

2026年中国大模型行业洞察： 从Token爆发到商业兑现——需求增长、 调用演进与AI应用新趋势

2026 China Large Language Model Industry

中国大型言語モデル産業

概览标签：人工智能、大模型

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

➤ 中国大模型市场的需求侧竞争，进一步聚焦有效调用、预算承接、场景价值密度与长期商业化质量。本报告围绕需求提出方、模型调用方与预算承担方之间的结构性分离，分析调用增长的场景来源、客户分层、ToC与ToB价值差异、开源与闭源选型，以及头部厂商竞争优势的形成机制，重点判断中国大模型调用增长如何在流量扩张基础上完成收入兑现与经营质量验证。

1. 需求结构：中国大模型需求由谁产生、由谁调用、由谁付费？

中国大模型需求侧的关键，是需求能否被转化为可计费、可预算、可续费的有效调用

中国大模型需求是由需求提出方、调用执行方和预算承担方共同构成。个人用户、企业员工、开发者和平台终端用户可以产生高频调用，收入质量却取决于预算是否能够被个人订阅、企业采购、平台代付或云资源消费承接。企业与平台型场景中，使用者、系统接入方和预算审批方经常分离，厂商需要同时解决任务价值、系统集成、安全治理和采购决策问题。

2. 调用增长：模型调用增长来自哪些场景，增长传导路径如何形成？

调用增长是从入口触发、场景激活、有效调用、付费转化到组织扩散的连续链条

调用增长主要来自搜索、交易、客服、办公、代码、内容生产、内部运营、行业垂直和端边设备等高频入口。入口决定调用规模，任务结果决定价值密度，定价机制决定收入确认方式，组织渗透决定增长持续性。ToC场景更依赖流量入口、产品留存和广告交易转化；ToB场景更依赖流程嵌入、ROI验证和预算归集；开发者与平台型客户则通过API、SDK、插件、模型路由和生态分发放大调用规模。

3. 商业化分层：ToC、ToB、开发者与平台型客户的价值密度有何差异？

高价值调用通常接近业务闭环、交易闭环或生产闭环，泛聊天和低门槛问答难以单独支撑高质量收入

消费者覆盖面最广，但单次调用的预算深度有限；企业客户采购链路更长，场景价值密度和续费稳定性更强；开发者不是最大付费主体，却能够通过产品封装和应用外溢放大调用需求；平台型客户掌握流量入口和二次分发能力，更容易形成批量调用和生态绑定。ToC的重点在于入口占位、习惯养成、数据反馈和多元变现，ToB的重点在于客服、代码、合同审核、经营分析和工作流自动化等可量化替代场景。

4. 厂商优势：为什么有些厂商做得更好，核心优势来自场景、产品、模型还是生态？

大模型厂商的长期竞争优势，将集中在场景占位、产品化转化、模型成本效率和生态控制力的综合匹配

做得更好的厂商，通常是在“调用—有效调用—付费—续费—毛利”链条中打穿关键瓶颈。场景决定厂商能否获得高价值需求，产品决定需求能否转化为稳定调用，模型决定结果质量、能力边界和单位成本，生态决定获客效率、迁移成本、复用半径和毛利修复空间。短期优势可能来自先发入口、补贴价格或流量分发，长期壁垒更依赖高价值场景占位、工作流嵌入、数据闭环、云与开发者生态、行业交付能力和成本控制能力。

内容目录

◆ 需求主体与调用增长机制	-----	5
• 中国大模型需求由谁产生、由谁调用、由谁付费？	-----	6
• 模型调用增长来自哪些场景，增长传导路径如何形成？	-----	8
• 消费者、企业、开发者与平台型客户的需求结构差异	-----	9
◆ ToC、ToB与高价值调用场景分层	-----	11
• ToC 场景的调用增长与消费级付费逻辑	-----	12
• ToB 场景的调用深化与企业级商业化逻辑	-----	13
• 哪些调用属于高频高价值，ToC 与 ToB 的价值密度有何不同？	-----	14
◆ 市场路径、模型选型与厂商竞争优势	-----	15
• 中国与海外市场在需求、付费、落地路径上的差异，以及开源/闭源选型逻辑	-----	16
• 为什么有些厂商做得更好，核心优势来自场景、产品、模型还是生态？	-----	17
◆ 头豹业务合作介绍	-----	19
◆ 方法论与法律声明	-----	20

名词解释

- ◆ **Token（词元）**：大模型处理信息的最小计量单位，可对应汉字、词语、标点、英文片段或代码片段。Token 是模型 API 计费、推理成本核算和调用强度评估的基础。
- ◆ **模型调用**：用户、企业系统、开发者应用或平台产品向大模型发起请求，并由模型完成理解、生成、推理、检索或工具调用的过程。调用规模决定算力消耗，但不必然等同于商业价值。
- ◆ **场景价值密度**：单次调用或单个应用场景能够创造的商业价值强度，主要体现为降本、增效、增收、风控或续费能力。任务边界清晰、结果可衡量、流程可嵌入的场景通常具备更高价值密度。
- ◆ **MaaS（Model as a Service）**：模型即服务，指模型厂商或云平台通过 API、控制台、模型网关、工作流编排和专属推理资源，将大模型能力标准化输出给企业、开发者和平台型客户。
- ◆ **API 调用**：客户通过标准化接口向模型平台发送请求并获取结果的使用方式。API 调用通常按输入 Token、输出 Token、上下文长度、并发能力和实时性进行计费。
- ◆ **Agent化**：大模型从单轮问答工具升级为能够理解目标、拆解任务、调用工具、访问知识库并执行多步骤流程的智能体形态。Agent 化使模型能力从内容生成进一步进入任务执行。
- ◆ **RAG（Retrieval-Augmented Generation）**：检索增强生成，指模型生成答案前先从企业知识库、文档系统或数据库中检索相关信息，再结合检索结果完成回答。RAG 主要用于降低幻觉风险、提升答案可追溯性。
- ◆ **Batch收费**：面向低实时性、大批量任务的异步推理计费模式。客户批量提交摘要、分类、抽取、评测等任务，由平台后台处理后返回结果，通常以较低价格换取较低实时性。
- ◆ **开源模型**：模型权重、代码或部分能力向开发者和企业开放的大模型。开源模型更适合成本敏感、数据敏感、需要本地部署或深度定制的场景。
- ◆ **闭源模型**：模型权重和核心训练体系不对外开放，客户主要通过 SaaS、API、企业版或专属托管方式使用的大模型。闭源模型更适合能力要求高、上线速度快、服务稳定性要求高的场景。
- ◆ **混合架构**：企业或平台同时使用开源模型与闭源模型，并按任务价值、数据敏感度、成本要求和能力上限进行分层调用的技术架构。
- ◆ **私有化部署**：客户将模型部署在本地数据中心、专有云、行业云或企业 VPC 环境中，以满足数据不出域、安全审计、权限控制和合规监管要求。

Chapter 1

需求主体与调用增长机制

- 中国大模型需求侧的核心矛盾是“需求提出方、调用执行方、预算承担方”经常分离

个人用户、企业员工、开发者和平台终端用户可以触发调用，但真正决定收入质量的是谁承担预算、谁控制系统入口、谁负责长期续费。

- 调用增长需要完成从任务痛点、模型调用、业务结果到预算支付的连续转化

仅完成调用转化，更多属于流量增长；进入业务结果验证，才具备有效调用基础；进入预算体系，才能支撑持续收入。

- 不同客户类型的增长机制不同

ToC依赖高频入口和产品留存，ToB依赖流程嵌入和ROI验证，开发者承担调用外溢功能，平台型客户通过流量入口和二次分发放大模型消耗。

中国大模型需求由谁产生、由谁调用、由谁付费？

- 大模型需求源于个人、企业、开发者和平台的任务缺口，调用消耗发生在终端用户、企业系统、API和平台入口，收入质量取决于预算承担方对有效业务结果的持续支付能力

需求、调用与预算的主体拆解

01 需求由谁产生：大模型需求本质上来自“任务缺口”

- 个人用户希望降低搜索、写作、学习、创作和决策成本；
 - 企业业务部门希望减少客服、销售、运营、审核、研发等环节中的重复劳动；
 - IT 部门希望在安全、权限、日志、数据治理和系统接入层面建立可控架构；
 - 开发者希望用较低成本调用模型能力并封装进产品；
 - 平台型客户希望把模型嵌入既有流量入口、软件产品或硬件终端，通过再分发扩大自身产品价值；
- 需求侧真正需要拆解是每一类需求能否完成三次转化；从任务痛点转化为模型调用，从模型调用转化为业务结果，从业务结果转化为预算支付。如果只停留在第一步，调用增长更多表现为流量和活跃；如果进入第二步，模型才开始产生有效产出；只有进入第三步，调用才具备收入质量和续费基础。

02 调用由谁产生：调用执行方决定成本消耗，需求提出方决定业务价值

- 调用增长并不等同于需求增长。需求提出方定义“为什么使用模型”，调用执行方决定“模型被如何消耗”。由于大模型每一次推理都会带来 Token、算力和延迟成本，调用执行方的变化会直接影响厂商成本结构和客户单位经济模型。
 - 个人用户在教育端输入问题，企业员工在业务系统中发起查询，开发者通过API调用模型，平台型客户将终端用户请求转发至模型服务，这些行为共同构成调用增长。调用执行方决定了成本发生的位置，包括输入长度、输出长度、上下文规模、调用频次、并发峰值和实时性要求，进而影响模型厂商的算力调度、计费方式和毛利水平。
- 调用执行方并不决定调用价值。需求提出方决定任务是否具备业务价值，调用执行方决定成本消耗规模；以企业知识库问答为例：

 业务部门提出“降低内部查询成本”需求

→

 员工在办公系统中高频调用模型

↺

 IT部门负责系统接入与权限治理

↻

 预算可能由数字化部门或业务部门共同承担

 - 该场景中，员工是调用执行方，但调用价值来自知识检索效率提升、重复咨询减少和组织知识复用。若调用只是泛聊天、开放式陪伴或低门槛创意生成，调用执行方可以带来大量消耗，但单次调用难以进入可核算收益链条，收入承接能力弱于生产性场景。

来源：头豹研究院

中国大模型需求由谁产生、由谁调用、由谁付费？

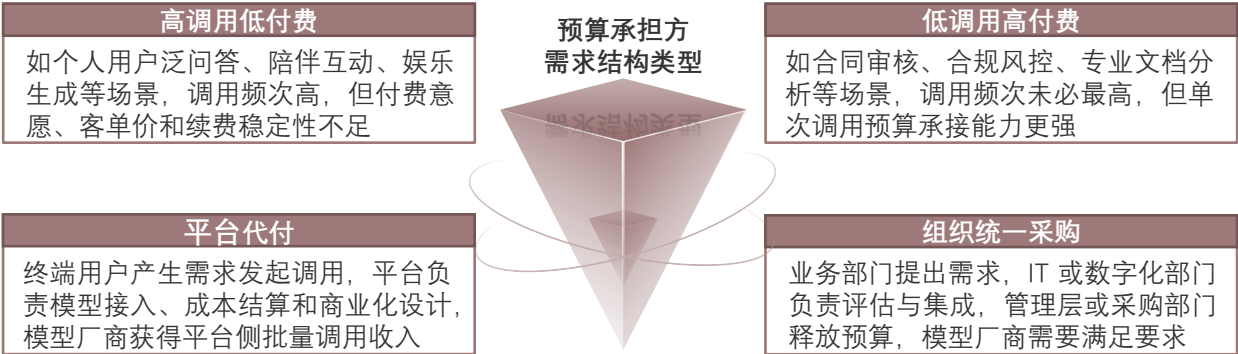
- 大模型付费主体应按预算归集能力识别；个人订阅覆盖轻量需求，企业采购与平台代付承载高质量收入，预算能否稳定进入结算与续费体系决定调用增长的商业化上限

付费主体与收入质量差异

03

付费由谁承担：预算承担方决定收入质量和客户经营方式

- 预算承担方是大模型商业化链条中最关键的角色，因为其决定调用能否从使用行为转化为收入；
- 个人用户付费以会员订阅、增值功能、算力包和内容权益为主，收入质量取决于留存、转化率和付费深度。
 - 企业客户付费以项目制、年框、私有化部署、API按量计费 and SaaS 订阅为主，收入质量取决于采购周期、系统嵌入深度和续费稳定性。
 - 平台型客户付费以API批量采购、平台代付、流量分成和生态绑定为主，收入质量取决于终端流量、调用留存和成本转嫁能力。



- 大模型市场核心矛盾是需求是否被转译成可计费调用
- 6.02亿用户和140万亿日均 Token证明需求已经真实存在，模型服务和调用模型能力的应用层正在快速分化，意味着越来越多调用是由应用和系统代为完成。
- 中国市场的“真正客户”是调用组织者和预算承担者
- 在消费侧，个人用户是需求提出方，但平台可以通过交易、商家工具和生态服务间接承接成本与收益；在企业侧，终端员工只是使用者，预算真正掌握在 IT、云、采购或业务负责人手中；在开发者生态中，开发者决定模型进场，但长期收入通常来自其后面的企业客户或平台客户。
- 三者分离最强的是企业和平台型场景
- 企业与平台更容易形成“需求提报-系统调用-统一结算”标准化链条，且容易用专属部署、保障型资源、预置吞吐、按量后付费等方式把模型能力纳入预算管理。

来源：头豹研究院

模型调用增长来自哪些场景，增长传导路径如何形成？

- 模型调用增长从“可试用”进入“必须用”，优先发生在搜索、电商交易、客服、办公、代码、内部运营、行业垂直和端边设备等高频、可重复、结果可核算的既有入口

模型调用增长的高频入口来源

场景来源与入口结构			
真正驱动调用增长的核心因素是模型能否嵌入已经存在、高频发生、可重复验证的行为入口 从中国市场当前需求结构看，大模型调用主要集中于六类场景，不同场景对应不同的增长速度、付费机制和价值密度。			
场景类别	典型需求	主要触发入口	频次/价值属性
ToC 消费	搜索问答、导购下单、写作、图像/视频生成、学习、陪伴	产品内嵌搜索、推荐流、聊天入口、会员权益、交易按钮	高频；若连接广告或交易则中高价值，脱离交易多为低价值
ToB 企业	客服、知识问答、办公写作、报表分析、代码、审批工单	SaaS 集成、企业 IM、协同办公、工作流自动化、知识库	中高频率高价值
开发者/平台			
内部工具/运营	■ 报告完整版/高清图表或更多报告：请登录 www.leadleo.com ■ 如需进行品牌植入、数据商用、报告调研等商务需求，欢迎与我们联系 联系邮箱：service@leadleo.com		

- ☐ ToC 场景是最容易放大调用量的来源，但并不天然等于高价值
只有当模型能把回答直接转化为搜索停留、广告点击、交易下单、内容生产时，调用才具备商业闭环。
- ☐ ToB 场景是价值密度最高的增长源
企业会为减人、提效、增收、控险、合规付费，因此企业调用的增长往往先从客服、知识库问答、文档处理、数据分析、代码开发等低风险场景起步，再进入审批、经营分析和自动化流程，腾讯ADP的产品能力说明清楚展示了这种路径。
- ☐ 开发者/平台型客户决定了增长的“外溢速度”
如GitHub Copilot，把开发者调用从“包年 seat”转成“按 AI Credits 消耗”，平台客户一旦进入生产环境，定价逻辑会迅速从订阅制转向使用量制。开发者生态的核心作用是把单一模型能力标准化、分发化、计量化。
- ☐ 内部工具/运营场景是最容易被低估但经常最先跑出体量的调用来源
其头部平台已经将大模型用在搜索、广告、内容审核和内部办公中；
- ☐ 行业垂直化与边缘/嵌入式场景体量不一定最大，但决定长期壁垒
HiLens的端云协同案例中，此类场景的核心诉求集中于低时延、高精度、远程更新，单轮聊天体验并非主要评价维度。

来源：头豹研究院

模型调用增长来自哪些场景，增长传导路径如何形成？

- 模型调用增长的关键在于入口触发真实任务后，依次完成场景激活、使用深化、价值验证、付费转化和组织扩散，只有穿透该链条的调用量才能转化为可持续收入

模型调用增长的链路传导机制



□ 调用增长是由入口触发、能力验证、定价结算、预算归集和组织扩散共同构成的传导链条

链条能否成立，关键在于模型调用能否从高频入口进入真实任务，再通过效果验证进入预算体系，并在组织内部形成持续扩散；任一环节中断，前端释放的调用量都可能转化为成本压力。

□ 入口决定调用规模，能力决定价值密度，定价决定收入确认方式，组织渗透决定增长持续性

中国市场在2025年至2026年推理侧加速，主要源于AI被嵌入搜索、交易、办公和工业流程，Token逐步承担可计量、可结算的商业单位功能，预算主体也从早期试用者转向业务部门和平台型客户。

□ 入口解决调用量问题，价值解决付费转化问题

只有当模型既能进入高频场景，又能推动点击、下单、流程完成或效率提升，调用才具备收入化基础。例如，搜索和即时零售本身具备高频流量属性，模型若能将回答延伸至点击或交易动作，调用频次和商业价值将同步提升；在企业流程中，模型则必须先达到可接受的准确率、稳定性和流程完成率，才可能进入财务预算和长期采购体系。

来源：头豹研究院

消费者、企业、开发者与平台型客户需求结构有何差异？

- 消费者贡献最广泛的调用入口，企业与平台型客户承担更高质量预算转化，开发者放大模型能力的外部分发半径，四类客户需求差异本质上由单次价值密度、系统嵌入深度、合规要求和采购复杂度共同决定

需求结构差异产生的机制

消费者和开发者都可能是高频群体，但两者高频的含义完全不同							
消费者的高频来自碎片化需求；开发者的高频来自构建与调试；平台型客户的高频则来自下游分发。大企业与平台型客户未必是“最高频用户”，但一定是更容易形成高价值调用池的客户。其原因是更容易把调用嵌入到既有流程、权限和预算体系里。中小微企业是最适合标准化打包的客户，其核心特征是对复杂实施、长周期交付和高运维成本的承受能力较低。							
客户类型	目标/任务类型	单次价值密度	偏好商业模式	集成复杂度	合规/安全/隐私要求	定制化需求	技术能力与采购流程
ToC 消费者	问答、学习辅导、写作、图片/视频创作、陪伴等	1-2/5	免费增值、会员订阅、按功能包、广告/交易导流	低	低	低	技术能力低；采购基本无链路，转化靠产品体验
ToB 中小微	客服、营销内容、商品描述、票据/文档处理等	3/5	按席位、包月、轻量项目、套餐封顶	中	中	中	技术能力中低；1-3 个决策人，采购周期短
ToB 大企业	代码辅助、合同/合规审核、风控、BI、工作流Agent等	4-5/5	企业订阅、项目制、私有化/专属资源、按调用/按价值混合	高	高	高	技术能力中高；多部门决策，周期长但扩张深
开发者	编码、原型验证、AI-native 应用后端、Agent 编排等	2-4/5	免费额度、按量计费、预付 credits、批处理折扣	中高	中	中高	技术能力高；采购摩擦低，切换快
平台型客户	内嵌Copilot、客户Agent、工作流自动化、渠道分发等	3-5/5	按席位+按动作/会话等；分层折扣与承诺消费常见	很高	很高	很高	技术能力高；产品、架构、财务、法务共同决策

差异产生的机制

成本结构决定了谁能为单次调用付高价

- 成本结构决定单次调用的付费上限；消费者市场并非缺乏需求，而是单次调用的可计费价值较薄。问答、陪伴、灵感生成主要产生体验价值，较难直接对应经营收益，因此ToC更依赖低推理成本、高留存和间接变现。企业、开发者工具和平台嵌入型产品则不同，单次调用可对应人力节省、周期缩短、错误减少或转化提升，预算承接能力更强。招股书显示，2025年前九个月AI-native products 毛利率仅为4.7%，Open Platform与其他AI企业服务毛利率达69.4%，核心差异在于调用是否进入可预算化的工作负载。

商业模式是客户差异的具体体现

- 需求是“清晰动作、清晰结果”，就更适合按 action、conversation或按价值收费；如果需求是“强治理、强预测”，就会出现席位+额度、承诺消费、资源包、Batch、缓存等混合定价。中国模型平台已进入分层计费阶段：以中国公开价格为例，阿里云百炼qwen-max为输入2.4元、输出9.6元/百万Token，并提供Batch半价、上下文缓存折扣和节省计划；百度ERNIE4.5Turbo价格为输入0.8元、输出3.2元/百万Token，则按上下文长度做阶梯计费并给出免费额度。海外闭源平台也在高性能模型基础上叠加Priority、Fast/Flex、Regionalprocessing等服务溢价；商业模式差异本质上是客户结构差异产品化后的结果。

来源：头豹研究院

Chapter 2

ToC、ToB与高价 值调用场景分层

- ToC场景具备高频调用优势，但价值密度受单用户付费深度约束

搜索问答、学习、陪伴和内容生成容易放大调用量，但只有连接创作成果、交易行为、会员权益或广告分发，才具备更强商业化承接能力。

- ToB场景的价值密度高于ToC，核心来自流程替代和结果可验证

客服、代码、合同审核、知识库、BI和工作流自动化能够对应人工节省、效率提升、风险降低和经营改善，更容易进入企业预算和续费体系。

- 高频高价值调用通常接近业务闭环、交易闭环或生产闭环

泛聊天、开放式陪伴和低门槛问答可以贡献流量，但单位付费深度有限；客服分流、代码辅助、经营分析、导购决策和专业文档处理更容易沉淀高质量收入。

ToC场景的调用增长与消费级付费逻辑

- ToC大模型应以高频入口放大调用规模，以重复成果、交易转化和重生任务提升付费深度，通过订阅、按量、捆绑与导流机制完成消费级收入承接

ToC调用增长的消费级收入承接逻辑



❑ 消费级大模型的第一原则，是先判断“用户为结果付费，还是为陪伴付费或省时间付费”
如果用户只是把AI当搜索替代或免费工具，最优策略是免费+增值。这种模式能最大化DAU，但如果增值权益过于抽象，ARPU会很弱，LTV只能依赖后续交叉销售或其他收入承接。

❑ 订阅模式只在“重复成果型”场景下真正稳固
以美图为例，2025年付费订阅用户达1,691万，订阅渗透率6.1%；其中生产力场景付费订阅用户216万，渗透率达9%，明显高于生活场景。因此可以得出结论，只有当AI输出能持续服务于图片修图、商品图制作、视频口播等重复任务时，月费才有合理性。

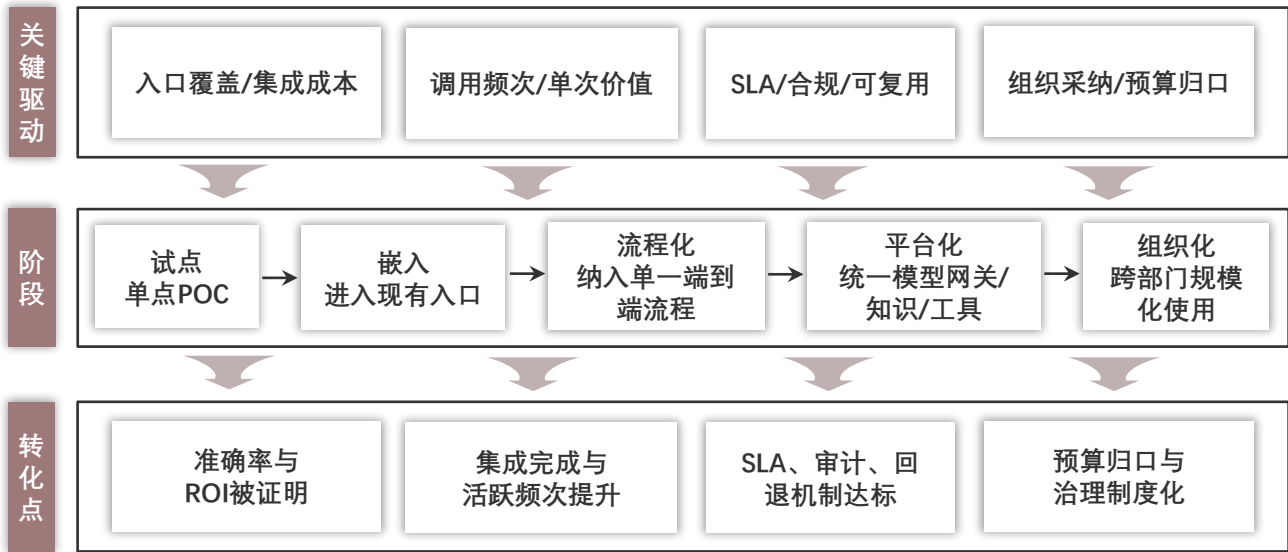
❑ 捆绑和广告/交易导流，是中国市场更容易被低估的收入逻辑
腾讯在官方材料中明确把元宝接入微信联系人，同时说明AI已用于广告定向、创意生成和转化优化；QuestMobile也指出，In-App AI直接跨过拉新门槛，能更快完成规模化渗透。换言之，中国大量“高频低付费”场景并不一定失败，它们可能以搜索、社交和内容平台的形式，转化为广告收入、交易佣金或会员捆绑。因此中国ToC最优定价模型是：泛助手靠免费入口和捆绑保渗透，垂直工具靠订阅保现金流，重生成任务靠按量保利润，流量型场景靠广告和交易保ARPU。

来源：头豹研究院

ToB场景的调用深化与企业级商业化逻辑

- ToB大模型调用深化以企业入口、端到端流程、治理机制和预算归口为核心，完成流程嵌入与价值核算的模型调用，才能从单点试用转化为可续费的企业级流程能力

ToB调用进入企业预算体系的阶段路径



- ❑ 企业级调用增长是流程渗透率提高的结果。从试点到嵌入最敏感的是入口与集成成本；嵌入到流程化最敏感的是调用频次和单次价值；流程化到平台化最敏感的是SLA、审计和可复用性；平台化到组织化最敏感的是预算归口和组织采纳。**调用量本身并不直接构成商业价值**，只有当调用嵌入企业的**稳定入口、核心流程与治理机制**，才具备转化为年费、席位费、专属并发费和续费合同的基础。企业级增长的关键在于将**高价值任务接入流程主干**，而非单纯将更多用户导向聊天框。

企业级商业化逻辑

- 1

➢ 按调用计费适合“需求不确定、研发主导、需要先压进入门槛”的阶段，其优势是低摩擦、易试错、便于比价，但弱点是预算不可预测、采购难以年化。中国头部平台公开价格处于非常低的水平；以豆包Seed2.0 Pro 为3.2元和16元/百万Token；GLM-5.1为6元和24元/百万Token。**价格之所以持续走低，其核心原因是API在ToB中承担的首先是“进入流程”的功能，而不是全部商业价值。**
- 2

➢ 按订阅、座席和功能付费，适合知识助手、办公Copilot、销售助手等入口固定、用户明确、预算可按人头核算的场景；**其定价单位是“可用性、安全、集成便利度和组织治理能力”**。中国市场已经出现从个人套餐向企业套餐延伸的智能体平台；公开信息显示，标准版498元/月起、旗舰版5,980元/月起；智能体开发平台则采用按月/按年订阅并叠加专属并发包的模式；海外则更成熟，企业版公开价为30美元/用户/月、按年付费，为定制报价。
- 3

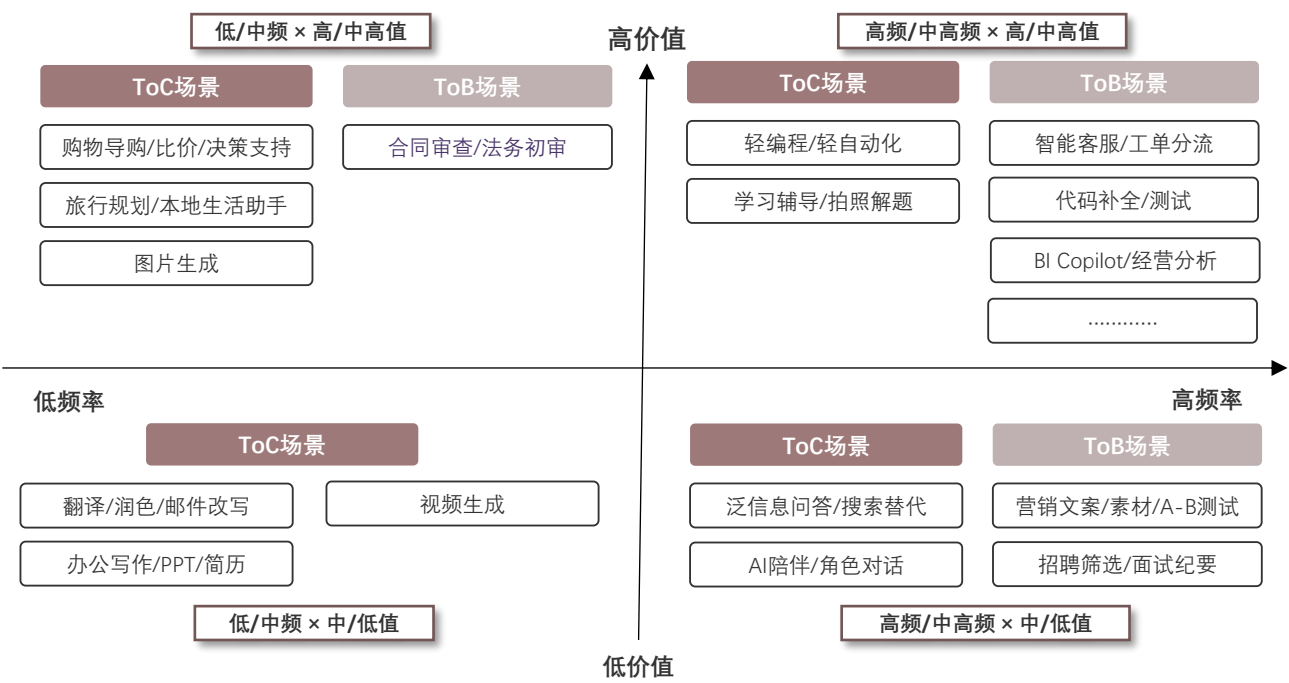
➢ 私有部署、专属推理和项目制定制，适合数据敏感、责任链重、并发稳定且价值密度高的场景，ARPU最高，销售周期和交付复杂度也最高。阿里云、百度智能云等已支持专享推理、企业VPC挂载、混合云打通和自部署模型调用，主要面向金融、政务、医疗、制造等主流场景，**因为客户买的是“可控、可审计、可归责”的运行环境。**

来源：头豹研究院

高频高价值调用及价值密度差异

- 高频高价值调用集中于交易决策、客服分流、代码辅助、BI经营分析和专业文档处理等结果可核算场景，ToC价值密度取决于连续行为转化，ToB价值密度取决于流程嵌入后的降本、增效、风控和增收能力

ToC与ToB高频高价值调用分布



□ ToC领域更具投入价值的方向，是能够将生成结果进一步转化为学习、创作、下单等连续行为的场景

MiniMax招股书披露，海螺AI每付费用户平均支出为56美元，Talkie/星野仅为5美元；同属消费市场，创作型调用的付费深度明显高于陪伴型调用。阿里巴巴披露，Qwen App 在春节期间促进近2亿笔购物、出行和娱乐相关订单，也验证了消费级调用只有贴近交易闭环，ToC流量才具备收入转化基础。

□ ToB 高价值调用的共同特征，是输出结果能够在业务流程中被量化

当一次调用能够减少客服人工、降低财务差错、压缩法务初审、替代开发者重复劳动时，预算责任主体和付费依据才会同时出现。公开披露显示，百度2025年AI Cloud Infra 收入约198亿元人民币，AI Applications收入超过102亿元；MiniMax开放平台付费账户从2023年约100家增至2025年前九个月约2,500家，但其招股书同时提示，平台注册账户无法可靠区分个人开发者和企业客户。因此，判断企业级价值不能只看账户数量，更应关注调用是否进入生产流程。

□ ToC场景以高频、日常化、轻量交互为主

用户决策链路短，产品能否留存主要取决于体验、人格化设计和工具便利性；其单次价值通常集中在学习、创作等场景，付费方式以订阅、点数、广告、佣金为主，价格天花板相对有限，增长更多依赖DAU、口碑传播和分发效率。ToB场景的调用频次不一定高于ToC，稳定性和可复制性更强，价值来源集中在客服、代码、法务、财务、BI等流程环节的效率提升和成本优化；其付费方式更适合按席位、调用量、项目实施、SLA、私有化部署等方式定价，预算来源更明确。

来源：头豹研究院

Chapter 3

市场路径、模型选型与厂商竞争优势

- 中国与海外市场的差异主要来自需求入口、预算结构和落地路径

中国更依赖平台入口、行业场景、本地合规和私有化部署；海外更依赖订阅、SaaS席位、API和企业版扩张；欧洲市场对合规、数据驻留和版权约束更敏感。

- 开源与闭源选型本质是能力上限、数据可控、成本效率和上线速度之间的权衡

高价值、低容错、快速上线场景更适合闭源模型；高吞吐、成本敏感、数据敏感和深度定制场景更适合开源或开放权重模型；多数企业长期会采用混合架构。

- 厂商能否跑出差异，取决于场景、产品、模型和生态是否匹配商业瓶颈

场景决定需求价值，产品决定调用稳定性，模型决定结果质量和成本边界，生态决定获客效率、迁移成本、复用半径和毛利修复空间。

中国与海外市场需求、付费及落地差异与开闭源选型逻辑

- 中国大模型市场的需求转化更依赖平台入口、行业场景、本地合规和私有化交付，美国市场更依赖订阅、SaaS席位、API与企业版扩张，欧盟市场以合规评估和数据治理约束落地节奏，开闭源选型应围绕能力上限、成本效率、数据可控和上线速度进行分层匹配

中国与海外市场的需求转化路径差异

从需求触发到付费转化的典型落地路径				
需求触发	维度	中国	美国	欧盟
选型评估	需求起点	场景牵引、平台入口、行业数字化改造	开发者先行、SaaS 工作流先行、独立AI产品成熟	企业采购和合规评估同步推进
POC	付费方式	会员、广告、交易费、云/API、项目制并存	个人订阅、团队席位、API、企业版阶梯清晰	企业合同、数据驻留、合规成本显著
试点	落地路径	平台嵌入、行业试点、私有化/一体机	自助试用→团队扩展→企业版	合规评估→有限部署→扩面
规模化	模型供给偏好	开源/开放权重与闭源混用，成本敏感	闭源前沿模型与开发者平台领先	更强调可解释、文档、数据与版权合规
付费转化				

- ❑ 消费者和开发者都可能是高频群体，但两者高频的含义完全不同。消费者的高频来自碎片化需求；开发者的高频来自构建与调试；平台型客户的高频则来自下游分发。大企业与客户未必是“最高频用户”，但一定是更容易形成高价值调用池的客户。 中小微企业是最适合标准化打包的客户，其核心特征是对复杂实施、长周期交付和高运维成本承受能力较低。
- ❑ 中国ToB需求主要来自政务、金融、制造、能源、医疗等数据敏感、流程复杂且具备既有IT预算的行业
- 购买决策通常由CIO、数字化部门、采购中心与业务条线共同参与；海外ToB需求触发逻辑更偏向知识工作与软件工程提效。OpenAI企业报告显示，75%受访员工认为AI提升速度或质量，活跃用户平均每天节省40–60分钟，高频价值集中在 IT、市场、财务等职能；也因此，海外企业更愿意用seat-based或API方式快速验证ROI，再扩大到流程重构。
- ❑ 开发者和平台型客户的差异更明显
- 中国开发者更看重性价比、开放权重、国产硬件适配与本地部署：阿里云百炼在中国内地的qwen-max输入/输出价格已降至每百万Token 2.4元/9.6元，Qwen开源家族到2025年4月累计超过3亿次下载、衍生模型超10万个；DeepSeek的折扣策略又进一步把价格战推到极致。海外开发者则更愿意为前沿闭源能力买单：OpenAIGPT-5.5 API定价为每百万输入/输出Token 5美元/30美元，Anthropic Team为20美元/席位起，Enterprise 则是席位费叠加API费；但与此同时，Meta Llama下载量已超过10亿次，说明海外开发者也在用开源模型压低边际成本。
- ❑ 平台型客户方面，中国平台更像“场景与流量持有者”
- 腾讯和百度都在把模型能力嵌入广告、搜索、内容与云；海外平台则更像“模型分发与结算层”，AWS Bedrock已服务全球超过10万家组织，核心卖点就是多模型接入、生产级安全与成本优化。

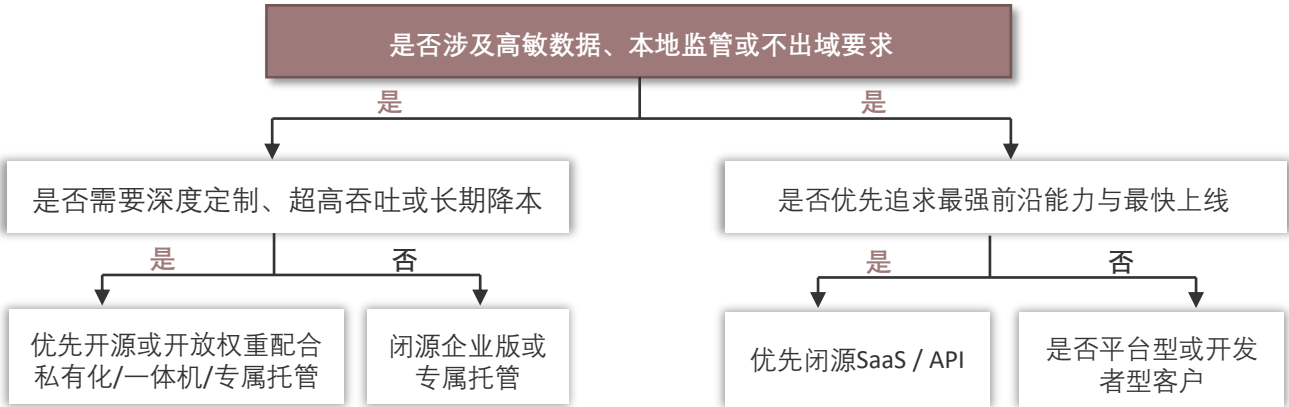
来源：头豹研究院

中国

中国与海外市场需求、付费及落地差异与开闭源选型逻辑

- 开闭源选型以任务价值密度、数据敏感度、调用规模和交付时效为决策轴：高价值、低容错、快速上线场景优先闭源，数据敏感、高吞吐、深度定制及长期降本场景优先开源，复杂企业场景更适合混合架构

中国与海外市场的开源/闭源选型逻辑



开源模型在2025–2026年进入主流选项，核心原因在于够用性能提升与边际成本下降同步发生。Stanford AI Index 2025显示，领先开源权重模型与闭源模型在Chatbot Arena上的差距已从8.04%缩小至1.70%；中美模型在MMLU、MATH等指标上的差距同步收敛。对于不要求前沿推理、复杂代理链路或最强多模态能力的场景，开源模型已可承接大量生产流量。

为什么中国更偏开源或混合？一是成本，API 价格战和本地推理使单位调用成本快速下降；二是合规，数据不出域、模型可审计、可部署在本地或专有云；三是定制化，可以围绕行业知识、工作流和国产硬件做微调与推理优化。Qwen 已形成3亿下载、10万衍生模型的生态，Llama下载量也已过10亿，说明开源的真正价值是把模型变成可以二次开发的基础件。

❑ 闭源模型优势集中于错误代价很高、需要最强工具链、需要企业级安全承诺、需要持续托管运维。如OpenAI Business/Enterprise提供 SAML、MFA、数据不参与训练、区域数据驻留和企业连接器；Anthropic Enterprise提供支出控制、审计日志、SCIM、合规API、网络级访问控制与 HIPAA-ready。闭源的核心是把模型、工具、运维、安全与更新节奏打包成一个可采购的服务等级协议。

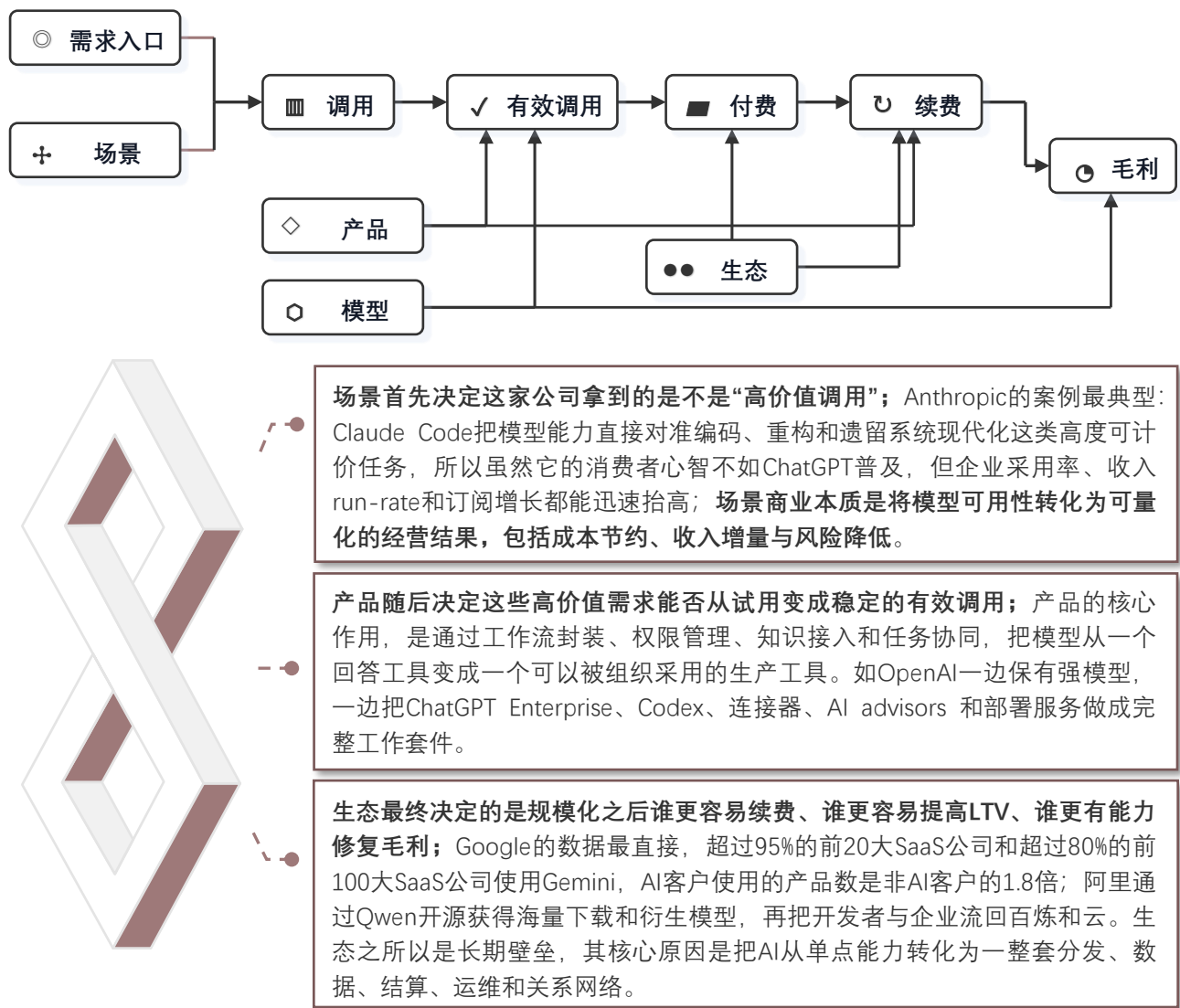
核心判断

❑ 选型逻辑应按“价值密度”而非“模型热度”来定。若任务价值高、容错低、上线快、团队缺MLOps，优先闭源；若任务量大、预算紧、数据敏感、需要深度改造流程，优先开源；若一个企业既有高价值少量流量，也有低价值海量流量，最优解通常是混合架构。这也是为什么AWS Bedrock把“model choice”作为核心卖点，Anthropic强调跨三大云的多平台可用，而欧盟对开源GPAI只做部分豁免：市场正在默认“多模型路由”会成为长期供给形态。

为什么有些厂商具备更强商业化能力与持续竞争优势？

- 厂商表现分化取决于场景、产品、模型与生态对商业瓶颈的匹配程度，能够把高价值场景转化为稳定调用、预算支付、客户续约和毛利修复的厂商，具备可持续竞争优势

厂商商业化表现及竞争优势的核心能力来源



核心判断

为什么厂商做的更好，其核心原因是它们在“调用→有效调用→付费→续费→毛利”的链条上，找到并打穿了最关键的瓶颈。场景决定需求是否高价值，产品决定调用能否被稳定使用，模型决定结果质量与成本上限，生态决定获客成本、迁移成本和复用半径；如果一家厂商只有高调用却没有高有效调用，它只是流量强；如果有高调用和高付费却续费弱，它只是短期项目强；如果能够把高价值场景嵌入workflow，并让客户在云、办公、搜索、广告、交易或开发者平台中持续复购，它才具备“做得更好”的结构性基础。

来源：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



梁超宇
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。