国金证券
SINOLINK SECURITIES

2026 年 7 月 17 日

产业研究深度|科技

Token 计费重塑 AI 产业链：你的企业定价权还在吗？

为企业客户提供全方位的“产业+金融”研究服务！

核心要点

我们认为，全球关于 AI 产业链的 Token 计费方式及相关商业模式的兴起，并非行业计费规则的简单迭代，而是对 AI 技术体系、商业分配机制、产业竞争格局的全链条底层重构。该商业模式将算力消耗、模型研发、场景服务等全产业链投入要素实现标准化、可量化、可交易，以 Token 调用量与单位成本为核心标尺，重塑行业价值计量与收益分配逻辑，推动 AI 产业摆脱零散定制化的项目制发展模式，正式迈入标准化、规模化、流量化的成熟发展阶段。

技术层面，AI 产业链的竞争逻辑实现根本性切换，从传统参数比拼转向全链路效能比拼。Token 经济落地前，行业技术评价聚焦模型参数规模、训练算力投入与通用性能跑分，重训练能力、轻落地效率。伴随 Token 模式普及，产业核心考核指标转变为单位算力 Token 产出、单 Token 推理成本、全链路调用时延及服务稳定性等工程落地指标，技术迭代核心目标从单纯追求性能领先，转向全链路生产效率提升与成本优化，推理优化、轻量化适配、国产算力兼容等工程能力，已然成为行业核心技术壁垒。

竞争格局层面，单一要素优势逐步失效，全链路整合能力成为核心壁垒。过往行业依靠算力、模型、场景等单一优势即可实现突围，而 Token 规模化落地后，行业竞争壁垒升级为算力整合、模型适配、渠道触达、场景变现、合规运维一体化的综合能力。头部运营商、云厂商凭借生态统筹、客户覆盖、合规结算、稳定交付的综合优势巩固核心主导地位，行业马太效应持续凸显。

整体而言，Token 计费模式的兴起标志着全球 AI 产业告别技术探索与小规模试点阶段，迈入商业化落地、规模化放量、生态化竞争的全新周期。此次由 Token 计费催化的产业链价值重构具备不可逆的趋势性。行业各类主体唯有重新审视技术路线、商业定位与生态策略适配全新产业逻辑，方能在新一轮产业周期中稳稳把握定价权和持续收获成长红利。

分析师及联系人

分析师：李忠宇
SAC:S1130524100002
lizhongyu01@gjzq.com.cn

更多免费行业报告 并购家 www.ipoipo.net

内容目录

第一章 Token 模式成为 AI 产业变现的核心抓手 4
1.1 Token 已从技术术语升级为产业核心计量基准 4
1.2 Token 产业爆发的三大核心支撑：算力、模型、场景 8
1.3 Token 商业模式出现前的 AI 产业链三大痛点 9
1.4 Token 计费是 AI 产业从技术导向到商业化导向的彻底转变 11
第二章 Token 五大商业模式：从基础计费到产业链重塑 11
2.1 API 计费：最成熟的商业化底层模式 12
2.2 订阅制：面向中小客户的标准化流量包模式 15
2.3 企业 Token 包：面向大客户的资源预付定制化模式 18
2.4 Token 产品金融收益分成化：行业价值循环的衍生配套模式 19
2.5 Token 工厂：整合产业资源的核心中枢 21
2.5.1 Token 工厂的四级演进模型：从 API 代理到企业 AI 网关 23
2.5.2 阿里 TokenFoundry：从 MaaS 到独立事业部的战略升级 25
2.5.3 全球视角：OpenRouter 与 Token 工厂的盈利能力验证 25
第三章 Token 正在重构 AI 产业技术竞争底座 27
3.1 核心技术竞争指标重塑 28
3.2 算力优化：从峰值能力到每瓦 Token 产出 29
3.3 模型效率：推理侧的降本增效革命 32
3.4 技术层冲击传导：从资源投入到价值产出的逻辑重构 33
3.5 Token 成本压缩的前沿突破：趋近普惠的生产力拐点 34
第四章 Token 正在重构 AI 产业的价值分配逻辑 35

4.1 价值分配重构：从资源采购到价值共生 36
4.2 生态协同重构：从线性链条到立体网格 38
4.3 竞争格局重构：产业链话语权的转移 40
4.4 商业层冲击传导：重新定义行业盈利模式 42
第五章 安全合规是 Token 商业模式落地的基线与护栏 43
5.1 AI 中转站风险显性化：从技术漏洞到国家安全层级隐患 43
5.2 企业合规落地路径：四层监管体系下的生存法则 45
第六章 中美对比 Token 商业模式趋同 or 背离？ 46
6.1 技术路线差异：开源与闭源的选择 46
6.2 商业模式差异：高端价值与普惠规模化的分化 47
6.3 生态组织差异：市场化交易与运营商主导协同的分化 49
6.4 监管取向差异：创新驱动与安全可控的分化 50
6.5 中美 Token 模式对比总结 51
第七章 Token 产业发展趋势展望与战略建议 52
7.1 Token 产业发展趋势展望 52
7.2 针对三类主体的战略建议 55
7.2.1 面向企业创始人/实控人的战略建议 55
7.2.2 面向一二级产业投资人的战略建议 56
7.2.3 面向行业政策制定者的战略建议 58
第八章 观点总结：Token 模式重构 AI 产业链底层逻辑与竞争格局 60
行业风险提示 62

更多免费行业报告 并购家 www.ipoipo.net

图表目录

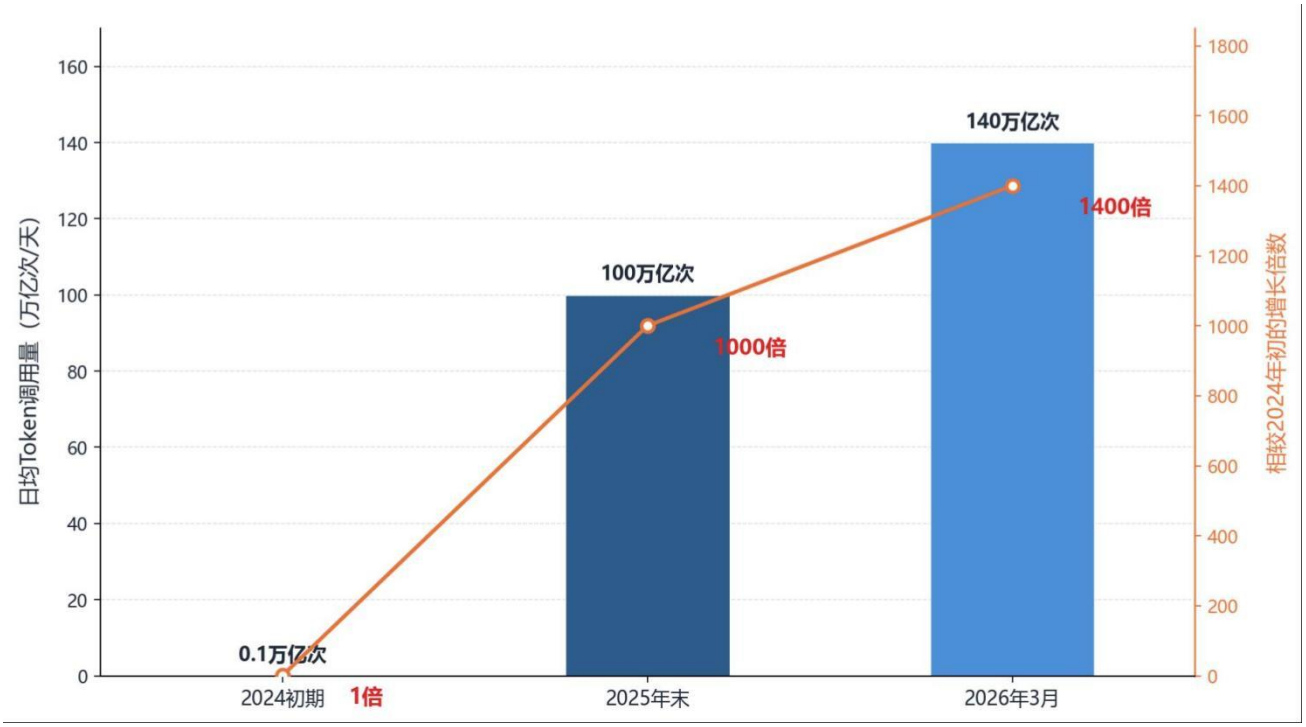
图表 1： 中国日均 Token 调用量增长趋势（2024—2026 年）.....4
图表 2： DeepSeek-V3 及其对标模型基准评测性能对比.....8
图表 3： 五大 Token 商业模式多维能力雷达图.....12
图表 4： 主流大模型 API 定价对比.....13
图表 5： 主流 AI 助手订阅制套餐对比（2026 年 5 月）.....15
图表 6： Token 全链路价值转换框架图.....22
图表 7： Token 工厂全栈服务能力要求.....24
图表 8： H100、B200 级 GPU 分别在按时租赁、TaaS 模式单卡年度收入测算.27
图表 9： NVIDIAGPU 架构历经六代架构迭代，推理能效提升 100 万倍.....30
图表 10： H100 GPU 在按小时计费模式和 Taas 模式下年收入对比.....31
图表 11： AI 推理单位 Token 成本持续下降趋势明显.....32
图表 12： IDC 数据显示 24-26 年国内智能算力规模持续高速增长.....34
图表 13： 以 Token 流动为核心的立体网状生态体系.....39

第一章 Token 模式成为 AI 产业变现的核心抓手

1.1 Token 已从技术术语升级为产业核心计量基准

❖ 自 2024 年以来，国内大模型的 Token 调用需求呈现指数级爆发增长，相关核心数据来自 2026 年 3 月 24 日国家数据局国新办新闻发布会官方披露口径。如图表 1 所示：**自 2024 年以来，国内大模型的 Token 调用需求呈现指数级爆发增长，相关核心数据来自 2026 年 3 月 24 日国家数据局国新办新闻发布会官方披露口径。如图表 1 所示：2024 行业发展初期全国日均 Token 调用量仅 0.1 万亿次；伴随企业级大模型、AI 智能体、MaaS 服务规模化落地，至 2025 年末国内日均 Token 调用量已攀升至 100 万亿次，相对 2024 年初规模扩张 1000 倍；截至 2026 年 3 月行业日均 Token 调用量进一步突破 140 万亿次，较 2024 初期实现 1400 倍增长。**

图表 1：中国日均 Token 调用量增长趋势（2024—2026 年）



来源：国家数据局、国金证券研究所

更多免费行业报告 并购家 www.ipoipo.net

- **海量 Token 调用需求直接带动算力租赁、模型服务、Token 经济产业链景气度持续上行，验证 AI 商业化落地进入高速放量阶段。**指数级扩张的词元调用规模，直观印证 Token 已经脱离单纯技术定义，率先成为国内 AI 产业落地可量化、可观测的核心运行指标。需求端持续爆发的同时，顶层政策与行业标准同步完成配套完善，自上而下完成 Token 概念的规范化定义、产业价值定性与交易模式指引，推动其从底层技术计量单位，升级为贯通算力、模型、数据、应用全链条的产业核心计量基准。
- **在技术层面，Token 是 AI 大模型处理信息的最小单位。**以主流中文模型为例，据阿里千问云官方文档粗略测算，单个中文字符平均对应 1.5 个 Token，高频通用汉字、固定成语可压缩至 1 个 Token，汉字 Token 消耗区间为 1－1.5 个单位；模型的每一次提问、每一次结果输出，都会被拆解为连续的 Token 序列进行统一计量。这一技术特性天然决定了其可量化、可拆分、可结算的商业属性，为后续行业标准化交易的落地提供了底层支撑。
- **从产业发展维度看，Token 的核心价值是为 AI 产业构建了统一的价值转化枢纽。**将算力的高强度物理消耗、模型的长时间技术研发投入、应用端的场景化服务输出，所有的技术、商业、场景投入，都能通过 Token 的调用数量、单位成本完成标准化计量，从而为产业商业化交付提供了真正的可交易基准度量。
- **Token 产业发展获得顶层政策与行业规范的双重背书。**2026 年 3 月 23 日，国家数据局局长刘烈宏在中国发展高层论坛上，正式明确 Token 官方中文定名为“词元”，并给出权威定位：**词元是智能时代的价值锚点，是连接技术供给与商业需求的结算单位。我们认为词元是数据要素与 AI 应用的关键转化载**

体。Token 的本次定名并非简单的文字翻译调整，而是由全国科学技术名词审定委员会组织多轮专业评审、凝聚行业共识后，面向全社会发布的人工智能领域优先试用名词，从顶层维度统一了行业基础概念，为后续词元计量、交易、标准化体系建设及产业合规发展筑牢了权威规范基础。

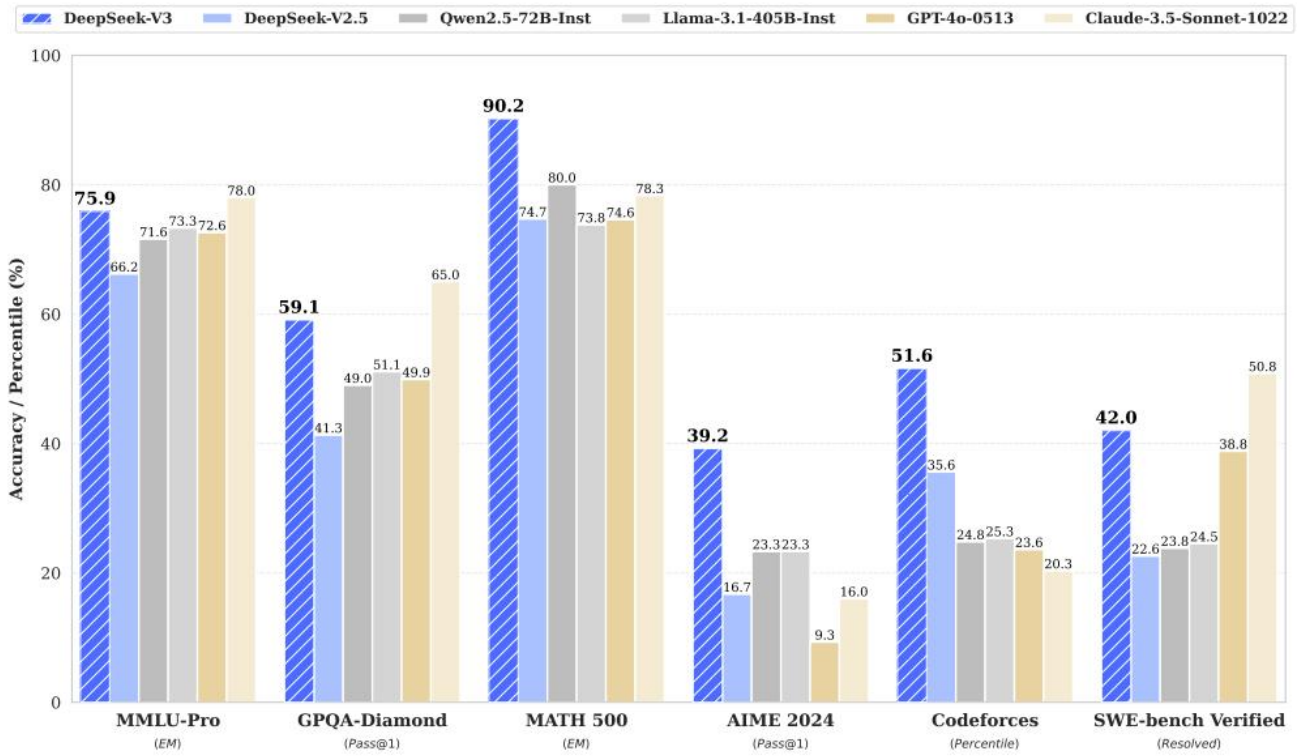
- 2026 年 4 月，国家数据局在《关于推进行业高质量数据集建设行动的实施方案（征求意见稿）》中进一步明确提出：**探索词元交易等新型交易模式，构建以词元为基础，可量化、可定价的数据集价值体系。**这一政策导向的落地，直接将 Token 从技术层面的计量工具，升级为驱动整个 AI 产业流通的价值枢纽，可见 Token 已不再是传统的技术术语，而成为贯穿算力、模型、数据、应用全链路的核心经济指标。
- **从词义演进维度看，Token 经历了三次具有历史意义的词义跃迁：第一次是计算机编程领域（1950-），**Token 指代词法分析（LexicalAnalysis）中的最小语义单元，是编译器将源代码拆解为可处理片段的基本单位；**第二次是区块链领域（2009-），**Token 演变为链上可流通的数字权益凭证，承载了资产确权与价值流转的功能；**第三次是 AI 大模型时代（2022-），**Token 升维为贯穿算力、模型、应用全链路的通用价值计量单位。这三次跃迁的本质，是 Token 从一个纯技术概念，逐步向经济概念、产业概念演化，最终成为 AI 产业的核心价值载体。
- **从技术实现层面看，当前主流大模型均采用 BPE（BytePairEncoding，字节对编码）算法进行 Token 化处理。**BPE 算法由 Sennrich 等人于 2016 年提出并广泛应用于 NLP 领域，其核心原理是：从字符级开始，统计语料中高频相邻字符对的共现频率，迭代合并形成子词级别（Subword）的 Token 单元。这

更多免费行业报告 并购家 www.ipoipo.net

一算法的核心优势在于，能够兼顾高频常见词的编码效率与低频罕见词的拆分灵活性，常见词被编码为完整 Token，罕见词则被拆分为多个子词 Token 组合表达。依据 OpenAI 官方开发者文档给出的通用测算标准，英文场景下约 0.75 个单词对应 1 个 Token，换算后单个英文单词平均消耗 1.33 个 Token。

- **从语义承载效率来看，**单个汉字语义密度高于英文单词，且国内大模型基于海量中文语料完成分词词表专项优化，成语、固定短句、高频通用词汇可直接编码为完整 Token，减少字符拆分带来的额外消耗。对标同等信息量文本可见，中文所需 Token 总量普遍低于英文，由此形成本土 AI 应用天然的成本优势。需补充说明，该结论仅针对国产自研中文优化大模型，海外通用模型词表以拉丁字符为核心，中文文本拆分损耗更高，不具备该特征。
- **从模型架构维度看，稠密 Dense 模型与 MoE 混合专家模型在 Token 经济学框架下呈现出根本性的商业化差异。**传统 Dense 模型在推理阶段需要全局激活全部网络参数，无论输入语义的复杂程度与计算负荷高低，均需完成全量参数计算，推理算力开销刚性强、冗余损耗高，整体单位调用成本偏高，海外 GPT-4、Claude 等闭源高端模型普遍采用该技术范式。与之不同，MoE 混合专家架构具备典型的稀疏激活特性，模型推理过程中仅根据输入语义特征，按需调度激活部分专家子网络参与计算，无需全员参数参与迭代，能够有效规避无效算力消耗，大幅优化推理端资源利用效率。当前 DeepSeek-V3、MiniMax-2.5 等国产头部模型均依托 MoE 架构的稀疏算力优势，实现推理成本的结构性优化，成为国内模型规模化、普惠化落地的重要技术底座。

图表 2：DeepSeek-V3 及其对标模型基准评测性能对比



来源：DeepSeek-V3 技术报告、国金证券研究所

整体而言，两类架构的推理机制差异，形成了行业模型成本体系与商业能力的天然分层。Dense 模型适配高精度、高复杂度、低并发的高端推理场景，具备性能稳定性优势但成本刚性较强；MoE 架构凭借高效的算力利用效率，更适配大规模、高并发、标准化的 Token 调用场景，具备更强的定价弹性与规模化变现潜力，直接决定了不同厂商的商业模式、利润空间与场景竞争优势。

1.2 Token 产业爆发的三大核心支撑：算力、模型、场景

- **Token 产业的爆发式发展，本质上是 AI 产业从技术验证走向商业化规模化落地的强信号体现。**Token 调用量快速增长的背后，是算力、模型、场景三大关键要素快速发展的重要体现。算力、模式、场景这三者业共同形成了驱动 Token 经济发展的铁三角支撑结构：
- **算力底座的大规模扩容是基础保障：**根据 IDC 联合浪潮发布的《中国人工智

更多免费行业报告 并购家 www.ipoipo.net

能计算力发展评估报告》公开数据，2026 年国内智能算力规模将达到 1460.3EFLOPS，较 2024 年的 725.3EFLOPS 实现规模翻倍，为 Token 的大规模生产提供了基础算力保障。

- **模型侧的推理降本取得持续突破：**从行业实际落地情况看，头部厂商在 Token 生成端的技术优化已经相对成熟，以英伟达为例，其 Hopper、Blackwell、VeraRubin 三代架构的连续升级，已经实现了单兆瓦算力的 Token 推理吞吐量超百万倍提升；在国内，弘信电子控股子公司燧弘华创 Token 工厂的实测数据显示，国产华为昇腾超节点集群的 Token 生成性能，较同规模传统英伟达集群高出 1.2 倍。这一系列技术突破，直接将行业 Token 级推理成本降到了商业化规模化的临界线以下。
- **AI 应用场景的全面渗透：**AI 技术在场景端的规模化落地是驱动 Token 爆发式增长的最直接动力。根据沙利文《中国 GenAI 市场洞察:企业级大模型调用全景研究（2025H2）》，国内头部厂商的 AI 应用场景，已从 2024 年的数十个垂直行业场景，大模型已广泛覆盖内容生产、知识获取、数据分析与研发支持等企业核心业务环节。其中，文本内容创作(19.4%)、问答助手(13.7%)、数据处理与分析(12.9%)、内部知识问答(12.6%)、代码生成与辅助(12.4%)，合计占比近 70%。这一场景侧的爆发式增长，直接拉动了上游算力、模型端的 Token 级资源需求。

1.3 Token 商业模式出现前的 AI 产业链三大痛点

在 Token 商业模式出现之前，全球 AI 产业链的商业化进程一直被三大难以破解的行业痛点掣肘。这些痛点本质上是由产业缺乏统一价值计量基准导致的，也构

成了 Token 经济模式崛起的底层动因：

- **AI 应用的价值无法量化：**算力厂商的资源投入以 GPU 卡时、功耗为基准，模型厂商的投入以参数规模、训练时长为基准，应用厂商的投入以场景化功能模块为基准，三者之间的价值计量基准完全无法对接，没有统一的标准来定义一次 AI 调用的技术成本与商业价值。这一情况下，上下游合作的成本核算、收益分配缺乏可量化依据，行业整体交易效率始终处于较低水平。
- **AI 服务的交付无法标准化：**行业内主流的项目制定制交付模式，存在严重的供需错配问题，部分中小微企业的轻量化 AI 需求，无法被标准化满足；而头部客户的大规模定制需求，则面临着交付周期长、技术方案不可复制、成本无法管控、上线后运维响应速度慢等一系列共性问题。这一模式下，AI 服务的交付质量、交付周期没有统一标准可约束，行业商业化进程始终无法形成规模化效应。
- **AI 产业链利益分配失衡：**传统算力租赁模式下，算力、模型、应用三方的利益分配机制，完全无法适配 AI 产业的实际价值流转逻辑，算力厂商按固定卡时收取租金，不关心模型的实际调用效果；模型厂商收取较高的软件授权费，却无法覆盖场景化落地的额外成本；应用厂商承担了最终的场景化运维风险，却只能收取固定的项目服务费。各方的风险收益结构完全失衡，行业缺乏可持续的价值循环机制。
- **Token 经济概念的出现本质上为破解这三大行业痛点提供了全新的解决方案：通过统一计量标准，**将算力、模型、应用三个环节的价值进行标准化锚定，以实际调用量作为结算基准，**这一逻辑直接重构了上下游产业链的价值分配逻辑，也重构了 AI 产业的商业底座。**

更多免费行业报告 并购家 www.ipoipo.net

1.4 Token 计费是 AI 产业从技术导向到商业化导向的彻底转变

从技术属性到商业属性的跃迁是 Token 真正成为 AI 产业核心的关键标志。这一过程中 Token 的商业价值维度逐步延伸到了 AI 行业商业化的各个环节：

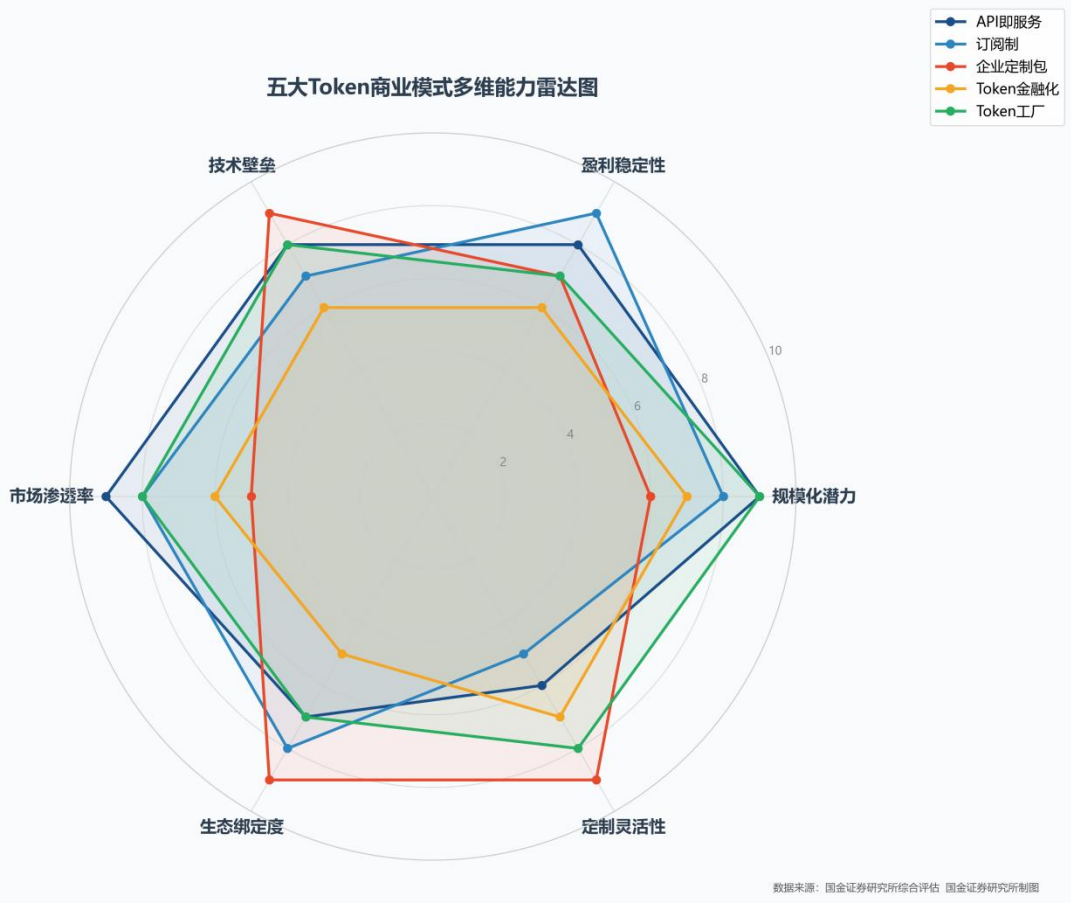
- **价值计量维度：**Token 成为唯一能够完整覆盖算力资源消耗、模型技术投入、场景化应用价值的全链路计量基准。厂商的算力投入、模型调用次数、用户实际使用量，都可以统一用 Token 进行量化核算。
- **交易定价维度：**此前行业由于缺乏统一计量基准，AI 服务定价乱象丛生，而 Token 的出现为行业提供了标准化的交易定价依据:2026 年行业已经形成了初步的 Token 定价体系，价格水平因模型能力的差异而呈现显著梯度。
- **结算分成维度：**Token 成为产业链上下游之间的核心结算纽带:从算力工厂到模型厂商，再到最终端的应用厂商，三方之间的所有交易、分成、利润结算，都可以围绕实际 Token 调用量展开。

这一商业化跃迁的本质，是 AI 产业从技术导向向商业化导向的彻底转变。随着 Token 在技术、商业、场景三个维度的价值被行业共识性认可，从而催化全球 AI 产业链的快速变革。

第二章 Token 五大商业模式：从基础计费到产业链重塑

Token 产业的出现不仅是创造了新的商业范式，而是重构了 AI 产业的整个商业底层逻辑，即**围绕 Token 的生产、流通、交易、结算全链路，AI 行业已经形成五大类典型的商业模式，共同支撑起 Token 经济的完整商业闭环。**

图表 3：五大 Token 商业模式多维能力雷达图



来源：麦肯锡全球研究院、国金证券研究所

2.1 API 计费：最成熟的商业化底层模式

这是当前全球 AI 产业最成熟、应用最广泛的 Token 商业模式，也是所有其他模式的技术基础即模型厂商将推理能力封装为可被直接调用的 API 接口，用户根据实际调用过程中产生的 Token 消耗量，支付对应的服务费用。

- **从行业实际落地情况看，这一模式的核心逻辑是按量精准计费，即用户只为实际消耗的 Token 买单，这在根本上区别于此前的固定包月、固定项目制付费模式；行业主流定价策略均采用阶梯计价+上下行分离计价逻辑，已经形成了相对稳定的价格梯度。**从公开定价数据来看，2026 年国内外头部厂商的主流 API 定价标准差异较大，具体对比如图表 4 所示：

更多免费行业报告 并购家 www.ipoipo.net

图表 4：主流大模型 API 定价对比

AI 大模型定价对比表					
模型	提供商	输入价格	输出价格	免费额度	定价策略特点
GPT-5.6Sol（旗舰）	OpenAI	$5.00/百万 Token	$30.00/百万 Token	新用户赠送 18 美元试用金，无永久免费额度	2026 年 6 月最新顶级旗舰，200 万 Token 上下文；大批量采购可洽谈阶梯折扣，企业级专属实例可定制定价
GPT-5.6Terra（均衡）	OpenAI	$2.50/百万 Token	$15.00/百万 Token	新用户赠送 18 美元试用金，无永久免费额度	现役主力均衡模型，对标 GPT-5.4 性能，同价新一代迭代；128K 标准上下文统一单价
GPT-5.6Luna（轻量）	OpenAI	$1.00/百万 Token	$6.00/百万 Token	新用户赠送 18 美元试用金，无永久免费额度	平价量产商用模型，面向高并发、低延迟场景；上下文无额外阶梯溢价
Claude5Sonnet（均衡主力）	Anthropic	$2.00/百万 Token（限时优惠）、$3.00/百万 Token（常规价）	$10.00/百万 Token（限时优惠）、$15.00/百万 Token（常规价）	无永久免费额度，新用户可申请企业试用	2026 年 6 月 30 日上线新款主力，限时优惠至 2026 年 8 月 31 日;200K 上下文内固定单价，超长上下文收取溢价
Claude4.8Opus（旗舰）	Anthropic	$5.00/百万 Token	$25.00/百万 Token	无永久免费额度，新用户可申请企业试用	旗舰模型大幅降价，硬件算力效率提升；面向深度推理、超长文档复杂任务，固定上下文阶梯定价
Gemini2.5Pro（旗舰）	Google	$1.25/百万 Token	$10/百万 Token	AIStudio 每日免费调用限额，无企业级永久免费	Pro 级模型已移除免费层，仅保留付费商用
Gemini2.5Flash（高性能轻量）	Google	$0.15/百万 Token	$0.6/百万 Token	AIStudio 每日免费调用限额，无企业级永久免费	GoogleI/O2026 发布 GA 版本，100 万 Token 上下文；定价介于旗舰与入门级之间，面向高性能商用场景

DeepSeek-V4-Pro	DeepSeek	¥0.025/百万 Token（缓存命中）¥3/百万 Token（缓存未命中）	¥6/百万 Token	注册赠送 500 万免费 Token，长期有效	对标 GPT-5.4 全能力；缓存命中输入成本降低 99%，百万超长上下文无额外加价
DeepSeek-V4-Flash	DeepSeek	¥0.02/百万 Token（缓存命中）¥1/百万 Token（缓存未命中）	¥2/百万 Token	注册赠送 500 万免费 Token，长期有效	新一代极致平价模型，缓存复用场景成本行业最低
文心 5.1（ERNIE-5.1）	百度千帆	¥6/百万 Token（推理 32K-128K）	¥22/百万 Token（推理 32K-128K）	新用户一次性免费额度，企业级可申请专属试用	长文本阶梯加价，支持飞桨全生态打通；批量调用可享阶梯折扣，私有化部署可定制定价
通义千问 3.7-Max（旗舰）	阿里云百炼	¥6/百万 Token	¥18/百万 Token	新用户 100 万免费 Token，长期有效	上下文超过 32K 阶梯涨价；批量推理半价、缓存复用专属优惠，阿里云生态捆绑可享额外折扣
通义千问 3-Flash（轻量）	阿里云百炼	¥0.8/百万 Token	¥3.2/百万 Token	新用户 100 万免费 Token，长期有效	极致平价商用模型，上下文无额外溢价；高并发场景可享批量阶梯折扣
Hunyuan-role-latest	腾讯云	¥2.4/百万 Token	¥9.6/百万 Token	全系共享 100 万/年免费额度	长文本阶梯加价，腾讯云生态全链路打通；批量调用可享专属折扣
Hy3（2026 新款推理旗舰）	腾讯云	¥1/百万 Token（缓存命中 ¥0.25）	¥4/百万 Token	全系共享 100 万/年免费额度	MoE 架构，Agent/代码推理优化，缓存复用大幅降低输入成本

数据来源：OpenAI、Anthropic、Google、百度智能云、阿里云、DeepSeek、腾讯云官方定价页面，数据更新至 2026 年 5 月；国金证券研究所整理汇总。注释：海外模型单位为美元/百万 Token，国产模型为人民币/百万 Token；汇率参考 2026 年 5 月央行中间价 1USD≈7.2CNY。

❖ **大模型 API 计费模式的核心优势，在于其精准匹配了用户的实际使用量与付费成本的需求，对于企业级用户而言，这一模式将此前的固定成本投入转变为可变成本投入，可根据实际业务流量灵活控制成本；对于厂商而言，这一模式的商业价值在于，其收入与模型实际调用效率直接挂钩厂商的技术优化效率越高，单位 Token 成本越低，利润空间就越充足。**因此，API 计费模式不仅是产业商业化的基础底层模式，更是驱动全行业提升 Token 效率的关键

更多免费行业报告 并购家 www.ipoipo.net

抓手。

2.2 订阅制：面向中小客户的标准化流量包模式

参考图表 5 所示，订阅制是 API 计费模式的重要补充，也是当前头部厂商争夺中小微企业客户、个人开发者流量入口的核心关键。其核心逻辑是固定周期费+免费 Token 额度+超额优惠：用户支付固定月度/年度订阅费，可获得对应档次的免费 Token 调用额度，超出免费额度的部分，可享受较标准 API 定价更低的折扣价。

图表 5：主流 AI 助手订阅制套餐对比（2026 年 5 月）

产品	免费版权益	专业版/月	企业版/团队版单价	核心权益	适用场景
ChatGPTPlus	GPT-3.5 限速有限访问	$20/月	ChatGPTBusiness 最低 2 席位，$20/席/年付	GPT-4o 优先调度、DALL-E 图像、长上下文	个人专业办公、内容创作
ClaudePro	每日消息限额基础访问	$17/月	ClaudeTeam 最低 5 席$20/席/年	5 倍消息额度、Projects 知识库、ClaudeCode	长篇文档分析、学术研究
Microsoft CopilotPro	基础网页 Copilot 免费	2025 年 10 月起停止新购，存量可用至 2026-08-01	Microsoft365Premium$19.99/月	Word/Excel/PPT 原生 AI、高限额 Designer 生图	Office 重度办公用户
文心一言专业版	基础对话有限额度	单月会员 59.9 元，连续包月 49.9 元，基础版免费。	B 端千帆 API 官方定价，根据企业实际使用分为三种套餐：149 元/月、419 元/月、989 元/月。	多模态生成、专属 API 额度、知识库	中小企业文案、营销
通义灵码	个人基础功能永久免费	个人社区版免费，每月有限的补全次数及体验资源额度；个人专业版 59 元/月，2000credits，更多次补全；	企业标准版 99 元/人/月，3000credits，1 人起购	IDE 插件、代码补全、单元测试生成	程序员研发提效

KimiPro	200 万字免费上下文	¥49/月、¥99/月	定制企业议价	超长文本、联网检索、本地文件解析	文献、财报深度研读

数据来源：OpenAI 官网、Anthropic 官网、Microsoft 官网、百度智能云官网、阿里云官网、Kimi 官网的定价页面，数据更新至 2026 年 5 月；国金证券研究所整理汇总汇率参考 2026 年 5 月央行中间价 1USD≈7.2CNY。

注：以上内容仅列举部分数据。需要特别说明的是 MicroSoftCopilotPro 是微软面向个人用户推出的独立付费 AI 订阅增值服务，仅适配 Microsoft365 个人/家庭版基础订阅。自 2025 年 10 月微软推出 MicroSoft365Premium 一体化套餐后，CopilotPro 已停止面向新用户售卖；存量 CopilotPro 订阅用户可正常使用至订阅周期结束，行业统一过渡期至 2026 年 8 月 1 日，到期后无法单独续费，微软推荐用户迁移至整合全部 Copilot 高阶能力的 MicroSoft365Premium 套餐。

- **从行业实际落地情况看，头部厂商均采用了多档次差异化的订阅套餐设计，以覆盖不同量级、不同场景的用户需求：**
- **运营商系套餐：**依托自身通信与渠道优势，将 Token 服务与传统流量、宽带服务进行捆绑，推出了行业门槛极低的普惠类套餐。以河北移动地方试点个人 Token 按量套餐为例，单次 15 元含 600 万 Token；多地分公司同步推出梯度包月产品（30 元/1200 万 Token、45 元/1800 万 Token）；中国电信推出的 9.9 元/月 Token 套餐，包含 1000 万 Token 调用额度，还叠加了额外的流量权益，这类套餐将 Token 门槛降到了极致，完全覆盖了普惠型用户的轻量化 AI 需求。
- **云厂商/模型厂商系套餐：**依托自身技术与生态优势，将 Token 服务与自身的其他云服务、行业化能力进行捆绑，推出了覆盖个人、团队、企业全层级的多档次套餐。以阿里云百炼平台为例，其推出的 TokenPlan 团队版套餐，分为三档标准配置：标准坐席 198 元/月（含 25000Credits，可兑换对应额度的

更多免费行业报告 并购家 www.ipoipo.net

Token）、高级坐席 698 元/月（含 100000Credits）、尊享坐席 1398 元/月（含 250000Credits）；超出套餐内额度的部分，可通过购买共享流量包进行抵扣，流量包换算标准为每 5000 元可兑换 625000Credits。这类套餐的核心竞争力在于，用户采购 Token 流量包后，可自由选择阿里云百炼平台上的通义千问、DeepSeek 等多款主流模型，无需单独为不同模型采购不同的调用资源包。

- **国际厂商系套餐：**面向海外企业级市场，采用高订阅费+高 API 折扣+高附加值服务的高端定价逻辑。以 ChatGPTBusiness 服务为例，按月订阅定价为 25 美元/坐席，年付优惠价 20 美元/坐席，套餐内提供无上限的 Token 调用额度，以及 SOC2 合规认证、数据隔离、专属技术支撑等企业级增值服务；面向百席以上大型企业的 ChatGPTEnterprise 无公开统一标价，需商务定制洽谈，中大型客户均可根据实际调用量，申请额外的批量级折扣优惠
- **从运营商套餐的具体定价结构来看，三大运营商在 Token 普惠化方面已形成高度差异化的竞争格局：**中国电信的 Token 套餐在运营商中定价 9.9 元/月套餐包含 1000 万 Token 调用额度，折合每百万 Token 仅 0.99 元，其定价逻辑是通过极低的门槛快速获取海量用户，再通过后续的叠加服务（如专属模型调优、数据安全保障）实现增值变现；此外，中国电信还推出了与 4G/5G 流量包捆绑的 Token 权益套餐，用户办理指定流量包后可免费获赠对应档位的 Token 调用额度。河北移动的个人版 Token 最低档套餐 15 元/月含 600 万 Token，折合每百万 Token 约 2.5 元，其定价逻辑是通过套餐的逐步梯度设计引导用户向高价值套餐升级。中国联通则明确将 Token 列为继语音、流量、宽带之后的第四大通信计量单位，并推出面向开发者社群的 CodingPlan 专项套餐，将 Token 服务与传统通信服务进行了深度绑定。

运营商的 Token 套餐策略，本质上是借助其覆盖全国的通信网络、线下营业厅渠道、成熟的计费结算系统，以及海量的存量用户资源，将 Token 服务从技术服务降维为通信基础服务进行大规模推广。这一策略的核心竞争力在于，运营商不需要从零构建用户触达体系，其已有的数亿级通信用户，就是 Token 服务的天然潜在客户。从行业演进规律看，运营商的 Token 套餐策略正在经历从流量捆绑到独立产品再到生态入口的三阶段演进：第一阶段以流量包赠送 Token 额度为主，目的是教育市场；第二阶段推出独立的 Token 月租套餐，形成独立收入来源；第三阶段将 Token 作为 AI 服务的通用入口，所有 AI 应用均通过 Token 额度进行统一消费。

从行业实际运营效果来看，订阅制的核心商业价值，在于它为行业提供了稳定的现金流基础。对厂商而言，订阅制可以提前锁定用户的长期调用流量，有效降低用户流失率；对用户而言，订阅制可以大幅简化采购流程，较直接按 API 付费的模式成本更低。这一模式，也成为头部厂商规模化收割下沉市场流量的关键抓手。

2.3 企业 Token 包：面向大客户的资源预付定制化模式

这是当前头部厂商面向中大型企业级客户、下沉渠道合作伙伴的主流销售模式，也是行业保障大规模资源供应稳定的关键抓手，其核心逻辑是提前锁定资源量、预支付采购费用、阶梯式结算成本：客户根据自身业务规划，提前向厂商采购一定数量的 Token 资源包，在约定的服务期内可按需调用；采购量越大、预付周期越长，单位 Token 的结算成本就越低。

从行业实际落地情况看，这一模式的核心是定制化，厂商会根据客户的场景需求、年度调用量规划、行业化适配难度，为客户定制专属的 Token 资源包方案；部分情况下，还会为客户提供专属的资源调度隔离服务，以保障客户的业务调用稳定

更多免费行业报告 并购家 www.ipoipo.net

性。以中国电信面向中大型企业客户推出的 Token 包为例，其将 Token 资源与网络连接、安全防护、行业化技术支撑能力进行了深度捆绑，提供了覆盖资源供应-场景调用-安全运维全链路的一体化解决方案；客户采购 Token 包后，可直接将资源用于自身的 AI 客服、工业级智能质检、智能文档处理等实际场景，无需额外配套安全防护、网络支撑资源。

这一模式的核心竞争力，在于其同时解决了供需双方的核心痛点：对客户而言，提前锁定了资源供应和成本价，规避了后续资源供应紧张或成本上浮的风险；对厂商而言，提前锁定了长期订单和现金流，还能将服务、运维、生态合作等附属价值进行捆绑变现，大幅提升单客价值贡献。**在行业实际落地中，企业 Token 包往往会和订阅制、API 计费模式进行组合使用，共同覆盖中大型企业客户从轻量化测试到大规模场景化落地的全链路需求。**

2.4 Token 产品金融收益分成化：行业价值循环的衍生配套模式

随着 Token 规模化交易体系的逐步成熟，行业内出现了 Token 收益分成化的衍生配套模式。**这一模式的核心，是将 Token 的生产、流通、交易各个环节与金融服务分成模式进行深度绑定，本质是产业发展到一定阶段的衍生配套服务，其价值在于支撑 Token 的生产与流通环节，降低产业链上下游的交易成本与风险。**

从行业实际落地情况看，当前主流的 Token 金融收益分成化模式，完全服务于实业、完全服务于供应链的真实需求：

- **支付结算类服务：**以支付宝 2026 年 5 月发布的 TokenPay 服务为代表：这是全球首个面向大模型产业公司的定制化支付解决方案，本质是为 Token 交易提供了专属的企业级资金清算与对账服务。这类服务解决了此前行业 Token 交易面临的资金流转效率低、对账流程复杂、合规性难以保障的核心痛点，

让模型厂商、算力厂商之间的 Token 级交易结算变得更加高效、安全，也为后续行业大规模标准化 Token 交易的落地提供了金融支撑基础。

- **收益分成类服务：**以商汤科技 SenseNova 算力计量（TokenPlan）方案为代表：其核心逻辑是将算力资源的实际生产能力锚定到标准化 Token 计量单位上，1 单位 Token 对应厂商的标准化算力资源单位。下游客户在采购了 Token 资源包后，可以在实际使用时自主选择将 Token 兑换为对应算力资源的模型调用服务，或直接兑换为对应金额的算力资源采购额度；在结算环节，算力厂商、模型厂商、应用厂商之间的收益分成，将基于 Token 的实际调用收入进行核算。这一方案，本质是将传统的固定算力采购交易模式，转变为按实际产出的 Token 调用量结算的弹性交易模式，精准匹配了上下游企业之间的风险收益需求。
- **供应链金融类服务：**部分头部厂商正在联合金融机构探索基于标准化算力资源凭证确权与真实贸易场景的供应链金融服务。上游算力厂商可以将已中标但尚未交付的算力订单、或是已经完成储备的标准化算力资源包作为质押物，向金融机构申请低成本订单融资，缓解算力固定资产投入带来的资金压力。这类服务依托标准化算力计量凭证的可核验交易属性，为算力企业开辟了灵活的动产融资渠道，打通了从算力投产到终端变现的全链路资金闭环。

➢ 国家数据局于 2026 年 6 月 8 日印发《关于推进行业高质量数据集建设行动的实施方案》，文件提出：**鼓励数据集在数据交易所（中心）等数据流通服务机构挂牌交易，发展“订阅模式”“商场模式”“定制模式”等多元服务形态，推动商业模式从基础数据包销售向 API 调用、模型化解决方案及全栈服务梯次跃升。探索词元交易等新型交易模式，构建以词元为基础，可量化、可定**

更多免费行业报告 并购家 www.ipoipo.net

价的数据价值体系，为词元要素市场化研究提供了政策支撑。**需要强调的是国内的词元（Token）要素市场化模式与虚拟货币有着本质区别，国内所有Token服务的前提是锚定真实算力资源、锚定真实场景需求、锚定真实交易背景，严格限定在产业闭环内的使用场景，严禁任何形式的炒作行为，完全服务于AI产业实际落地需求。**

2.5 Token工厂：整合产业资源的核心中枢

- **Token工厂是2026年产业热度最高、落地规模最大的Token商业模式，也是串联算力、模型、应用三大环节的核心中枢，其本质是算力调度平台、模型适配平台、Token交易平台的三合一集合体。**它将底层的算力资源，与中游的各种行业模型、场景化模型进行深度适配，并通过标准化的API接口将算力与模型资源封装为可被直接调用的Token服务，最终面向上层应用侧实现标准化交付.
- **从实际落地情况看，Token工厂的核心突破在于它将算力资源从传统的硬件租赁资产，转化为可被直接调用、可被精准计量、可被标准化交易的Token流量产出，这是整个Token经济能够规模化落地的关键前提。**如图表6所示，Token工厂的运营方不再是单纯的算力供应商或模型服务商，而是整个Token经济的资源整合方与价值分配方，其核心工作是将算力厂商的资源、模型厂商的技术能力，与应用端的场景化需求进行精准对接，完成Token的标准化产出、流转、变现全链路调度。

图表 6：Token 全链路价值转换框架图

来源：中国工业互联网研究院、国金证券研究所

这一模式的巨大商业价值，直接体现在头部厂商的落地规模上：

中国电信 Token 工厂项目：2026 年 5 月，中国电信宁夏分公司发布的 Token 工厂生成能力服务集采项目，是国内首个以 Token 工厂命名的百亿级规模化集采项目，项目整体含税规模达 174.38 亿元，不含税规模 164.51 亿元，共划分 11 个标包，服务期长达 5 年。这一项目的核心意义，在于它正式确立了 Token 工厂的标准化采购机制，运营商将上游算力、模型、技术服务商的资源整合后，直接向下游应用端交付标准化的 Token 服务，行业正式进入 Token 规模化交付阶段。

中国移动 Token 工厂项目：在 2026 移动云大会上，中国移动正式发布 Token 运营生态体系，宣布将 Token 经营作为公司的核心主业之一；同时开放了万亿级 Token 规模的服务体验包，面向全行业开放 Token 生产能力。在具体落地层面，中国移动的 Token 工厂项目，以移动云的算力资源底座为支撑，整合了行业内超 300 款主流大模型，形成了覆盖从轻量化场景到超大规模场景的全链路 Token 供

更多免费行业报告 并购家 www.ipoipo.net

应能力；北京、江苏、广东等多地省公司，已同步推出了面向行业客户的 Token 专属套餐包，将 Token 资源与行业化场景应用方案进行了捆绑交付。

行业头部厂商 Token 工厂项目：除运营商外，头部科技厂商也在加速推进 Token 工厂落地。2026 年 5 月，阿里云将旗下 MaaS 模型服务正式升级为 TokenFactory 业务板块，与阿里云、淘天平级，由集团 CEO 吴泳铭亲自挂帅统筹；在技术层面，阿里云依托自研的神龙弹性算力集群、飞天操作系统、以及自研的模型推理调度优化引擎，实现了对通义千问、DeepSeek 等主流行业模型的适配支撑，为行业客户提供标准化 Token 交付服务。同月，弘信电子旗下燧弘华创 Token 工厂正式落地无锡高新区，首期部署了 4 台华为昇腾 384 超节点服务器，单台服务器搭载 384 卡算力，整体形成 1536 卡超大规模国产算力集群；经实测，该算力集群的 Token 生成性能，超传统英伟达同规模算力集群 1.2 倍。在技术架构层面，燧弘华创 Token 工厂采用了华为昇腾的超节点架构，以及液冷散热、高密算力调度等一系列先进技术，为后续行业大规模 Token 生产提供了可复制的标杆样本。

从产业实际落地情况看，Token 工厂已经成为头部厂商整合行业资源、抢占生态话语权的核心赛道，它串联起了算力、模型、应用三个环节，也成为决定 Token 经济未来发展走向的关键核心。

2.5.1 Token 工厂的四级演进模型：从 API 代理到企业 AI 网关

从产业演进角度看，Token 工厂已经形成了清晰的四级演进模型。**参考图表 7 所示，我们可以做如下分级：**

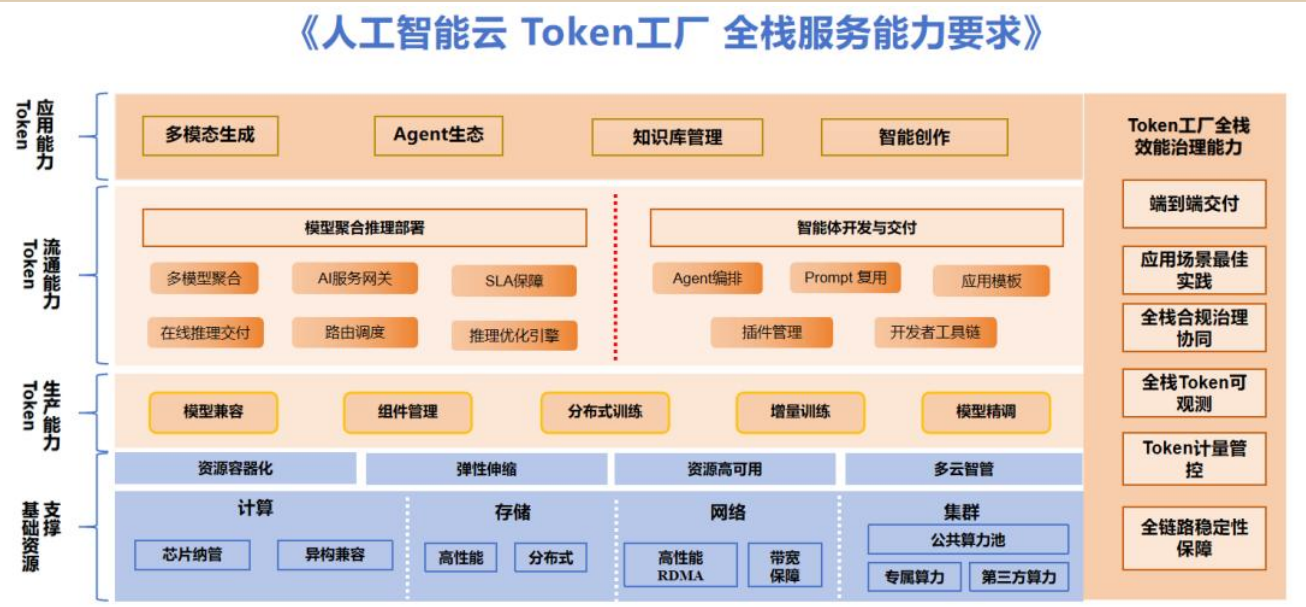
第一级（1.0）是 API 代理模式：企业批量采购模型厂商的 API 调用额度后，以加价方式转售给下游客户，赚取渠道价差，技术含量最低、利润空间最薄，模式类似通信行业的虚拟运营商。

第二级（2.0）是聚合平台模式：将多家模型厂商的 API 汇聚到统一平台，向下游提供一站式多模型接入服务，解决客户需要对接多个模型厂商的痛点，典型代表为海外的 OpenRouter 及国内的一站式 API 聚合平台。

第三级（3.0）是推理基础设施模式：企业自建或租赁大规模算力集群，在自有基础设施上部署并运行开源模型（如 DeepSeek、Qwen、Llama 等），实现从"转售第三方 API"到"自主生产 Token"的质变，这是 Token 工厂毛利率跃升的关键转折点。

第四级（4.0）是企业 AI 网关模式：面向中大型企业客户，提供安全审计、合规管控、智能模型路由、成本优化、Token 全生命周期管理等全套企业级 Token 网关服务。

图表 7：Token 工厂全栈服务能力要求



来源：中国信息通信研究院、国金证券研究所

当前，国内市场的 Token 工厂大多处于 2.0 向 3.0 过渡的关键阶段，少数头部厂商（如阿里云百炼平台、火山引擎方舟平台）已率先进入 3.0 阶段；海外头部平台如 FireworksAI、TogetherAI 和 OpenRouter 已进入 3.0-4.0 阶段。正如黄仁

更多免费行业报告 并购家 www.ipoipo.net

勋演讲所述，2025 年全球 AI 基础设施资本开支在 2025 年已超 6000 亿美元。我们认为 Token 工厂作为这一巨额投资的核心价值变现通道，其战略地位将持续提升。

2.5.2 阿里 TokenFoundry：从 MaaS 到独立事业部的战略升级

2026 年 5 月，阿里云将百炼 MaaS 平台升级为 TokenFactory 独立业务板块，层级与阿里云智能、淘天集团平齐，由集团 CEO 吴泳铭直接统筹，AI 业务正式上升为集团核心战略。在 2026 财年 Q4 财报电话会上，吴泳铭表示阿里 AI 已告别投入期、进入商业化回报阶段：目前阿里 AI 模型与应用服务年化 ARR 已突破 80 亿元，截止 6 月季度 ARR 预计突破 100 亿元、年末突破 300 亿元，并强调该业务高利润率优势凸显，将成为集团高质量增长核心支柱。高速增长的营收规模与商业化兑现能力，充分印证 Token 经济模式已成为头部云厂商 AI 商业化的核心底座与战略主线。

- **阿里云将其 TokenFactory 定位为“AI 时代的云计算操作系统”**，依托自研的神龙弹性算力集群、飞天分布式操作系统、以及自研的模型推理调度优化引擎，构建了覆盖模型训练、推理部署、Token 交易、安全合规在内的全栈服务能力。在生态层面，阿里百炼平台支持通义千问、DeepSeek、Llama、Qwen 等多模型的一站式调度与统一 Token 计量。其在企业级市场的核心竞争力在于，将 Token 服务与阿里云已有的钉钉企业协同生态、瓴羊数据智能平台、以及 1688 产业带资源进行了深度绑定，形成了差异化竞争壁垒。

2.5.3 全球视角：OpenRouter 与 Token 工厂的盈利能力验证

OpenRouter 是全球头部第三方 Token 聚合交易平台，截至 2026 年 6 月平台已接入 400 余款 AI 模型、合作 60 余家模型供应商，服务全球 800 万+开发者，月度

Token 处理量达 100 万亿级别，折算日均 Token 交易量超 3 万亿。

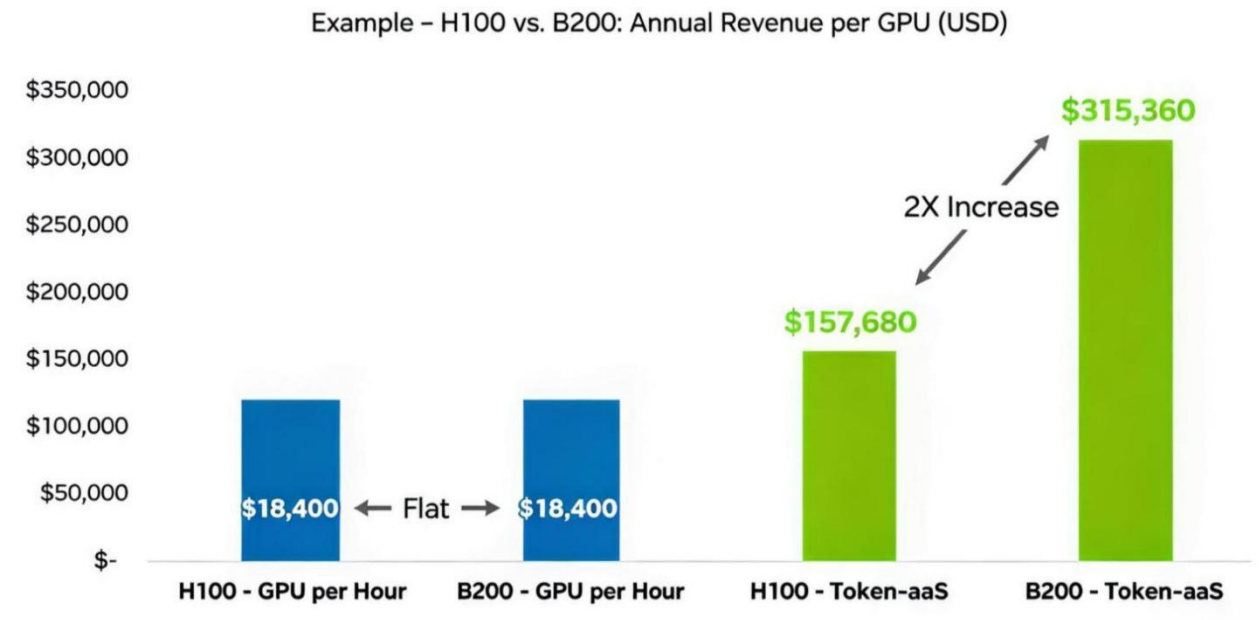
OpenRouter 的商业模式是 Token 工厂 3.0 成熟形态标杆：技术层面依托全球分布式边缘调度基础设施，通过智能路由优化推理成本，对外交付的 Token 调用价格普遍低于厂商官方直连 API；商业层面统一标准化 API 接口与 Token 计量结算规则，一站式消除企业对接多模型厂商产生的多账号、多计费碎片化成本。

据 Sacra 机构产业测算，平台 2026 年初年化经常性收入已突破 5000 万美元，充分验证 Token 标准化聚合交易具备独立、高增长的商业赛道空间。

- **从盈利能力维度对比，Token 工厂模式相比传统算力硬件租赁具备更强的利润弹性。**以美国头部 Token 聚合厂商 FireworksAI 为参照，平台依托自研推理调度体系部署开源模型，公开披露综合毛利率稳定维持 50%以上；反观传统 GPU 机架租赁业务，行业平均毛利率仅落在 20%-30%区间，盈利水平存在明显断层。**二者盈利差距核心来自资产利用效率的质变：Token 工厂依靠模型量化、批量推理、算力智能路由、动态阶梯定价等技术运营手段，相同硬件投入下 Token 产出规模数倍于传统租赁模式，完成算力资产从“按时出租”的租赁经济向“按产出变现”的 Token 生产经济转型，重构算力单位盈利模型。NVIDIA2026 年官方技术博客给出两代 GPU 的收入测算结果。**
- **图表 8 直观对比 H100 与 B200 在两种经营模式下的年度收益：两款硬件传统小时租赁年收入均为 18400 美元，差异极小；但进入 TaaS 服务模式后差距明显，H10 全年 Token 相关收入 157680 美元，新一代 B200 凭借更高推理能效年收入可达 315360 美元，相对自身租赁收入实现两倍增长。可以看到硬件架构升级叠加 Token 计费模式，会持续放大算力资产的盈利上限。**

更多免费行业报告 并购家 www.ipoipo.net

图表 8：H100、B200 级 GPU 分别在按时租赁、TaaS 模式单卡年度收入测算



来源：NVIDIA2026 年 5 月官方开发者博客，国金证券研究所

与此同时，Token 工厂催生出全新细分产业角色，Token 专业运营服务商。该类主体不持有自有算力硬件，核心输出算力调度优化、多模型适配封装、Token 输出质量管控、算力成本精细化核算等配套服务，赋能 IDC、算力厂商、模型企业提升 Token 生产效率。**该细分赛道的落地标志着 Token 产业链专业化分工持续细化，也为中小技术服务商、创业团队开辟差异化市场空间。**

第三章 Token 正在重构 AI 产业技术竞争底座

Token 商业模式的规模化落地，彻底重构了 AI 产业的技术竞争逻辑与价值评价体系。行业竞争重心已从传统的模型参数、训练性能比拼，转向 Token 全链路生产效能竞争，核心评价指标从硬件峰值算力、单卡理论性能，迭代为每瓦 Token 吞吐量、单位 Token 推理成本、全链路交付效率。**高效、稳定、低成本的 Token 生产、调度与交付能力，已成为 AI 技术研发与算力产业竞争的核心主线。**

3.1 核心技术竞争指标重塑

Token 的标准化计量特性为 AI 产业搭建了统一可量化的技术评价体系，摒弃了过往偏重硬件参数、模型规模的粗放式评估模式，构建起以词元商业产出效能为核心的全新考核标准，所有核心指标均依托权威实测与行业通用测算口径贴合商业化落地需求。

- **单位算力 Token 产出是算力资产定价与集群效能评估的核心指标，衡量单位算力的有效词元生产能力，实现硬件参数与商业价值的直接挂钩。**英伟达 2026 年官方实测数据显示，全新 VeraRubin 液冷集群的单位算力 Token 产出效率，较上代 Blackwell 架构提升最高 10 倍；国内弘信电子燧弘华创 Token 工厂落地华为昇腾 384 液冷超节点，同等算力规模下的 Token 生产效率，较传统英伟达算力集群高出 1.2 倍，国产化算力效能跻身行业第一梯队。
- **单位 Token 推理成本决定模型厂商的定价空间与盈利水平，是衡量 AI 推理商业化效率的核心指标，指代单位规模词元推理产生的综合算力、电力与调度成本。**极限纯净推理场景下，VeraRubin 架构单位 Token 推理成本较 Blackwell 架构下降 90%；国内是石科技全栈优化方案，可在同等算力条件下实现 Token 推理吞吐量提升 30%-50%，同步压降单位 Token 成本 40%。
- **Token 流转链路效率关乎终端服务体验与产品竞争力，覆盖词元生产、调度、传输、交付全流程损耗。**清程极智依托赤兔推理引擎、AIPing 评测路由平台、八卦炉智能软件栈搭建的全链路解决方案，通过智能路由择优、任务智能编排，实现推理服务成本降低超 37%、吞吐提升超 90%、推理延迟降低超 20%，大幅优化商业化交付能力。

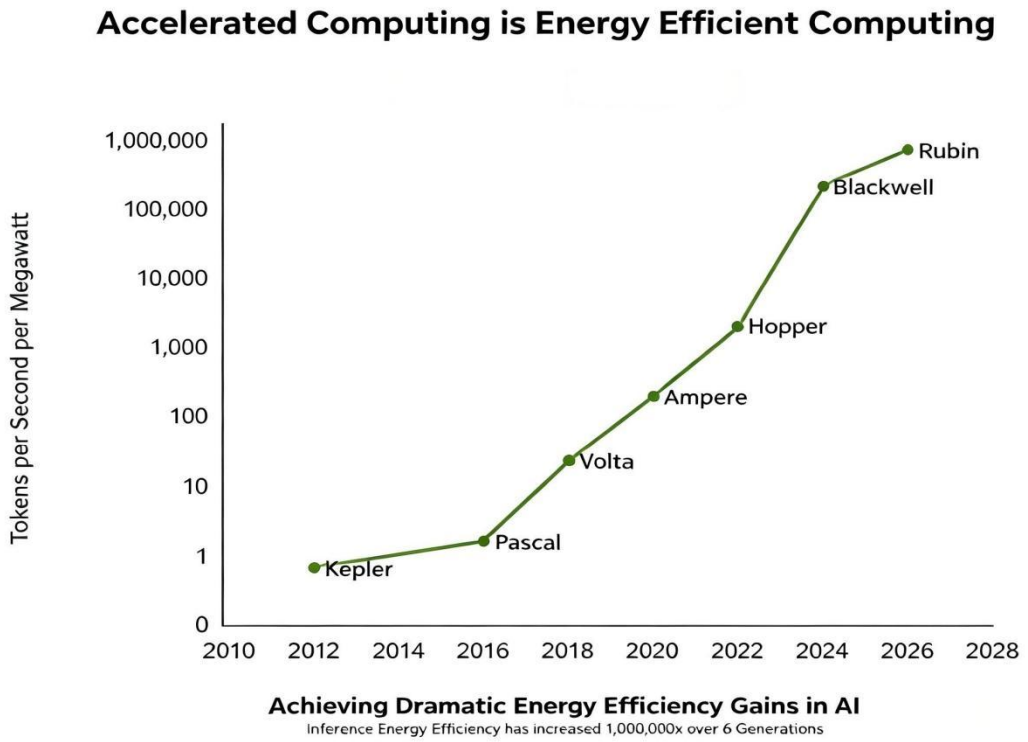
更多免费行业报告 并购家 www.ipoipo.net

- **Token 生成稳定性是政企采购的核心硬性指标，以服务有效可用时长占比为核心评判依据。**目前国内 Token 工厂行业已形成统一基准，AI 推理服务基础可用性 SLA 不低于 99.9%，成为行业规模化交付的最低准入标准。
- **整体来看，AI 产业技术研发已完成硬件参数导向、模型规模导向向商业价值导向的彻底转型，行业研发核心不再是纸面参数领先，而是聚焦 Token 全链路效能优化，最大化释放算力与模型的商业化价值。**

3.2 算力优化：从峰值能力到每瓦 Token 产出

- **算力是支撑 Token 产业的底层根基，算力的能效、单位 Token 生产成本，直接决定各类 AI 服务的产能上限与盈利空间。**在 AI+行业商业化落地阶段，算力比拼早已不再单纯比拼硬件标称峰值，核心考核指标已转向单位算力 Token 产出、单 Token 推理成本。所有软硬件优化最终都围绕有限电力与硬件资源展开，进而实现有效 Token 产出最大化。目前行业降本增效主要依靠硬件架构升级、算力智能调度、算力与模型协同适配三条路径。
- **硬件架构升级是算力提效的根本抓手，更是 Token 降本的核心驱动。**英伟达 2026 年推出 VeraRubin 平台，搭载新一代 NVLink6 互联与原生液冷方案，对比 Blackwell 架构，单瓦推理吞吐量最高提升 10 倍，单位 Token 生产成本下降九成。回溯 2012 年 Kepler 到 2026 年 VeraRubin 六代 GPU 迭代，整体推理能效实现百万级提升，推理能耗从毫瓦级压缩至纳瓦区间；其中 Hopper 迭代至 Blackwell 一代，推理吞吐、能效、综合成本三项指标同步实现 50 倍、50 倍、35 倍优化。国内这边，华为昇腾液冷超节点同样完成多轮迭代，弘信电子落地的昇腾 384 算力集群，实测 Token 产出效率优于同规格英伟达传统集群。

图表 9：NVIDIAGPU 架构历经六代架构迭代，推理能效提升 100 万倍



来源：NVIDIA2026 年 3 月官方开发者博客、国金证券研究所

从图表 9 曲线能够直观看到，英伟达六代 GPU 迭代过程中单位兆瓦算力可产出 Token 数量呈指数增长，硬件底层能效的持续突破也是 TaaS 模式能大幅拉高硬件收益的核心前提。

算力调度优化打通硬件峰值性能与实际产能的壁垒，通过异构算力匹配、任务合并、动态调度，解决算力空耗、资源错配问题。VeraRubin 配套自研调度系统，可按需动态分配算力、批量合并同类任务，最大化释放整机柜效能；是石科技智能调度方案，通过多模型任务精细化编排，实现同等算力下 Token 吞吐提升 30%-50%、成本下降 40%，适配规模化商业运营。

算力与模型适配优化依托软件技术实现精细化降本，行业普遍采用算子融合、推理引擎重构、INT8/INT4 量化压缩技术，在无损推理效果的前提下削减冗余算力

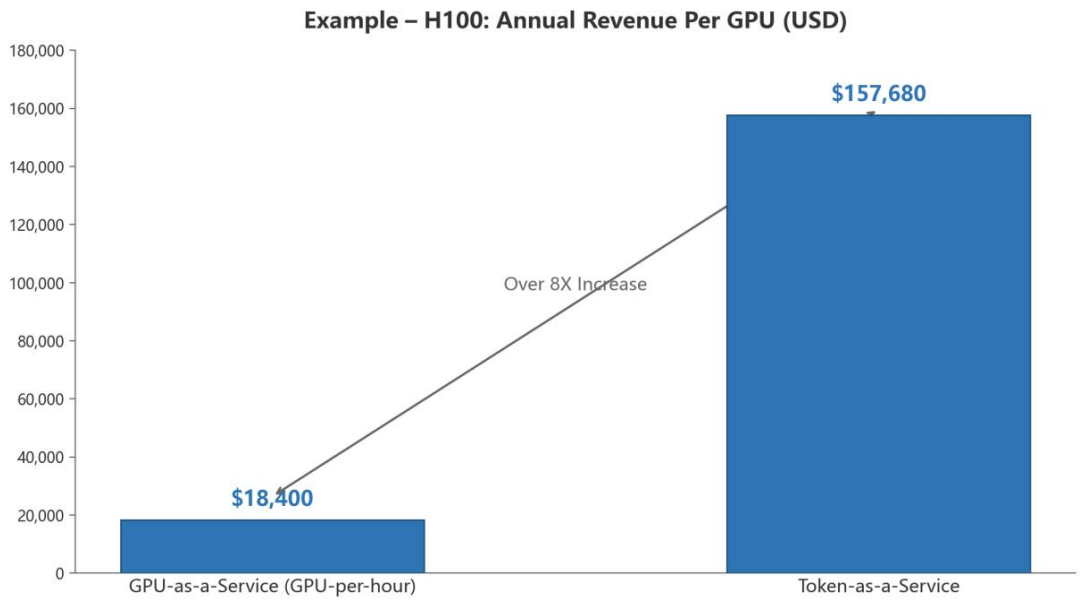
更多免费行业报告 并购家 www.ipoipo.net

开销。阿里云、华为云通过定制化算子优化与链路适配，有效压降量产场景下的单位 Token 综合成本，实现算力资源高效利用。

现在行业评判算力综合竞争力，不再只看硬件规模而是算力规模、单位 Token 产出、单 Token 成本三者结合的综合产能性价比，这也是当前算力厂商拉开差距的核心。

英伟达基于多年技术迭代构建了 Token 工厂经济学体系，明确每 Token 成本是衡量 AI 基础设施价值的核心指标，彻底替代传统硬件评价标准。其独创的 TaaS 按量计费模式颠覆了传统 GPU 时长租赁逻辑，解决了硬件利用率偏低、迭代红利流失的问题，可使同等 H100 集群年收入提升 8.5 倍，最大化释放硬件商业价值。

图表 10：H100 GPU 在按小时计费模式和 Taas 模式下年收入对比

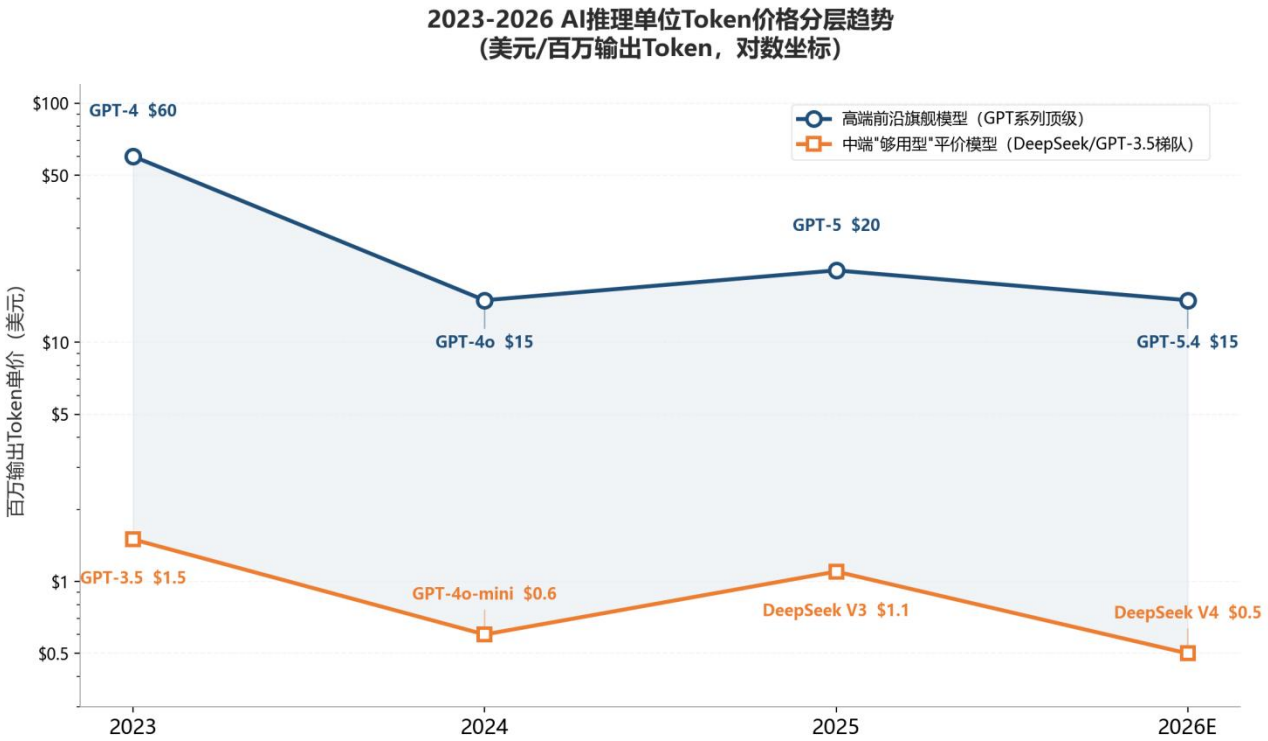


来源：NVIDIA2026 年 5 月官方开发者博客、国金证券研究所

图表 10 直观对比两种经营模式的收入差距，清晰印证依托 Token 计量的 TaaS 调度模式，能够充分利用硬件迭代带来的能效红利，大幅放大算力资产商业回报。

3.3 模型效率：推理侧的降本增效革命

图表 11：AI 推理单位 Token 成本持续下降趋势明显



来源：TokenMix.ai-《AIPricingTrendsHistory》，国金证券研究所

注：图表 11 中全部价格数据统一为百万输出 Token 计价，区分输入/输出分段定价；2026 年数值为 TokenMix.ai 基于硬件、模型轻量化技术推演预测值；数据库覆盖 300+海内外大模型厂商历年调价记录，仅统计公开商用 API 标价。

从图表 11 所表达的内容，我们可以做如下总结：

- **模型是 Token 生产的核心载体，模型竞争逻辑已从训练性能、参数规模比拼，转向推理效率、单位 Token 成本、场景适配能力的商业化综合竞争，优化路径集中在轻量化压缩、引擎调度、场景适配三大维度。**
- **模型轻量化通过量化、蒸馏、结构裁剪等成熟技术，在保障输出质量的前提下降低推理算力消耗，**阿里云、华为云均通过自研轻量化与推理优化技术，显著压降主流大模型的推理能耗与单位 Token 成本。
- **推理引擎调度优化聚焦算子融合、批处理机制升级，大幅提升算力利用率。**

更多免费行业报告 并购家 www.ipoipo.net

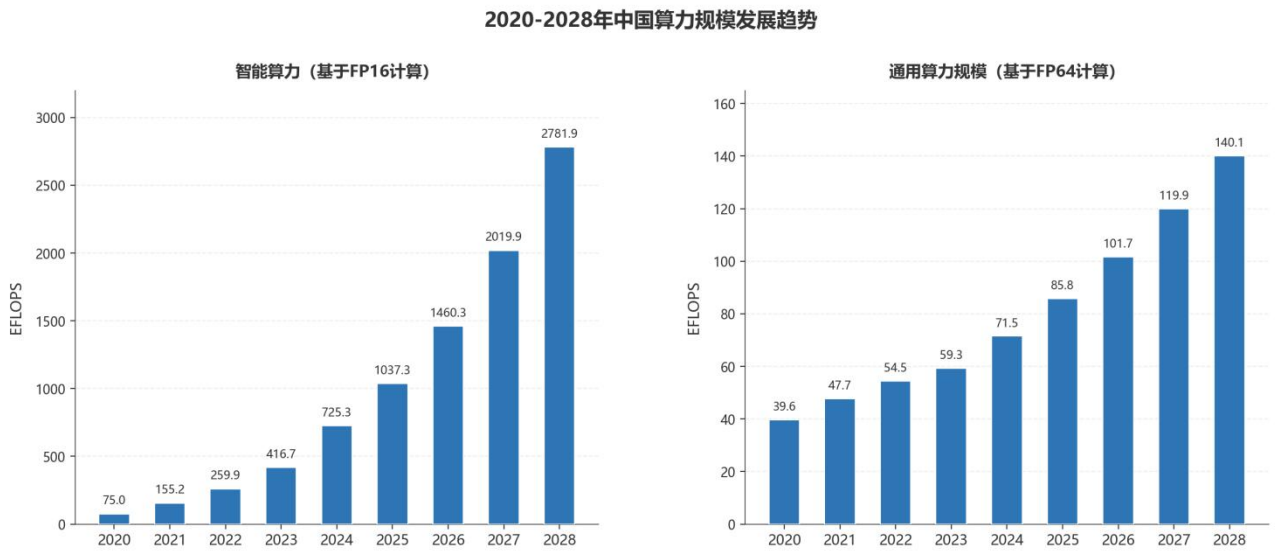
英伟达 TensorRT 引擎可智能精简计算链路、批量处理任务，显著提升推理吞吐；国内清程极智推理引擎通过同类技术优化，有效提升量产场景下的 Token 推理效率，缩小国内外技术差距。

- **场景化适配通过差异化定制模型能力，匹配不同场景的调用频率、时延、上下文需求，规避通用模型的算力浪费，是厂商打造差异化成本优势、提升商业化落地能力的关键。**
- **模型领域研发重心已从技术突破转向商业落地，单位 Token 生产效率、成本控制能力与场景适配能力，取代参数规模与训练指标，成为模型厂商的核心竞争壁垒。**

3.4 技术层冲击传导：从资源投入到价值产出的逻辑重构

- **Token 商业模式构建了完整的产业正向循环：技术效率提升→单位 Token 成本下降→应用调用放量→规模效应增强→产业盈利提升，实现技术优化与商业落地的双向赋能。**
- **IDC 联合浪潮的行业数据（参考图表 12）验证了这一传导逻辑：**2024-2026 年国内智能算力规模持续高速增长，叠加全链路技术迭代，Token 生产效率大幅提升，推动行业 Token 成交价格较 2024 年大幅下降。**技术迭代创造成本优势，支撑服务规模化普及；规模化流量增长反向倒逼技术持续优化，形成良性循环，成为当前 AI 产业持续增长的核心底层逻辑**。成本下探激活了千行百业的智能化需求，带动终端 Token 调用量爆发式增长，充分释放产业规模化红利。

图表 12：IDC 数据显示 24-26 年国内智能算力规模持续高速增长



来源：IDC&浪潮信息 2025 年中国人工智能计算力发展评估报告、国金证券研究所

3.5 Token 成本压缩的前沿突破：趋近普惠的生产力拐点

2026 年上半年 AI 推理领域涌现两大里程碑式技术路线，从输入侧、推理侧双向压降 Token 边际成本，推动 Token 向基础设施级普惠生产要素迭代。

- **输入侧语义压缩技术可大幅削减无效 Token 消耗。**开源 Headroom 框架能够在不损伤输出质量的前提下，智能精简 Prompt、对话历史、RAG 文本等内容，实现 Token 消耗量降低 60%-95%，对文档分析、智能 Agent 等长文本输入场景降本效果突出。
- **推理架构革新破解长文本推理核心瓶颈。**2026 年 6 月 arXiv 公开的 QKV 共享投影架构，可在保障推理精度的基础上，将模型 KV 缓存占用降低 96.9%，仅保留 3.1%的缓存开销，大幅减少长上下文场景的内存、带宽损耗，从底层压低推理成本。

两大技术叠加将实现 Token 成本的极致压缩，回看产业发展，**AIToken 生产成本已从 2020 年 GPT-3 时代的每百万 Token60 美元，降至 2026 年的 1 美元以内，六**

更多免费行业报告 并购家 www.ipoipo.net

年降幅超 98%。随着前沿技术规模化落地，**我们预测未来两年国内 Token 边际成本有望持续下探，有望成为水电般普惠的基础数字生产力，彻底重构 AI 产业的成本结构、定价体系与竞争格局。**

第四章 Token 正在重构 AI 产业的价值分配逻辑

- **我们认为技术层面的模型压缩、KV 缓存优化、推理架构革新等效率迭代，最终会系统性传导至产业商业端重塑 AI 行业底层运行规则。Token 商业模式对产业最核心的变革，并非单纯降低推理成本而是结构性重构 AI 产业链的价值分配逻辑、生态协作模式与行业竞争格局，而是从根本上改写全行业的盈利范式。**
- 市场对 Token 产业变革普遍存在认知误区：**多数从业者并未真正理解，为何 Token 时代会倒逼整条 AI 产业链进行全面价值重塑，以及为何下游应用场景会直接决定 Token 的商业价值天花板。**我们认为核心底层逻辑是 **Token 本身并不强制锁定产业链固定分成比例，而是 AI 产业诞生以来，第一个能够贯通算力、模型、运营、应用全链路的标准化价值计量、成本核算与商业结算单元。真正重塑产业利益格局、决定各环节收益权重的核心要素，并非 Token 数量而是下游场景的稀缺性与商业化变现能力。**
- 传统 AI 产业体系下，算力、模型、应用三大环节的成本、收益、核算体系完全割裂、互不打通。算力厂商按照 GPU 卡时收取固定租金，模型厂商依靠软件授权与定制项目获利，应用厂商赚取项目实施服务费。各环节定价口径完全独立、没有统一标尺，行业长期处于“资源投入定收益、项目包干定利润”的固化格局，上下游利益互不绑定、产业无法形成正向循环。
- Token 体系落地后，这一格局正逐渐被结构性颠覆。从上游电力能耗、硬件折

旧、集群调度，到中游模型推理、算法优化，再到下游场景落地、客户变现，产业链所有生产损耗、技术投入、服务产出，均可统一折算为 Token 量化计量。原本封闭、割裂、黑箱化的传统利润体系逐步失效，行业需要基于统一 Token 标尺，重构一套“按价值贡献分配收益、按风险承担匹配利润”的全新商业规则。

- 与此同时，Token 仅承担“成本标尺、产能单位、结算载体”的功能，并不决定最终商业价值。**Token 的生产成本由算力、电力、运维等重资产投入决定，构成产业链收益分配的刚性底盘；但同一单位 Token 的商业溢价、最终收益空间，完全由终端应用场景决定。**通用闲聊、基础文案等低附加值场景，Token 变现倍率偏低，利润更多向上游算力、模型环节集中；而金融风控、法律尽调、工业智能、医药研发等高壁垒垂直场景，能够将同等 Token 资源放大数十倍、数百倍的商业价值，场景溢价抬升应用端话语权，反向调整整条产业链的利润权重。
- 综上，**Token 时代产业链全面价值重塑的根本原因在于：Token 统一了全链路成本与产出的计量标准，打破了传统产业固化的利益边界；而场景稀缺性与变现能力，最终定义各环节真实价值权重。在此底层逻辑之上，AI 产业完成了价值分配、生态协同、竞争格局、盈利模式的全方位商业重构。**

4.1 价值分配重构：从资源采购到价值共生

- **传统 AI 产业链的合作模式的本质是资源采购逻辑，各环节收益与终端落地效果完全脱钩。**算力厂商按照 GPU 卡数、算力时长收取固定租金，只核算资源投放规模，不关注模型调用效果与用户体验；模型厂商主要依靠软件授权、定制开发项目获利，营收与场景流量、终端转化效果无关；应用厂商以系统

更多免费行业报告 并购家 www.ipoipo.net

集成、项目实施为主，收入多为固定服务费。**这种割裂的盈利模式，导致上下游风险、收益完全不匹配，产业链难以形成持续正向的商业循环。**

- **Token 商业模式落地后，行业价值分配规则被彻底改写。**产业链收益不再依据前期硬件、研发的投入体量划分，而是可以统一锚定终端真实 Token 调用营收。算力服务商、Token 工厂、模型厂商、应用开发商、生态运营方，依据各自的硬件投入、技术壁垒、风险承担和场景落地价值，按贡献参与收益分成，全链路利益深度绑定。上游通过技术迭代持续降低 Token 生产成本、提升生产效率，能够直接带动下游调用规模增长，让全链条参与者同步增收，真正形成产业共创、收益共享的良性循环。
- **目前 Token 产业链已逐渐形成相对市场化的分成体系，各环节收益权重完全取决于其在 Token 生产、流转、变现全链路中的核心价值与不可替代性。**
- **从 Token 工厂与模型厂商的合作来看，Token 工厂承担算力集群搭建、资源调度、日常运维等重资产刚性成本是 Token 规模化稳定供给的硬件基础；**模型厂商提供底座技术、推理优化、场景适配能力，是 Token 核心竞争力的技术支撑。通用场景下，重资产算力投入是核心壁垒，收益更多向 Token 工厂倾斜；而具备高稀缺性、高技术壁垒的头部模型厂商，凭借更强的技术溢价和市场议价权，能够获得更高的分成比例。
- **从算力方与 Token 运营方的分工来看，产业落地中二者多为独立主体。**算力资源方负责提供硬件算力与能耗支撑，承担设备折旧、算力闲置等刚性经营风险；运营方多为运营商、平台机构，负责整合模型与渠道资源，对接终端客户，完成交付、结算与全流程运维，承担市场化运营风险。双方最终分成比例，根据资源投入规模、风险承担水平与运营贡献综合核定，实现风险与

收益匹配。

- **从应用端价值分配来看，应用厂商是 Token 场景落地、客户转化与商业变现的核心载体，直接决定 Token 的终端价值。**行业普遍遵循“场景稀缺性决定收益权重”的逻辑：普通标准化场景下，应用端获取基础比例的 Token 流水分成；手握独家政企资源、垂直行业数据、深度定制服务能力的优质应用厂商，场景壁垒更高、价值更大，分成比例显著提升。这一机制真正实现了上游算力、模型资源价值与下游终端场景价值的对等绑定。
- **整体而言，Token 模式的出现扭转了 AI 产业链的利润分配重心，行业收益不再向重资源投入端倾斜，而是完全转向价值创造端。**在 Token 全链路中，价值贡献更大、资源投入更多、风险承担更高、场景更为稀缺的环节，能够占据更高的产业收益份额，彻底重构了行业底层商业规则。

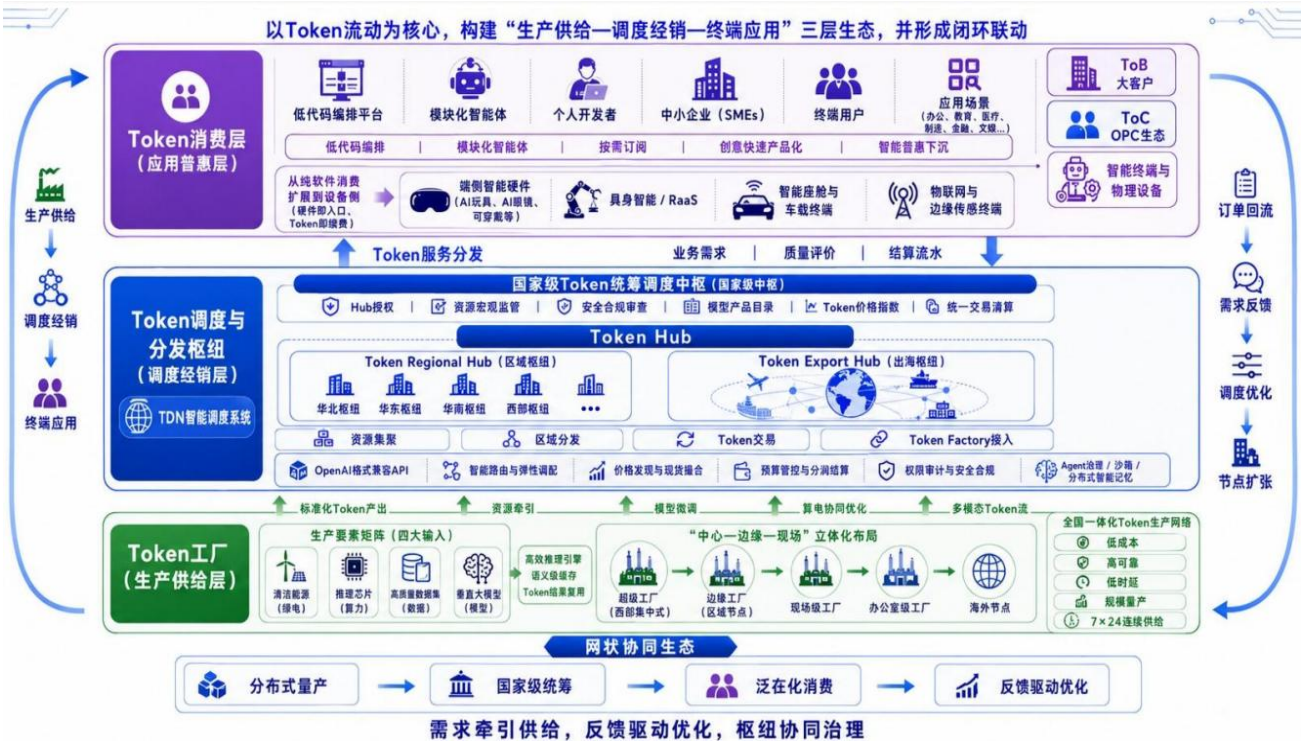
4.2 生态协同重构：从线性链条到立体网格

- 传统 AI 产业的生态协作是简单的线性交付链条，整体路径为“算力厂商—模型厂商—应用厂商—终端用户”。这种模式下，上下游合作多为碎片化、定制化的一对一谈判，缺乏统一标准接口，交易流程繁琐、成本高昂，同时存在严重的信息不对称问题，产业协同效率偏低，无法支撑大规模、标准化的行业落地。
- **Token 商业模式打破了传统线性协作模式，构建起以 Token 流动为核心的立体网格生态。如图表 13 所示，Token 工厂作为产业核心枢纽，向上整合算力资源与模型技术能力，向下为应用厂商提供标准化 Token 接口与调用服务，最终通过场景应用触达终端用户。生态内所有交易、交付、计量、结算均基于统一的 Token 标准完成，接口、计量、结算规则全面标准化，大幅消解了**

更多免费行业报告 并购家 www.ipoipo.net

传统产业的碎片化交易成本，产业整体协同效率实现明显提升。

图表 13：以 Token 流动为核心的立体网状生态体系



来源：中国工业互联网研究院、国金证券研究所

❖ 国内运营商的 Token 工厂落地案例，可直观体现这一生态变革逻辑。以中国电信宁夏 Token 工厂集采项目为例，平台采用开放生态运营模式：算力端整合多家主流算力服务商资源；模型端全面接入通义千问、DeepSeek、智谱等市面主流大模型；应用端开放标准化 API 接口，可适配上百类行业场景落地。中国电信作为核心运营主体，不直接自研算力硬件与底层模型技术，专注于生态整合、客户导流、全链路结算与运维支撑，通过统一的行业标准精准匹配上下游资源与需求。

❖ **这种平台化协同模式让各环节厂商得以聚焦自身核心能力建设：算力厂商专注提升算力利用效率、优化 Token 单位产出；模型厂商专注迭代推理技术、持续压降 Token 调用成本；应用厂商深耕垂直场景，实现 Token 价值落地变**

现；运营商持续完善生态体系、拓宽渠道覆盖、优化交易结算机制。

❖ 总体来看，Token 生态重构解决了行业长期存在的碎片化交易痛点，产业合作从点对点的零散定制交易，升级为“平台聚合、标准交付”的网格协同模式，交易效率、落地规模显著提升，为 AI 产业规模化普及筑牢了生态基础。

4.3 竞争格局重构：产业链话语权的转移

❖ **Token 模式的全面落地改写了 AI 产业链的竞争格局，行业核心话语权逐步从传统的算力资源端、模型技术端，转移至具备全链路整合、规模化交付能力的 Token 运营端。**

产业发展初期，算力资源、模型技术是行业核心壁垒，掌握硬件与技术的厂商占据产业链主导地位。但随着产业逐步成熟，算力供给持续扩容、模型技术快速普及，资源与技术逐步趋于标准化、同质化。此时，行业核心竞争壁垒不再是单一环节的资源或技术优势，而是全链路整合与规模化交付能力，即能否高效聚合算力、模型、渠道与场景资源，低成本量产 Token、大规模服务终端客户，而这正是运营商、头部云厂商的天然核心优势。

❖ **当前行业格局已清晰成型，各类厂商依托自身资源禀赋，在 Token 产业链中形成了明确的差异化定位。**

❖ **运营商成为产业核心主导者。依托全国骨干网络、稳定电力储备、海量政企与个人客户资源、成熟的计费运维体系，运营商具备 Token 规模化生产、流通、变现的完整基础能力，同时借助 Token 业务突破传统管道业务增长瓶颈。**

目前三大运营商已系统性布局 Token 经营赛道，例如中国电信落地宁夏 Token 工厂大型集采项目，重点深耕政企与产业端标准化 Token 服务；中国移动发布 Token 生态运营体系，依托全域算网底座，接入数百款主流模型，分层开

更多免费行业报告 并购家 www.ipoipo.net

放 Token 调度资源；中国联通在 2026 世界电信日明确提出，Token 将成为继语音、流量、宽带之后新一代通信基础计量品类，在多省落地个人与政企融合 Token 套餐，打造“Agent+Token+AI 云”一体化服务范式。

- **算力与头部模型厂商转型为配套供应商。**传统 AI 体系中，算力、模型厂商是产业链核心主体，但在 Token 生态下，算力资源、模型技术仅为 Token 生产的基础组件，无法直接完成客户触达与商业变现，必须依托运营方的渠道、生态与结算体系实现落地。这也导致算力、模型厂商的行业议价权有所弱化，主要获取 Token 分成的协商固定份额，从产业主导者转变为标准化配套供给方。
- **头部云厂商、互联网厂商为核心生态参与者。**这类企业具备成熟的技术平台、算力模型资源与客户运营能力，核心发展路径为整合自身技术与资源，搭建标准化 Token 资源池，再联动运营商渠道完成终端交付。例如阿里云将 MaaS 服务升级为 TokenFactory 专属业务，通过生态联动实现技术资源与渠道资源的精准匹配，成为产业重要的技术支撑方。
- **应用厂商成为核心价值放大器与主要 Token 消耗方。**场景资源是 Token 生态的稀缺核心资产，直接决定 Token 的商业变现价值，因此应用厂商的产业地位显著提升。手握独家垂直场景、优质政企客户资源的头部应用厂商，具备更强的议价能力，能够协商更高比例的 Token 流水分成，部分优质企业可在合作谈判中掌握主动权，是 Token 价值落地与规模扩容的关键核心环节。

整体来看，Token 产业的竞争逻辑已完成迭代：从单一的资源、技术竞争，升级为全链路整合能力的竞争。算力整合、模型适配、渠道覆盖、场景变现的综合能力，成为行业核心壁垒，具备全链条统筹能力的主体，逐步占据产业链核心话语

权。

4.4 商业层冲击传导：重新定义行业盈利模式

由 Token 商业模式导致的 AI 产业链价值分配、生态协同、竞争格局的全面重构，最终落地为行业盈利模式的根本性变革。AI 产业逐步告别传统的资源投入变现逻辑，进入以 Token 规模化调用为核心的全新盈利周期，各环节的收入结构、利润弹性与商业逻辑均被重塑。

- **算力端盈利逻辑全面升级。**传统算力租赁模式下，厂商收入来自固定 GPU 卡时租金，收益上限由算力硬件规模决定，无论下游调用量高低，收入基本固定，利润弹性偏低。Token 模式下，算力厂商收益与终端 Token 真实调用量深度绑定，收入由调用规模、单价、分成比例共同决定，不再单纯受限于硬件存量规模。同时，算力集群的折旧、电力、运维成本多为固定成本，随着 Token 调用规模持续提升，边际成本无限趋近于零，新增调用量可更多转化为净利润，算力行业的盈利弹性被打开。
- **模型端摆脱项目制盈利桎梏。**传统模式下，模型厂商核心收入来自软件授权与定制化开发费用，交付成本高、项目复用性差、规模化扩张难度大。Token 标准化交付模式落地后，模型能力以通用 API 接口形式对外输出，无需重复投入定制化交付成本，厂商可依托单次模型研发投入，实现海量、持续的 Token 调用增收，低成本完成规模化扩张，盈利弹性显著提升。
- **应用端实现从一次性收入到持续性流水的转型。**传统应用厂商依赖项目制实施服务费，项目周期长、人力成本高、收入不可持续。Token 生态下，厂商依托自有场景产生持续 Token 调用，获取长期流水分成，单次场景落地即可实现长期收益，优化了收入结构，提升盈利稳定性与增长空间。

更多免费行业报告 并购家 www.ipoipo.net

- **运营端打开全新增长曲线。**运营商告别了单一流量、宽带管道的盈利模式，通过整合全产业链资源、搭建 Token 交易与运营体系，获取 Token 生态全链路分成收益，打破了传统通信业务的增长天花板，培育出全新的数字化服务增长引擎。

整体而言，Token 模式让 AI 产业完成商业逻辑切换，行业从“资源有价”转向“场景有价、调用有价”。**产业链所有主体的收入与利润，不再取决于自身资源投入规模，而是由 Token 真实场景调用量与变现价值决定，Token 正式成为定义 AI 产业盈利水平的核心标尺。**

第五章 安全合规是 Token 商业模式落地的基线与护栏

- **安全合规是 Token 商业模式规模化落地的核心底线。随着 Token 成为全行业通用结算标的，部分无资质第三方平台瞄准价差套利空间，催生了 AI 中转聚合、无证 API 倒卖、Token 违规分销等中间业态。**这类脱离真实场景增值、纯依靠 Token 流转倒卖的经营模式，不仅扰乱了行业正常价值分配秩序，更滋生了数据泄露、模型篡改、数据跨境传输等一系列高危安全隐患。**因此，Token 产业的规模化发展，必然伴随严格的安全合规治理，合规体系成为规范产业价值秩序、肃清行业乱象、保障良性循环的核心护栏。**

5.1 AI 中转站风险显性化：从技术漏洞到国家安全层级隐患

目前行业内存在第三方 AI 中转聚合平台（AI 中转站），主要承接模型聚合、Token 转售、API 接口分发等业务，处于终端用户与官方大模型服务商的中间链路。这种中间人架构，在缺乏标准化监管约束和常态化安全防护的情况下，极易存在各类安全漏洞，成为整个 Token 产业链安全最薄弱的环节。

针对第三方中转平台的系统性安全风险，已有加州大学圣塔芭芭拉分校（UCSB）

安全实验室 2026 年专项学术研究实证佐证，研究团队对全网 428 个主流 AI 中转节点开展全覆盖安全审计，发现多类普遍性安全问题：部分节点存在恶意代码植入、用户凭证窃取行为，大量平台在未获取用户授权的前提下，私自留存完整对话 Prompt 数据，并归纳出三类高隐蔽性攻击路径，普通用户难以自主甄别。

- **模型替换套利（无感知偷梁换柱）：**部分中转平台为赚取上下游差价，后台私自替换用户指定的付费商用模型，用低成本、低精度的基础模型替代 GPT-4、ClaudeOpus 等头部境外模型对外输出。高低阶模型输出差异相对细微，普通用户不易分辨，平台在用户不知情下持续套利，同时拉低终端服务质量。
- **Prompt 指令劫持（长期信息操控）：**中转节点可在用户原始指令传入大模型前，静默植入隐蔽恶意指令，诱导模型推送钓鱼链接、采集用户隐私数据，或输出倾向性内容；长期多轮对话可渐进植入偏见信息，形成隐蔽的认知干预风险。
- **流式输出注入：**依托 Token 逐帧流式输出特性，攻击者可在合法回复中穿插错误信息或恶意代码，实时滚动形式难以即时察觉，隐蔽性极强。

随着行业无序扩张，AI 中转站风险已从普通网络安全问题，上升至数据安全、国家安全层面。2026 年 6 月 8 日，国家安全部官方公众号发布专项风险提示，明确界定四大核心风险：用户对话数据明文存储与未授权泄露、私自替换模型服务套利、后台植入恶意程序、用户敏感数据违规跨境传输。官方警示：部分违规中转平台存在海量用户对话 Token 数据泄露问题，内容包含商业机密与个人隐私；另有平台遭黑客利用漏洞，被盗取大额 Token 资源，造成经济与数据安全双重隐患。

更多免费行业报告 并购家 www.ipoipo.net

5.2 企业合规落地路径：四层监管体系下的生存法则

当前国内 Token 产业已形成国家法律、部门规章、地方政策、行业自律四层叠加的立体化监管体系，监管规则持续细化、治理力度持续升级。行业企业想要实现稳健可持续经营，不能依靠单点整改、临时规避，必须搭建适配多层监管要求的全链路合规体系。

- **第一层为国家法律底线。**以《网络安全法》《数据安全法》《个人信息保护法》《生成式人工智能服务管理暂行办法》为核心，划定了数据处理、模型调用、服务交付、跨境传输的基础法律红线，所有 Token 业务开展必须严格遵守，无变通与豁免空间。
- **第二层为行业部门细则。**网信、工信、国家数据局等主管部门出台专项规范性文件，针对 AI 新业态明确细化合规要求，涵盖 ICP/EDI 电信资质、大模型算法备案、数据安全评估、跨境传输审核等硬性条件，全面覆盖 Token 聚合、中转、分销、运营全流程，是企业日常合规经营的核心执行标准。
- **第三层为地方属地政策。**各省市结合区域数据安全治理与产业发展需求，出台地方性补充规则，在数据本地存储、算力调度、业务准入、服务定价等方面，提出多项高于国家通用标准的细化要求。企业跨区域布局 Token 业务时，必须因地制宜调整合规方案，适配属地监管规则。
- **第四层为行业自律标准。**行业协会、产业联盟制定的技术规范、安全准则虽无强制法律效力，但已成为头部运营商、云厂商生态准入的核心前置条件。不符合行业安全规范的中小平台与供应商，将直接被排除在核心商业生态之外，丧失主流合作资源。

在行业强监管、严执法的长期趋势下，合规投入并非单纯的经营成本，而是企业

的战略性核心投资。完善的合规体系，既能帮助企业规避刑事、行政风险，守住基本经营盘面，也能提升企业在头部生态中的准入资质与议价能力，获取高端客户信任，在行业新一轮洗牌中构筑差异化竞争壁垒。

第六章 中美对比 Token 商业模式趋同 or 背离？

作为全球 AI 产业的两大核心主体，美国和中国在 Token 商业模式的发展路径上，基于各自的产业基础、市场场景、政策环境、资源禀赋差异，走出了完全不同的逻辑，两国在技术路线、商业策略、生态组织、监管取向上的差异，本质是不同产业禀赋、不同市场需求下的差异化竞争策略，也共同构成了全球 Token 经济的完整发展格局。

6.1 技术路线差异：开源与闭源的选择

在技术路线的选择上，中美两国的差异相对清晰，这一差异本质是基于各自产业基础的优势选择：

- **美国路线：以闭源技术为核心，打造技术壁垒，通过技术优势掌握行业的核心话语权。**美国头部厂商的技术路线逻辑，是在技术的各个环节上，采用高度闭源的技术架构，从算力芯片的架构设计，到模型的推理引擎技术、量化压缩技术，再到 Token 全链路的调度管理系统，全部采用自研闭源技术架构，形成了从底层算力到上层模型再到场景 Token 接口的全链路技术闭环。

这一技术路线的核心优势是，厂商可以对全链路的所有技术环节进行深度定制与协同优化，最大化提升 Token 的生产效率；同时，通过闭源技术构建起极高的技术壁垒，锁定高端客户，在全球产业中占据技术主导位置。以英伟达为例，其从

更多免费行业报告 并购家 www.ipoipo.net

GPU 硬件架构、到 CUDA 核心计算库、再到 TensorRT 推理引擎，全部采用闭源技术，形成了全链路的技术闭环；这一技术壁垒，确保了其在高端算力 Token 生产环节的绝对竞争优势。

- **中国路线：以开源技术适配为核心，重点打造国产化算力的体系化融合能力，通过协同优化，规避单一环节上的技术短板。**中国头部厂商的技术路线逻辑，是在模型、算力、调度系统等各个环节，采用国产化技术栈，以及适配主流开源模型的技术架构；在模型侧以开源的主流大模型为基础，进行针对性的轻量化、推理优化适配；在算力侧以国产化算力芯片的技术特性为基础，进行定制化的算子优化；在调度系统侧以国产化算力的调度特点为基础，进行针对性的 Token 生产任务调度适配。**这一技术路线的核心优势是，通过国产化全链路适配，将不同厂商的不同技术资源，进行高效整合与优化配置，最大化提升 Token 的生产效率。**以中国电信的 Token 工厂为例，其在技术层面，没有完全限定使用某一家厂商的技术资源，而是通过对不同国产化算力、不同国产化模型的整合适配，构建了全链路的国产化 Token 生产能力；这一技术路线，完全规避了国内在单一技术环节上的短板。

6.2 商业模式差异：高端价值与普惠规模化的分化

在商业策略层面，中美两国的 Token 商业模式，基于各自的成本优势、场景禀赋、市场定位差异，也呈现出显著分化，本质是不同成本、不同场景优势下的价值选择：

- **美国路线：高端定价、高价值、小众场景的溢价路线。美国头部厂商的 Token 商业模式逻辑是依托自身的技术性能优势，服务于全球顶尖科研机构、大型头部企业、政府机构的高端 AI 需求。**这类客户的核心诉求是技术性能、服务

稳定性、隐私安全合规，对 Token 的价格敏感度极低；厂商通过提供高价值、高稳定性的 Token 服务进行溢价化定价以实现单位 Token 价值的最大化，美国路线的核心是依托技术壁垒获取高端价值利润。

- **中国路线：低价、普惠、海量场景、以规模为核心的路线。中国头部厂商的 Token 商业模式逻辑是依托自身的算力成本优势、海量场景优势，覆盖全球范围内的中小微企业、普通个人用户、以及下沉市场的轻量化 AI 需求。**这类客户的核心诉求是成本可控、调用稳定，对技术性能的敏感度相对较低；厂商通过极致的成本控制，提供低价、普惠型的 Token 服务，以实现 Token 调用规模的最大化。国内的海量下沉场景调用数据，虽然单位 Token 的利润极低，但凭借超大规模的调用量，完全可以支撑起整个产业的营收增长空间。中国路线的核心通过极致的成本优化，实现 Token 规模的快速扩张，以规模效应放大产业价值。
- **结合 OpenRouter 平台分周期统计数据公开测算结果，国产大模型在全球第三方开发者市场的 Token 调用规模、增长持续性均具备明确优势，中美商业化路径差异持续印证。**
- 从 2026 年 5 月 4 日－5 月 10 日统计周期来看，平台中国模型周 Token 调用量为 7.94 万亿、美国为 3.76 万亿，国内调用总量达到美国的 2.11 倍；当期中国模型调用规模环比基本持平，美国模型环比上涨 14.41%，在美国当期显著回暖的背景下，国内模型相对增长动能仍约为美国两倍，该高增速优势为该周期阶段性特征。
- 进入 5 月 18 日－5 月 24 日统计周期来看，全球 Token 市场需求持续回暖，全平台周总调用量达 28.9 万亿 Token，环比增长 7.4%。当期中国模型周调用量

更多免费行业报告 并购家 www.ipoipo.net

提升至 9.223 万亿 Token，环比增长 19.89%；美国模型周调用量为 4.93 万亿 Token，环比增长 16.27%，中国调用规模为美国的 1.87 倍。该周期中美增速差距小幅收敛，但国产模型依旧保持更高增长斜率。

- 进入 6 月 29 日－7 月 5 日最新统计周期，全球 Token 调用总量达 46.7 万亿，整体环比基本持平；其中国产模型单周调用量 23.45 万亿 Token，环比增长 15.01%，美国模型仅 4.28 万亿 Token、环比增速 4.28%，国产模型调用规模进一步扩大至美国的 5.48 倍，规模优势持续固化。
- 从头部产品结构来看，6 月 29 日－7 月 5 日周期内，全球 Token 调用榜单前六名全部为国内厂商模型。其中 DeepSeek-V4-Flash 连续七周位居平台调用首位，小米 MiMo-V2.5、MiniMaxM3、腾讯 Hy3preview、智谱 GLM-5.2、阶跃星辰 Step3.7Flash 稳定位居前列，多款全新开源国产模型上线后快速实现规模化放量。

我们认为，国产模型调用规模持续扩容的核心支撑来自国内差异化的商业化定价体系。依托海量中小企业、海外开发者等下沉应用场景，国内产业持续摊薄单 Token 算力投入与研发固定成本，在单位 Token 利润偏低的前提下，依靠超大规模调用总量持续支撑整体产业营收增长，形成区别于美国高端溢价路线的规模化商业路径。

6.3 生态组织差异：市场化交易与运营商主导协同的分化

在生态组织层面，中美两国的 Token 商业模式由于产业协同的主体、资源整合的难度差异也呈现出显著不同的特征，本质是不同产业结构下的协同逻辑差异：

- **美国路线：完全市场化的松散合作生态，由头部厂商通过技术标准构建产业协同体系。美国的 Token 产业生态是完全市场化的分散交易生态，生态内的**

算力、模型、应用厂商，都是独立的市场主体，相互之间通过签订市场化商务协议，完成 Token 资源的采购、供应、结算；没有单一主体能够主导整个生态的资源整合，厂商之间的协同关系完全由市场需求决定。以英伟达为例，其通过技术标准输出，将算力资源厂商、模型厂商的技术资源，整合到自己的生态平台中，实现技术层面的统一适配；但在商业交易层面，各方之间的结算、分成逻辑，完全是一对一市场化谈判的结果。这一生态模式的核心优势是，充分发挥了市场的竞争机制；但缺点是整合成本极高，难以形成规模化的产业协同。

- **中国路线：以运营商/头部云厂商为核心的高度协同生态。中国的 Token 产业生态，是由运营商、头部云厂商作为核心主导的强协同生态，运营商、头部云厂商凭借自身的网络、渠道、客户、生态资源，以及统一的标准化计量、结算规则，将上游算力、模型、技术服务商的资源进行集中整合，再统一向下游应用端客户交付标准化的 Token 服务。**整个生态内的技术适配、结算分成、运维支撑，都由运营方作为核心节点来统筹完成；上下游厂商之间，不需要单独进行碎片化的谈判、交易、结算。这一生态模式的核心优势是极大降低了产业内的碎片化交易成本大幅提升了产业协同效率；缺点是生态的发展高度依赖运营方的资源整合能力。

6.4 监管取向差异：创新驱动与安全可控的分化

在监管取向上，中美两国基于不同的产业发展阶段、数据安全需求、产业治理逻辑，也走出了完全不同的路径，这一差异是决定两国 Token 商业模式落地边界的核心变量：

- **美国路线：行业自主驱动，以企业自律、市场化合规为核心的宽松监管模式。**

更多免费行业报告 并购家 www.ipoipo.net

美国监管层的核心逻辑，是不干预产业的商业化创新进程，监管的重点是放在 Token 服务的合规性、用户数据隐私保护、市场交易行为的合规性上；对于 Token 本身的业务模式、交易场景、技术落地形态，没有设置任何额外的行业准入限制。这一监管导向释放了行业的创新活力，让头部厂商可以快速推进 Token 的商业化落地，充分挖掘高端场景的价值红利。

- **中国路线：政策与产业协同共振，把安全合规、标准可控放在首位的审慎监管模式。中国监管层的核心逻辑，是在产业创新发展的同时，必须严格保障行业的安全可控、合规有序、健康发展。**政策层面为 Token 产业划定了明确的落地边界：首先，所有 Token 业务必须完全锚定真实算力资源、真实场景需求，严格禁止任何形式的炒作、投机、虚拟交易相关行为；其次，Token 的生产、流转、交易的全链路环节，必须完全纳入国家的相关行业监管体系，符合国家的相关技术标准、合规标准、数据安全保护规范；第三，行业的准入门槛较高，仅有具备资质的基础电信企业、头部云厂商、头部模型厂商，才能参与 Token 产业的相关落地工作。这一监管导向为后续产业的长期可持续发展提供了坚实的制度基础，预期将中长期助力产业链规范有序且持续创新发展。

6.5 中美 Token 模式对比总结

- 综合来看，中美两国在 Token 商业模式上的所有差异，都是基于自身的产业基础、资源禀赋、应用场景、政策环境做出的现实选择，各有自己的独特优势与发展边界。**美国的技术优势决定了其可以走高端化路线获取行业的高价值利润；中国的成本与场景优势决定了其可以走普惠化路线获取规模效应下的利润空间。两国之间的技术竞争本质是高端技术壁垒与规模化场景红利之**

间的竞争；而两国的 Token 商业模式并非完全对立、非此即彼，而是存在着巨大的互补空间，美国厂商需要中国的下沉市场场景机会来放大自身的 Token 变现规模；中国厂商需要美国的高端技术能力来提升自身的 Token 全链路生产效率。

- 从实际产业落地情况看，两国的 Token 商业模式已经形成了明确的分工与互补：美国主导高端技术与高端价值场景，中国主导普惠化技术与大规模垂类下沉场景，这一格局将在未来很长一段时间内长期存在；而两国在 Token 技术、生态、商业层面的互促竞争，也成为驱动全球 Token 经济发展的核心动力。

第七章 Token 产业发展趋势展望与战略建议

7.1 Token 产业发展趋势展望

综合国内算力基础设施建设进度、全球大模型商业化落地数据及头部厂商迭代进展，未来 3-5 年国内 Token 产业将告别单纯的技术迭代驱动，进入商业规模化落地、场景价值兑现、生态格局定型的成熟发展阶段。我们预判未来 5 年，国内 Token 产业整体呈现如下发展趋势，行业竞争逻辑与商业价值体系将迎来系统性重构。

- **算力产能高速扩容，Token 产业长期增长空间明确：国内智能算力基建持续加码落地，算力供给规模稳步扩张，同时模型轻量化改造、推理调度优化、缓存技术迭代持续提升单位算力产出效率。**供需两端共同驱动下，国内 Token 调用总量进入长期高增通道，行业规模化底座持续夯实。中长期来看，Token 产业整体规模是未来 AI 基础设施核心的增量赛道之一。
- **行业价格中枢稳步下移，价格出现结构性分化：随着国内算力产能持续释放、模型推理工程能力迭代优化、行业供给持续扩容，通用场景下 Token 的生产**

更多免费行业报告 并购家 www.ipoipo.net

成本与市场定价将长期稳步下行，低价普惠是国产 Token 产业规模化发展的核心特征。价格下行并非行业恶性内卷，而是 AI 基础设施商品化、规模化成熟的必然特征。同时行业呈现明显结构性分化：标准化、通用型 Token 持续降价，但面向超长上下文、复杂智能体推理、行业深度定制等高附加值服务，仍具备技术溢价与涨价空间，整体呈现“通用普惠、高端溢价”的分层定价格局。

- **行业竞争逻辑切换，场景价值成为核心壁垒：当前通用算力、基础模型能力已进入同质化阶段，单纯依靠算力规模、模型参数已无法构建长期竞争壁垒。**行业竞争重心已从“技术能力比拼”转向“落地价值比拼”，能否深度绑定垂直行业、沉淀优质下沉场景、输出可落地的定制化解决方案，成为决定企业 Token 变现质量、客户粘性与长期议价权的核心关键。
- **行业资源加速聚拢，头部集中格局持续强化：Token 业务属于典型重资产、重运维、强合规的全链路业务，对算力整合、稳定交付、客户运营、合规结算的综合能力要求极高。**多数中小厂商受限于资金、渠道、运维体系，难以独立完成完整商业闭环。未来行业订单、算力资源、客户流量将持续向运营商、头部云厂商集中，行业集中度稳步提升，马太效应持续凸显。
- **技术壁垒持续弱化，国产开源适配成为主流路线：全球大模型技术迭代趋于成熟，中外模型综合性能差距持续收窄，海外闭源模型的技术垄断壁垒逐步消解。**行业竞争维度发生迁移，相较于传统的模型训练能力，基于国产算力的适配优化、推理降本、工程落地、场景微调能力愈发关键，国产化开源技术栈已成为国内产业落地的主流选择与安全底座。
- **运营商核心枢纽地位稳固，产业主导力持续凸显：在 Token 商业化体系中，**

算力厂商、模型厂商侧重生产供给，而运营商掌握稀缺的政企渠道、全域客户、网络传输、统一结算与合规运维能力，是产业规模化、合规化落地的核心枢纽。未来行业各类市场主体的大规模商业化变现，基本都需依托运营商生态落地，运营商的产业统筹与主导地位将持续夯实。

- **产业金融配套有序落地，赋能实体算力产业循环：随着 Token 产业规范化发展，适配实体算力交易的配套金融服务逐步成型、有序落地。**行业所有金融创新均基于真实算力产能、真实 Token 交易场景展开，涵盖供应链金融、产能预售、分级结算等工具，核心用于帮助产业链主体对冲价格波动风险、优化资金周转效率、平滑扩产现金流压力，全程服务实体算力产业，与虚拟货币炒作业态严格区分。
- **定价体系迭代升级，多模式叠加构建成熟商业体系：行业 Token 定价模式已摆脱单一按量计费的初级形态，进入多模式叠加的成熟阶段。**基础按量计费筑牢行业交易底盘，结果导向的效果计费实现价值精准匹配、改善客户付费感知，产能金融化定价实现风险对冲与成本锁定。三类模式并非替代关系，而是互补共存、分层适配，推动行业商业模式从简单资源售卖，转向精细化、价值化、稳固化的成熟业态。
- **中美产业路径分化，双轨竞争格局长期固化：全球 AI Token 产业已形成稳定的差异化双轨格局，且中长期难以逆转。**美国头部厂商依托长期技术积累、闭源壁垒、高端政企客户资源，聚焦高附加值复杂场景，维持高定价、高 ARPU 的高端溢价模式；国内厂商依托充沛且低成本的算力基座、海量下沉实体经济场景，走规模化、普惠化路线，以调用体量和落地广度构建核心竞争力，双方形成错位竞争、互补发展的全球产业格局。

更多免费行业报告 并购家 www.ipoipo.net

- **行业标准逐步统一，安全合规成为基础准入基线：国内 Token 产业正从野蛮生长走向规范发展，行业早期碎片化的技术、计量、结算、合规规则持续整合，国家层面统一的标准体系加速落地，有效降低产业协同成本与交易摩擦。**同时，数据安全、隐私隔离、合规审计、AI 安全等监管要求全面常态化，合规能力不再是企业的加分优势，而是参与行业竞争的基础准入门槛，不合规、不规范的市场产能将持续出清。

7.2 针对三类主体的战略建议

基于 Token 商业模式对 AI 产业链的技术、商业、格局重构冲击，以及中美差异化的发展路径，我们国金证券产业研究中心团队面向企业经营者、产业投资人、行业政策制定者三类不同主体分别给出战略建议：

7.2.1 面向企业创始人/实控人的战略建议

企业端未来的核心竞争逻辑不再是“做大参数、堆叠算力”，而是找准生态位置、控制推理成本、落地真实场景、实现稳定现金流。我们建议企业经营层从五大方向调整战略打法：

- **精准卡位 AI 产业链定位，形成专属不可替代能力。**行业同质化竞争加剧后，全链条通吃的企业将越来越少。算力厂商应聚焦推理效率与成本控制，打造规模化交付优势；模型厂商需放弃盲目训练竞赛，深耕轻量化、国产化适配与场景微调能力；应用厂商重点沉淀垂直行业场景与客户资源，提升生态议价权；头部云厂商与运营商则依托整合能力搭建生态壁垒，巩固平台核心地位。
- **深度融入头部生态，避免独立碎片化发展。**当前 Token 商业化的核心壁垒是全链路交付与合规运营能力。中小厂商独立获客、运维、结算、合规的成本

极高，很难形成规模化收益。企业应主动接入运营商、头部云厂商的成熟生态，依托平台渠道、客户资源与合规体系快速放量，降低试错成本，实现可持续商业化落地。

- **研发重心由训练侧全面转向推理与工程落地。**随着基础模型能力趋同，未来技术差距主要来自推理效率、调度优化、显存压缩、国产化算力适配等工程能力。企业需将研发资源向降本、提速、场景适配倾斜，通过技术精细化迭代持续降低单位 Token 成本，构筑长期性价比优势。
- **从资源售卖转向场景化一体化方案交付。**单纯的算力、模型资源已不具备持续溢价能力，只有结合客户真实业务场景才能形成稳定复购与商业价值。企业需以行业需求为导向，打包算力、模型、应用、运维、合规的一体化服务，降低客户使用门槛，提升单客价值与客户粘性。
- **理性扩产，平衡长期投入与短期现金流。**Token 产业重资产、长周期、强波动的特征明显，盲目扩产容易造成算力闲置与现金流压力。企业应根据长期订单储备、生态合作质量、自身资金能力制定扩产节奏，在保证生存现金流的前提下稳步扩张，规避周期波动风险。

7.2.2 面向一二级产业投资人的战略建议

我们建议产业投资人尽早把握 Token 产业链的投资主线，识别高壁垒、高成长潜力的投资标的，获取长期投资收益。基于这一目标，建议按以下方向布局：

- **从关注算力/模型参数竞赛，转向关注 Token 全链路效能竞赛标的：**在 Token 商业模式下，衡量企业竞争力的核心指标，不再是算力的规模、模型的参数大小，而是 Token 的全链路效能水平，即单位算力的 Token 产出、单位 Token 的推理成本、全链路 Token 流转时延、Token 服务稳定性这四类核心技术指标。

更多免费行业报告 并购家 www.ipoipo.net

投资标的筛选中，应优先选择在这四类核心指标上，具备行业级领先优势的企业，这类企业的技术壁垒、盈利弹性、成长空间，远高于行业内的其他企业。

- **重点布局全链路整合环节的头部企业：**在 Token 商业模式下，产业链的核心利润环节，集中在具有整合交付能力的头部企业手中，这类企业掌握着生态的渠道、客户资源，以及上下游的分成议价权，是产业链中价值占比最高、成长确定性最强的环节。投资布局中，应重点关注这类拥有算力资源+模型技术+渠道覆盖+场景价值全链路整合能力的头部企业，以及在分成环节中，占据优势议价权的核心生态企业；这类企业的业绩增长确定性，远高于产业链上下游的其他企业。
- **重点卡位不可替代的核心环节，掌握分成主动权：投资布局中，应重点卡位 Token 产业链中不可替代的核心环节，即产业链中资源稀缺性程度较高、议价权较高、长期价值分成占比较高的关键环节。**这类环节的核心特征是，资源的供应方数量极少，在生态整合中的价值贡献度极高。例如支撑 Token 高效生产的高密算力核心零部件厂商、关键场景资源的头部应用厂商、核心的算力调度技术厂商；这类企业在产业链中，拥有较强的分成议价权，投资价值远高于普通环节的企业。
- **关注标的的真实变现能力，规避概念性标的：**Token 产业的核心价值，是真实可量化的 Token 实际变现能力；没有实际变现能力的概念型技术、资源，没有任何长期投资价值。**在投资标的筛选中，要特别关注企业的实际客户资源、以及可量化的 Token 实际调用规模收入，重点布局已经具备成熟的规模化 Token 交付能力，且拥有稳定、高质量的长期订单支撑的企业；规避只有**

技术概念、没有真实商业化落地支撑，或缺乏头部生态渠道支撑的企业。

- **重视合规性风险，淘汰不合规标的：**Token 产业是强监管、强合规的行业；合规资质是行业的基础准入门槛，也是后续企业实现规模化变现的核心前提。在投资标的筛选中，要将合规性资质水平，作为标的筛选的前置条件；**优先选择已经具备完备的行业合规资质，且与头部运营商、头部云厂商建立了稳定的生态合作关系的企业；淘汰合规性有瑕疵、没有明确头部生态合作支撑的企业，规避行业监管政策变化带来的投资风险。**

7.2.3 面向行业政策制定者的战略建议

行业政策制定者的核心目标是构建健康、可持续发展的 Token 产业生态，释放产业红利，同时保障行业安全可控的健康持续发展。基于这一目标，建议按以下方向施策：

- **加快完善 Token 产业的顶层标准体系，建立统一的行业基准：**建议在国家层面，加快制定覆盖 Token 全链路环节的统一技术标准、计量标准、结算标准、合规标准接口，建立行业统一的可量化、可溯源、可验证的标准基准；通过行业标准的统一，降低行业内的碎片化交易成本，激活产业的协同共创价值。同时，依托运营商的网络资源、行业合规运营资质及技术落地基础，在全国范围内打造一批标准化、可复制、可推广的 Token 示范项目；通过标杆项目的示范引导，推动全行业的标准化落地。
- **大力推动国产化技术栈的适配工作，构建安全可控的产业生态：**建议针对国内产业链的技术现状，重点扶持国产化算力、国产化模型的全链路适配优化工作，通过设立专项产业基金、出台专项财税激励政策等方式，激励算力厂商、模型厂商、技术服务商，针对国产化算力的技术特性，进行定制化的模

更多免费行业报告 并购家 www.ipoipo.net

型推理适配、调度系统优化、场景化技术适配支撑；提升国产化技术栈的 Token 生产效率，构建安全可控的国产化 Token 产业生态。

- **引导完善产业链内的分成生态机制，形成可持续的产业价值循环：**针对产业链内分成机制可能存在的不合理问题，通过行业政策指导、产业联盟协调等多种方式，引导产业链上下游企业构建起基于价值贡献度的合理分成机制；平衡好算力资源方、模型技术方、应用场景方之间的利益分配关系，让各方的资源投入、技术贡献、场景价值，都能得到匹配的收益分成，形成可持续的产业价值循环。同时，通过产业引导，避免行业内的恶性降价竞争，保障行业的长期可持续发展。
- **以扩大行业应用为牵引，推动 Token 产业和实体经济的深度融合：**要明确需求牵引、场景落地的产业发展思路，将 Token 产业的规模化落地，与工业、农业、服务业等实体行业的智能化需求深度绑定，通过将 Token 的标准化服务能力，引入到实体行业的各类细分场景中，以实体行业的实际场景需求牵引 Token 技术服务的规模化落地；同时，以 Token 的标准化服务能力，赋能实体行业的智能化转型升级，实现 Token 产业与实体经济的双向互促发展。
- **构建包容审慎、灵活弹性的行业监管机制，守住安全底线：**在监管层面，要在充分吸收借鉴国际成熟经验的基础上，结合国内产业发展的实际情况，构建包容审慎、灵活弹性的行业监管机制。一方面，为产业的技术创新、模式创新留出足够的空间，在严格守住不炒作虚拟货币、不进行过度金融化的行业安全底线前提下，给予行业足够的创新试错空间；另一方面，明确行业的合规经营底线，建议将 Token 生产、流通、交易、结算的全链路环节，纳入到行业监管体系中，建立健全全流程的行业合规监管机制，强化企业的合规

经营主体责任，保障行业健康有序发展。

第八章 观点总结：Token 模式重构 AI 产业链底层逻辑与竞争格局

我们认为全球 Token 商业模式的兴起并非行业计费方式的简单迭代，而是一次覆盖 AI 技术体系、商业分配机制、产业竞争格局的全链条底层重构。Token 体系通过将算力消耗、模型研发、场景服务等所有产业链投入要素标准化、可量化、可交易，以调用量与单位成本为核心标尺，重塑了行业价值计量与收益分配规则，推动 AI 产业从零散的定制化项目时代，正式迈入标准化、规模化、流量化的产业成熟阶段。

- **技术端：行业竞争从参数比拼转向全链路效能比拼。**在 Token 经济落地前，行业技术评价体系以模型参数规模、训练算力投入、通用能力跑分作为核心标准，重“训练能力”轻“落地效率”。随着 Token 模式普及，技术竞争逻辑彻底切换，产业核心关注点转向单位算力 Token 产出、单 Token 推理成本、全链路调用时延与服务稳定性等工程化、落地化指标。技术迭代的核心目标不再是单纯的性能领跑而是实现全链路生产效率提升与成本持续优化，推理优化、轻量化适配、国产算力兼容等工程能力，正式成为技术壁垒的核心载体。
- **商业端：价值分配从资源投入导向转向真实产出导向。**传统 AI 产业以项目制、定制化交付为主，价值分配高度依赖资源投入规模与渠道壁垒。Token 模式彻底颠覆原有商业规则，产业链收益分配不再取决于资本与资源的堆叠，而是依托各环节在 Token 真实流转、调用、变现中的实际价值贡献。算力供给、模型优化、场景落地、生态运营各环节的收益水平完全市场化、透明化，构

更多免费行业报告 并购家 www.ipoipo.net

建起“价值贡献决定收益分成”的全新商业秩序。

- **竞合端：单点优势时代落幕，全链路整合能力定义新壁垒。**过往行业竞争依靠单一优势即可突围，包括算力资源、独家模型技术、垂直场景资源等。Token 规模化落地后，单一要素壁垒持续弱化，行业核心竞争壁垒升级为算力整合+模型适配+渠道触达+场景变现+合规运维的全链路综合能力。头部运营商、头部云厂商凭借生态统筹、客户覆盖、合规结算、稳定交付的综合优势，确立了产业核心主导地位，行业马太效应持续强化。
- **全球端：中美双轨差异化路径长期固化，形成错位竞争格局。**基于产业基础、资源禀赋、市场结构与政策环境的差异，中美 AI 产业形成两条长期并行、错位发展的成熟路线。美国厂商依托闭源技术积累与高端客户资源，深耕高附加值、高技术溢价的高端市场；国内厂商依托充沛低成本算力底座与海量下沉实体经济场景，走规模化、普惠化、高调用体量的发展路径。两条路线相互竞争、相互渗透、彼此迭代，构成未来全球 Token 产业发展的核心驱动力。
- 整体来看，**Token 经济的崛起标志着全球 AI 产业彻底告别早期技术探索与小规模试点阶段，进入商业化落地为主、规模化放量为核心、生态竞争为关键的全新周期。本次由 Token 驱动的产业链重构具备不可逆的趋势性。行业所有参与主体均需重新审视自身的技术路线、商业定位与生态策略，主动适配新的产业价值与竞争逻辑，才能在新一轮产业周期中稳稳把握定价权和持续收获成长红利。**

行业风险提示

当前国内 Token 产业处于高速发展与规范迭代的关键阶段，行业技术落地、商业变现、监管体系及外部产业环境均存在不确定性，产业发展过程中面临多重潜在风险。其一，Token 行业整体商业化落地、生态扩张进度存在不及预期的可能；其二，行业监管体系持续完善，存在政策超预期收紧的风险；其三，全球科技竞争格局持续演化，外部产业扰动或对国内技术迭代与产业落地造成冲击；其四，算力基础设施重资产投入属性突出，建设及运营成本存在超预期上行压力；其五，AI 技术迭代、场景适配进度存在不确定性，或拖累产业商业化节奏；其六，产业规模化应用过程中，存在隐性成本抬升、自主智能体运行合规漏洞、行业统一标准缺失引发的系统性运营风险；其七，Token 产业配套金融创新落地过程中，存在潜在的体系性风险。

更多免费行业报告 并购家 www.ipoipo.net

特别声明：

国金证券股份有限公司经中国证券监督管理委员会批准，已具备证券投资咨询业务资格。

本报告版权归“国金证券股份有限公司”（以下简称“国金证券”）所有，未经事先书面授权，任何机构和个人均不得以任何方式对本报告的任何部分制作任何形式的复制、转发、转载、引用、修改、仿制、刊发，或以任何侵犯本公司版权的其他方式使用。经过书面授权的引用、刊发，需注明出处为“国金证券股份有限公司”，且不得对本报告进行任何有悖原意的删节和修改。

本报告的产生基于国金证券及其研究人员认为可信的公开资料或实地调研资料，但国金证券及其研究人员对这些信息的准确性和完整性不作任何保证。本报告反映撰写研究人员的不同设想、见解及分析方法，故本报告所载观点可能与其他类似研究报告的观点及市场实际情况不一致，国金证券不对使用本报告所包含的材料产生的任何直接或间接损失或与此有关的其他任何损失承担任何责任。且本报告中的资料、意见、预测均反映报告初次公开发布时的判断，在不作事先通知的情况下，可能会随时调整，亦可因使用不同假设和标准、采用不同观点和分析方法而与国金证券其它业务部门、单位或附属机构在制作类似的其他材料时所给出的意见不同或者相反。

本报告仅为参考之用，在任何地区均不应被视为买卖任何证券、金融工具的要约或要约邀请。本报告提及的任何证券或金融工具均可能含有重大的风险，可能不易变卖以及不适合所有投资者。本报告所提及的证券或金融工具的价格、价值及收益可能会受汇率影响而波动。过往的业绩并不能代表未来的表现。

客户应当考虑到国金证券存在可能影响本报告客观性的利益冲突，而不应视本报告为作出投资决策的唯一因素。证券研究报告是用于服务具备专业知识的投资者和投资顾问的专业产品，使用时必须经专业人士进行解读。国金证券建议获取报告人员应考虑本报告的任何意见或建议是否符合其特定状况，以及（若有必要）咨询独立投资顾问。报告本身、报告中的信息或所表达意见也不构成投资、法律、会计或税务的最终操作建议，国金证券不就报告中的内容对最终操作建议做出任何担保，在任何时候均不构成对任何人的个人推荐。

在法律允许的情况下，国金证券的关联机构可能会持有报告中涉及的公司所发行的证券并进行交易，并可能为这些公司正在提供或争取提供多种金融服务。

本报告并非意图发送、发布给在当地法律或监管规则下不允许向其发送、发布该研究报告的人员。国金证券并不因收件人收到本报告而视其为国金证券的客户。本报告对于收件人而言属高度机密，只有符合条件的收件人才能使用。根据《证券期货投资者适当性管理办法》，本报告仅供国金证券股份有限公司客户中风险评级高于 C3 级(含 C3 级）的投资者使用；本报告所包含的观点及建议并未考虑个别客户的特殊状况、目标或需要，不应被视为对特定客户关于特定证券或金融工具的建议或策略。对于本报告中提及的任何证券或金融工具，本报告的收件人须保持自身的独立判断。使用国金证券研究报告进行投资，遭受任何损失，国金证券不承担相关法律责任。

若国金证券以外的任何机构或个人发送本报告，则由该机构或个人为此发送行为承担全部责任。本报告不构成国金证券向发送本报告机构或个人的收件人提供投资建议，国金证券不为此承担任何责任。

此报告仅限于中国境内使用。国金证券版权所有，保留一切权利。