而是需要一套可部署、可管理、可审计、可合规的AI服务。因此，MaaS平台和云平台将成为模型能力商业化的主要承接节点。云厂商的优势在于算力资源、企业客户、平台工具、安全能力和行业交付体系。阿里云、百度智能云、腾讯云、华为云等厂商可以将大模型与云资源、数据库、安全、办公和行业解决方案打包销售。
- **应用入口和行业解决方案将获得更强的终端定价权，其直接掌握用户、数据和业务结果。**在C端和行业市场，用户并不直接感知底层模型差异，而是感知产品体验和任务完成效果。豆包、元宝、通义千问等AI应用的月活用户增长，说明AI入口竞争已经开始。拥有内容生态、社交入口、电商场景和搜索入口的厂商，可以更低成本获得用户，并沉淀交互数据。B端方面，金融、制造、能源、医疗等行业客户更关注模型能否嵌入业务流程，真正具备付费能力的知识管理、客服质检、风控审核等具体解决方案。
- **总的来说，大模型产业成熟后，通用模型供给增加和API价格战会压缩模型层单位利润，模型厂商必须依靠规模化调用和推理成本优化维持盈利。**相比之下，算力基础设施、MaaS平台、应用入口和行业解决方案更可能成为利润沉淀节点。未来厂商能否获得长期利润，取决于能否掌握价值链中更强定价权的位置，并把这种定价权转化为可持续的收入和毛利。

来源：企业官网、Reuters、头豹研究院

头豹业务合作

全球视野 · 本土洞察 · 助力企业把握市场先机

核心业务



行业数据 API

开放原创报告与研究数据接口，
支持企业知识库、系统平台及AI
应用高效接入和调用



KNIT解决方案

构建企业可信内容体系，提升品
牌在AI搜索与问答中的可见度、
准确性与转化效果



报告会员账号

可阅读全部原创报告和百万数据，
提供PC及移动端，方便触达平台
内容



定制报告/白皮书

对产业及细分行业进行现状梳
理和趋势洞察，输出全局观深
度研究报告



商业尽调

面向投资并购和商业决策，评
估标的公司的商业前景、价值
及风险



招股书引用

研究覆盖国民经济19+核心产
业，内容可授权引用至上市文
件、年报

业务咨询



客服电话：
400-072-5588



官方网站：
www.leadleo.com

报告作者



袁栩聪
首席分析师



廖子烨
行业分析师



service@leadleo.com

办公地点



深圳办公室

广东省深圳市南山区粤海街道华
润置地大厦E座4105室
邮编：518057



上海办公室

上海市静安区南京西路1717号
会德丰国际广场2701室
邮编：200040



南京办公室

江苏省南京市栖霞区经济开发
区兴智科技园B栋401
邮编：210046

2026年 AI环境下 营销最佳应用实践

企业征集正式启动

聚焦AI驱动下的营销创新实践，遴选年度标杆案例与优质服务商

两大征集方向



AI环境下 最佳营销应用实践案例

面向具备AI营销落地实践的企业案例，
关注真实应用、业务增长与可量化价值

覆盖行业



零售快消



跨境电商



房产家居



泛互联网



教育培训



汽车出行



文旅



工业制造



金融



医疗健康



AI环境下 最佳营销服务商TOP10

面向AI营销相关服务机构，关注AI能力、
解决方案成熟度、市场表现与成长潜力

覆盖对象



AI营销服务商



MarTech厂商



广告营销技术平台



内容智能化服务商



数据智能服务商



客户运营服务商

重要时间



一轮申报截止

2026年6月22日



二轮材料提报截止

2026年7月3日



报告发布

2026年8月5日

扫码参与征集



✉ 邮箱: oliver.yuan@leadleo.com

🌐 官网: www.leadleo.com

☎ 电话: 400-072-5588

方法论

- ◆ 头豹研究院布局中国市场，深入研究19大行业，持续跟踪532个垂直行业的市场变化，已沉淀超过100万行业研究价值数据元素，完成超过1万个独立的研究咨询项目。
- ◆ 研究院依托中国活跃的经济环境，研究内容覆盖整个行业的发展周期，伴随着行业中企业的创立，发展，扩张，到企业走向上市及上市后的成熟期，研究院的各行业研究员探索和评估行业中多变的产业模式，企业的商业模式和运营模式，以专业的视野解读行业的沿革。
- ◆ 研究院融合传统与新型的研究方法，采用自主研发的算法，结合行业交叉的大数据，以多元化的调研方法，挖掘定量数据背后的逻辑，分析定性内容背后的观点，客观和真实地阐述行业的现状，前瞻性地预测行业未来的发展趋势，在研究院的每一份研究报告中，完整地呈现行业的过去，现在和未来。
- ◆ 研究院密切关注行业发展最新动向，报告内容及数据会随着行业发展、技术革新、竞争格局变化、政策法规颁布、市场调研深入，保持不断更新与优化。
- ◆ 研究院秉承匠心研究，砥砺前行的宗旨，从战略的角度分析行业，从执行的层面阅读行业，为每一个行业的报告阅读者提供值得品鉴的研究报告。

法律声明

- ◆ 本报告著作权归头豹所有，未经书面许可，任何机构或个人不得以任何形式翻版、复刻、发表或引用。若征得头豹同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“头豹研究院”，且不得对本报告进行任何有悖原意的引用、删节或修改。
- ◆ 本报告分析师具有专业研究能力，保证报告数据均来自合法合规渠道，观点产出及数据分析基于分析师对行业的客观理解，本报告不受任何第三方授意或影响。
- ◆ 本报告所涉及的观点或信息仅供参考，不构成任何投资建议。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。在法律许可的情况下，头豹可能会为报告中提及的企业提供或争取提供投融资或咨询等相关服务。本报告所指的公司或投资标的的价值、价格及投资收入可升可跌。
- ◆ 本报告的部分信息来源于公开资料，头豹对该等信息的准确性、完整性或可靠性不做任何保证。本文所载的资料、意见及推测仅反映头豹于发布本报告当日的判断，过往报告中的描述不应作为日后的表现依据。在不同时期，头豹可发出与本文所载资料、意见及推测不一致的报告和文章。头豹不保证本报告所含信息保持在最新状态。同时，头豹对本报告所含信息可在不发出通知的情形下做出修改，读者应当自行关注相应的更新或修改。任何机构或个人应对其利用本报告的数据、分析、研究、部分或者全部内容所进行的一切活动负责并承担该等活动所导致的任何损失或伤害。