

# 2026年中国Token经济产业研究

Token成为新信息单位，谁将掌握AI时代的产业控制点？

**概览标签：Token经济、AI算力、大模型、人工智能**

China's Token Economy Industry

中国のトークン経済産業

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

➤ AI产业正从模型竞赛进入应用规模化阶段，产业当前关注重点由模型可用性转向任务成本可控性和稳定性。回顾电报、电话和互联网三次信息产业演进，字符、通话分钟和流量等新信息单位一旦被标准化并进入商业计价体系，通常会带来单位成本下降、需求规模扩张和产业价值迁移。Token正在成为AI时代连接模型推理、资源消耗、API计费和任务执行的基础单位，因此有必要将其放入百年信息单位商业化规律中理解。本报告重点回答Token经济是否正在复现这一产业规律、AI产业价值将如何从算力与模型供给迁移至效率技术、模型调度和场景入口，以及中国企业在推理侧、端侧和工作流场景中的结构性机会。

## 1. 研究起点：为什么现在需要研究Token经济？

AI竞争正在从训练能力转向推理经济性，Token成为连接模型能力、资源消耗和业务ROI的核心口径。

- 模型能力仍是基础门槛，但企业客户开始关注单位任务成本、响应时延、任务成功率和服务稳定性。
- Token把模型调用、算力消耗、API计费和应用成本放入同一核算框架，使AI支出具备可比较、可拆解和可优化的产业属性。

## 2. 规律验证：Token经济是否正在复现百年产业规律？

已具备第一项条件：Token进入计价体系，第二项条件（单位任务成本持续下降）和第三项条件（低成本推理激活高频需求）正在形成但尚未完全兑现。

- 电报的“字符”、电话的“分钟”、互联网的“流量”都经历了计价标准化、单位成本下降和需求规模扩张过程。
- Token作为AI时代的认知生产单位，正在把模型推理、工具调用和任务执行转化为可计量、可收费、可优化的商业活动。

## 3. 价值迁移：哪些环节先兑现收入，哪些环节掌握长期价值？

短期收入先由Token生产层兑现，长期价值将向效率层、调度层和消费层迁移。

- GPU、AI服务器、智算中心、云平台和高速互联最先承接Token调用放量带来的基础设施需求。
- 随着Token供给趋于充裕，推理优化、模型路由、Agent平台和企业工作流入口将成为更关键的产业控制点。

## 4. 竞争变化：AI企业分化是否会从模型能力转向单位任务成本？

模型能力趋近后，单位任务成本将成为企业采购和厂商分层的重要指标。

- Agent和企业工作流会拉长调用链条，Token消耗、任务成功率、重试次数和人工复核成本共同决定真实成本。
- 具备模型压缩、缓存复用、模型路由、Fallback和结果评估能力的企业，更容易在效率层和调度层形成差异化。

## 5. 中国机会：非对称格局下，哪些方向更容易受益？

中国Token经济的结构性机会更可能集中在低成本推理、端侧入口、企业工作流和垂直Agent。

- 训练侧受高端算力、先进芯片和资本开支约束，推理侧更依赖工程优化、模型压缩、端侧部署和场景适配。
- 中国具备互联网用户规模、政企数字化需求、制造业场景和终端生态优势，更适合通过高频任务和行业落地地放大Token消费。

# 内容目录

|                                 |    |
|---------------------------------|----|
| ◆ Token经济研究起点：为什么现在需要研究？        | 5  |
| • 产业信号：AI竞争焦点正在从训练能力转向推理经济性     | 6  |
| • 百年锚点：从字符、分钟、流量到Token          | 7  |
| ◆ 历史规律复盘：新信息单位如何重构产业价值？         | 8  |
| • 电报时代：字符计费与远程信息市场形成            | 9  |
| • 电话时代：通话分钟与运营商定价权              | 10 |
| • 互联网时代：流量扩张与平台入口迁移             | 11 |
| • 百年规律总结：成本下降、需求迁移与价值迁移         | 12 |
| • 周期中的潮起潮落：每轮信息单位重新定义都是企业座次的重排  | 13 |
| ◆ Token经济验证：AI时代是否正在复现百年规律？     | 14 |
| • 商业计量信号：Token是否已正式进入AI商业计价体系？  | 15 |
| • 成本下行信号：Token是否进入单位成本下降周期？     | 16 |
| • 需求扩张信号：Token降价是否正在放大使用需求？     | 17 |
| • 竞争变化信号：产业竞争是否从模型能力转向单位任务成本？   | 18 |
| ◆ Token经济价值链拆解：产业价值向哪里迁移？       | 19 |
| • 第一层：Token生产层先兑现基础设施红利         | 20 |
| • 第二层：Token效率层解决规模化使用的成本瓶颈      | 21 |
| • 效率层企业：DeepSeek、月之暗面、智谱等的差异化定位 | 22 |
| • 第三层：Token调度层在多模型环境中形成资源配置价值   | 23 |
| • 第四层：Token消费层决定需求规模和价值留存能力     | 24 |
| ◆ 中国视角：非对称格局下的产业位置              | 25 |
| • 中国AI产业坐标：训练侧约束与推理侧机会          | 26 |
| • 中国企业分层：谁在生产Token，谁在组织Token需求  | 27 |
| • 五年关键变量：端侧落地、平台定型与监管边界         | 28 |
| ◆ 最终判断：Token经济是否复现百年产业规律？       | 29 |
| • 核心结论：Token经济已经出现百年产业规律的早期特征   | 30 |
| ◆ 头豹业务合作介绍                      | 31 |
| ◆ 方法论与法律声明                      | 32 |

# 名词解释

- ◆ **Token:** 词元，大模型处理文本、代码、图片转写、音频转写等信息时使用的基础计量单位。Token既是模型推理过程中的技术单位，也是API计费、成本核算和任务消耗评估的重要口径。
- ◆ **输入Token:** 用户问题、提示词、上下文、文档、知识库材料、图片或音频转写内容进入模型时形成的Token。输入Token影响模型理解成本，尤其在长上下文、企业知识库和多轮Agent场景中占比较高。
- ◆ **输出Token:** 模型生成文本、代码、推理结果、摘要、回复或结构化内容时产生的Token。输出Token直接影响生成成本，通常与回答长度、任务复杂度和模型生成策略相关。
- ◆ **缓存Token:** 模型对重复上下文、历史会话或固定知识库内容进行复用形成的Token计费口径。缓存Token可以降低重复输入成本，是长文档、多轮对话和企业工作流控制AI支出的关键手段。
- ◆ **任务Token:** 完成一次业务任务所消耗的总Token，包括输入、输出、缓存、工具调用、多轮修正和重试等环节。任务Token更接近企业采购视角，适合衡量单位任务成本和AI应用ROI。
- ◆ **单位任务成本:** 企业完成一次具体AI任务所需支付的综合成本，包括Token费用、模型调用费用、重试成本、人工复核成本、系统集成成本和失败回退成本。该指标比单纯Token单价更能反映AI应用的商业化可行性。
- ◆ **上下文窗口:** 模型在一次推理中能够接收和处理的信息长度。上下文窗口越长，模型越适合处理长文档、代码库、复杂对话和多步骤任务，但也会提高输入Token消耗和推理资源压力。
- ◆ **KV Cache:** 大模型推理过程中保存历史Key和Value状态的缓存机制，可减少重复计算并提升生成效率。在长上下文、多轮对话和Agent任务中，KV Cache占用会影响显存、时延和单位Token成本。
- ◆ **模型路由:** 根据任务类型、成本约束、质量要求和响应时延，在不同模型之间动态分配请求的机制。模型路由可以让简单任务调用低成本模型，复杂任务调用高能力模型，从而降低整体单位任务成本。
- ◆ **Agent:** 能够围绕目标进行任务拆解、计划生成、工具调用、结果校验和多轮修正的AI系统。Agent会显著拉长Token调用链条，也会提升模型调度、成本监控和任务成功率的重要性。
- ◆ **MaaS (Model as a Service) :** 以API、模型平台或云服务形式向开发者和企业提供模型调用、部署、调度和管理能力的服务模式。MaaS平台通常连接多个模型、开发工具和企业客户，是Token调度层的重要载体。

# Chapter 1

## Token经济研究起点：为什么现在需要研究？

---

- 产业信号：AI竞争焦点正在从训练能力转向推理经济性
- 百年锚点：从字符、分钟、流量到Token

# Token经济研究起点：为什么现在需要研究？

- AI产业正在从训练能力竞赛进入推理规模化落地阶段，Token成为衡量模型调用成本、资源消耗和任务ROI的核心口径，单位任务成本开始决定AI商业化边界

## 产业信号



- 

**训练能力决定进入门槛，推理经济性决定商业化边界**  
大模型发展早期，竞争围绕训练算力、参数规模、数据质量和榜单成绩展开，这一阶段决定的是企业能否进入基础模型供给体系。当通用模型能力趋近、头部差距收窄后，单一的能力指标已不足以解释采购决策。进入规模化落地阶段，企业购买的是客服响应、代码生成、文档处理、知识检索等可验收结果，采购指标随之转向单位任务成本、响应时延、任务成功率与服务稳定性——这一转向，构成Token经济研究的产业前提。
- 

**Token进入计价体系后，AI支出开始从模型报价下沉到任务成本**  
Token进入计价体系后，成为连接技术能力、资源消耗与业务回报的核算单位。输入Token对应问题、上下文、文档与知识库，输出Token对应生成结果，缓存Token对应重复上下文复用，任务Token对应一次业务流程的总消耗。四类口径叠加，使AI成本从笼统的“模型报价”下沉为可拆解的“任务成本”，企业由此能够横向比较不同模型、不同上下文长度、不同缓存与路由方案下的真实支出。
- 

**推理降本激活大量任务需求，从而改变AI应用的需求边界**  
推理降本正在改变AI应用的需求边界。推理成本较高时，AI应用更容易停留在单次问答、文案生成等低频辅助环节；单位Token成本下降后，多轮Agent、长上下文、企业知识库、代码任务流和多模态交互才具备规模化部署基础。因此，研究应关注低成本推理能否把潜在任务需求转化为高频、可付费的实际调用，并观察企业预算、续费和任务ROI提升。

■ Token经济的研究起点，是AI竞争从训练供给能力转向推理经济性与单位任务成本。Token作为统一核算口径，使AI成本从模型报价下沉到任务消耗层面，具备可比较、可拆解和可优化的产业属性。

来源：OpenAI、DeepSeek、Google、头豹研究院

百年锚点：从字符、分钟、流量到Token

- Token作为AI时代的新信息单位，为理解模型调用成本、资源配置与商业化路径提供了分析锚点，其研究意义是理解价值从算力与模型供给向效率、调度与任务入口迁移的机制

百年锚点

从字符、分钟、流量到Token：新信息单位商业化会重排产业控制点

|      | 核心计量单位    | 单位背后的产业能力        | 效率优化方向                     | 价值沉淀位置         |
|------|-----------|------------------|----------------------------|----------------|
| 电报   | 字符/词数     | 电缆、电报网络、通信线路     | 编码压缩、线路复用                  | 电报局、新闻社、远程信息入口 |
| 电话   | 通话分钟      | 交换网络、长途线路、号码体系   | 多路复用、自动交换                  | 运营商、号码资源、用户关系  |
| 互联网  | 数据包/字节/流量 | 光纤、IDC、CDN、云基础设施 | 压缩算法、缓存、内容分发               | 搜索、社交、电商、短视频平台 |
| 人工智能 | Token     | 算力、模型、推理服务       | 模型压缩、KV Cache、模型路由、Agent编排 | 效率层、调度层、场景入口   |

研究背景：Token经济应放在百年信息单位商业化规律中理解，基础设施需求通常先释放

电报时代，字符和词数把远程文字传输变成可计价服务；电话时代，通话分钟把实时连接变成持续收费服务；互联网时代，字节、数据包和流量支撑内容访问、平台分发和交易转化。每一轮都先释放基础设施需求，如电缆与电报局、交换网络与号码体系、光纤与IDC/CDN，相应环节最先兑现收入，待单位成本下降、需求边界扩大后，价值重心向更低成本处理信息、控制入口、组织高频需求的一方迁移。

研究边界：Token与前几轮信息单位的可比性在于商业化规律，降本机制需单独判断

字符、分钟、流量本质都是“传输单位”，其降本由网络带宽与复用技术驱动，受物理链路和资本开支硬约束；Token是“生成/计算单位”，其降本由算力供给、模型架构与厂商竞争共同驱动，更接近持续叠加的工程曲线。这一差异意味着：一、Token降本的速度可能快于历史上的传输单位；二、Token降本受算力供给与电力等实体约束。由此，Token与字符、分钟、流量的可比性主要体现在商业化规律：新计量单位进入计价体系后，会重排产业控制点；其降本速度和工程约束仍需按照生成/计算单位的属性单独判断。

研究重点：Token经济关注认知生产单位商业化后的成本下降、需求扩张和价值迁移

Token把模型调用、推理资源消耗、API计费、应用成本和任务结果放在同一口径下，使AI产业可以围绕单位任务成本展开竞争。若Token成本持续下降并带来任务需求扩张，产业价值预计沿Token生产能力、Token效率优化、Token资源调度和Token场景消费逐层迁移。

- Token经济应放在百年信息单位商业化规律中理解，其产业含义在于推动AI价值从模型与算力供给，向推理效率、资源调度和任务入口迁移。

来源：GBPS、AT&T、Alibaba Cloud、头豹研究院

## Chapter 2

# 历史规律复盘：新 信息单位如何重构 产业价值？

---

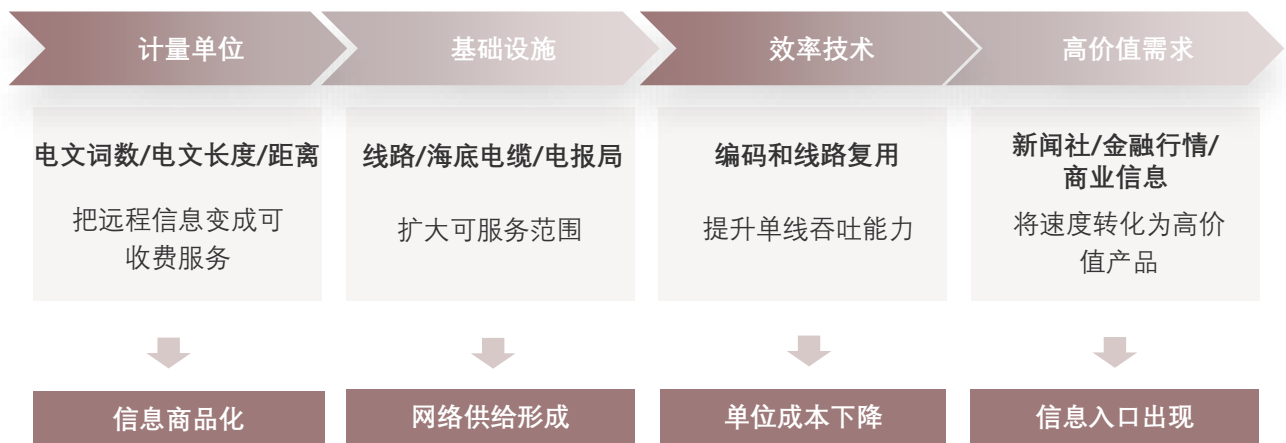
- 电报时代：字符计费与远程信息市场形成
- 电话时代：通话分钟与运营商定价权
- 互联网时代：流量扩张与平台入口迁移
- 百年规律总结：成本下降、需求迁移与价值迁移
- 周期中的潮起潮落：每轮信息单位重新定义都是企业座次的重排

## ■ 电报时代：字符计费与远程信息市场形成

- 字符计费把远程通信变成可结算市场，电报产业由此形成基础设施先行、效率技术降本和信息入口获取溢价的早期结构。这为理解Token经济提供了第一层历史参照

### 电报时代

字符计费把远程通信变成可结算市场，效率技术和信息入口开始分配产业价值



新信息单位进入计价体系后，需求会先集中在高价值场景，随后由降本打开市场

电文字符、词数、消息长度和传输距离成为计费依据后，远程文字传输从通信能力转化为可定价、可结算、可规模扩张的商业服务。高价格决定了早期需求集中在新闻、金融行情、商业指令与政府通信等对速度高度敏感的场景；价格门槛之下，电报并未立即成为大众工具。这说明新信息单位进入计价体系的第一阶段，需求由少数高价值场景承接。

基础设施先兑现收入，效率技术决定资产产能释放水平

陆上电报线、海底电缆和电报局扩张后，远程通信覆盖范围扩大，基础设施企业率先受益。随着网络覆盖提升，产业关注从线路建设量转向单线路信息承载能力。摩尔斯码、复式电报和四复式电报提高单线吞吐能力，降低单位信息传输成本，使同一网络资产可以承载更多电文需求。

信息入口把通信速度转化为可出售的高价值产品

新闻社、金融信息服务与商业通信机构把电报的时效转化为标准化信息产品，掌握客户关系与分发渠道，获取超额收益。映射到Token经济：算力与模型供给会先兑现收入，但只有当单位任务成本下降、并被办公、客服、代码、研发等高价值任务吸收，Token需求才会从试用走向规模化。

- 电报时代的启示是，新信息单位只有进入计价体系，才能把技术能力转化为效率红利与入口溢价。对Token经济而言，算力与模型供给是起点，低成本Token能否进入高价值任务场景，将决定规模化空间。

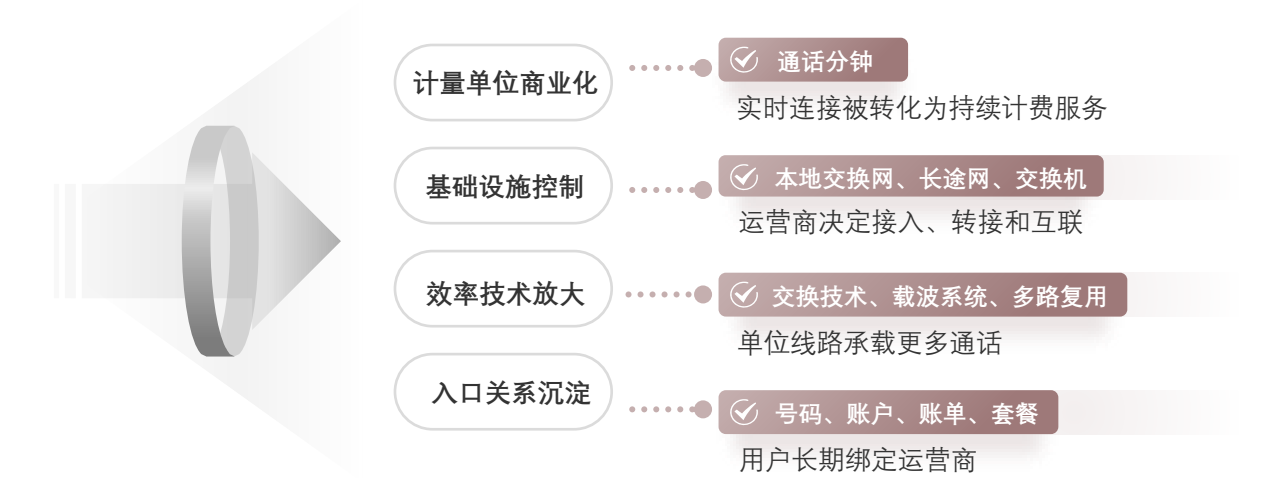
来源：NBER、The Journal of Economic History、ETHW、Rutgers、Reuters、The Associated Press、头豹研究院

## 电话时代：通话分钟与运营商定价权

- 通话分钟把实时连接转化为持续计费服务，运营商因掌握网络、号码、账户和账单关系获得稳定定价权。对应AI时代，持续任务关系将比单次调用更具长期价值

### 电话时代

通话分钟把实时连接变成持续收入，运营商因控制网络、号码和账单获得定价权



电话时代的关键变化，是信息单位从离散消息升级为持续连接时长

通话分钟成为计量单位后，通信服务从一次性信息传递变为持续占用网络资源的实时连接，运营商可围绕分钟、距离、套餐、月租与账单形成重复收入，收入模型较电报时代更稳定。计费方式的升级，本身就改变了价值沉淀的位置。

运营商获得定价权，来自网络控制、效率放大和用户入口的叠加

本地交换网、长途线路、交换机和号码体系构成重资产壁垒，运营商决定用户接入、转接、互联和计费方式。自动交换、载波传输和多路复用提高同一线路承载能力，进一步放大网络规模经济。号码、账户和账单关系则把用户长期绑定在运营商体系内，使运营商从网络供给方升级为连接入口方。单一环节的技术领先无法复制这种地位，控制点的组合才是关键。

对Token经济的启示在于，持续任务关系比单次模型调用更有价值

当AI使用从单次问答走向企业工作流、Agent任务和长期助手，账单关系、任务记录、权限体系 and 结果交付将变得更重要。模型单价可以下降，但掌握用户账户、企业流程、任务入口和成本账单的企业，更可能获得稳定议价权。

- 电话时代的长期价值来自持续连接、号码关系和账单体系。对应Token经济，当AI使用从单次调用走向持续任务流，掌握账户、任务入口、成本账单和客户关系的企业将获得更强议价权。

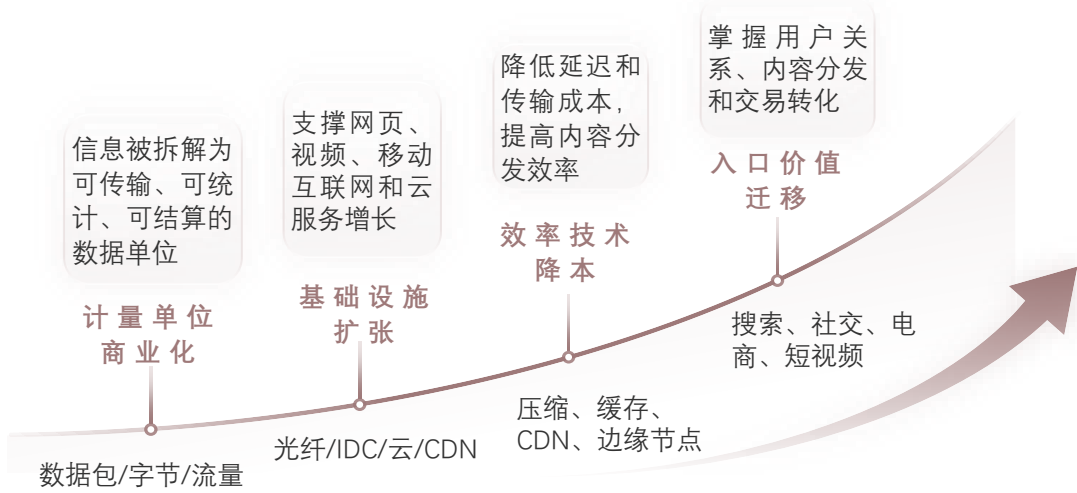
来源：Federal Communications Commission、ETHW、头豹研究院

## 互联网时代：流量扩张与平台入口迁移

- 互联网流量扩张率先带动光纤、IDC、CDN和云基础设施需求，但长期利润向用户入口、内容分发和交易闭环迁移。Token经济也将经历类似的价值迁移过程

### 互联网时代

流量降本释放内容分发需求，长期利润向用户关系和交易闭环迁移



基础设施承接流量扩张，长期利润会向用户关系、内容分发和交易闭环迁移

数据包、字节和流量成为核心计量单位后，网页、搜索、视频、移动应用和云服务具备规模扩张基础。光纤、宽带、IDC、云和CDN首先受益于用户、设备和内容接入数量增长，但这些环节主要解决承载能力，并不直接决定最终利润归属。

流量降本改变需求结构，使互联网从信息访问扩展到商业转化

压缩算法、缓存、CDN和边缘节点降低单位内容分发成本后，流量需求从网页访问扩展到搜索、社交、电商、直播、短视频和移动应用。分发半径扩大，也抬升了广告、交易、会员、云服务等商业模式的可行性。在此，降本的作用是把“可用性”推向“高频可负担”，从而打开新的消费场景。

平台企业获得长期利润，关键在于把流量转化为可变现关系

搜索平台把用户意图转化为广告收入，社交平台把用户关系转化为内容分发效率，电商平台把访问流量转化为交易佣金和商家服务。映射到Token经济：基础设施会承接Token需求放量，长期利润取决于企业能否把Token消耗转化为用户关系、企业 workflows、数据闭环与业务结果。

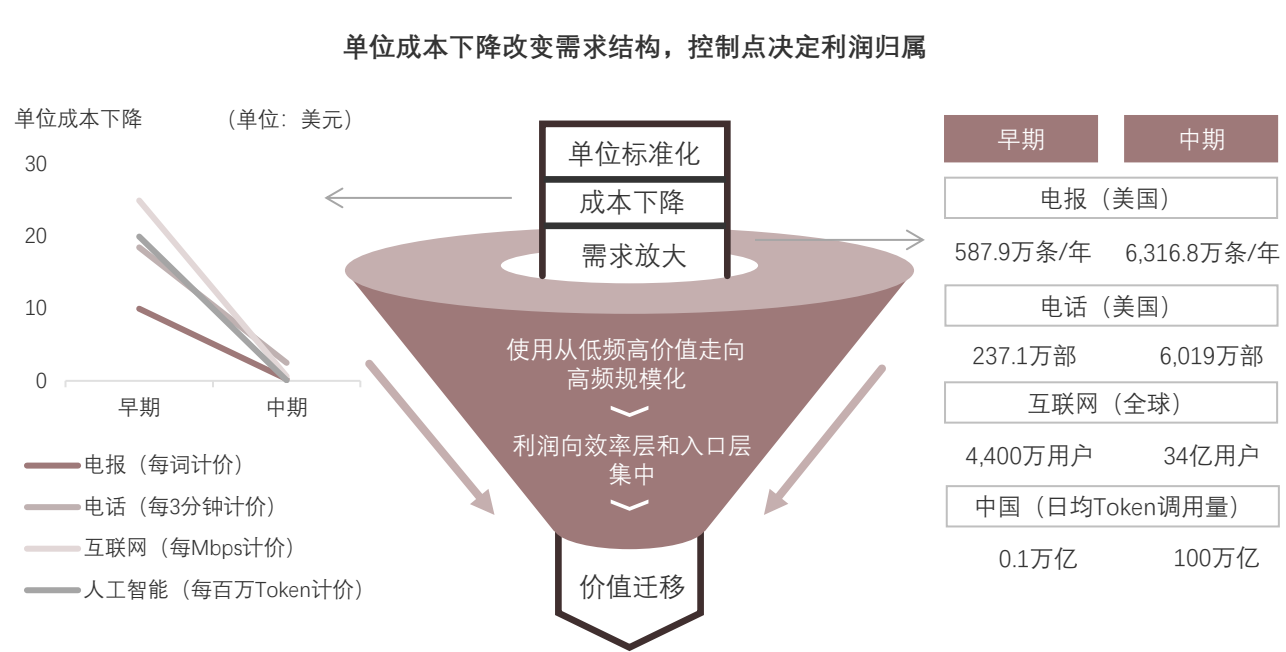
- 互联网时代表明，基础设施支撑流量规模扩张，长期利润会向掌握用户入口、内容分发和交易转化的平台层迁移。对应Token经济，基础设施先承接放量，长期利润取决于能否把Token消耗转化为 workflow 嵌入与商业转化。

来源：RFC、ITU、W3C、SEC、头豹研究院

百年规律总结：成本下降、需求迁移与价值迁移

- 百年信息产业规律显示，新信息单位商业化通常带来单位成本下降、需求规模扩张和价值迁移。Token经济能否成立，核心在于这三项条件是否在AI产业中持续出现

百年规律总结



百年技术革命揭示，新信息单位进入商业计价体系后，产业会沿“标准化-降本-需求放大-价值迁移”展开

字符、通话分钟和流量分别推动电报、电话和互联网形成可收费、可结算、可扩张的市场。基础设施在初期最先受益，因为新单位需要网络、设备和服务能力支撑；随后单位成本下降扩大需求边界，使用场景从少数高价值客户进入更高频、更广泛的应用场景。

“杰文斯悖论”机制下，成本下降的产业作用将取决于需求的价格弹性

当单位成本下降，若需求富有弹性，使用量的增长会超过单价的下降，总支出反而上升。电报降价后远程信息从政府、金融、新闻扩展到广泛商业通信；电话接入与长途成本下降后通信进入家庭与企业；互联网传输与分发成本下降后流量从网页访问扩展到视频、社交、直播与移动应用——历史均印证了高弹性下“降价做大市场”的规律。若需求弹性不足，价格下降会更多体现为供给方利润压缩，难以推动行业收入池扩张。

Token经济是否复现百年规律，需要同时观察三项条件

第一，Token是否已经成为AI商业计量单位；第二，单位Token或单位任务成本是否进入下降周期；第三，低成本Token是否正在激活更高频、更复杂、更可付费的任务需求。若三项条件持续成立，产业价值预计从Token生产层逐步向效率层、调度层和消费层迁移。若调用增长主要来自试用、补贴或低价值流量，且企业付费、续费和任务结果未同步提升，Token经济将停留在需求启动阶段，价值兑现会出现折扣。

- Token经济成立的关键，在于低成本Token能否进入高频、复杂、可付费任务并转化为企业预算与可验收结果；调用量增长本身只是前置信号，需求弹性决定百年规律能否在AI时代复现。

来源：微软、日本经济研究所、FRED、Census、Our World in Data、Stanford HAI、头豹研究院

■ 周期中的潮起潮落：每轮信息革命中企业座次的重排

- 每轮信息单位切换都会重新定义企业核心资产，旧周期优势需要在新周期重新验证。Token经济下，成本效率、模型调度和任务入口将成为企业座次重排的关键控制点

信息革命中企业座次的重排

信息单位切换会重写核心资产，新周期赢家往往掌握效率或入口控制点

| 信息单位切换         | 上升企业案例                                                                                                                                                                                                                                                                                                                                                                                                                                                 | 下行或失速企业案例         | 座次变化原因                                     |
|----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|--------------------------------------------|
| 电文/电报<br>↓     |  AT&T  中国电信                                                                                                                                                                                                                                                                          | Western Union电报主业 | 价值点从电文传输转向实时连接、号码、交换网络和账单                  |
| 通话分钟<br>↓      |  Google  Alibaba<br> Meta  微信  抖音 | Aol.    yahoo!    | 价值点从网络连接、接入服务和门户网站，迁移到搜索、社交关系、交易、推荐分发和移动应用 |
| 通话分钟/门户流量<br>↓ |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                   |                                            |
| 平台与算法入口<br>↓   |                                                                                                                                                                                                                                                                                                                                                                                                                                                        |                   |                                            |
| 流量入口<br>↓      |  OpenAI  deepseek                                                                                                                                                                                                                                                                    | 待验证               | 价值点可能从模型能力转向单位任务成本、模型调度和场景入口               |
| Token任务入口      |  ANTHROPIC  智谱                                                                                                                                                                                                                                                                   |                   |                                            |



- 报告完整版/高清图表或更多报告：请登录 [www.leadleo.com](http://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](mailto:service@leadleo.com)

来源：ETHW、Axios、The Guardian、头豹研究院

# Chapter 3

## Token经济验证： AI时代是否正在复 现百年规律？

---

- ❑ 商业计量信号：Token是否已正式进入AI商业计价体系？
- ❑ 成本下行信号：Token是否进入单位成本下降周期？
- ❑ 需求扩张信号：Token降价是否正在放大使用需求？
- ❑ 竞争变化信号：产业竞争是否从模型能力转向单位任务成本？

商业计量信号：Token是否已正式进入AI商业计价体系？

- Token已经进入AI商业计价体系，输入、输出、缓存和任务Token构成企业成本核算基础。AI支出由此具备可计量、可定价、可比较和可优化的颗粒度

商业计量信号

Token支出开始具备可核算、可比较、可优化的颗粒度



- 可计量

Token已经成为模型调用和企业使用成本核算的基础单位

Token商业化的第一步是可计量。主流模型服务商已在API响应、价格页与账单系统中披露输入、输出、缓存Token用量，企业能在调用前估算、调用中追踪、调用后复盘。AI成本由此从难以拆解的服务费，转化为可计量的底层资源消耗。
- 可定价

输入、输出、缓存Token已构成主流模型服务的价格结构

Token商业化的第二步是可定价。国内外平台普遍以每百万Token列示价格：OpenAI区分Input、Cached input、Output；DeepSeek区分Input Tokens的Cache Hit、Cache Miss与Output Tokens。
- 可比较

跨厂商Token口径趋同，模型采购形成横向成本比较基础

Token商业化的第三步是可比较。模型、上下文长度、缓存命中率与输出长度的差异，会使同一任务呈现不同成本结构，企业采购逻辑随之从"比模型"转向"比同一标准任务在不同模型组合下的总成本"。需要说明的是，各模型的Token切分方式与质量表现不一致，单纯Token单价不足以作横向比较，更合理的对象是标准化任务包成本。
- 可优化

缓存、Batch和模型路由使Token成本进入运营管理环节

Token商业化的第四步是可优化。缓存、Batch、模型路由、提示词压缩和上下文管理，使Token成本进入运营管理环节。长文档、多轮 Agent、知识库问答和企业工作流场景中，缓存命中率、重试率、模型选择和上下文复用都会影响最终账单。Token由此成为可优化的运营指标。
- Token已进入AI商业计量体系，并具备可计量、可定价、可比较、可优化四类产业属性。这验证了百年产业规律中的第一项条件，即新的信息单位已具备商业化结算基础。

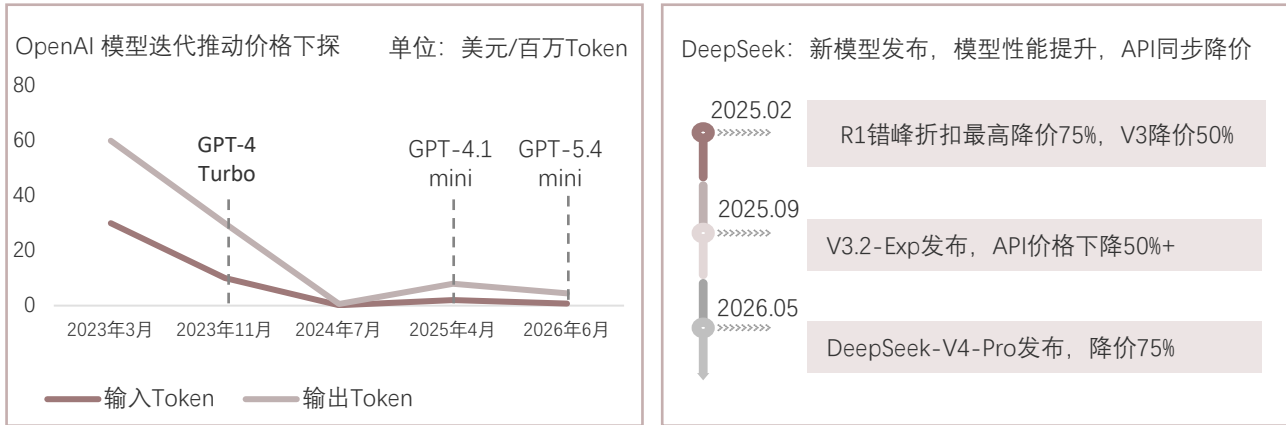
来源：OpenAI、Gemini、Claude、DeepSeek、腾讯、头豹研究院

## 成本下行信号：Token是否进入单位成本下降周期？

- Token成本下降已从API报价下探，扩展到同等能力推理效率提升和模型分层供给。收入池能否扩大，取决于任务频次、复杂度和付费规模能否跑赢单价下降

### 成本下行信号

Token 成本下降已从模型降价扩展到同等能力推理效率提升



#### Token成本下降的两个层面观察：模型服务报价下探和同等能力推理成本下降

模型服务报价下降体现厂商价格竞争，同等能力水平下的任务完成成本决定企业能否扩大部署。同等能力的推理成本变化可作直接证据：Stanford AI Index 2025显示，达到GPT-3.5水平系统的推理成本在约18个月下降超过280倍，不同任务下LLM推理价格年降幅可达9至900倍。

#### 模型分层正在把不同复杂度任务放入不同价格区间

GPT-4 Turbo相较GPT-4实现输入价格约3倍下降、输出价格约2倍下降；GPT-4o mini以低价切入高频调用场景。到2026年，GPT-5.4 mini价格为0.75/4.50美元/百万Token，适用于高频、子任务和Agent分工场景。厂商通过分层供给，使高价值任务调用能力更强的模型，高频简单任务由轻量模型、缓存与Batch承接，从而压低整体单位任务成本。

#### 国内模型价格同向加速，倒逼单位任务成本下移

2025年2月，DeepSeek推出错峰API折扣；2025年9月，V3.2-Exp发布并同步宣布API价格下降50%+；2026年5月，DeepSeek推出V4-Pro并降价75%。对企业客户而言，低价API和开源模型扩散降低了接入门槛；对模型厂商而言，价格下降会压缩单纯API利润，迫使竞争继续转向推理效率、缓存计费、模型分层、任务成功率和服务稳定性。收入池扩张需要任务频次、任务复杂度和付费规模同步提升，以覆盖单价下降带来的利润压力。

- Token成本下降已经从API报价下探，延伸到同等能力下的任务完成成本下降。价格下行会压缩单纯模型API利润，也会扩大高频调用需求，推动竞争焦点转向推理效率、任务成功率和单位任务成本。

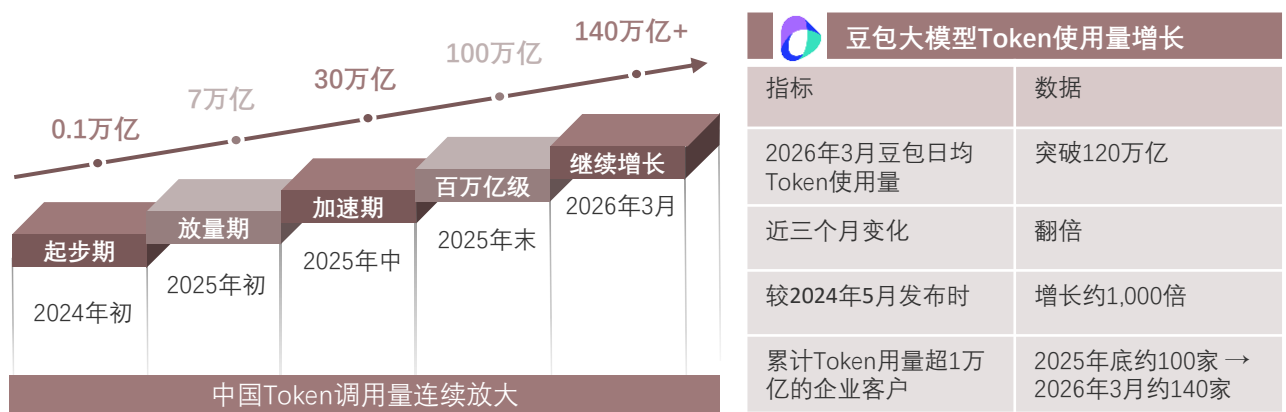
来源：Stanford AI Index、OpenAI、Reuters、DeepSeek、头豹研究院

# 需求扩张信号：Token降价是否正在放大使用需求？

- 中国Token调用量连续放大，说明低成本推理正在激活更高频、更广泛的AI使用需求。后续判断重点应从Token调用量，转向企业付费、续费和可验收任务结果

## 需求扩张信号

调用量高增验证需求被激活，长期价值仍要回到可付费任务结果



01

需求激活

中国Token调用量快速增长，是低成本推理进入规模化使用阶段的前置信号

国家数据局披露，全国日均Token调用量从2024年初约0.1万亿增长到2025年底约100万亿，并在2026年3月突破140万亿。《全国数据资源调查报告（2025年）》显示，2025年全年Token调用量约21,100万亿。说明产业资源消耗重心正从模型训练转向应用调用。需求由早期探索进入跨平台、跨行业、跨任务的规模化调用阶段。

02

动力来源

平台级数据进一步显示，Token消耗增长正在由消费端、企业端和多模态共同拉动

火山引擎披露，截至2026年3月，豆包大模型日均Token使用量突破120万亿，三个月内翻倍，较2024年5月发布时增长约1,000倍；平台上累计Token使用量超过1万亿的企业客户，从2025年底约100家增至2026年3月约140家。用户的日常交互在消费端产品贡献了核心使用频次与Token消耗；企业端在客服、营销、办公等多元场景中提升调用强度；而多模态能力的迭代升级，进一步拓宽了Token消耗的场景维度与技术支持空间。

03

价值转化

Token需求已经被激活，价值兑现仍需验证高消耗能否转化为可付费任务结果

Token调用量快速增长证明需求已经启动，这反映大模型应用走向更高频、更广泛的产业落地。对企业客户而言，Token消耗越大，预算管控、任务成功率和ROI评估越重要；模型路由、缓存复用、轻量模型和成本监控将成为控制单位任务成本的关键手段。后续产业观察重点，应从Token总量转向可计价、可验收、可复用的任务结果。

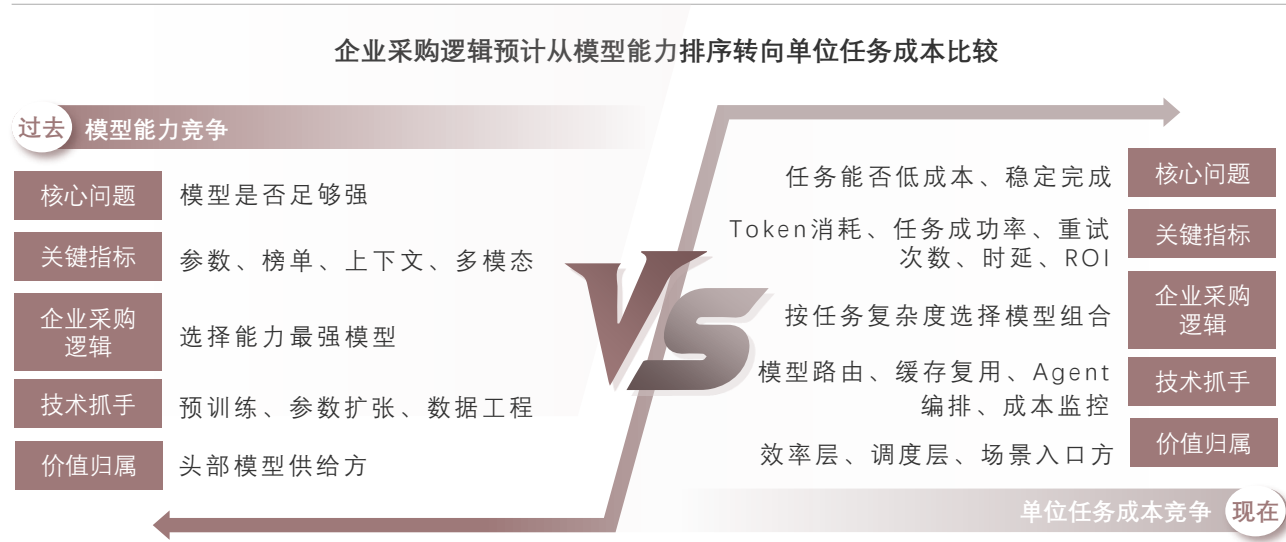
Token调用量增长说明需求被激活，但商业价值兑现仍取决于企业是否为任务结果付费。观察指标应从调用量转向付费客户、席位、续费和任务ROI。

来源：国家数据局、新浪财经、头豹研究院

## 竞争变化信号：产业竞争是否转向单位任务成本？

- 模型能力仍是基础门槛，但企业采购逻辑正在从能力排序转向单位任务成本比较。Agent和企业工作流拉长调用链条后，任务成功率、重试率和人工复核成本将共同拉开厂商差距

### 竞争变化信号



**模型能力仍是基础门槛，单位任务成本开始决定规模化部署能力**

通用模型能力仍然重要，但模型能力差距正在缩小。Stanford AI Index 2025显示，开源模型与闭源模型在部分基准上的性能差距从8%缩小至1.7%，能力趋近与推理降本同时发生。当模型能力逐步成为基础门槛后，成本效率会成为更重要的选择标准。客户将更关心模型是否能在可控成本下稳定完成业务任务。

**Agent和企业工作流拉长任务链条，会放大单位任务成本差异**

Agent与企业工作流放大了单位任务成本的差异。单次问答主要消耗一次输入与输出Token；Agent任务包含任务拆解、计划生成、工具调用、数据库检索、结果校验、多轮修正与异常处理，每一步都带来新的Token消耗、时延与失败风险。一个Token单价较低的模型，若任务成功率不足、需要多次重试或大量人工复核，最终任务成本仍可能偏高。因此，单价低与总成本低，在Agent链条下并不等价。

**单位任务成本将成为下一阶段产业分化指标，价值将向效率层、调度层和场景入口迁移**

具备模型压缩、推理优化、缓存管理和服务调优能力的企业，更容易在效率层获得价值；具备模型路由、成本监控、Fallback和评估体系的企业，更容易在调度层获得价值；掌握客户流程、行业数据和任务结果交付能力的企业，则更可能把Token消耗转化为持续收入。

- 模型能力仍是基础门槛，单位任务成本开始决定规模化部署能力。随着Agent和企业工作流拉长调用链条，Token消耗、任务成功率、重试率和人工复核成本将共同决定厂商分化。

来源：中国信通院、Stanford AI Index、头豹研究院

# Chapter 4

## Token价值链拆解： 产业价值向哪里迁移？

---

- ❑ 第一层：Token生产层先兑现基础设施红利
- ❑ 第二层：Token效率层解决规模化使用的成本瓶颈
- ❑ 效率层企业：DeepSeek、月之暗面、智谱等的差异化定位
- ❑ 第三层：Token调度层在多模型环境中形成资源配置价值
- ❑ 第四层：Token消费层决定需求规模和价值留存能力

■ 第一层：Token生产层先兑现基础设施红利

- Token生产层最先承接调用放量带来的基础设施需求，AI芯片、服务器、智算中心和云平台率先兑现收入，单位Token产能效率将成为生产层企业穿越周期的核心变量

Token生产层先兑现基础设施红利

Token生产层最先兑现收入，长期议价权取决于单位Token产能效率



Token生产层最先兑现收入，应用调用会直接传导为算力和数据中心需求

应用层每增加一次模型调用，底层都会形成芯片算力、显存、网络传输、存储访问和电力消耗。《数字中国发展报告（2025年）》显示，2025年底全国在用算力设施机架数超过1,373万标准机架，建成万卡智算集群42个，智能算力规模达到159万PFlops（FP16）。这构成Token经济的第一环：应用能否收费、Agent能否盈利仍待时间验证，算力设备、云资源与数据中心容量的投资需求已先行形成。

高并发推理正在推动算力、网络、存储和能源系统协同升级

高并发推理、长上下文、多模态和Agent任务会同时抬升芯片算力、显存带宽、高速互联、存储访问、数据中心电力和散热需求。AI芯片、HBM、AI服务器、高速交换机、光模块、液冷、数据中心与云平台共同决定Token生产的基础产能，单一环节的扩张无法独立释放产能，协同建设的重要性随之上升。

全球云厂商和芯片厂商已经验证生产层红利，但也提示红利的周期性

全球头部企业的财务数据已经验证基础设施层率先受益。NVIDIA FY2026数据中心业务全年收入达到1,937亿美元，同比增长68%；Microsoft FY2026 Q3资本开支达到319亿美元，其中约三分之二用于GPU、CPU等短周期资产。算力供给紧张时，芯片、服务器与云资源议价能力更强；供给改善、模型效率提升或客户预算收紧后，生产层可能面临价格下行与利用率压力。因此，生产层的长期竞争力将回到单位Token产能效率：同样的芯片、机柜与电力投入，谁能实现更高利用率、更低时延与更稳定服务，谁更能抵御资本开支周期波动。

- Token生产层收入兑现更早，但周期属性更强。能够穿越周期的生产层企业，需要把算力资源转化为更高吞吐、更低时延和更稳定的单位Token产能。

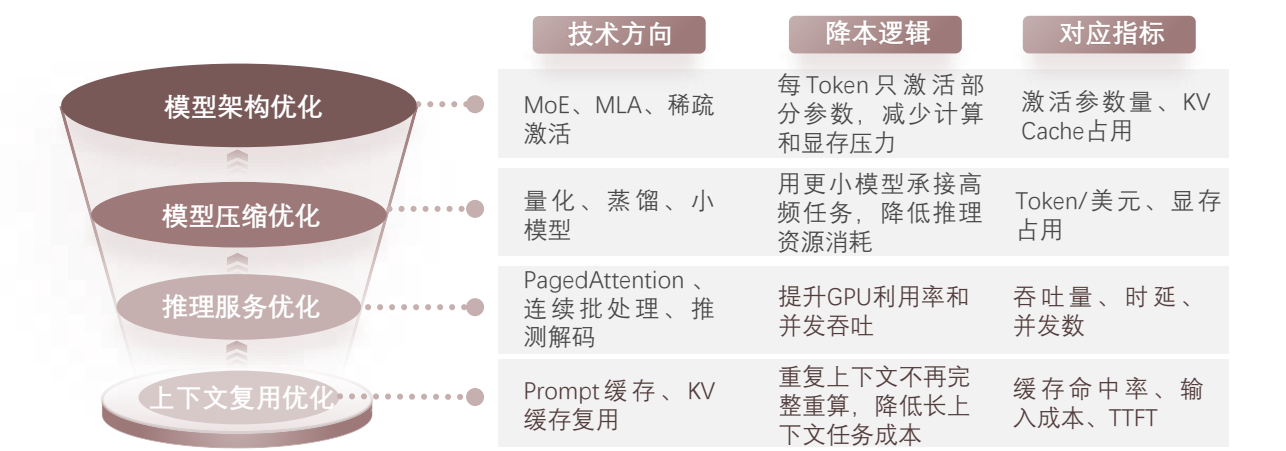
来源：国家数据局、工信部、NVIDIA、Microsoft、头豹研究院

## ■ 第二层：Token效率层解决规模化使用的成本瓶颈

- 效率层通过模型压缩、缓存复用、推理加速和服务工程降低单位任务成本，是AI应用规模化落地的关键瓶颈。该层决定算力产能能否转化为可盈利、可规模化的Token供给

### Token效率层解决规模化使用的成本瓶颈

效率层把算力供给转化为可盈利Token，是规模化落地的前置条件



效率层把算力产能转化为低成本、可盈利、可规模化的Token供给，是应用落地的前置条件。企业客服、代码生成、知识库问答和Agent任务一旦进入高频使用，成本会随调用次数、上下文长度、工具调用和重试次数快速累积。以DeepSeek V4为例，V4-Pro采用MoE架构，参数规模达1.6T，每Token激活49B参数。在100万Token上下文场景下，V4-Pro相较V3.2仅需约27%的单Token推理FLOPs和约10%的KV Cache。该案例说明，效率层通过模型架构和推理系统优化，持续降低长上下文、高频任务和复杂调用的单位Token成本。

效率层的产业价值在于能否用工程实现企业可感知的成本下降和服务体验的提升

Token效率提升是模型架构、推理系统和服务工程的组合优化。模型侧，MoE通过稀疏激活降低每个Token实际参与计算的参数量，MLA通过压缩KV Cache降低长上下文推理中的显存和带宽压力；服务侧，PagedAttention、连续批处理和推测解码提升GPU利用率和并发吞吐。这些优化最终须落到Token/美元提升、显存占用下降、吞吐提升、时延缩短、并发增强与缓存命中率提高等可测指标上，否则难以支撑企业采购。

效率层预计成为模型厂商差异化核心竞争环节，单点技术优势存在被内化的风险

当通用模型能力逐渐接近、API价格持续下降后，效率层将成为模型厂商分化的重要来源。模型架构优化、缓存复用和推理系统能力可能被头部模型厂商、云平台和开源生态吸收。具备相关能力的企业，需要继续向API平台、开发者生态、企业部署、模型调度或垂直场景延伸，才能把技术效率转化为客户留存和商业壁垒。

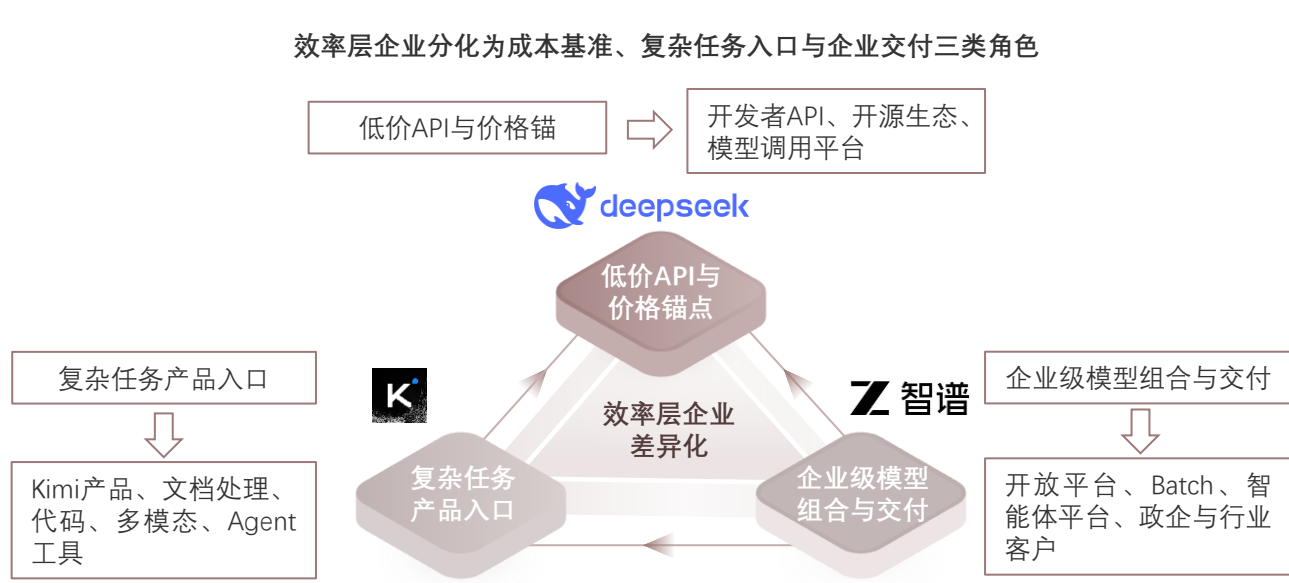
- 效率层决定AI规模化使用的成本边界，单点技术优势容易被头部模型厂商和云平台吸收，长期竞争力来自效率能力的产品化、平台化和场景嵌入。

来源：DeepSeek V4模型卡、arXiv:2606.19348、头豹研究院

■ 效率层企业：DeepSeek、月之暗面、智谱等的差异化定位

- 效率层企业正在分化为成本基准、复杂任务入口和企业交付三类角色。低价API、长上下文能力和行业交付能力，分别决定不同企业的价值兑现路径

DeepSeek、月之暗面、智谱等的差异化定位



第一类是成本基准型，以DeepSeek为例，“低价+开源”改变开发者对高能力模型的认知

DeepSeek用低价API与开源扩散降低开发者与中小企业的接入门槛，带来调用量与生态占有的提升，调用量与生态规模又使其成为同行API定价的横向参照。基准一旦确立，便反向倒逼竞品在推理效率、缓存计费与模型分层上持续投入。该类企业长期需要关注算力成本控制、开发者留存和企业调用质量。

第二类是任务入口型，以月之暗面为例，把模型能力组织成可连续使用的生产力工具

月之暗面更偏向长上下文、文档处理、代码、深度研究、多模态和Agent任务场景。Kimi K2.6支持文本、图片与视频输入，以及256K上下文，Kimi K2.7 Code则进一步面向代码生成和开发者任务。任务入口型企业的价值在于把Token消耗组织成连续任务流，并通过上下文承载、交互体验和任务连续性积累用户入口。该类企业更适合在个人效率、开发者工具和复杂内容生产场景中形成使用粘性。

第三类是企业交付型，以智谱为例，把模型能力转化为可采购、可集成、可运维的服务

智谱等厂商通过不同规格模型、Batch、智能体平台、私有化部署和行业项目交付，满足政企、金融、办公和行业场景对成本、合规、服务和稳定性的综合要求。企业交付型厂商的价值来自客户关系、项目交付能力、生态适配和行业数据沉淀，规模化复制能力决定其收入质量。

- 效率层企业的竞争将从模型能力延伸到商业化路径。成本基准型企业着重成本控制和价格策略，任务入口型企业考验用户留存，企业交付型企业关注规模化复制和收入质量。

来源：DeepSeek、Reuters、月之暗面、智谱、头豹研究院

第三层：Token调度层在多模型环境中形成资源配置价值

- 多模型供给扩大后，企业部署重点将从接入模型转向治理模型池。模型路由、成本监控、Fallback和结果评估， 将把分散模型供给转化为企业AI资源控制面

Token调度层在多模型环境中形成资源配置价值

多模型供给扩大后，调度层从模型接入工具升级为企业AI资源控制面

|      |          |                                                                             |
|------|----------|-----------------------------------------------------------------------------|
|      | 多模型供给增加  | 国内：通义开源300余个模型，全球下载量突破6亿次，衍生模型17万个<br>海外：OpenRouter连接300+模型、60+模型提供方        |
| 趋势信号 | 统一入口调用放量 | 国内：阿里云百炼模型日均调用量过去一年增长15倍<br>海外：OpenRouter从约10万亿Token/年增长至2025年中超过100万亿Token |
|      | 路由降本被验证  | 模型调度层案例：小度AI助手使用模型路由后，成本下降99%，响应速度提升200%；复杂场景仍可节省30%成本                      |



多模型供给扩张后，企业需要从接入模型转向治理模型池

缺少统一治理时，企业容易出现两类资源失配：简单任务过度调用高价模型，复杂任务误用低成本模型导致失败与重试。调度层通过统一接入、权限管理、模型选择、调用监控和成本账单，把分散模型供给转化为可配置的企业AI资源。以阿里云百炼平台为例，其提供千问系列以及DeepSeek、Kimi等第三方大模型，企业可在同一平台调用不同能力、不同成本和不同模态的模型；海外OpenRouter连接300+模型与60+提供方，承担同类模型池入口角色。

统一调用入口正在形成模型资源治理能力

阿里云披露，百炼平台模型日均调用量过去一年增长15倍；OpenRouter研究显示，其平台Token处理量在2025年末出现单日处理超过1万亿Token的规模，平台服务超过500万开发者。调用放量说明调度层开始具备平台属性，掌握模型接入、调用日志、成本账单、路由策略与Fallback机制后，平台具备成为企业AI资源控制面的条件，可把分散的模型调用和Token支出，转化为可监控、可分摊、可优化的资源成本。

模型路由验证了调度层的经济性，价值核心从“选模型”转向“控成本”

模型路由已验证降本价值。如百度智能云千帆提供模型路由能力，可按任务类型自动分配合适模型；小度AI助手使用模型路由后，相比全部任务调用旗舰版文心大模型，成本下降99%，响应速度提升200%。该案例说明，模型路由由能够直接影响单位任务成本和服务体验。调度层的核心价值，正在从连接模型转向治理模型池、控制任务成本和保障结果质量。

- Token调度层的产业价值在于治理模型池：在不同能力、价格、时延和任务质量要求之间配置模型资源，使企业AI调用从分散接入转向成本可控、质量可验收的资源管理。

来源：新华网、a16z、头豹研究院

■ 第四层：Token消费层决定需求规模和价值留存能力

- 消费层决定Token需求上限，价值留存来自高频任务、付费关系、数据闭环和结果交付。办公、客服、代码和行业 workflows 等高频入口，将决定Token消耗能否转化为持续收入

Token消费层决定需求规模和价值留存能力

消费层决定Token需求上限，真正议价权来自高频任务和付费闭环



中国互联网络信息中心报告显示，截至2025年12月，生成式AI应用C端用户规模达6.02亿，普及率42.8%，应用从聊天问答扩展到生活、学习、办公与生产场景。底层算力扩容和模型效率提升只能降低Token的供给成本，真正拉动Token消耗的是用户是否愿意持续使用、企业是否把AI嵌入工作流程、场景是否扩展到长期任务。

头部AI应用正在形成多入口竞争，消费层议价权来自用户频次和场景沉淀

QuestMobile数据显示，2026年3月豆包、千问、DeepSeek月活分别达到3.45亿、1.66亿和1.27亿，三者一季度平均活跃率分别为33.5%、17.1%和21%。模型和算力可以被替换，调度平台也可能被集成，但高频任务、组织账户、业务流程、行业数据与结果验收一旦沉淀在应用入口，客户迁移成本显著上升，入口的黏性由此形成。

消费层价值正在从流量延伸到付费关系和任务结果

消费层厂商的价值留存，来自将Token消耗绑定到高频、刚需、可验收的任务结果。当AI产品进入文档处理、代码开发、知识库、营销素材和行业 workflows 后，企业客户才会形成持续预算。消费层的关键不只是带来调用量，更重要的是把调用量转化为可复用任务、可计费结果和持续付费关系。

- 消费层直接连接客户任务和付费结果，是Token经济价值最容易留存的环节。高频任务、业务数据和结果交付能力，决定Token消耗能否从单次调用转化为持续业务支出。

来源：中国互联网络信息中心、OpenAI、QuestMobile、企业年报、头豹研究院

# Chapter 5

# 中国视角：非对称格局下的产业位置？

---

- ❑ 中国AI产业坐标：训练侧约束与推理侧机会
- ❑ 中国企业分层：谁在生产Token，谁在组织Token需求
- ❑ 五年关键变量：端侧落地、平台定型与监管边界

## 中国AI产业坐标：训练侧约束与推理侧机会

- 中国Token经济的结构性机会更集中在推理侧。训练侧受高端算力和先进芯片约束，推理侧更依赖工程优化、端侧部署和场景适配，更符合本土产业条件

### 训练侧约束与推理侧机会

训练侧资源约束强化头部集中，推理侧场景密度打开中国结构性机会

| 维度        | 训练侧                 | 推理侧                      |
|-----------|---------------------|--------------------------|
| 核心竞争      | 模型能力上限、参数规模、训练集群    | 单位任务成本、响应时延、任务成功率        |
| 主要约束      | 高端GPU、HBM、先进制程、资本开支 | 工程优化、部署架构、场景适配           |
| 中国机会      | 集中于头部云厂商、模型厂商和算力平台  | 分布于云厂商、行业软件、终端厂商、Agent企业 |
| Token经济意义 | 决定Token生产能力         | 决定Token消费规模和商业回报         |

#### 训练侧的瓶颈来自资源约束的叠加

大模型训练需要高端GPU、HBM、高速互联、先进制程和大规模数据中心协同，单次训练成本高、迭代周期长，硬件供给、能耗、集群调度和资本投入都会放大不确定性。训练侧仍由少数具备算力资源、模型团队、资金实力和生态入口的企业承担，是中国AI能力底座的重要门槛。

#### 推理侧更符合中国Token经济的资源禀赋和商业化路径

通过小模型、端侧模型、模型路由、缓存复用、量化、蒸馏与Agent拆解，可把不同复杂度任务分配给不同成本结构的模型，以更低成本、更稳时延、更高成功率完成更多任务。这一转向已从数据结构上显现：2025年中国用于AI训练与推理的数据总量达199.48EB，同比增长42.86%，其中推理数据量首超训练数据量、达101.34EB，说明产业资源的消耗重心正从训练转向应用调用与任务执行。

#### 中国优势：场景密度、工程转化和终端生态支撑Token快速消耗

中国具备大规模互联网用户、5G网络、物联网终端、政企数字化需求与完整制造业体系，天然带有多轮交互、流程自动化、知识检索、内容生成、视觉识别与设备控制等消费入口。推理侧更容易进入高频场景，并通过任务完成效果形成付费。云厂商可依托模型平台、算力调度和企业客户沉淀推理需求；垂直软件企业可围绕行业流程嵌入AI能力；终端厂商可推动端侧推理和本地智能助手；AI原生企业则可围绕模型压缩、Agent开发、任务编排和工具链形成差异化。

- 中国Token经济的结构性机会更集中在推理侧。工程优化、端侧部署和场景适配，更能发挥中国互联网、政企、制造业和终端生态的资源禀赋。

来源：人民日报、头豹研究院

# 中国企业分层：谁在生产Token，谁在组织Token需求

- 中国Token经济将围绕平台型、模型效率型和场景入口型企业形成分层竞争。平台企业提供AI基础平台，效率型企业压低单位任务成本，场景入口企业把Token消耗嵌入真实 workflow

## 生产和组织Token

中国企业将围绕Token生产、效率优化和需求组织形成三类分层



**平台型企业：提供AI基础平台，竞争重点将转向企业数据、任务流和行业方案**  
平台型企业能力基于算力资源、模型平台、API服务、企业客户和开发者生态。阿里云、百度智能云、火山引擎和腾讯云等企业，既能通过云资源和模型服务生产Token，也能通过MaaS平台、模型路由和Agent开发工具调度Token。随着开源模型扩散和API价格下降，单纯模型调用容易陷入价格竞争。平台企业的长期价值在于把云资源、模型平台和企业客户关系进一步连接到行业数据、任务流和解决方案交付中，从模型供给方升级为企业AI基础平台。

**模型效率型企业：用低成本推理能力影响任务成本和开发者选择**  
模型效率型企业同样参与Token生产，但竞争重点在于用更低成本、更高质量或更强任务能力完成推理。DeepSeek、月之暗面、智谱、MiniMax等AI原生企业，通过模型架构优化、长上下文、推理加速、开源生态和Agent能力，降低企业完成一次任务的Token消耗和部署门槛。这类企业的挑战在于模型效果可能被追赶，低价API也可能被平台企业吸收。后续需要通过产品入口、开发者生态、企业部署或垂直场景，把效率优势转化为稳定客户关系。

**场景入口型企业：把Token消耗嵌入办公协同、企业管理和行业 workflow**  
场景入口型企业的核心能力在于掌握高频任务、业务流程、组织账户、权限体系和行业数据。WPS、钉钉、飞书、企业微信等平台未必拥有最强模型，但更接近文档、会议、审批、研发、客服、制造和专业服务等Token持续消耗场景。一旦AI能力嵌入这些 workflow，Token消耗会从单次模型调用转为持续业务支出。该类企业的竞争边界在于能否把客户关系、业务数据和结果交付能力沉淀下来，并形成稳定的企业付费关系。

- 中国Token经济的竞争分层围绕算力能力、推理效率、客户流程和结果交付展开。越接近企业业务系统和付费关系，越容易形成价值留存能力。

来源：阿里云、腾讯云、企业年报、头豹研究院

■ 五年关键变量：端侧落地、平台定型与监管边界

- 未来五年，端侧入口、平台调度、行业采购和监管合规将决定Token经济能否从调用放量走向商业闭环。持续收入需要由企业续费、席位扩张、任务ROI和合规部署共同验证

端侧落地、平台定型与监管边界



端侧落地：扩大Token消费入口

AI手机、AI PC、车载助手和IoT端侧模型若能规模化落地，将推动AI从云端调用进入本地助手和高频工作流，扩大轻量任务和即时交互场景。

平台定型：确定Token调度权归属

Agent平台、MaaS平台和企业工作流入口若形成稳定分工，云厂商、办公平台和行业软件企业将围绕模型调用、任务拆解、工具执行和数据回流争夺Token调度权。

行业采购：影响商业闭环速度

办公、客服、代码、制造等场景需要把Token消耗转化为降本、增效、转化率提升或研发提效，企业续费、席位扩张和行业解决方案收入是关键验证指标。

监管边界：决定高价值场景开放节奏

金融、医疗、政务、教育和内容平台等场景，需要明确数据安全、模型备案、内容标识和责任边界，合规能力强的平台型和行业型企业更容易获得大客户订单。

- 未来五年观察重点应从调用量转向商业闭环。企业续费、席位扩张、任务ROI与合规部署，将成为区分三种情景的核心指标。

来源：国务院、工信部、网信办、头豹研究院

# Chapter 6

## 最终判断：Token 经济是否复现百 年产业规律？

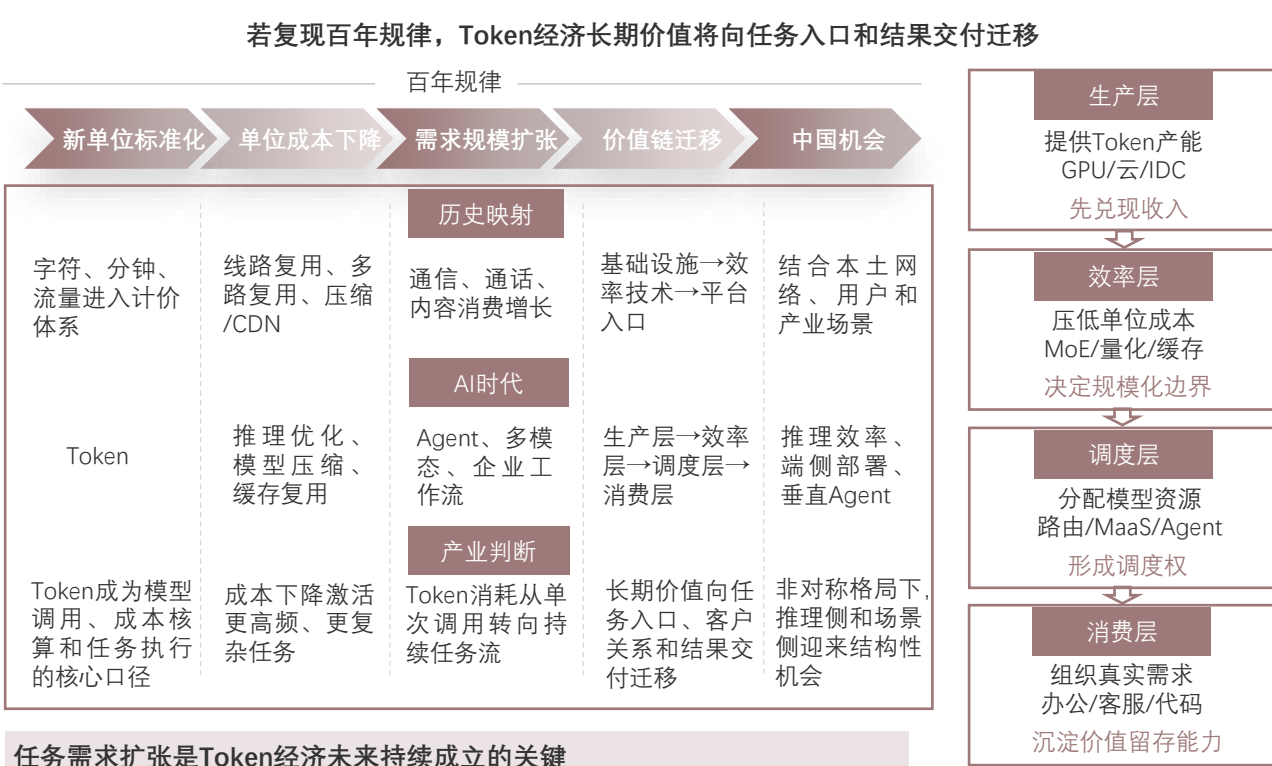
---

□ 核心结论：Token经济是百年产业规律在AI时代的延续

核心结论：Token经济已经出现百年产业规律的早期特征

- Token经济已经出现百年信息单位商业化规律的早期特征。其后续成立程度，取决于低成本Token能否持续激发高频、复杂、可付费任务需求，并推动价值向效率、调度和消费层迁移

核心结论



任务需求扩张是Token经济未来持续成立的关键

Token经济是否复现百年产业规律有明确的成立条件，须同时满足：Token进入商业计价体系（已确立）、单位任务成本持续下降（正在发生）、低成本推理激活更高频更复杂更可付费的任务需求（待产业验证）。三项同时成立，Token消耗将从单次调用转向持续任务流，AI支出有望进入更多行业预算；若调用增长无法转化为付费、续费与任务结果，价值兑现将出现折扣。

产业价值将沿生产、效率、调度、消费逐层迁移

若百年规律在Token经济中复现，短期，GPU与数据中心等基础设施率先受益；中期，推理优化与模型压缩成为竞争焦点，多模型调度与Agent平台掌握任务分发入口；长期，决定价值归属的是掌握办公、客服、研发等工作流与客户需求的应用层企业。随着Token供给趋于充裕，价值重心从"生产Token"转向"组织与消费Token"，任务入口、客户关系与结果交付将成为价值留存的关键。

中国机会集中在低成本推理、企业工作流、端侧部署和垂直Agent

在高端算力与先进芯片受限的非对称格局下，中国企业更需要通过工程优化、模型压缩、云边端协同与行业适配扩大推理侧优势。具备长期竞争力的，将是能把低成本Token嵌入真实工作流与垂直场景、并形成客户关系、数据闭环与商业收入的参与者。

Token经济已经出现百年信息单位商业化规律的早期特征。短期收入先由生产层兑现，长期价值更可能留存在任务入口和结果交付环节。

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](http://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](mailto: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](mailto:oliver.yuan@leadleo.com)

🌐 官网: [www.leadleo.com](http://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。