

词元(Token)经济的内涵特征、基本规律与发展策略

欧阳日辉¹,黄芝萱²

(1. 中央财经大学中国互联网经济研究院;2. 中国科学院大学经济与管理学院)

摘要:词元(Token)作为生成式人工智能处理信息的最小语义单元,正从技术概念演变为智能经济的核心价值载体。基于技术-经济范式理论,界定了词元经济的概念内涵与技术-经济特征,系统梳理了其从萌芽、爆发到扩散的发展脉络。研究表明,词元经济的有序运行遵循四大基本规律:其一,边际成本递减与需求价格弹性共同驱动的超增长规律,形成“成本下降-降价-需求激增-调用量指数增长-进一步摊薄成本”的正反馈循环,驱动词元调用量呈现超指数级扩张。其二,规模经济与范围经济叠加的超线性收益规律,批处理效率提升、集群调度优化与固定成本摊薄带来显著的规模经济,而多模态任务共享底层计算资源与运维体系则产生范围经济,二者深度耦合使收益增速远超要素投入增速。其三,长尾应用驱动的词元效率分化规律,不同场景的单位词元所创造的用户价值显著分化,进而推动价格分层、提示词工程兴起和企业预算管理模式转变。其四,数据-词元-价值的循环增强规律,形成“更多调用→更多交互数据→更好模型→更高词元效率→进一步刺激调用”的飞轮效应。四大规律相互嵌套、深度耦合,共同构成词元经济发展的完整动力学系统。当前,中国词元调用规模已达全球领先水平,但面临定价偏低、安全合规等突出问题,特别是受数据回流断裂、基础设施错配、品牌认知缺失三大结构性制约。为实现词元经济健康可持续发展,须以数据要素为核心抓手,筑牢数据基础设施支撑体系,构建词元效率评价与数据激励机制,培育全球算力调度与成本优化生态,建立健全全链条治理体系,建立数据互信与出海保障机制,塑造中国词元国际竞争新优势,推动词元经济从超增长向高质量发展跃迁。

关键词:词元;词元经济;词元出海;数据要素;数据跨境流动;智能经济

一、问题的提出

作为智能时代的价值锚点与连接技术供给和商业需求的“结算单位”,词元(Token)经济正重塑数字经济格局。国家数据局局长刘烈宏在中国发展高层论坛2026年年会上指出,词元“不仅是智能时代的价值锚点,更是连接技术

供给与商业需求的‘结算单位’,为商业模式的落地提供了可量化的可能。”^[1]当前,全球词元经济正高速扩张,集中体现为词元调用量的爆发式增长和市场格局的深刻变化。英伟达CEO黄仁勋强调,词元生成本质是能源与算力的转化,是AI产业可量化的产出载体^[2],且单位词元成本应作为核心评估指标^[3];OpenAI也

引用本文:欧阳日辉,黄芝萱.词元(Token)经济的内涵特征、基本规律与发展策略[J].河海大学学报(哲学社会科学版),2026,28(3):127-142.

基金项目:北京市社会科学基金重点项目(24JCB027);国家自然科学基金专项项目(72442026)

作者简介:欧阳日辉(1973—),男,教授,博士,主要从事数字经济、数据要素与数字金融研究。E-mail:ouyangcass@163.com

指出,词元统计受字符、标点、语种等因素影响,非英语文本词元占比往往更高。与此同时,《人民日报》正式将“Token”中文译名为“词元”,全国科学技术名词审定委员会同步发布相关公告,标志着该译法在国内正式确立。国家数据局数据显示,我国日均词元调用量从2024年初的1 000亿次,飙升至2025年底的100万亿次,2026年3月突破140万亿次,两年增幅超千倍^[4]。OpenRouter数据显示,2026年3月中国大模型周调用量远超美国,全球前五热门模型中我国占据四席^[5]。当前,中美AI调用量交替领先,行业竞争持续加剧,同等水平大模型对话成本18个月内下降至原来的1/280^①。词元调用量的爆炸式增长,既反映智能经济活动的空前活跃,也导致价格的断崖式下跌,推动词元从技术计量单位转变为衡量AI产品产销的核心经济标尺,成为研判全球AI产业发展的重要先行指标。然而,中国词元经济在发展过程中却暴露出“规模超增长但价值锚定不足”的结构性矛盾:词元调用量规模已居全球领先地位,凭借庞大的词元生成规模占据市场,但单位词元的定价显著偏低,且受到数据回流断裂、基础设施错配、品牌认知缺失三类结构性制约。探索这一深层悖论不仅关乎词元经济的健康可持续发展,更涉及数字经济时代国家竞争力构建的核心命题。

针对词元经济的健康发展,我国政府展开了一系列战略性布局和顶层设计,并以此构筑坚实的政策框架:一是依托示范应用激活数据要素价值。国家数据局等十七部门联合印发《“数据要素×”三年行动计划(2024—2026年)》,明确2026年底前打造300余个标杆场景,拉动数据产业年均增速超20%、实现数据交易规模倍增^[6]。二是完善数据流通底座以支撑产业发展。可信数据空间规划提出,2028年将建成百个以上可信数据空间^[7]。三是推动数据产业向规模化、服务化升级。相关指导意见设定2029年产业年均增速15%以上的目标,重点扶持模型即服务等新兴业态^[8]。目前,政策成效逐步显现:截至2025年底,全国优质数据集超10万个、总量达890 PB;2026年3月,日均词元调用量突破140万亿次,较2024

年初增长千倍以上^[4]。一系列政策部署表明,我国正从概念规范、基础设施到交易规则等多维度系统搭建词元经济治理框架。

目前,学界对词元经济的研究已从概念引入步入了机制探究阶段,但仍处于理论初创期,现有研究成果集中在概念内涵、经济属性、价值度量与产业趋势四大方向:概念界定方面,黄群慧将词元经济视作智能经济的重要组成部分,明确其以词元为核心载体,关键是要解决智能经济量化难题^[9];属性研究方面,陈波指出,剖析词元经济属性关乎解读增长动能、把握智能时代经济规律^[10];价值度量方面,有学者认为词元是智能经济的核心价值标尺,可为AI产业提供可量化、可交易的价值体系^[11];产业趋势方面,多家研究机构认为词元将成为算力、模型等领域的通用计量标准^[12],行业报告数据也印证了词元经济高速增长,如2025—2030年我国词元调用量将保持高速扩容态势^[13]。

总体来看,现有研究为理解词元经济奠定了基础,但研究多停留于概念解读与现象梳理层面,缺少对词元经济内在运行机制的系统性剖析,尤其是未形成词元从技术单位向经济单位转化路径的完整理论体系,且对词元经济发展规律的研究零散,针对词元出海相关现实问题的探讨较为薄弱。聚焦词元经济内在发展规律,本研究从概念内涵、技术经济特征、运行基本规律、出海制约因素与政策路径展开研究,重点回应三大核心问题:一是词元经济的核心内涵与技术经济属性;二是词元经济快速扩张的内在运行规律;三是低价竞争、合规压力及数据回流断裂、基础设施错配、品牌认知缺失三大深层因素对词元出海的制约作用。基于上述问题,立足技术-经济范式,系统阐释词元从AI基础技术单元,演变为算力、服务与价值核算核心尺度的过程,并从超增长规律、超线性收益规律、词元效率分化规律及数据价值循环规律4个维度搭建词元经济理论分析框架,同时剖析词元出海现实阻碍,围绕数据要素、交易规则、

①数据来源:AI Index Steering Committee. Artificial Intelligence Index Report 2025[R/OL]. (2025-04-08)[2025-04-25]. https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf.

基础设施、国际合规等方面提出优化政策建议。

二、词元经济的内涵特征与发展历程

随着生成式人工智能规模化普及,词元已从技术单元演化成为智能经济核心范畴,词元经济作为新型智能经济形态,其发展特征与产业体系日趋清晰。立足既有研究,明确词元与词元经济的定义、内涵及意义;依托技术-经济范式理论,剖析词元经济关键生产要素、基础设施、产业扩散及治理特征,并划分词元经济的发展历程,梳理其从技术概念向经济形态的质变路径。

1. 词元与词元经济的定义与内涵

词元是大模型处理信息的最小单元,具有可计量、可定价、可交易等属性,是人工智能服务计费、结算和价值核算的重要基础^[14]。但这一“最小单元”并不是静态概念,早期词元主要指自然语言处理中的文本片段,随着多模态大模型发展,词元的范围由文本扩展到多模态信息。随着大模型进入智能体和具身智能等场景,词元还能用于衡量任务复杂度、资源消耗和服务价值。因此,本研究讨论的 Token 为大模型文本词元,特指对文本、多模态内容切分后的基础计算单元,既包括狭义的语言模型分词单元,也包括广义上可被模型编码、调用、生成和结算的智能信息单元,与区块链领域的数字通证 Token 分属两类完全独立的概念。

当前学界与产业界对词元经济的定义尚未统一,但已形成四类主流视角:一是计量单位视角,以词元作为 AI 应用的核算与定价基础,实现智能服务量化落地^[15];二是产业模式视角,指出词元经济已取代传统流量经济,形成以智能生产力为核心的按量计费商业范式^[16];三是生产要素视角,认为词元是新型核心生产要素,赋能认知经济发展^[17];四是价值度量视角,将词元视作智能经济的通用价值标尺,认为其构建了可交易的量化价值体系^[11]。梳理既有研究可知,词元经济的界定经历了从技术计量到经济范畴、从单一功能到系统框架的深化过程,各类定义的切入点虽不同,但核心共识一致:词元已由技术概念升级为经济范畴,词元经济是以词元为基础,开展 AI 服务生产、交易定价与

价值分配的新型经济形态。然而,现有界定仍存在不足:多为局部视角,缺少整体系统性定义;未充分解释词元具备经济属性的底层逻辑;词元经济与算力经济、智能经济等关联形态的边界与关联也未得到清晰界定。综合上述既有研究的贡献与不足,本研究提出以下定义:词元经济是指以生成式人工智能中的词元(Token)为基本计量单位和价值载体,围绕模型推理服务的生产、定价、交易、消费与价值分配而形成的一套资源配置方式与经济运行体系。词元经济既是一种新型产业组织形态,也是智能经济时代价值衡量与资源跨主体配置的基础性制度安排。从概念层级看,数字经济、智能经济与词元经济具有包含关系:数字经济是最广义的经济形态,智能经济是数字经济在人工智能阶段的深化形态,词元经济则是智能经济中以词元为基本计量单位、价值载体和结算尺度的核心组成部分。

结合上述定义,词元经济的内涵可从以下 5 个层面加以阐释:第一,词元是 AI 时代基础计量单元。大模型推理生成均会消耗词元与算力,相较于按次、按时计费,按词元核算更贴合任务复杂度与调用强度。第二,词元计费构成新的商业逻辑。OpenAI、DeepSeek 等官方平台均按输入、输出词元差异化定价,词元已成为 AI 服务定价的核心基准。第三,词元是智能经济的价值度量衡。词元兼具信息产出与算力消耗双重属性,可同步衡量投入与产出,为 AI 服务搭建可比的经济核算体系。第四,词元具备新型生产要素特质。区别于静态数据,词元是模型加工后的动态知识成果,凭借可拆分、可计量、可交易的特性支撑精细化资源调配。第五,词元经济推动智能服务生态协同优化。词元需求扩张倒逼数据集、算力、模型推理能力持续优化,形成数据、模型、价值联动的良性循环。从更宏大的历史视野看,词元经济正推动信息经济向认知经济跃迁,这一进程正在深刻改变知识生产、内容生成、产业组织和价值实现方式,重构数据要素价值释放路径,加速制造、农业、医疗等传统行业的智能化升级。因此,词元经济的深层意义在于,其使 AI 服务从难以度量的技术能力转化为可核算、可交易、可治理的生产要素。人工智能不再只是“能做什么”的技术

问题,也成为“做了多少、成本多少、效率多高、价值多大”的经济问题,这正是词元经济作为智能经济新范式的重要价值。

2. 词元经济的技术-经济特征

技术-经济范式理论是演化创新经济学的核心分析框架^[18],该理论指出重大技术革命会催生适配的经济运行模式,主要涵盖关键生产要素、新型基建、主导产业、生产组织与制度体系。在这一框架下,关键生产要素普遍具备成本下行、供给充足、跨域通用的特征。技术革命从新的关键要素出现,经由动力部门(生产关键要素的产业)和支柱部门(大规模应用关键要素的产业),最终催生引致部门(在关键要素带动下兴起的新兴产业),完成从技术突破到经济范式转换的全过程。有别于数据要素、算力经济、平台经济等经济形态,词元经济使 AI 服务从不可度量的技术能力转化为可核算、可交易、可治理的经济对象。依托技术-经济范式,系统剖析词元经济的技术经济特征,重点是要凸显其与既有经济形态的质的差异,明确其对既有数据要素理论的增量贡献。

首先,词元已具备关键生产要素的典型特征。与数据要素理论强调“价值实现依赖场景赋能”不同,词元将 AI 技术标准化、可量化拆解,形成独立的价值衡量体系:一是成本持续下行。区别于数据要素“成本高、变现难”的特点,国内大模型词元单价近年来降幅较大,并且伴随技术优化将持续降低,按需计费模式倒逼企业良性竞争。二是供给快速扩张。词元产出依托能源、算力、数据、算法协同支撑,供给扩张速度与协同效率远超单一数据要素,突破了“供给分散、难规模化协同”的困境。三是通用性与可计量性突出。数据要素缺乏统一计量标准,相比之下,AI 场景均可量化为词元消耗,多模态技术能拓展词元计量范围。技术-经济范式转换必然伴随新型基础设施的构建,词元经济正在形成“能源-算力-服务”的完整基础设施链条,与数据要素、算力经济的基建体系存在本质差异:第一层是能源基建。区别于数据要素基建中能源的辅助定位,依托“东数西算”工程,我国将西部绿电转化为算力优势,直接支撑词元规模化供给与成本控制,为词元经济提供底层能源

支撑。第二层是算力基建。由 GPU 集群等构成的词元生产系统,以“词元产出效率”为核心,突破了算力经济“重规模、轻效率”的倾向。第三层是平台基建。阿里、腾讯等头部企业围绕词元重组模型、平台和应用能力,核心功能是词元计量、交易与协同而非单纯数据或算力调度。第四层是应用生态基建。智能体平台、API 聚合服务等推动词元渗透千行百业,实现“可交易性”落地,解决数据要素“场景化变现难”的问题。

其次,词元经济的核心是以词元为基础的价值传导机制。词元经济产业扩散同样遵循技术-经济理论的三层结构规律,但扩散逻辑与既有形态存在本质不同,词元经济的动力部门是模型训练、推理服务提供商等词元生产者,核心竞争力聚焦于词元成本控制。智谱 MaaS 平台 2025 年营收同比提升 60 倍,API 业务毛利率提升近 5 倍^①,模型服务商业化潜力凸显,这是数据要素等形态难以实现的。支柱部门是词元规模应用者,其核心是通过词元实现 AI 价值可衡量,破解“投入高、效果难量化”的困境。据火山引擎披露,2026 年累计 Token 使用量超万亿的企业客户从 100 家增至 140 家^②,非 AI 原生企业也加速接入模型服务。引致部门则是由词元经济催生的新兴产业,涵盖智能体开发、词元效率优化、提示词工程等。2026 年被产业界普遍视为“智能体(AI Agent)规模化应用元年”,这些引致部门的涌现标志着词元经济产业生态的全面繁荣,突破了数据要素“难催生全新产业”的理论局限。

最后,词元经济推动生产组织与治理机制变革。不同于数据要素治理侧重安全隐私、平台经济侧重垄断规制,词元治理主要围绕计量准确性与交易公平性展开:一是按词元计费成为模型服务商业化基础模式,解决 AI 服务“定价模糊”痛点,实现从技术服务到商品交易的转变。二是分级定价与精细化资源配置强化,

① 数据来源:北京智谱华章科技股份有限公司. 2025 年年度报告[R/OL]. (2026-03-31) [2026-04-25]. https://www.zhipuai.cn/investor_relations/2025%E5%B9%B4%E5%A0%B1.pdf.

② 数据来源:谭待. 豆包大模型日均词元调用量破 120 万亿[N]. 证券日报, 2026-04-03(B03).

不同模型能力、任务复杂度对应不同价格层级,推动算力向高价值场景倾斜,这能有效解决数据要素“价值难分级、配置低效”的问题。三是多维度动态价格机制完善词元交易体系,推动其成为真正的经济对象,OpenAI 官方价格体系已形成区分输入输出、服务层级的多维度动态价格结构。四是词元计量与交易等治理需求凸显,建立可信、透明、可审计的计量和治理体系,能为 AI 时代经济治理提供新方向。进一步分析发现,词元经济的价值分配关系也随生产组织的变革而变化。与传统产业按物质生产分配收益不同,词元经济分配主要取决于调用规模、模型能力、服务品质与交互数据,其分配逻辑为:算力决定基础成本,模型赋予技术溢价,场景带来服务溢价,用户数据反哺模型迭代。收益若过度向算力、模型端集中,会挤压应用开发者利润;若终端定价长期偏低,模型厂商难以支撑训练、推理与安全治理成本,易引发行业低价恶性竞争。

总体而言,词元经济并非既有经济形态的简单延伸,而是人工智能从技术可行走向经济可行、从模型能力走向产业组织的关键环节。词元经济的出现也补充并拓展了数据要素理论:词元使 AI 产业获得了类似电力经济中“千瓦时”、石油经济中“桶”的基本计量单位,也使智能经济的成本核算、价值分配与治理有了清晰的分析对象。

3. 词元经济基于技术-经济特征的发展历程

词元经济作为生成式人工智能走向大规模商业化应用的产物,经历了从技术概念到新兴经济形态的演变。基于词元调用规模、商业模式成熟度与产业扩散广度等变化,将词元经济的发展历程划分为萌芽期(2017 年底—2024 年初)、爆发期(2024 年初—2025 年底)和扩散期(2026 年初至今)共 3 个阶段。

萌芽期,相关技术突破催生了词元经济新范式。词元经济并非凭空出现,而是大模型商业化的结果,2017 年 Transformer 以注意力机制重塑自然语言处理,2022 年底 ChatGPT 推动大语言模型向公众普及,词元从分词编码环节逐步进入产业视野。作为模型处理文本的基本单

元,词元主要承担计量输入输出与计算成本的功能。2022 年底至 2024 年初,API 服务尚未广泛规模化,词元仍是“隐形耗材”,尚未独立成为可定价、可交易的经济对象。此阶段的词元作为关键生产要素的条件逐步成型:价格因技术垄断偏高,但算力基建、模型优化带来的供给潜力开始显现;基础设施体系开始搭建,GPU 集群和数据中心为词元生产奠定物质基础;大模型厂商构成的“动力部门”初步形成,但支柱和引致部门尚未壮大。

爆发期,词元实现了从成本单元到计价单元的转变。2024 年以后,生成式 AI 渗透多场景,词元调用量高速增长,模型服务价格快速下探,如字节跳动豆包主力模型 2024 年 5 月给出 0.000 8 元/千 Tokens 的企业定价^①,推动大模型价格进入“厘时代”。词元由此从后台成本转为前台计价基准,按量计费、API 调用和平台化分发共同构成 MaaS 商业闭环。智谱 2025 年开放平台及 API 收入由 0.49 亿元增至 1.90 亿元,平台注册用户超 400 万,覆盖 218 个国家和地区^②,表明词元计费已具备持续收入能力。此阶段,词元作为关键生产要素特征集中显现:词元价格短期内下降两个数量级,供给弹性、普适性显著提升;各大科技公司加速基建布局,模型厂商(动力部门)规模化产出词元,规模应用企业(支柱部门)快速崛起。

扩散期,大模型向智能体演进,服务拓展为 TaaS 范式。2026 年以来,智能体推动大模型从单轮问答转向长链路任务执行,推理、工具调用等大幅增加词元消耗,使其成为智能生产的消耗品与产出物。郑纬民指出,产业竞争正由 MaaS 转向 TaaS,重点比拼词元产能、成本等效率指标^[19]。同时,国家治理开始介入,如国家数据局将 Token 定名为“词元”并探索词元交易等新型模式;2026 年政府工作报告亦提出要

① 数据来源:谭待. 字节跳动豆包大模型发布,火山引擎全栈 AI 服务助力企业智能化升级[EB/OL]. (2024-05-21) [2026-04-25]. <https://www.volcengine.com/docs/6360/1264663?lang=zh>.

② 数据来源:北京智谱华章科技股份有限公司. 2025 年年度报告[R/OL]. (2026-03-31) [2026-04-25]. https://www.zhipuai.cn/investor_relations/2025%E5%B9%B4%E5%A0%B1.pdf.

打造智能经济新形态、深化拓展“人工智能+”。此阶段,词元经济特征系统耦合,价格持续处于“厘时代”;供给弹性因“东数西算”工程和算力基础设施的持续扩张进一步增强,普适性因全模态模型的普及达到新高度;应用生态基础设施进入爆发期,动力、支柱部门持续壮大,词元效率优化服务等引致部门逐步涌现。

正如《中国经济时报》所指出的,词元调用量的超快速增长是一种特殊的经济现象,其凸显了对词元经济进行深入研究的必要性和迫切性^[20]。词元经济的演进远未完成,未来的发展轨迹将取决于技术创新、市场竞争和政策治理三者的持续博弈并趋向动态平衡。

三、词元经济发展的基本规律

词元本质上是数据要素的新表现形式,各种非结构化语言通过处理转化成可计量、可交换的标准化单元,实现了信息价值的可度量性与可交易性^[21]。目前,词元经济正处于由规模高速扩张向高质量价值创造深度转型的特殊阶段,有别于传统产业与平台经济的发展规律,支撑词元经济有序运行的四条相互关联的基本规律——超增长规律、超线性收益规律、词元效率分化规律和数据—价值循环增强规律,构成了嵌入我国数字经济结构性转型进程的核心动力学体系,既解释了词元调用量的爆发逻辑,也指明了行业向价值优化演化的方向。

1. 边际成本递减与需求价格弹性驱动的超增长规律

超增长规律是词元经济爆发的核心驱动力之一。相较于传统制造业产能约束下的成本趋稳、互联网平台经济流量见顶后的成本刚性,依托算力迭代、算法优化与批处理复用,词元经济呈现出无物理产能硬约束、成本下行空间更大且价格弹性放大效应更强的特殊形态。其运行逻辑主要由边际成本递减与需求价格弹性放大效应共同主导,推动词元调用量实现指数级增长。

(1)生成式 AI 推理单位词元边际成本持续快速下降

生成式 AI 推理单位词元的边际成本的持续快速下降是超增长规律的核心机制,这一趋

势由三大因素驱动:一是硬件迭代赋能。升级的新架构 GPU 能显著提升词元处理效率,单位美元可购买的算力每 18—24 个月实现翻倍,如英伟达 Blackwell B200 GPU 的词元/s 吞吐量能达到 H100 GPU 的 3.7 至 4 倍^①,大幅提升了硬件利用效率。二是算法优化创新。INT8 量化、稀疏化等压缩技术能够降低模型内存需求、推理延迟和计算开销,并用于 LLM 部署^[22],其中 INT8 量化在一定精度损失可控条件上能有效降低内存占用、提升处理速度。三是批处理规模经济释放,同一 GPU 并行处理更多请求,能够有效摊薄内核启动、显存搬运等固定开销,由此能进一步压缩单位词元边际成本,边际成本下降直接传导至 API 定价,使国内主流大模型每百万词元价格在两年内实现大幅下降。截至 2026 年 4 月,DeepSeek 等头部厂商每百万词元输入缓存命中价格已降至 0.02 元,与 2024 年初相比降幅显著^②,印证了成本下降向定价端的传导效应。

(2)需求价格弹性的放大效应

词元服务的高需求价格弹性进一步放大了超增长效应,其区别于传统商品与互联网服务的核心特征是无刚性消费门槛、海量潜在场景及价格阈值敏感:一方面,实时摘要、代码补全、客服机器人等大量潜在场景,在词元价格跌破阈值后会集中转化为实际消费;另一方面,词元消费不存在固定投入门槛,用户可根据价格波动灵活调整使用量,实现消费规模的平滑扩展。这种高弹性使价格每下降一个百分点,需求量的上升幅度远超过一个百分点,结合边际成本持续下降的反馈机制,可形成“成本下降→降价→需求激增→词元调用量指数增长→进一步摊薄固定成本→继续降价”的正循环,推动超增长规律持续发挥作用。

(3)超增长规律的经济实证

我国词元调用量的爆发式增长,为这一规

① 数据来源:NVIDIA. NVIDIA blackwell sets new standard for generative AI in MLPerf inference debut[EB/OL]. (2024-08-28) [2026-04-26]. <https://blogs.nvidia.com/blog/mlperf-inference-benchmark-blackwell/>.

② 数据来源:DeepSeek. DeepSeek-V4 系列模型 API 价格说明[EB/OL]. (2026-04-28) [2026-04-26]. https://api-docs.deepseek.com/zh-cn/quick_start/pricing/.

律提供了强有力的实证支撑。根据国家数据局公布的数据,2024 年初我国日均词元调用量约为 1 000 亿,至 2026 年 3 月已突破 140 万亿,两年间增长超千倍。同期,每百万词元价格下降约两个数量级,词元调用量的增长与价格下降呈现高度吻合的态势,进一步验证了边际成本递减与需求价格弹性驱动超增长规律的合理性。从经济含义来看,这一规律表明当前词元经济仍处于规模扩张向价值提质的过渡期。未来,随着边际成本逐渐逼近电力、硬件等基础成本下限,词元调用量增长率将逐步转为线性增长或进入 S 型饱和阶段,倒逼行业从追求调用量转向挖掘单位词元价值,契合我国智能经济转型的必然要求。

2. 规模经济与范围经济叠加的超线性收益规律

超线性收益规律根植于经典的规模经济与范围经济逻辑。区别于传统制造业单一品类规模报酬有限、互联网平台业务边界固化的特征,依托大模型技术同质性、多模态任务共享基础设施,词元经济形成了规模与范围经济深度耦合、收益增速远超要素投入增速的态势,催生了“垄断倾向与竞争并存”的独有格局。

(1) 超线性收益规律的规模经济效益

词元推理领域的规模经济,本质是“规模越大、成本越低”,其优势主要来源于 3 个方面:一是批处理效率的提升。GPU 可并行处理多个请求,隐藏部分内存搬运和计算等待时间,从而提高吞吐量^[23]。二是集群调度优化的赋能。大规模词元服务提供商可通过混合调度策略,合理分配不同优先级、不同延迟要求的请求,充分填满 GPU 的空闲时隙,避免硬件资源闲置,这与亚马逊 AWS、阿里云等公共云“多路复用、削峰填谷”的核心逻辑一致。三是固定成本的摊薄效应。模型训练、技术研发、数据中心建设等固定投入可通过更大规模的词元调用量进行分摊,规模越大,单位词元承担的固定成本越低,由此,头部厂商在单位词元成本上形成明显优势,中小厂商则面临较高的规模壁垒。

(2) 超线性收益规律的范围经济效益

范围经济侧重于“多任务的资源共享”,即同一基础设施可支撑多种词元相关任务的高效

运行。相较于传统行业多品类经营需新增固定投入、边界受限的特点,词元服务的范围经济具备底层技术同源、多模态拓展边际成本趋近于零的独特优势,其可行性主要源于底层计算逻辑相似,即不同任务虽在输入输出编码上存在差异,但核心计算过程均依赖矩阵运算(如 Attention 机制、FFN 网络)^[24]。具体而言,范围经济的收益主要体现在 3 个层面:一是共享硬件与显存。不同类型的任务可在时间上交错填充,提升硬件利用率。二是共享工程与运维。一套监控、调度、容灾系统可服务于所有模态的词元任务,无需为不同任务单独搭建运维体系,大幅降低运维成本。三是共享模型基座。GPT-4o、Gemini、Qwen3.6 等多模态模型均采用同一参数集处理不同输入类型,边际新增任务时仅需优化输入输出编码,无需重复投入模型研发成本。

(3) 超线性收益形成市场集中趋势

超线性收益规律的规模经济效益与范围经济效益的双重叠加,使得词元供给市场呈现明显的寡律分布,即极少数头部巨头占据大部分市场份额的“市场向头部集中”趋势。头部厂商凭借规模经济实现单位词元成本大幅降低,进而可设定更低的服务价格,吸引更多用户与词元请求,从而进一步摊薄固定成本、提升批处理效率,强化规模优势。同时,范围经济让头部厂商能够以极低的边际成本拓展多模态服务,丰富服务场景,增强用户黏性,进一步扩大市场份额。因此,词元经济具有自然垄断或寡头竞争倾向。但这并不意味着市场完全丧失竞争活力,开源有望成为垄断的重要突破口。开源模型的代码、参数和工具链开放后可被多主体重复使用,这使中小企业、科研机构和个人开发者能够以较低成本参与词元生产。开源社区通过代码优化、数据清洗、模型评测插件开发,扩大词元供给主体,削弱少数闭源模型对底层能力的垄断。开源协同还具有正外部性,单个开发者的优化成果能被整个社区共享,提高整体词元生产效率。2026 年政府工作报告提出的“支持人工智能开源社区建设,促进开源生态繁荣”表明开源社区已成为智能经济发展的重要支撑,助力词元经济形成“垄断倾向与有效竞争并存”的特殊格局。这一规律说明,词元经

济既需发挥头部企业的规模效率,也要通过开源生态和差异化竞争防止低水平内卷,推动行业从规模扩张转向技术创新和价值深耕。

3. 长尾应用驱动“词元效率”的分化规律

长尾应用的大规模崛起催生了“词元效率”分化规律,这一规律深刻影响词元市场的价值分配、定价策略与发展走向。其核心逻辑在于,词元价格的持续下降让低价值长尾应用得以规模化落地,进而导致不同应用场景的词元效率差异凸显,最终引发一系列连锁性经济后果,成为词元经济从高速增长向高质量分化演进的重要标志。

(1)“词元效率”的定义

词元效率本质上是单位词元所创造的用户价值,这种价值具体体现为经济收益、时间节省、用户满意度提升等多元形式,且不同应用场景的词元效率存在天壤之别。从实际应用来看,词元效率呈现出清晰的等级分层,具体类型与其创造的价值估算如表 1 所示。

表 1 词元效率的应用类型及其创造的典型价值梳理

应用类型	每百万词元创造的典型价值
医疗诊断辅助、法律合同分析	数百至数千美元
代码生成、学术论文润色	数十至数百美元
客服对话、邮件撰写辅助	数美元至数十美元
游戏 NPC 闲聊、娱乐故事生成	不足 1 美元
无意义重复对话、垃圾内容生成	接近零甚至负值

资料来源:作者整理。

(2)长尾应用的膨胀机制

词元效率分化的核心驱动力是长尾应用的规模化膨胀,而这一膨胀过程的关键前提是词元价格的持续下降。与互联网平台长尾需求具备基本商业价值不同,词元经济的长尾应用大量集中在低价值、弱效用领域。随着成本下降,大量原本不具备经济可行性的低效率应用得以落地,最典型的案例便是游戏 NPC 闲聊场景:当每百万词元价格从 10 美元降至 0.5 美元时,若一个游戏 NPC 每小时闲聊消耗 10 万词元,成本便从 1 美元降至 0.05 美元,这种成本降幅让开发者愿意大规模部署智能 NPC,丰富游戏交互场景。此外,自动回复“谢谢”“好的”等礼貌性生成场景,虽然单次词元消耗创造的价值极低,但由于应用场景广泛、使用频率极高,其

累计词元消耗量十分可观。值得注意的是,这种低词元效率的长尾应用往往数量极多、使用频率极高,会逐渐吞噬词元总消耗量的主要份额,导致词元经济总调用量中,高价值应用占比不断下降,长尾低价值应用占比持续上升。这一趋势与我国日均 140 万亿词元调用量的结构分布高度吻合,其中大量调用集中在低价值的基础交互场景。

(3)分化规律的经济后果

长尾应用驱动下的词元效率分化,推动行业向价值创造转型:其一,定价策略呈现明显的价格分层。词元服务商根据应用场景的词元效率差异,实施差异化定价,如对于医疗、金融等高词元效率的实时应用,服务商收取一定溢价,重点保障低延迟、高可靠性,满足专业场景的核心需求;对于闲聊、日志总结等低词元效率的批处理应用,则采取极低定价策略,最大化利用算力资源。其二,提示词工程兴起与词元效率优化成为趋势。为降低成本、提升价值回报,高词元效率场景的用户(如企业、专业从业者)会主动投入精力优化提示词,通过结构化、精准化的指令,用更少的词元达成同等甚至更优的效果,由此还催生了“词元效率顾问”这一新兴职业,帮助企业和个人提升词元利用效率。其三,企业词元预算管理模式发生转变。企业在采购 AI 服务时,不再单纯关注总词元量的多少,而是更加注重“有效词元”占比,即能够真正创造价值的词元消耗比例,这种需求推动词元服务商提供更细粒度的价值度量服务。这些变化倒逼市场摆脱粗放式竞争,转向单位价值挖掘与资源高效配置,契合智能经济高质量发展的导向。

4. 数据-词元-价值的循环增强规律

数据-词元-价值的循环增强规律突破了传统平台“用户-数据-服务”的简单循环,形成以词元交互为核心、反馈数据精准反哺模型迭代、价值与调用双向强化的形态。该规律不仅构筑了头部厂商的核心数据壁垒,进一步强化市场集中趋势,还反映出词元经济要想长期发展必须从规模竞赛转向价值提升。

(1)循环增强规律的闭环机制

词元交互的闭环机制是这一循环的起点,每一次“用户输入-模型输出”都会留下可复用

的优质数据资产,其中既包括用户点赞、点踩等明确的显式反馈,更涵盖了大量隐含的行为反馈信号:用户复制生成内容,意味着对输出结果的认可,构成正面反馈信号;用户选择重新生成,表明输出未满足需求,属于负面或不确定反馈信号;用户对生成内容进行编辑修改,则形成了针对性的修正信号,这些反馈共同构成模型优化的重要来源。经过处理后,交互数据可用于模型微调、强化学习和检索增强等环节:模型微调使模型更适应特定领域和用户偏好;强化学习借助人类反馈提升输出的安全性和可用性;缓存与检索增强则可减少重复推理,降低单位词元成本;模型优化后,同样的词元消耗能够产生更高质量的输出,或者用更少词元完成同等任务。由此,用户获得更高价值回报,调用意愿也会增强,闭环得以持续运转。

(2) 闭环机制形成的飞轮效应

闭环机制的持续运转,最终形成了强大的飞轮效应,构建起“更多调用→更多交互数据→更好模型→更高词元效率→进一步刺激调用”的正反馈飞轮,且循环增长规律与前文所述的超增长规律形成双重加速效应,共同推动词元经济实现指数级扩张。其中,超增长规律从供给侧降低了词元服务的门槛,让更多用户能够负担词元消耗,为飞轮启动提供了基础;数据-词元-价值的循环则从需求侧提升了词元服务的吸引力,通过持续优化模型质量、提升词元效率,让用户愿意投入更多词元消费,进而推动

飞轮加速转动。

(3) 循环增强规律构建数据壁垒

值得注意,数据-词元-价值的循环增强规律还会催生明显的数据壁垒与先发优势,其强度远超传统平台。由于这一循环的核心依赖真实用户的交互数据,早期实现大规模部署的词元服务提供商能够率先积累起海量、高质量的反馈数据,这些数据具有极强的稀缺性^[25];新进入者即便拥有参数规模相近的模型,也可能因缺少高质量反馈数据而难以快速提升词元效率。这种数据优势会随着循环的持续运转不断扩大,形成“强者愈强”的格局,进一步强化超线性收益规律中“市场向头部集中”的趋势。这一规律说明,词元经济的竞争本质上不仅是算力的竞争,更是数据飞轮的竞争,头部厂商凭借早期积累的数据优势,会不断加宽竞争“护城河”,巩固市场主导地位,而这一现象折射出我国智能经济“高质量数据集供给不足、中小主体数据积累门槛高”的结构性制约。对此,国家数据局相关征求意见稿提出要探索词元交易等新型交易模式、构建数据集价值体系,正是从政策层面补齐短板,引导行业跳出规模内卷,转向数据价值深挖,推动词元经济完成结构性跃迁。

5. 四条基本规律的耦合关系

词元经济的四大基本规律并非孤立存在,而是相互嵌套、相互赋能、深度耦合的有机整体,其共同构成了词元经济发展的完整动力学系统。图 1 系统梳理了四条规律的耦合关系,

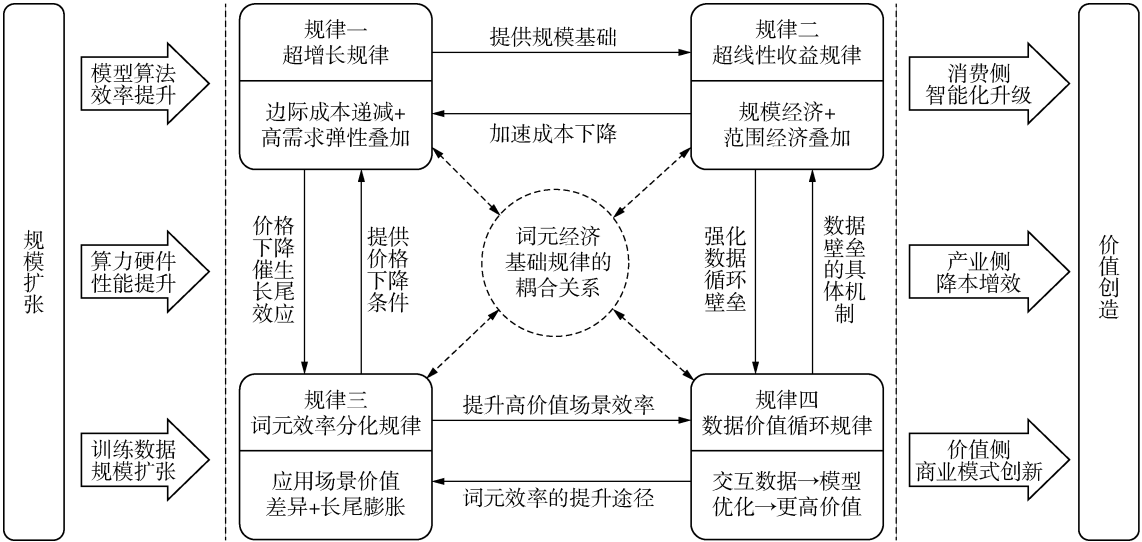


图 1 四条基本规律的耦合关系

其中,超增长规律以边际成本递减和高需求弹性为驱动力,推动词元调用量指数增长,为超线性收益规律提供规模基础,也为词元效率分化规律提供价格条件。超线性收益规律的核心驱动力是规模经济与范围经济,其带来的市场集中与成本优势又会加速边际成本下降,并强化数据-价值循环的数据壁垒。词元效率分化规律源于场景价值差异与长尾应用膨胀,其形成依赖超增长规律导致的价格下降,且催生的效率优化需求为数据-价值循环提供了方向。最后,数据-价值循环通过数据交互推动模型优化,进而创造更高价值,其产生的飞轮效应使头部企业具有先发优势,这既是超线性收益规律数据壁垒的具体机制,也是提升词元效率的关键途径。四大规律联动发力,共同支撑词元经济从“规模扩张”向“价值创造”深度转型。

四、词元跨境流通的商业模式与主要挑战

国内 AI 市场竞争日趋激烈,各大模型厂商为突破增长瓶颈、寻找新的发展空间,纷纷加快布局步伐,词元出海也已从最初的趋势探讨逐步进入大规模落地实践阶段。结合我国当前词元出海的实际探索情况,将其主流商业模式归纳为 4 种,同时详细阐释在出海进程中,定价偏低的结构性问题、安全合规方面的多重挑战及三大深层结构性制约如何成为限制词元经济进一步发展的三大关键阻碍。

1. 词元出海的主流模式

我国词元出海已形成多元发展格局,衍生出多种成熟度不同的主流模式,主要包括 API 服务输出、云平台出海、智能体生态输出及垂类 B 端解决方案等。

其一,API 服务输出模式是当前最成熟、规模最大的模式。该模式以标准化 API 为媒介,向全球开发者输出大模型推理服务,按词元消耗量计费,核心是将国内“电力-算力-词元”全链条成本优势转化为跨境数字服务。第三方平台 OpenRouter 数据显示,2026 年 2 月中国大模型周词元调用量达 5.16 万亿,3 月升至 7.359 万亿,远超美国的 3.536 万亿,总占比达 61%,全球调用量前五模型中中国独占四席^①。

其二,云平台出海模式是 API 服务的基础

设施延伸。该模式以阿里云、腾讯云等为代表,通过海外自建数据中心或智算中心,整体输出中国云服务能力,为词元出海提供直接的算力接入点。此外,部分独立智算服务商采用“轻资产协同”模式,以智算操作系统和调度能力为核心,与海外本地主体合作建设运营,具备轻资产、快复制、合规灵活的优势。

其三,智能体生态输出模式是词元经济进入扩散期后的新兴出海路径。2026 年被视为智能体规模化应用元年,智能体应用对词元消耗较传统应用提升百倍,为出海带来新增长空间。如月之暗面 Kimi 推出智能体一键部署功能,无需硬件和编程,精准满足海外用户需求,登顶 OpenRouter 调用量周榜,海外收入超国内;稀宇 MiniMax 更是实现了 73% 的收入来自海外,业务覆盖全球超 200 个国家和地区^②。

其四,垂类 B 端解决方案模式聚焦产业纵深。这一模式为金融、医疗等特定行业提供定制化 AI 解决方案,将词元能力嵌入海外企业的业务流程,以项目制或 SaaS 订阅实现价值变现。智谱在此领域探索最为深入,73.7% 收入来自本地化定制化模型部署^③,还在马来西亚落地国家级主权 AI 平台,结合本地数据主权要求,探索出“主权 AI”出海新路径。

2. 词元出海定价偏低制约我国词元经济发展

数据资产定价的原则维度较为多元,主要包括:价格可以真实反映买家效用、买方收入最大化、隐私保护和计算效率等^[26]。作为一种数据资产,词元的定价方法体现了以上原则的取舍与融合。当前,我国词元出海定价偏低,表面上词元经济的成本优势能快速抢占全球市场,但从词元经济四大基本规律来看,这种持续性

① 数据来源:OpenRouter. Real-Time Model Usage Rankings [EB/OL]. [2026-04-27]. <https://openrouter.ai/rankings>.

② 数据来源:MiniMax Group Inc. 截至 2025 年 12 月 31 日止年度的年度业绩公告[R/OL]. (2026-03-02) [2026-04-26]. <https://filecdn.minimax.chat/public/f464310a-6b55-4375-9845-723b82332ca5.pdf>.

③ 数据来源:北京智谱华章科技股份有限公司. 2025 年年度报告[R/OL]. (2026-03-31) [2026-04-25]. https://www.zhipuai.cn/investor_relations/2025%E5%B9%B4%E5%A0%B1.pdf.

低价正从深层次制约行业长期健康发展与国际竞争力提升。

一是对超增长规律的制约。超增长规律的核心是“边际成本下降→价格下降→需求激增→进一步摊薄成本”的正反馈循环,但当价格下降速度快于边际成本下降速度或低于可持续盈利水平时,循环将难以为继:一方面,我国词元出海定价仅为海外同类产品的 1/20 至 1/10,相同调用量下,收入大幅低于海外企业,直接限制研发投入,悬殊差距导致国内企业在算力、人才、基础研究上投入受限;另一方面,国内价格战惯性延续至海外,企业以低价为核心竞争手段,行业整体利润率下降,陷入“增收不增利”的困境,难以积累资金用于长期发展。

二是对超线性收益规律的制约。超线性收益依赖规模与范围经济形成的正向循环,但定价偏低从两方面制约这一过程:一是海外算力基建投入不足,中等规模海外智算中心需数亿美元投资,低价导致企业难以支撑如此规模的海外基建投入,智谱、DeepSeek 等已出现服务延迟问题,规模效应无法充分发挥。二是低价抑制服务分层,企业缺乏溢价空间构建差异化定价体系,难以在高利润的垂类 B 端领域突破,范围经济潜力未能释放。

三是对词元效率分化规律的制约。合理定价能够引导资源向高词元效率场景流动,但定价偏低降低了效率优化激励,用户及开发者缺乏动力精简提示词,导致低效使用词元。同时,低价还吸引了大量低价值、高消耗应用,稀释了高价值场景占比,形成“劣币驱逐良币”的局面,降低行业整体效率。

四是对数据-价值循环增强规律的制约。数据-价值循环的关键是高质量数据驱动模型优化,但定价偏低易吸引大量低质量用户,产生的噪声数据难以用于模型优化,甚至造成“数据污染”。同时,低价使企业与用户仅维持浅层次交易,难以获取高质量反馈数据,削弱模型迭代速度,进而影响飞轮运转。

低价策略虽能在短期内抢占市场份额,但从长期看正在对词元经济的健康发展形成系统性制约。当前,词元出海正处于从“规模扩张”向“价值创造”转型的关键时期,能否走出低价

竞争的路径依赖尤为重要。值得关注,智谱 2026 年一季度上调 API 价格 83%后,调用量仍增长 400%,这印证了高质量服务可实现“量价齐升”,即当模型能力足够强、服务质量足够高时,用户愿意为更好的体验支付更高价格。因此,走出低价依赖,转向以技术创新和服务差异化为核心的价值竞争,才是我国词元经济实现高质量出海的发展之道。

3. 词元出海的核心安全与合规问题

在商业模式的支撑之下,词元出海面临一系列不容忽视的挑战,其中安全与合规问题尤为突出。

其一,词元出海本质是数据跨境流动,需承受国内外双重合规压力。国内层面,《中华人民共和国网络安全法》《中华人民共和国数据安全法》《中华人民共和国个人信息保护法》构建了完善的出境监管体系,明确了个人信息、重要数据出境需符合合规路径。但词元调用中,用户提示词、模型输出可能含敏感信息或商业秘密,且常伴随技术出境,合规风险突出。国外层面,各国数据保护立法差异显著,欧盟《通用数据保护条例》(GDPR)强调合法透明原则、美国侧重行业自律与消费者保护、东南亚地区推进数据本地化,不同司法辖区对同一行为认定不同,形成“一地一策”的碎片化合规环境,增加了企业的适配成本。

其二,算力硬件供给不足也是词元出海的重要阻碍。数据作为生产要素并非单独发挥价值,而与算力、存储等实体介质的资源分配息息相关^[27]。美国对我国实施先进计算芯片出口管制,导致国内企业难以获取高端 GPU,这种“卡脖子”问题直接制约词元出海服务能力:一方面,海外调用量激增需要充足的海外算力节点支撑,但芯片限制导致智谱、DeepSeek 等企业出现服务延迟等问题,算力节点建设速度直接决定了国产模型增长上限;另一方面,高端算力不足也限制了高价值词元的规模化输出,影响企业盈利。

其三,日益上升的地缘政治风险亦给词元出海带来挑战。随着中国模型在全球 API 调用市场份额攀升,技术竞争易被政治化解读,“数据安全威胁”“技术依赖”等议题可能引发

数据本地化升级、API 服务限制等连锁反应。从更深层次来看,不同司法管辖区的制度摩擦使得企业应对各类合规要求需投入高额且重复的成本,这既会增加负担,又会降低商业效率,成为词元经济全球化扩展的结构性瓶颈。

其四,词元出海面临严峻的结构性挑战。产业链层面,独立大模型厂商在算力成本竞争中处于劣势,大厂凭借完善的算力基建掌握定价权,依赖 API 服务出海模式的独立厂商长期盈利能力受到严重挤压。同时,部分企业海外收入占比极高,对海外市场依赖度较强,易受单一市场政策、竞争格局变动影响,市场多元化布局成为词元出海企业在扩张过程中亟待解决的战略命题。

其五,词元出海还面临服务本地化能力不足的挑战。当前,词元出海多停留在技术输出层面,仅向海外开发者提供模型 API,而在服务交付环节存在明显不足,如海外算力节点布局不够、本地化技术支持体系不完善、多语言与多文化适配能力不足等。在进入欧美成熟市场时,这些短板更为突出,海外用户对服务响应速度、稳定性要求更高,而国产模型尚未完全适配当地服务标准与用户习惯,难以建立可持续的市场竞争力。

4. 词元出海的三大深层结构性制约

超越定价和安全合规层面,我国词元出海还面临三大更深层次的结构性制约,这些制约与词元经济核心规律深度关联,是词元高质量出海的关键瓶颈,也是实现全球化突破必须跨越的障碍。

一是数据回流与飞轮效应断裂,这与数据—价值循环增强规律直接相关。该规律的核心是词元交互数据通过微调等机制优化模型,形成自增强飞轮,需依托高质量、适配场景的海量数据。词元出海后,模型部署于海外,用户的交互数据(如提示词、生成内容)受欧盟 GDPR、美国隐私法等法规限制,无法合法回流国内用于模型训练,即便匿名化处理,也存在流程复杂、成本高、政策风险大等问题,导致数据—价值循环断裂。尽管我国模型在 OpenRouter 平台占比超 61%,但海量调用数据无法反哺模型,迭代只能依赖国内数据,与海外场景适配不

足;而 OpenAI 等海外头部企业可依托本土数据形成闭环,迭代更快、适配性更强,我国模型不仅面临数据获取障碍,还在海外长尾场景中效率难以提升,在高价值场景渗透方面受到阻碍。

二是海外基础设施与成本结构错配,这关联词元经济的“新型基础设施”特征与超线性收益规律。国内词元成本优势依托“东数西算”、西部绿电等形式形成成本洼地,超线性收益的规模经济与范围经济也基于此,但出海后成本优势大幅削弱:国内数据中心电价低至 0.3 元/度,而欧美地区电价达到了 0.6~1.2 元/度,单位词元电力成本或上升 2~4 倍;同时,高端 GPU 因美国出口管制获取困难、成本偏高,海外数据中心建设、国际带宽等成本也明显更高,这导致海外规模经济难以显现,叠加国内低价策略,企业面临“越出海、越亏损”的倒挂风险,部分企业探索的“轻资产协同”模式也存在利润让渡、依赖合作伙伴的局限。

三是品牌认知与价值锚定缺失,这与词元“价值载体”属性紧密相关。词元价格是技术、品牌、合规的综合体现,用户支付的价格中包含着信任溢价,现阶段我国模型性能虽已跻身全球第一梯队,但海外用户尤其是企业级客户存在“信任赤字”,即部分用户对中国技术产品存在固有偏见。反观海外的头部企业则绑定了成熟生态,有标杆客户背书。此外,我国企业在 SOC 2、ISO 27001 等国际认证上投入滞后,形成进入高价值市场的隐性门槛。上述多种原因导致中国词元被锚定为“低价替代品”,即便智谱 GLM Coding Plan 已获海外用户认可,其性能优势也未转化为普遍溢价,难以渗透海外企业级市场。

五、词元经济健康可持续发展的政策建议

词元经济的高质量发展,本质上是使数据要素完成从使用价值到交换价值再到国际竞争力价值的三重跃迁。目前,我国词元经济正处于从“规模扩张”向“价值创造”转型的关键时期,要破解当前存在的定价偏低惯性、安全与合规问题及三大结构性制约,关键在于以数据要素为核心抓手,建立一套系统前

瞻的政策体系,这是因为数据不仅是词元经济运行的核心支撑,更是打通国内国际双循环、重塑价值锚定、优化资源配置的重要政策着力点。

1. 前瞻布局筑基,筑牢数据基础设施支撑体系

适度超前建设数据基础设施,为词元经济落地提供坚实的支撑。面对人工智能加速演进、数据要素加快流通、数字经济深度重构的新形势,数据基础设施建设不能满足于“跟着需求走、围着问题转”,而应坚持前瞻布局、统筹谋划、梯次推进,切实把基础能力建设做在前面、把制度规则完善在前面、把关键支撑储备在前面:一是强化顶层设计,按照国家数据基础设施“三步走”安排,围绕云网数算协同,加快形成覆盖行业、区域、企业的一体化布局。二是发挥重大项目牵引作用,聚焦基础、公益、战略等方向,推动算力网络、数据流通平台、可信数据空间等重点设施落地,筑牢大规模数据流通和高频模型调用的底座。三是加快标准体系建设,完善技术架构、接口规范等领域制度供给,推动设施互联互通、互信互认;坚持试点先行,加快形成可复制、可推广的经验。四是统筹生态培育与开放合作,调动政企研等多方力量,构建协同创新的发展格局,避免基础设施滞后于产业变革、技术迭代,从而把“超前布局”真正转化为支撑词元经济发展的“先手优势”和“长板优势”。

2. 提质增效引领,构建词元效率评价与数据激励机制

建立词元效率评价体系与高质量数据激励体系,为破解词元经济发展瓶颈提供制度保障。面对定价偏低导致的低质量数据泛滥、飞轮数据质量下降、词元效率分化及资源配置扭曲等突出问题,不能仅停留在被动应对,而应切实把评价标准立起来、激励机制建起来、数据质量提上来。一是发布“词元效率指数”并纳入行业考核,由国家数据局联合工信部牵头,明确单位词元产出、任务完成度等核心指标,定期发布指数;对效率领先企业,在算力分配、政府采购中予以倾斜,树立行业标杆。二是建设“高价值词元交互数据集”国家资源库,聚焦医疗、法

律、金融等高效场景,整合脱敏后的高质量数据,以“数据沙箱”形式向合规企业开放,助力模型微调与词元效率优化。三是实施“数据质量反馈补贴”政策,通过税收抵扣、算力券等方式,补贴提供高质量反馈数据的用户和企业,激励其参与数据质量优化;建立反馈数据分级定价机制,明确区分普通使用数据与高价值反馈数据,形成科学合理的市场价格信号。

3. 数智协同赋能,培育全球算力调度与成本优化生态

培育以数据驱动算力调度与成本优化的生态,助力摆脱词元出海成本困境、释放规模经济效应。针对国内成本优势不可迁移、海外基础设施成本高企、规模经济难以实现等问题,需坚持数据赋能、协同发力、精准施策,切实通过技术创新、机制完善、资源整合等途径,降低海外运营成本。一是建设“全球词元算力调度大脑”,依托国家算力平台,构建跨境智能调度系统,实时监测全球节点算力、电价等核心指标,动态分配词元请求至成本最优节点,优先在“一带一路”沿线的电力成本优势国家布局协同节点。二是推动“算力-数据-词元”联合报价的标准化合约,借鉴电力市场经验,设计词元期货、捆绑合约等金融工具,锁定海外成本并降低风险;依托国家数据交易所设立“词元交易专板”,试点组合产品挂牌交易,提升流动性。三是建立出海企业“算力成本共担基金”,由政策性银行、产业基金和头部企业出资,以低息贷款、股权投资等方式支持企业共建海外算力节点,并结合数据分析实施差异化支持政策,优先资助高效、合规企业,全方位优化词元出海成本。

4. 安全合规护航,构建词元经济全链条治理体系

构建词元安全与合规的治理体系,应对词元经济的新兴风险。随着词元由技术计量单元加速走向结算单位和价值载体,其外延不断扩展、应用持续渗透,相关风险也由单一技术风险演化为交织数据安全、金融秩序与国家安全的复合性风险。一是防范概念混淆引发的市场失序,明确 AI 词元、区块链代币与网络身份令牌的边界,规范市场认知,防止技术概念被炒作和

异化利用,遏制诈骗、传销等非法活动。二是筑牢全流程技术安全防线,利用加密传输、数字签名、权限最小化等机制,防范词元泄露、伪造、篡改造成账号被盗、数据外流等问题,守护产业布局、技术路线等核心信息。三是健全协同治理体系,推进产业标准、交易规则、风险提示机制建设,强化执法监管协同,加大对非法融资、数据窃取等行为的打击力度,推动技术防护、企业合规、政府监管与公众认知多方共振。四是建立词元跨境流通的知识产权合规机制,推动建设模型输出的来源记录、授权证明、使用边界和争议处理规则,降低跨境服务中的版权侵权和商业秘密泄露风险。

5. 跨境融通破界,建立数据互信与出海保障机制

建立支撑词元出海的跨境数据合规与互信机制,推动词元高效、安全、合规出海。针对海外交互数据难以回流、飞轮效应断裂的突出问题,需通过机制创新打通数据跨境流通堵点,实现数据安全与价值挖掘的双向兼顾。一是探索“数据主权对齐”的跨境数据信托机制,在国家数据局指导下,与欧盟、东南亚等重点市场建立双边或多边数据信托框架,由第三方托管海外用户数据,经隐私计算处理后回传统计特征或梯度信息,实现“数据可用不可见”。二是设立“词元出海数据回流合规试点”,在海南自贸港、上海临港等先行区开展试点,给予合规企业评估数据出境安全的“绿色通道”或“白名单”,试点成熟后形成可复制的合规指引。三是推动建立国际互认的“AI数据护照”标准,联合国贸易伙伴制定元数据标准与溯源体系,通过国际组织或双边协议实现数据护照跨境互认,为数据合法回流筑牢技术基础。

6. 品牌标准引领,塑造中国词元国际竞争新优势

实施数据驱动的品牌建设与国际标准引领工程,为破解中国词元“低价替代品”定位构建国际信任背书提供有力支撑。针对中国词元缺乏品质信号与信任背书的突出问题,需通过多维举措传递品质价值,提升国际话语权。一是发布“中国词元质量认证”体系及公开数据仪表盘,由国家认监委指导、行业协会牵头,建立

涵盖性能、安全、隐私等维度的质量认证标准,认证企业使用“中国词元质量”专属标识;强制企业公开延迟、吞吐量等关键指标,以透明化数据披露建立海外用户信任。二是建设“中国词元出海标杆案例数据库”,收集高价值应用案例,形成多语言、可量化案例库,通过国家渠道推介,同时鼓励国际审计机构核验,增强案例可信度。三是主导或深度参与国际标准制定,依托我国词元调用量优势,在ISO、IEEE等组织发起标准提案,将技术优势转化为标准话语权。四是建立“数据-词元”跨境信任联盟,联合“一带一路”沿线及金砖国家,制定共同隐私准则与质量互认协议,简化成员间的数据回流审批、共享算力资源,从而降低中国词元海外信任门槛。

参考文献:

- [1] 刘烈宏. 深化数据赋能人工智能发展,加快培育智能经济新形态——在中国发展高层论坛 2026 年年会上的演讲(3月23日)[EB/OL]. (2026-03-23)[2026-04-23]. https://www.nda.gov.cn/sjj/jgsz/jld/llh/llhldhd/0323/20260323202204680553721_mobile.html.
- [2] HUANG J. AI is a 5-Layer cake[EB/OL]. (2026-03-10)[2026-04-23]. <https://blogs.nvidia.com/blog/ai-5-layer-cake>.
- [3] KOPARKAR S. Rethinking AI TCO: why cost per Token is the only metric that matters[EB/OL]. (2026-04-15)[2026-04-23]. <https://blogs.nvidia.com/blog/lowest-token-cost-ai-factories>.
- [4] 王云杉. 我国日均词元调用量突破 140 万亿[N]. 人民日报,2026-03-24(08).
- [5] 罗云鹏. 中国 AI 大模型调用 Token 量何以连续霸榜[N]. 科技日报,2026-03-24(02).
- [6] “数据要素×”三年行动计划(2024—2026 年)[EB/OL]. (2023-12-31)[2026-04-23]. https://www.cac.gov.cn/2024-01/05/c_1706119078060945.htm.
- [7] 可信数据空间发展行动计划(2024—2028 年)[EB/OL]. (2024-11-21)[2026-04-23]. https://www.nda.gov.cn/sjj/zwgk/zcfb/1122/20241122164142182915964_pc.html.
- [8] 国家发展改革委等部门关于促进数据产业高质量发展的指导意见[EB/OL]. (2024-12-28)[2026-04-23]. <https://www.gov.cn/zhengce/>

zhengceku/202412/content_6995430. htm.

[9] 黄群慧. 词元经济的三问三答[N]. 人民政协报, 2026-04-10(11).

[10] 陈波. 一份关于词元经济研究的问题清单[N]. 中国经济时报,2026-04-21(A03).

[11] 李隼. 拥抱机遇、直面挑战,迎接 Token 新经济时代[N]. 人民邮电报,2026-04-17(01).

[12] 中国信通院. Token(词元)经济视角下的算力互联网价值体系再思考[EB/OL]. (2026-04-13) [2026-04-25]. <https://thinktank.miit.gov.cn/column-detail?articleId=69037&columnId=3>.

[13] 艾媒咨询. 2026 年中国词元经济产业链全景分析报告 [EB/OL]. (2026-04-07) [2026-04-25]. <https://www.iimedia.cn/c400/110545.html>.

[14] 孙庭阳,姚坤. 词元(Token)和词元经济[J]. 中国经济周刊,2026(6):108-109.

[15] 刘继为. 何为“词元经济”[N]. 学习时报,2026-04-22(02).

[16] 李方. 词元重构 AI 时代商业生态[N]. 经济日报,2026-04-20(11).

[17] 张鑫. 词元:新生产要素助力认知经济崛起[EB/OL]. (2026-04-07) [2026-04-25]. <https://www.yicai.com/news/103117264.html>.

[18] PEREZ C. Technological revolutions and financial capital: the dynamics of bubbles and golden ages [M]. Cheltenham; Edward Elgar Publishing, 2002.

[19] 郑纬民. 从模型服务走向词元服务,是智能体时代对基础设施的必然要求[EB/OL]. (2026-04-15) [2026-04-25]. <http://www.nbd.com.cn/articles/2026-04-15/4341338.html>.

[20] 陈波,元宝. 词元经济的崛起、挑战与治理策略[N]. 中国经济时报,2026-04-13(A01).

[21] 张骞. 用好我国的词元优势[N]. 经济日报,2026-04-22(05).

[22] DANTAS P V, CORDEIRO L C, JUNIOR W S S. A review of state-of-the-art techniques for large language model compression [J]. Complex & Intelligent Systems, 2025(11): 407.

[23] AHN J, SON Y B, KIM D, et al. Dynamic micro-batch and Token-budget scheduling for IoT-scale pipeline-parallel LLM inference[J]. Sensors, 2026, (26): 1101.

[24] XU P, ZHU X T, CLIFTON D A. Multimodal learning with transformers: a survey [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(10): 12113-12132.

[25] DON-YEHIYA S, BURTENSHAW B, ASTUDILLO R F, et al. The future of open human feedback[J]. Nature Machine Intelligence, 2025, 7: 825-835.

[26] 熊巧琴,汤珂. 数据要素的界权、交易和定价研究进展[J]. 经济学动态,2021(2):143-158.

[27] 谢丹夏,魏文石,李尧. 数据与算力经济:一个数字经济的一般均衡理论[J]. 管理世界,2026,42(3):37-56.

Token Economy: Connotation Characteristics, Basic Laws and Development Strategies /

OUYANG Rihui¹, HUANG Zhixuan²(1. China Center for Internet Economy Research, Central University of Finance and Economics; 2. School of Economics and Management, University of Chinese Academy of Sciences)

Abstract: Token, as the smallest semantic unit for information processing in generative artificial intelligence, is evolving from a technical concept into a core value carrier of the intelligent economy. Based on the techno-economic paradigm theory, this paper defines the conceptual connotation and techno-economic characteristics of the Token economy, and systematically reviews its development trajectory from germination, explosion, to diffusion. This paper finds that the orderly operation of the Token economy follows four fundamental laws: The super-growth law driven by decreasing marginal cost and price elasticity of demand. It forms a positive feedback loop of “cost reduction-price decline-demand surge-exponential growth in invocation volume-further cost amortization”, driving a super-exponential expansion of Token invocation. The super-linear returns law resulting from the overlapping effects of economies of scale and scope. Significant economies of scale arise from improved batch processing efficiency, optimized cluster scheduling, and fixed cost amortization, while economies of scope emerge from multimodal tasks sharing underlying computing resources and operation and maintenance systems. The deep coupling of the two makes the growth rate of returns far exceed that of factor inputs. The differentiation law of “Token efficiency” driven by long-tail applications. The user value created by a unit Token varies significantly across different scenarios, thereby promoting price stratification, the rise of prompt engineering, and a transformation of enterprise budget management models. The circular reinforcement law of the data-token-value cycle, forming a flywheel effect of “more invocations-more interaction data-better models-higher Token efficiency-further stimulation of invocations”. The four fundamental laws are interwoven, mutually reinforcing, and deeply coupled, jointly forming a complete dynamic system for the development of the Token economy. Currently, China has achieved a globally leading scale of Token invocation, but faces prominent problems such as low pricing, security, and compliance issues, particularly three major structural constraints: fragmented data backflow, infrastructure mismatch, and lack of brand recognition. To achieve healthy and sustainable development of the Token economy, it is necessary to take data elements as a core lever, build a robust data infrastructure support system, establish Token efficiency evaluation and data incentive mechanisms, foster a global ecosystem for computing power scheduling and cost optimization, establish a whole-chain governance system for the Token economy, build data mutual trust and overseas expansion guarantee mechanisms, and shape new international competitive advantages for Chinese Tokens, thereby promoting the Token economy’s transition from super-growth to high-quality development.

Key words: Token; Token economy; Token going global; data elements; cross-border data flows; intelligent economy