头豹
LeadLeo

行业研究报告

AI业务与应用场景&商业化洞见

从 Token 爆发走向场景兑现，
AI 的商业价值取决于场景价值密度

AI产业市场洞察

China AI Industry Market Insights

中国のAI産業市場動向

头豹研究院

研究目的与观点摘要

- **在AI应用场景加快分化、商业化路径逐步清晰的背景下，AI业务正由模型能力竞争转向场景价值验证与收入兑现能力竞争。本篇报告围绕AI业务与应用场景的商业化逻辑，从价值定义、Token机制、调用与收入转化、To C与To B变现模式、付费场景判断以及价格、成本与收入匹配关系等维度展开分析，主要回答AI企业业务如何通过高价值场景、有效调用和商业化转化机制，将用户规模与调用增长真正转化为可付费、可续费、可盈利的持续收益，对AI应用的商业价值来进行系统梳理与深入研究。**

1. 商业价值：AI商业价值的核心判断标准正在发生怎样的变化？

AI商业价值的判断标准，正由用户规模、活跃度和调用热度，转向场景价值密度与结果兑现能力

AI业务的商业价值，已不再主要取决于用户规模、活跃度或调用热度，而更取决于单次调用能否进入真实业务场景、形成明确产出并创造可验证价值。其核心不在于用户使用数量，而在于每一次使用能否对应降本、增效、增收或风险控制。这一市场风向的转变表明，AI商业化竞争正在从流量获取转向场景深耕，从产品普及转向价值兑现。未来能够获得更高估值和更强商业承接能力的AI产品，往往不是用户规模最大的产品，而是能够持续嵌入高价值场景、形成稳定付费与复用需求的产品。

2. Token消耗：Token高消耗与商业价值兑现的内在关联是什么？

Token高消耗首先反映AI需求扩张，但高价值取决于单位调用是否产生有效产出

Token规模的快速增长，说明市场中AI需求真实存在且持续扩张，但Token消耗本身并不天然等同于高质量商业价值。不同模型、不同场景在单位Token产出能力上存在明显差异，部分高消耗更多体现为用户规模扩大、使用频次增加、上下文拉长和任务链条复杂化，而非高效结果输出。判断Token增长质量的关键，不在于总量是否快速放大，而在于单位调用能否带来更高任务完成率、更强结果质量和更明确的收入转化。只有当Token增长同步推动付费、续费、效率提升或收入改善时，才可被视为高质量增长。

3. To C端：抢占用户资源的核心价值体现在哪里？

To C端的核心价值不在短期订阅利润，而在用户入口、数据反馈和多元化变现能力

在中国市场，消费级AI产品难以单纯依靠订阅模式形成持续高利润，其更重要的价值在于占据用户入口、培养使用习惯、沉淀交互数据，并向广告分发、交易撮合、会员服务、API调用及企业业务持续导流。To C产品的战略意义，不仅是形成用户规模，更在于构建长期流量资产、数据反馈闭环和生态协同能力。随着训练与推理成本逐步下降，To C业务的单位经济性有望改善，其价值也将从单纯的用户获取入口，进一步延伸为广告、交易和平台分发的新型商业化入口。

4. To B端：AI商业化的关键兑现因素是什么？

To B端商业价值主要取决于场景攻克能力，而不是单纯的模型技术提升

企业客户真正愿意付费的，并不是抽象意义上的更强模型能力，而是能够嵌入流程、替代重复劳动、形成可量化结果的场景化解决方案。To B商业化的关键，在于识别高频、标准化、规则清晰且ROI可验证的具体任务，并通过工作流嵌入、Agent执行和系统协同实现持续交付。相较于一味追求模型参数、上下文长度或推理能力提升，围绕真实业务需求进行场景攻克，才是To B价值兑现和收入放大的更关键路径。未来To B竞争将更强调行业理解、交付能力、流程适配和成本控制，而非单一技术指标领先。

内容目录

名词解释

- **Token（词元）：**大模型处理信息的最小计量单位，可对应单个汉字、词语、标点、英文片段、代码或数学符号等。模型在接收输入、生成输出、调用工具和处理上下文时，均会消耗Token。
- **Agent化：**AI产品从单轮问答工具升级为能够理解目标、调用知识库或插件、执行多步骤任务并与业务流程集成的智能体形态。Agent化的核心不只是回答问题，而是通过系统提示词、RAG知识库、插件调用和API集成，将模型能力转化为可执行任务能力。
- **工作流应用：**将复杂任务拆分为一系列有序步骤，并通过大模型、API、函数计算、插件等节点组合完成任务的应用方式。其价值在于降低系统复杂度、提升执行效率。
- **Batch收费：**面向低实时性、大批量任务的异步推理计费模式。用户通常以文件或任务队列形式批量提交请求，平台在后台异步处理并返回结果。
- **上下文缓存**：平台对多次请求中重复出现的公共前缀、系统提示词、长文档内容或历史上下文进行缓存，以减少重复计算、降低延迟和压缩输入成本的机制。
- **场景价值密度：**指单次AI调用或单个AI场景能够创造的商业价值强度，通常体现为是否能够降本、增效、增收、降低风险或形成持续付费。高场景价值密度的应用往往具有任务边界清晰、结果可衡量、流程可嵌入和复用频率较高等特征，是判断AI商业价值的重要指标。

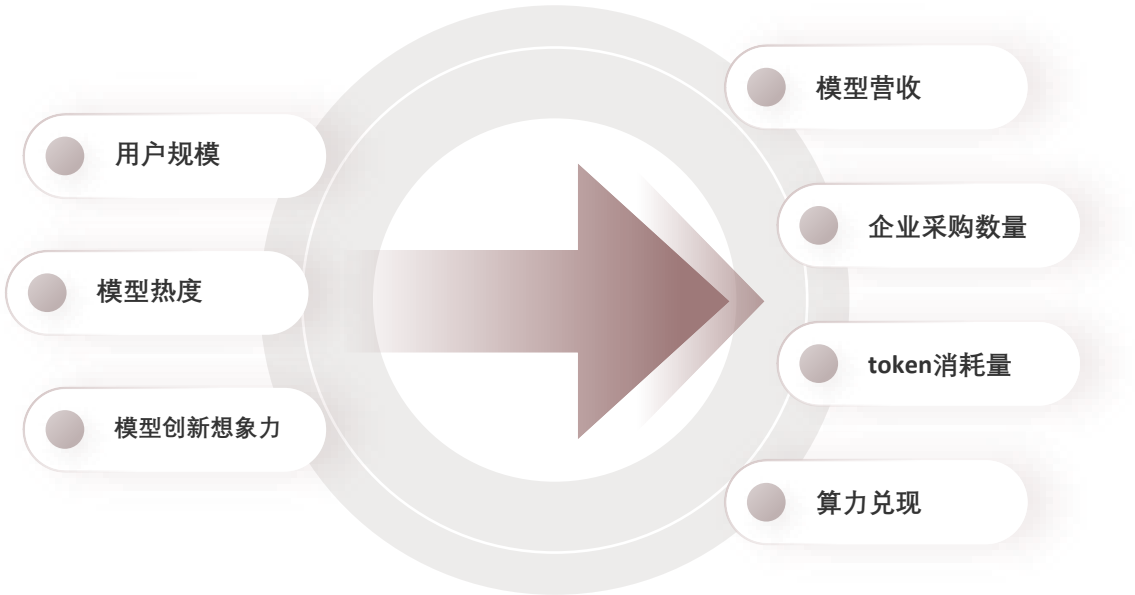
Chapter 1
AI价值定义与场景判断

- 用户规模体现流量入口、品牌认知和分发效率，但真正决定收入质量、估值支撑与可持续性商业价值的是AI是否嵌入高价值场景，形成高强度、可付费、可续费的工作负载，为客户带来切实收益
- AI的商业价值不取决于调用频次，而是由单次调用的价值兑现密度决定。高价值场景需满足四大条件：任务边界清晰、工作结果可衡量、深度嵌入业务流程与易形成付费和续费需求。缺乏业务闭环的高频应用，商业转化依然有限

为什么用户规模不等于收入质量

- 用户规模体现流量入口、品牌认知和分发效率，但真正决定收入质量、估值支撑与可持续性商业价值的是AI是否嵌入高价值场景，形成高强度、可付费、可续费的工作负载，为客户带来切实收益

AI估值关注转变



场景价值密度比用户规模更能决定AI商业价值

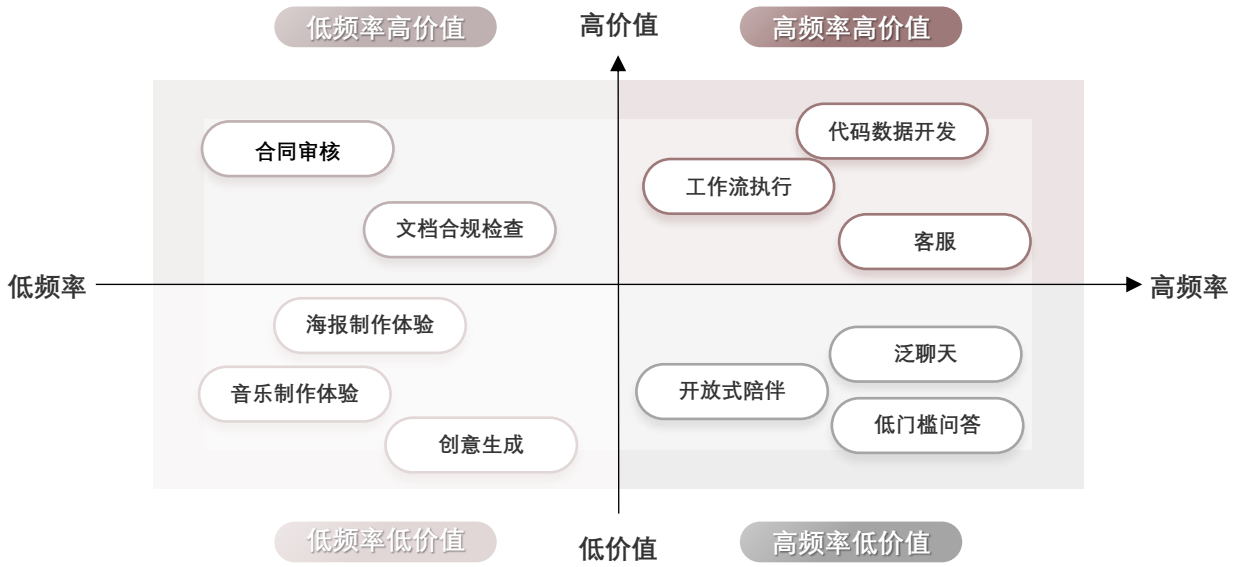
- 行业初期，用户规模代表了更高的流量入口和品牌覆盖，在行业发展前期，较高的影响力成为了企业竞争的关键要素。而如今，市场关注发生了阶段转变，效率改善、企业营收与投资回报兑现成为投资者关注重心。ChatGPT拥有约9亿周活跃用户、超过5000万订阅用户，截至2026年2月，OpenAI年化收入超过250亿美元，而Anthropic在更小用户基础下，2026年初年化收入已达到约90亿美元，收入差距正在缩小。这表明AI能否形成商业价值，关键不在于有多少用户，而在于其是否进入高价值、可复制、可持续付费的业务场景。
- 中国头部厂商的收入结构也反映出这一趋势。百度披露，2025年文心助手月活跃用户达到2.02亿，同期AI Cloud Infra收入约198亿元，同比增长34%，AI Applications全年收入超过102亿元，其中AI加速基础设施的订阅收入同比增长143%。阿里则表示，过去三个月百炼MaaS平台Token消耗规模提升6倍，并预计商业化MaaS收入将成为阿里云最大的收入产品。相较于泛用户增长，企业订阅、模型服务和工作流调用正在成为更核心的收入承接方式。用户规模决定入口，场景价值密度决定收入质量。
- **核心原因在于，不同场景的“价值密度”差异很大。**泛聊天、轻陪伴、低门槛问答虽然容易做大用户规模，但单次调用对应的付费意愿、业务结果和预算承接能力普遍较弱。相反，编码、客服自动化、文档处理、企业流程辅助等场景，虽然用户基数未必最大，却更容易形成持续调用、明确ROI和长期续费。

来源：Reuters，百度，阿里巴巴，头豹研究院

为什么高频调用不一定代表高价值

- AI的商业价值不取决于调用频次，而是由单次调用的价值兑现密度决定。高价值场景需满足四大条件：任务边界清晰、工作结果可衡量、深度嵌入业务流程与易形成付费和续费需求。缺乏业务闭环的高频应用，商业转化依然有限

AI场景频率与价值差异



高价值关键在于单位调用能否产生有效产出

- AI场景运用的价值差异，不在于调用频率与调用次数，而在于调用后是否能够产生有效的业务场景效应。分场景分析，客服、代码辅助、工作流执行等任务型场景，往往同时具备高频使用和明确产出。将现有技术和模型嵌入到既有流程中，将可重复流程形成自动化接管，完成具体任务再接入到系统之中；合同审查、合规审核、关键文档分析等场景调用的频次不高，但能在单次调用过程中实现时间节省与风险控制，较大程度地提升工作流中的效率，实现专业替代价值；这一类高价值场景的共性在于：能够拥有明确的效率输出路径与能够切实地提升生产效率。相较之下，开放式聊天、泛陪伴、一次性创意生成等场景，更容易形成活跃度和体验价值，作为产品接入C端体验的有效窗口，往往停留在交互层面，难以形成体系化效率提升和效用转化。以MiniMax招股书披露的数据看，星野（Talkie/Xingye）截至2025年9月30日，平均MAU为2,005.1万，付费用户为139.04万；MiniMax主应用平均MAU为142.9万，付费用户为1.03万。但在平均每付费用户支出上，MiniMax为73美元，显著高于星野的5美元，约为后者的14.6倍。该对比表明，陪伴互动类产品可以拥有更高活跃度和更高付费渗透，但其单位付费深度相对有限。判断AI场景价值不能只看调用频次和用户活跃度，更应关注单次调用所承载的任务复杂度、结果强度和付费深度。
- 高价值场景通常具备四个共同特点：一是**任务边界清晰**，围绕客服应答、合同审核、数据处理等具体目标执行；二是**结果可衡量**，能够对应处理时长、人工替代、风险识别或流程完成率；三是**能够嵌入现有工作流程**，不是停留在独立工具层，而是可被接入企业业务链条持续调用；四是**更容易形成付费与续费**，因为其价值可以被业务部门直接验证。

来源：MiniMax，头豹研究院

Chapter 2
AI Token商业逻辑

- 中国AI Token定价体系正由基础输入/输出计费，逐步延伸至长上下文、缓存与Batch等多元化模式。其核心逻辑是通过差异化定价匹配不同任务的算力占用方式，增加AI厂商企业收入，降低算力成本
- 围绕企业实际应用，Batch、长上下文、缓存及资源包等模式正加快铺开，本质上是根据时效性、上下文长度与复用程度对算力成本进行再分配。不仅降低企业单位调用成本、提升高价值场景的使用频次，也有助于平台提高低峰时段资源利用效率
- Token增长需求繁荣已成现实，反映AI需求的真实扩张，用户规模提升、产品普及和场景渗透是总量放大的主要驱动因素。但调用尚未同步转化为高质量结果，Token消耗与模型产出并不构成稳定线性关系，不同模型产出效率上存在显著差异
- 未来Token调用的核心增量预计仍将主要来自企业侧。To B场景能够通过工作流嵌入、Agent执行和多场景复制，将模型持续纳入标准化流程，更稳定地形成高频、可持续的调用需求。To C端增量仍将存在，但更多来自既有用户使用习惯深化，而非新增用户的高速扩张

AI Token主流调用模式多元分层的价值是什么

- 中国AI Token定价体系正由基础输入/输出计费，逐步延伸至长上下文、缓存与Batch等多元化模式。其核心逻辑是通过差异化定价匹配不同任务的算力占用方式，增加AI厂商企业收入，降低算力成本

中国各企业AI产品计费对比

企业	代表模型系列	调用/接入模式	计费模式结构	典型定价参考（元/百万Tokens）
阿里云	Qwen-Max/Plus/Turbo/Long	API/SDK、Batch异步、控制台、工作流编排	输入/输出分离、按上下文长度阶梯定价、上下文缓存命中折扣、Batch半价	Qwen3-max输入/输出分开计费；示例价：0K-32K输入2.5、输出10；32K-128K输入4、输出 16；128K-252K输入7、输出28
百度	ERNIE 5.0/ERNIE 4.5 Turbo	API、Web控制台、Fine-tuning、企业专属网关	输入/输出分离、按量付费、新用户免费额度、企业阶梯协议、支持批量推理	ERNIE 4.5 Turbo输入/输出分开计费；在线推理示例价：输入0.8、输出3.2
腾讯	HY 2.0 Think/HY 2.0 Instruct	API、微信生态集成、企业微信插件、SaaS网关	输入/输出分离、基础免费额度、用量阶梯折扣、生态内调优	HY 2.0 Think输入/输出分开计费；0K-32K输入3.975、输出15.9；32K-128K输入5.3输出21.2
智谱AI	GLM-5.1 / GLM-5-Turbo	API开放平台、批量推理、微调、开源本地部署	输入/输出分离、按上下文窗口定价、开发者免费额度、Batch优惠	GLM-5.1：0K-32K输入6、输出24；32K+输入8、输出28；GLM-5-Turbo：0K-32K输入5、输出22；32K+输入7、输出26
月之暗面(Moonshot)	Kimi K2.5	API、开发者控制台、长文档专用接口	输入/输出分离、按上下文长度阶梯、长窗口溢价、支持Batch Api价格为标准模型60%	输入区分缓存命中/未命中；K2.5示例价：缓存命中输入0.70、未命中输入4.00、输出21.00
MiniMax	MiniMax-M2.7 / M2.5	API、企业专属通道、多模态语音捆绑	输入/输出分离、超低价策略、免费额度、语音+文本打包计费	M2.7输入2.1、输出8.4
火山引擎	Doubao-Seed-2.0-Pro	API（火山引擎方舟）、SDK、控制台、批量推理、端云协同	输入/输出分离、按上下文窗口阶梯、激进低价策略、开发者高额免费额度、缓存折扣	输入/输出分开计费；0-32K输入3.2、输出16；32K-128K输入4.8、输出24；128K-256K输入9.6、输出48

Token收费机制由单一按量计费走向多元分层，助力AI厂商企业降本增效

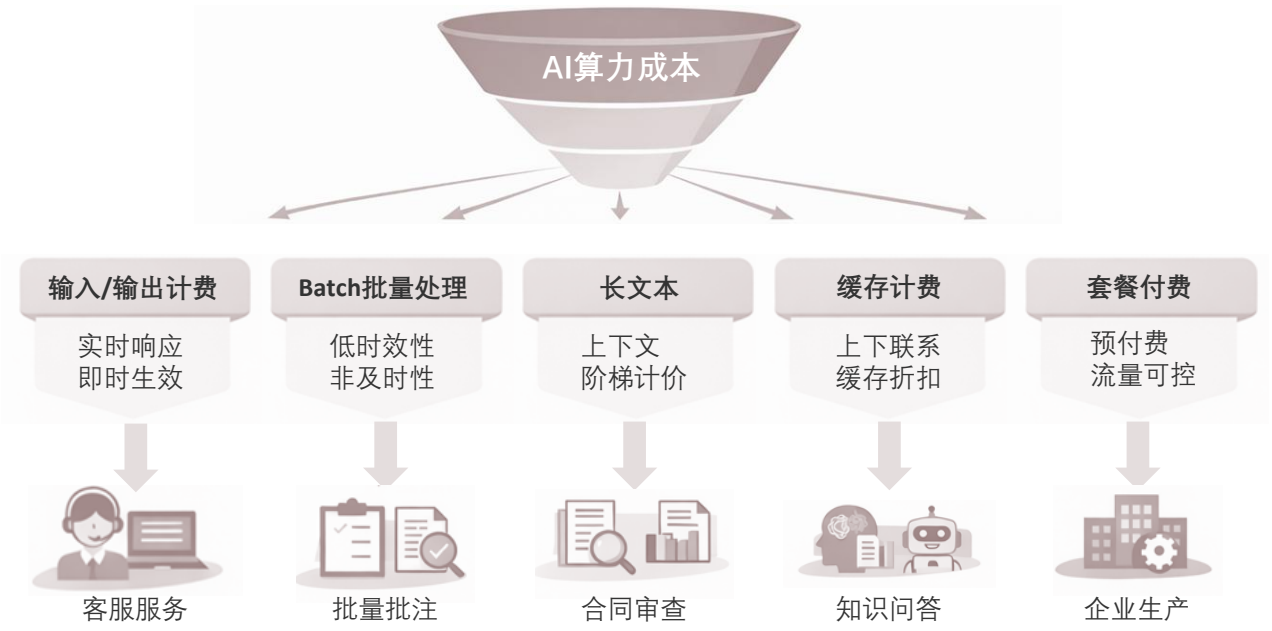
- 以阿里云、百度、腾讯、MiniMax的核心产品基础文本档位为样本，输入价格均值约为2.4元/百万Tokens，输出价格均值约为9.6元/百万Tokens。当前中国主流平台的定价重心在输出侧，输出侧价格约为输入侧的4倍。平台间差异已不主要体现为是否按Token收费，而主要体现为长上下文溢价、缓存折扣力度及异步处理优惠程度的不同。从成本结构看，长上下文会显著增加模型推理过程中的计算资源、显存占用和调度压力，因此多数平台对较长上下文区间设置更高价格档位；缓存机制则通过复用系统提示词、历史上下文或知识库内容，减少重复计算并压缩输入侧成本；Batch模式牺牲实时性为代价，将批量任务转入异步处理流程，提高低峰时段算力利用率并降低单位调用成本。因此，Token收费的多元化，对应的是不同场景下厂商成本差异，核心在于降低AI厂商成本。

来源：阿里云，百度智能云，腾讯云，智谱，月之暗面，MiniMax，火山方舟，头豹研究院

多元层次计费如何匹配场景效益

- 围绕企业实际应用，Batch、长上下文、缓存及资源包等模式正加快铺开，本质上是根据时效性、上下文长度与复用程度对算力成本进行再分配。不仅降低企业单位调用成本、提升高价值场景的使用频次，也有助于平台提高低峰时段资源利用效率

AI算力计费方式与适用场景



主流计费以输入/输出Token为主，场景化优化向Batch、长上下文与缓存扩展

- 当前中国大模型商业化的基础计费框架，仍以输入/输出Token分开计费为主，围绕特定商业场景的优化计费正在加速丰富：**Batch批量推理已从补充功能逐步演变为主流配套方案**，阿里云将Batch价格设为实时调用的50%，Kimi Batch API为标准价格的60%，百度千帆亦单列在线推理与批量推理价格，说明厂商正通过"低时效换低价格"的方式承接批量摘要、分类、抽取、评测与夜间处理等任务，以降低企业单位成本、提升可承受调用频次；从平台侧看，百度千帆已公开推出"闲时调度训练免费、推理部署低至3折"等政策，反映出错峰定价正被用于吸收非高峰时段需求、提高闲置资源利用率。而**长上下文与上下文缓存更适用于合同审查、长文档研读、知识库问答等信息密集型场景**：阿里云与腾讯对32K以上输入设置阶梯加价，Kimi、MiniMax与阿里云则对缓存命中提供折扣，表明平台正按照"上下文长度—复用率—实时性"重构收费逻辑。**输出Token消耗扩张已是明确趋势**，阿里云最新Qwen模型已对思考模式和超长上下文单独计价，腾讯TokenHub列示的主流模型中，Kimi-K2.5最大输出为256k、MiniMax-M2.5为192k、GLM-5为128k、Tencent HY 2.0 Think为64k，反映出推理、Agent和长程代码任务正在系统性抬升单次生成的Token预算。短期内输入/输出Token计费仍将是行业主流框架，但Batch、缓存、TPM/资源包、订阅套餐及搜索增强等附加收费模式预计将继续铺开，形成更适配企业工作流与平台资源调度的混合计费体系。**企业调用场景的多样化推动计费体系由单一按量计费向场景化、精细化定价演进。**

来源：腾讯TokenHub，百度千帆，阿里云，Kimi，头豹研究院

Token增长如何区分真实需求扩张与低效消耗

- Token增长需求繁荣已成现实，反映AI需求的真实扩张，用户规模提升、产品普及和场景渗透是总量放大的主要驱动因素。但调用尚未同步转化为高质量结果，Token消耗与模型产出并不构成稳定线性关系，不同模型产出效率上存在显著差异

AI模型Token产出效率对比

模型名称	Humanity's Last Exam			AIME 2025基准考试		
	成绩（%）	Token消耗（百万）	Token产出效率	成绩（%）	Token消耗（万）	Token产出效率
Kimi K2.5	29.4	56	0.53	94.7	70	1.35
Qwen 3.6 Max preview	28.9	48	0.60	82.3	21	3.92
MiniMax-M2.7	28.1	54	0.52	82.7	48	1.72
GLM-5.1	28.0	82	0.34	95.0	60	1.58
Deepseek V3.2	22.2	40	0.56	92.0	47	1.96
Doubao Seed Code	13.3	23	0.58	79.3	40	1.98
GPT-5.4（xhigh）	41.6	36	1.16	99.0	40	2.48

Token产出效率=测试成绩/总Token消耗，即每万/百万Token消耗获得的成绩点数
Token消耗包括输入标记（提示词）、推理标记（用于推理模型）和答案标记（最终反应）

Token增长本质上体现了需求扩张，但高质量产出转化仍显不足

- **Token快速放量已成为当前AI市场需求扩张的直接体现。**Token消耗的快速上升反映了真实需求扩张，而非单纯的统计噪音。国家数据局披露，2026年3月中国日均Token调用量已超过140万亿，较2025年底的100万亿增长约40%。火山引擎披露，豆包大模型日均Token使用量已突破120万亿，三个月内翻倍；Meta内部30天Token使用量超60万亿。数据共同表明，Token放量已成为国内外AI工具加速渗透后的现实结果。

- **Token消耗增加与产出效果之间并不构成稳定的线性对应关系。**从模型测试结果看，Token投入与成绩提升并非严格同步。Humanity's Last Exam表现中，Qwen 3.6 Max preview、Kimi K2.5、MiniMax-M2.7与GLM-5.1的成绩均集中在28.5%，但Token消耗分别为48、56、54和82百万，投入差异明显；AIME 2025测试中，Qwen以21万Token取得82.3%的成绩，Token产出效率显著高于多数国产模型，而Kimi和GLM虽取得更高分数，但对应消耗也明显更高。这表明，Token增长只能说明调用规模扩大，不能直接等同于模型效率或结果质量同步提升。

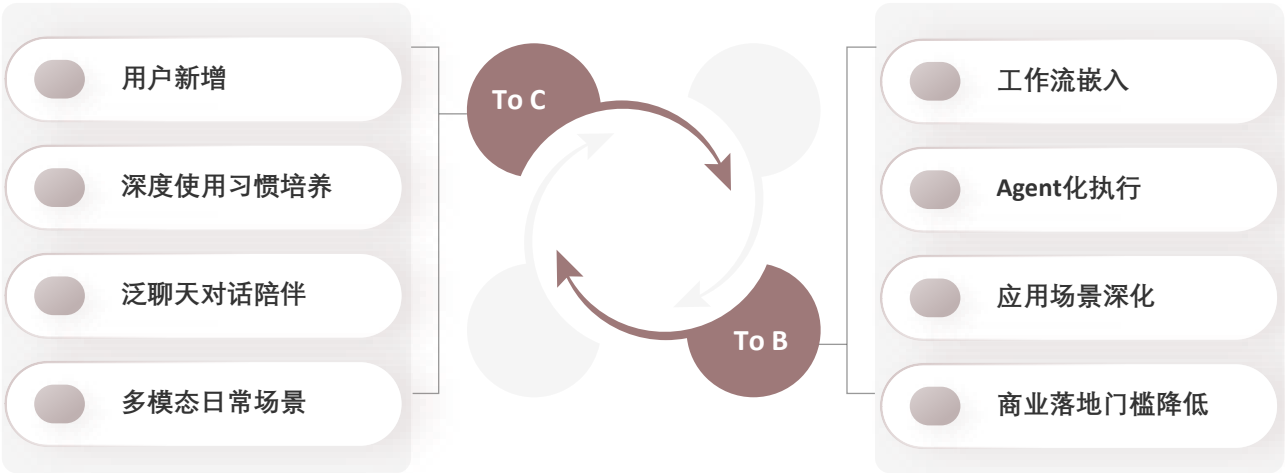
- **Token消耗激增主要来自用户基数扩大与产品普及，高质量产出占比仍待提升。**当前市场Token总量的快速抬升，更多是由使用者规模扩张、产品渗透加深与商业习惯形成所推动，而非厂商主动制造"Token繁荣泡沫"。以豆包为例，其规模优势极为突出；结合其测试Token消耗数量，可以判断其高消耗更主要来源于用户基数扩大和产品铺开。同时，按照测试结果，豆包在人类最后考试和AIME2025中的成绩分别为13.3%和79.3%，均处于样本末位，但其Token产出效率并非最低。表明，豆包的高消耗并不是由极端低效调用所致，而是由大规模、中等效率调用所支撑；但其模型能力上限与高质量输出能力仍明显弱于头部模型。当前Token增长为需求繁荣的表现，但需求尚未有效转化为高质量产出。

来源：Artificial Analysis，国家数据局，火山引擎，头豹研究院

未来Token调用的增长来源主要是哪里

- 未来Token调用的核心增量预计仍将主要来自企业侧。To B场景能够通过工作流嵌入、Agent执行和多场景复制，将模型持续纳入标准化流程，更稳定地形成高频、可持续的调用需求。To C端增量仍将存在，但更多来自既有用户使用习惯深化，而非新增用户的高速扩张

未来Token调用增长源



To C端：未来Token增量将由新增扩张转向使用深化

- **新增用户仍有增量，但增长节奏放缓。**To C端用户规模仍将扩张，但高增阶段已基本过去。2026年春节期间，豆包、千问、元宝均通过集中投放快速拉升日活，但节后出现不同程度回落，表明个人市场已由单纯拉新转向留存竞争。**使用习惯固化将成为个人侧更稳定的增量来源。**用户将搜索、写作、规划、购物决策等行为迁移至AI界面，单用户调用频次和调用链路有望持续拉长。千问在节后保持相对更稳的留存，并通过Agent功能促成近2亿商品订单，未来个人侧Token增长，更多将来自既有用户的深度使用。**泛聊天与陪伴型交互将贡献高频Token消耗。**MiniMax旗下星野陪伴式AI截至2025年9月30日的九个月中平均MAU为2,005.1万，且用户平均日使用时长超过70分钟。此类场景单次会话长度长、互动频次高，能够稳定贡献大规模Token消耗。**多模态日常化将进一步提升个人调用密度。**AI调用场景从文字输入扩展至语音、图像和视频。阿里云Qwen系列已推出Omni与Realtime模型，支持文本、音频及图像视频输入，随着实时语音与视觉理解普及，个人侧调用将更高频、更碎片化。

- **To B端：未来Token增量将主要由工作流嵌入与Agent执行驱动**

- **工作流嵌入与Agent执行将共同推动企业侧Token需求增长。**工作流嵌入强调将客服、审批、文档处理、数据分析等重复性环节流程化，Agent通过任务拆解、工具调用和结果回传提升自动化深度。本质上服务于降本增效，替代重复人工、压缩处理时长和边际成本，提升任务闭环能力与执行效率，使Token调用由零散问答转向持续、可预算的生产性消耗。未来**场景深化与多样化将持续拓宽企业侧需求边界，同时落地门槛下降将提升企业部署速度与复制能力。**商业化成熟度提升，将使企业侧Token增长更具可复制性。部署方式由重开发转向模块化与低门槛接入后，企业内部扩散速度将加快，Token需求也将更快释放。

来源：MiniMax，头豹研究院

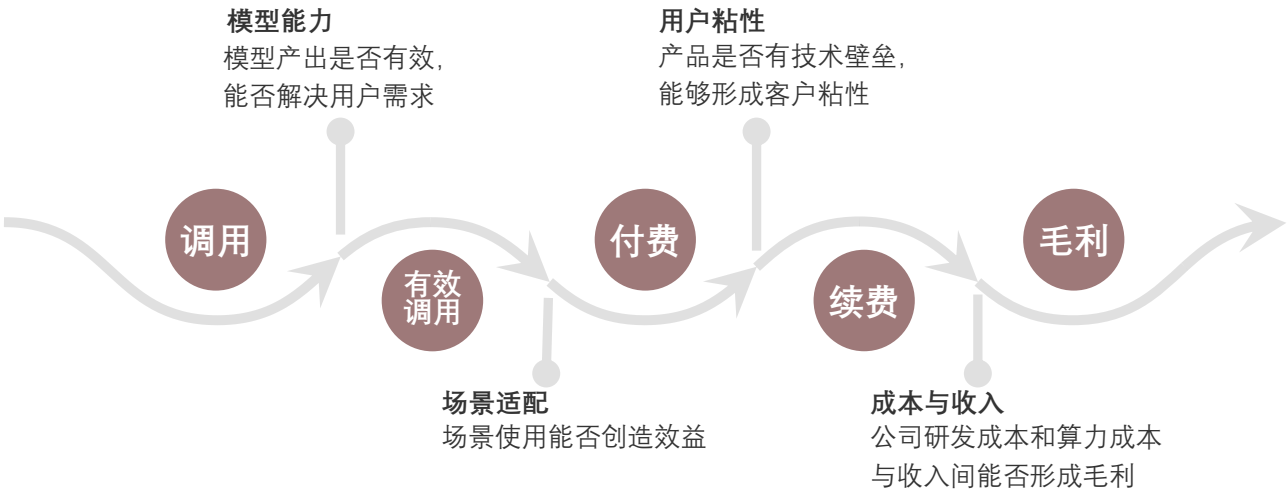
Chapter 3
AI产品商业化与收入模式

- AI产品调用向收入兑现的关键，在于将流量型调用转化为可付费、可续费、可盈利的有效需求。其核心链条包括：模型能力与任务匹配决定有效调用，场景价值决定付费转化，流程嵌入与替换成本决定续费稳定性，推理效率与成本控制决定毛利水平
- To C平台厂商追求的核心价值是通过占据用户入口、培养使用习惯和沉淀真实交互数据，建立长期流量与品牌壁垒；同时，将消费端流量持续导向API、企业服务、广告分发和交易撮合等更高质量收入环节
- To B客户当前更愿意为能够替代重复、规则化工作且ROI可验证的AI场景付费，典型包括智能客服、合同审核、报表分析和流程自动化。其核心考察点在于能否降本、增效、增收并控制风险。未来替代空间仍将沿财务共享、供应链协同、人力资源与运营控制等方向持续扩展
- 价格、成本与收入之间存在错位，根源在于消耗逻辑与价值逻辑并不一致。价格通常围绕Token消耗、上下文长度和输出规模设定，但收入兑现取决于场景是否能够形成可验证价值、客户是否愿意持续付费以及产品能否实现稳定续费

AI产品调用与收入间如何兑现

- AI产品调用向收入兑现的关键，在于将流量型调用转化为可付费、可续费、可盈利的有效需求。其核心链条包括：模型能力与任务匹配决定有效调用，场景价值决定付费转化，流程嵌入与替换成本决定续费稳定性，推理效率与成本控制决定毛利水平

AI产品调用到收入转化流程



AI产品收入兑现的核心在于提升有效调用率，并修复单位经济性

- **调用→有效调用：以“模型-场景”匹配度为核心。**收入兑现的前提并非调用规模，而是调用结果能否直接嵌入业务流并被终端采纳。有效调用取决于模型的可用性、稳定性与场景适配性，本质是技术能力向可交付结果的转化。当前企业端需求已从“体验验证”转向“结果可验证”，如百度2025年AI应用收入突破百亿、AI云基础设施订阅收入同比增143%，印证了在流程可嵌入、结果可量化的场景中，有效调用已具备规模化承接能力。
- **有效调用→付费转化：商业场景价值验证（ROI）。**付费意愿强弱取决于AI能否在降本、增效或增收等核心经营指标上形成明确回报。若仅停留在体验优化，预算释放缓慢；一旦切入客服、数据治理、合规审查等可量化场景，商业化进程显著加速。当前市场瓶颈不在流量获取，而在“有效调用能否支撑明确的付费决策与预算立项”。
- **付费→持续续费：由工作流嵌入深度与替换成本决定。**单次付费体验不等于可持续收入，续费稳定性取决于产品是否深度嵌入客户核心工作流、是否形成数据资产沉淀及替换成本壁垒。对比MiniMax 2025年前九个月数据：Open Platform等企服/开发者业务营收占比28.9%，毛利率达69.4%；而AI原生C端产品占比71.1%，毛利率仅4.7%。表明企服场景虽当前占比有限，但凭借高嵌入性，其客户生命周期价值（LTV）与收入质量显著优于消费端。
- **规模收入→毛利修复：取决于单位经济模型（UE）的优化速度。**商业化第二道坎在于“收入增速能否跑赢推理与交付成本增速”。高调用规模不必然转化为高利润，需持续优化算力调度、模型压缩与工程交付效率。以智谱2025年云端API为例，收入同比增292.6%至1.9亿元，毛利率仍处18.9%的爬坡期；结合MiniMax C端产品4.7%的毛利率表现，进一步验证单位经济性修复是当前盈利能力的核心制约。

来源：百度，智谱，MiniMax，头豹研究院

To C难以形成高利润的背景下，厂商追求核心价值点是什么

- To C平台厂商追求的核心价值是通过占据用户入口、培养使用习惯和沉淀真实交互数据，建立长期流量与品牌壁垒；同时，将消费端流量持续导向API、企业服务、广告分发和交易撮合等更高质量收入环节

To C价值核心

入口价值

- 抢占平台入口，培养用户习惯
- 竞争焦点转向入口、品牌和粘性
- 锁定未来高频调用和流量分发主动权

中国生成式AI用户：6.02亿（渗透率约40%）

导流变现

- 向API、会员与企业服务持续导流
- 消费端流量带来高质量收入转化入口
- 高毛利来自企业服务

MiniMax Open Platform 即企业服务毛利率69.4%

数据反馈

- 获取交互反馈与使用数据
- 获取提示词、场景偏好、多模态数据
- 支撑模型优化、产品迭代与场景识别

To C 是持续优化的重要反馈池

生态变现

- 多元化变现替代单一订阅逻辑
- 广告、交易与生态价值上升
- AI入口向商业推送、广告投放延伸

Qwen 春节期间促成近2亿商品订单

To C难以形成高利润背景下，厂商更重视入口资产沉淀、导流转化与反馈闭环

- **当前To C布局的核心，在于沉淀入口价值并向高质量业务导流。**在中国市场，消费级AI产品的首要目标是率先占据高频入口、培养用户习惯，并通过真实交互持续积累品牌影响力、使用数据与场景反馈。原因在于，基础模型能力差距正在收敛，价格竞争持续强化；入口位置、使用习惯与平台认知更难在短期复制。截至2025年末，中国生成式AI用户已达6.02亿，渗透率升至约40%，头部平台已具备足够规模的真实反馈池。对厂商而言，消费端产品的现实价值主要体现在三方面：一是占据流量入口，形成未来分发主导权；二是向API、会员及企业服务导流，推动流量向高毛利业务转化；三是通过高频交互沉淀提示词、失败样本与场景偏好，反哺模型与产品迭代。

- **未来To C变现的重点，在于多元化模式探索与成本端持续改善。**在用户付费订阅难以成为中国市场主导模式的背景下，消费端AI的下一阶段重点，将是探索广告、交易分发、生态合作等多元化变现路径，并同步推动模型成本持续下降。腾讯总裁刘炽平明确表示，中国市场较难复制美国AI工具以用户付费为主的模式；与之相应，腾讯2025年第四季度在线广告收入同比增长17%，公司将增长部分归因于AI增强的广告定向能力，这表明AI入口正在具备广告与商业分发价值。同时，成本端的改善正在为To C商业化打开新的空间。MiniMax-M1其完整训练周期约为3周、基于512张H800的租赁成本约53万美元，并在长输出场景下显著降低推理FLOPs；MiniMax消费端业务毛利率已达到4.7%，尽管仍处于较低水平，但若后续训练与推理成本持续下降，消费端业务的单位经济性有望逐步修复。在此基础上，To C产品有望由“流量入口”进一步演变为兼具广告分发、交易撮合与独立收入贡献能力的新型商业入口。

来源：MiniMax，腾讯，新华社，头豹研究院

To B客户愿意为哪些具体的AI场景付费，核心考察点是什么

- To B客户当前更愿意为能够替代重复、规则化工作且ROI可验证的AI场景付费，典型包括智能客服、合同审核、报表分析和流程自动化。其核心考察点在于能否降本、增效、增收并控制风险。未来替代空间仍将沿财务共享、供应链协同、人力资源与运营控制等方向持续扩展

To B付费场景

智能客服与外呼
- 高频使用、标准化、替代对象清晰
- 直接对应降本、提效与转化

高频承接

合同与文本审核
- 非结构化文本转化
- 规划明确、可审核、可追责、人工可复查，确定审核质量

结构化处理

报表分析与洞察
- 接入经营决策链条
- 效益直接被时间、转化和成本验证

经营分析

流程自动化
- 替代重复、规划清晰的流程任务
- 稳定提效、降低人工与出错成本

流程替代

To B场景下客户更愿意为“可量化替代”的AI能力付费

- **当前已形成付费的AI场景，主要集中在客服、文档、分析与流程自动化。**现阶段企业付费意愿较强的AI场景，主要集中在四类：一是智能客服与外呼，核心价值在于替代高频人工服务并提升转化效率；二是合同审核、票据识别、文档解析等结构化处理场景，核心价值在于提升处理效率并降低合规风险；三是报表分析与经营洞察，核心价值在于压缩分析周期并提升经营决策效率；四是流程自动化与数字员工，核心价值在于替代重复、规则清晰的操作性工作。阿里云Quick BI、阿里云RPA、百度千帆合同审核与百度智能客服等案例均表明，企业当前更愿意为“结果可验证、流程可嵌入、替代对象明确”的AI能力付费。
- **企业付费的核心考察点，在于ROI可验证、流程可嵌入与风险可控制。**To B客户的付费判断取决于该能力能否形成可核算的业务价值。其核心考察标准主要包括三点（1）是否能够**直接对应降本、增效或增收**；（2）是否适合**替代高频、重复、规则化任务**，并嵌入既有系统与流程；（3）是否具备**准确性、可解释性与风险可控性**。企业真正采购的并非泛化AI能力本身，而是能够进入预算体系、形成可复制部署并持续交付结果的场景化解决方案。
- **未来仍具替代空间的方向，将继续向财务、供应链、法律延伸。**未来可进一步形成付费的场景，仍将集中在规则边界清晰、数据结构化程度较高、人工处理链条冗长的环节，重点包括财务审核与对账、采购与供应链协同、人力资源筛选与培训、内部办公助手与法律医疗等高价值专业场景等。这些场景的共同特征在于：流程标准化程度高、人工成本相对刚性、替代收益较易量化，因此更具商业落地与规模复制条件。To B付费场景的扩展方向，仍将沿着“重复劳动替代—流程自动化—多步骤协同执行”的路径持续推进。

来源：阿里云，百度，头豹研究院

价格、成本与收入之间为什么会存在错位

- 价格、成本与收入之间存在错位，根源在于消耗逻辑与价值逻辑并不一致。价格通常围绕Token消耗、上下文长度和输出规模设定，但收入兑现取决于场景是否能够形成可验证价值、客户是否愿意持续付费以及产品能否实现稳定续费

错位本质是成本收入与商业回报不对等

消耗逻辑与价值逻辑不一致，Token消耗、场景价值与付费能力不同步

- **用户侧：高成本调用未必能够转化为高价值结果。**在用户端，成本与收入的错位主要体现在“高消耗场景未必形成高收益”。对于长上下文、长输出、多轮推理及仍需人工复核的任务，Token消耗和调用成本往往显著上升，但其结果未必能够直接转化为收入提升或效率改善。若模型输出仍需大量人工修正，其边际价值将明显下降，进而导致使用成本与转化收益之间出现偏离。
- **平台侧：高成本能力供给未必对应客户真实付费需求。**在平台端，错位主要表现为高性能模型的成本投入与客户实际需求之间的不匹配。高端模型通常对应更高的训练、推理和迭代成本，但企业客户的核心诉求往往是任务稳定、流程可嵌入和成本可控制，而非持续为更强性能支付溢价。若平台以高成本能力承接低价值或标准化任务，便容易出现成本上升快于收入释放的情况。
- **高Token消耗并不天然对应更强付费意愿与用户黏性。**Token消耗反映的是使用强度，而非商业质量。用户是否愿意持续付费，取决于产品是否解决明确问题、是否形成流程依赖，以及是否具备稳定的替代价值。若高消耗主要来自低价值交互、重复试错或低效率调用，其结果往往是成本增加，而非收入沉淀。价格通常围绕Token消耗、上下文长度、输出规模和实时性要求设定；收入兑现则取决于场景价值、付费意愿、续费能力与单位经济性。二者衡量标准不同，因此价格、成本与收入之间天然存在错位。AI商业化的关键，不在于单纯扩大调用规模，而在于提升高价值场景占比，使成本投入、产品定价与收入兑现形成更稳定的匹配关系。

来源：头豹研究院

业务合作

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/词条

行企研究多模态搜索引擎及数据库，募投可研、尽调、IRPR等研究咨询

定制白皮书

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

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

市场地位确认

对客户竞争优势进行评估和证明，助力企业价值提升及品牌影响力传播

行研训练营

依托完善行业研究体系，帮助学生掌握行业研究能力，丰富简历履历

报告作者

袁栩聪
首席分析师

麦嘉昊
行业分析师

- service@leadleo.com

业务咨询

- **客服电话：400-072-5588**
- **官方网站：www.leadleo.com**

商务咨询与深度合作

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

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