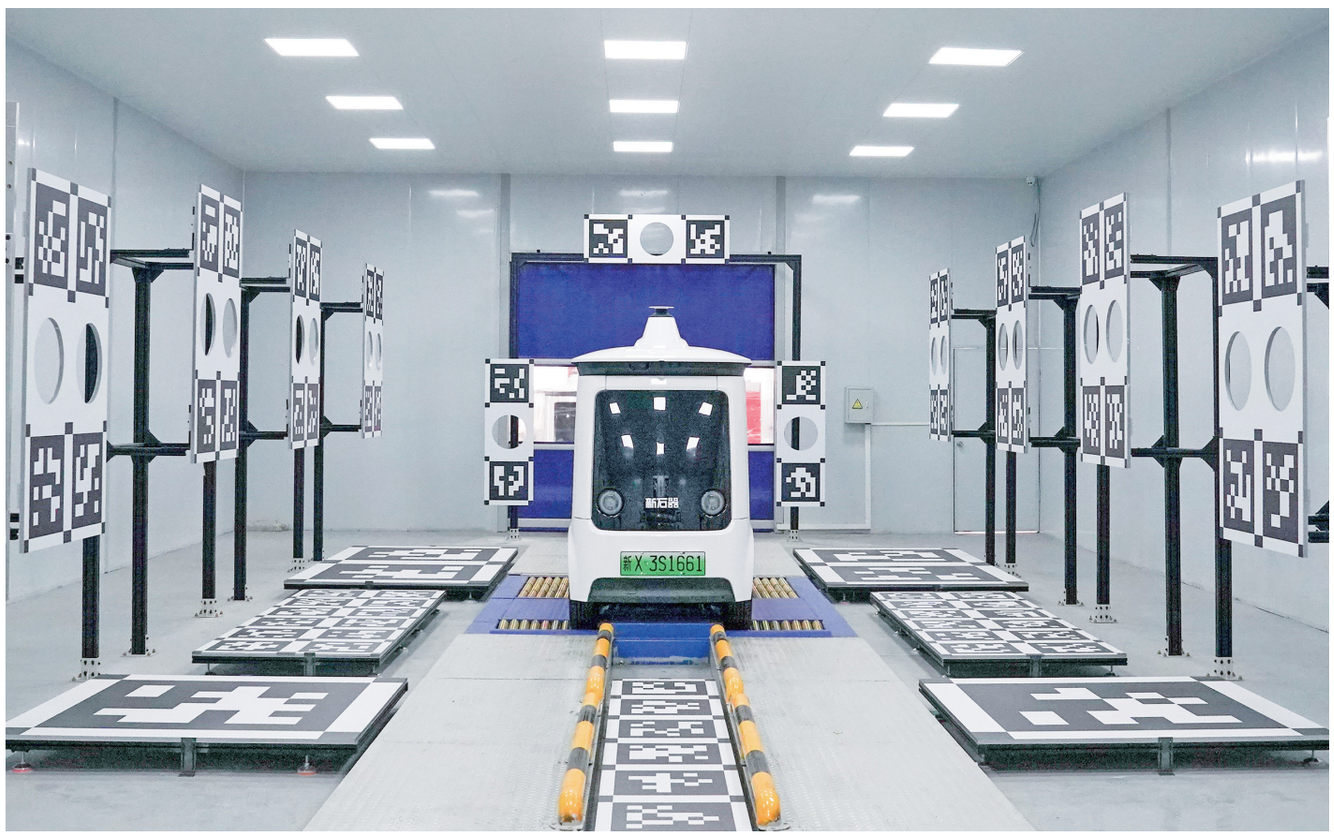
词元经济如何塑造智能经济新生态

□ 本报记者 杜 壮

近日,国家数据局正式将Token的中文译名定为“词元”。随着国家层面完成对“词元”(Token)的官方定名,一个以词元为核心的价值计算与流通体系正从理论走向产业化。这一基础概念的厘清将如何影响AI与数据要素市场?将面临哪些现实挑战?又将催生怎样的经济生态?



在新石器(盐城)智能制造有限公司无人车生产线,一辆无人车在进行调试。 (新华社/图)

调用量两年激增超千倍

词元(Token),是AI模型处理信息的最小单位,每一次提问、生成、交互与推理,都需要消耗词元。

日前,在中国发展高层论坛2026年年会上,国家数据局局长刘烈宏表示,Token“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。

国研新经济研究院创始院长、智能经济首席专家朱克力告诉本报记者,国家层面对“词元”完成官方定名,终结了过往概念表述杂乱、认知口径不一的局面,为相关理论研究与产业实践筑牢了基础支撑。此前,Token相关探索多停留在行业圈层,缺少统一规范,行业发展容易出现零散化、碎片化问题,甚至因概念模糊引发市场误解与合规隐患。定名之后,词元有了权威定义,产学研用、监管部门能够形成统一话语体系,为计量标准制定、市场交易流转、行业合规治理扫清认知障碍,也让行业发展有了明确的制度衔接基础。

随着“词元”定名在政策、标准、产业等多个层面形成共识,企业侧与地方实践也已快速跟进,涌现出一批标志性案例。

在城市平台层面,广州于2026年4月上线全国首个基于“词元”级调度的城市综合算力运行服务平台,以词元为统一计量基准,目前已接入14家单位并管理超过16000P的算力资源,将逐步全域纳管,规模超60000P。在地方服务层面,合肥、芜湖等地的算力公共服务平台也相继推出“词元服务”,面向开发者提供一站式、低门槛的大模型API调用能力,有效降低了中小企业的创新成本。此外,多个算力节点城市和新兴数据要素集聚区,也已将“词元”作为规划、招商与产业统计的核心指标之一,例如乌兰察布正依托“东数西算”战略机遇,加速打造“Token之都”。

国家数据局相关数据显示,2024年初,中国日均词元(Token)调用量为1000亿;至2025年底,跃升至100万亿;今年3月,已突破140万亿,两年增长超千倍。

在朱克力看来,叠加市场需求的快速增长,词元不再只是技术层面的计算单位,已经成为智能经济时代承载价值、核算消耗、链接供需的核心载体。这意味着Token经济学正式从小众理论探索,迈入产业化落地的全新阶段,不再局限于企业内部成本核算,而是成为数据要素市场化配置、AI产业提质增效的重要抓手。

“整体来看,官方定调叠加市场赋能,推动词元这一领域从无序生长转向规范发展,从单一技术应用转向体系化经济生态构建,为长期健康发展锚定了清晰方向,也为对接全球智能经济规则打下了坚实基础。”朱克力说。

产业链条清晰协同

随着国家层面完成“词元”定名,一个以这一基本单位为价值核心的新经济体系正在快速形成,引发产业界广泛讨论与深入实践。业内专家指出,“词元经济学”已超越单纯的技术概念,正演变为一套覆盖生产、流通、分配、消费与治理全链条的完整经济框架,其对应的产业链条也已清晰显现,展现出稳健的发展势头。

朱克力认为,词元经济学以词元为核心纽带,构建起覆盖生产、计量、定价、流转、分配、治理的完整经济体系,其内涵覆盖多个关键板块。既有词元标准化界定、精准计量规则、质量分级标准这类基础支撑内容,也包含生产成本核算、能效管控、产能调配等经营内核,同时涵盖市场化定价机制、跨平台供需交易模式,以及上下游算力方、模型方、应用方价值分配逻辑,还配套合规监管、知识产权界定、数据安全防控、风险预警等治理相关内容,形成全链条闭环发展框架。

实际上,这一涵盖全链条的经济学框架并非空中楼阁,它正迅速催生并塑造出一条清晰且协同的实体产业链。

朱克力分析,上游是词元生产底座,涵盖AI芯片、算力集群、供电散热、数据中心建设等基础设施,核心比拼生产能效与供给稳定性。中游承担词元调度、定价匹配与供需衔接,以大模型企业、算力服务平台、云服务商为主体,搭建价值流转的中间枢纽。下游覆盖千行百业应用场景,从日常智能办公、智能客服到工业智能制造、专业医疗、智慧城市等领域,是词元消耗与价值落地的关键环节。

“目前产业链整体协同性持续增强,发展重心逐步从单纯产能扩张,转向能效优化、场景深耕与商业落地,国产配套能力不断提升,产业生态愈发成熟。”朱克力说。

提升词元复用效率

词元调用量的本质是真实使用量,背后反映的是AI大模型在真实场景中的渗透深度、应用的频率和广度。调用量越高,意味着模型被用得越多,而支撑这一切的,是算力。

算力基础设施是支持词元生产的主要底座。今年3月,包括腾讯云、阿里云、百度智能云在内的多家头部云厂商,相继上调了AI算力及大模型相关产品的价格,最高涨幅超过460%。

3月,在英伟达GTC大会上,创始人兼CEO黄仁勋抛出一个重塑行业认知的新概念——Token工厂。在他看来,数据中心正在经历一场根本性的角色转变:不再是过去那个用于存储文件与数据的“电子仓库”,而是一座日夜运转的智能生产线。这座工厂输入的是电力和数据,产出的则是Token。但从商业角度看,目前有很多AI公司可能并不赚钱,陷入了“用户越多、成本越高”的怪圈。

“眼下行业热议‘词元工厂’理念,把数据中心视作词元生产车间,但现实经营中不少AI企业陷入用户越多、成本越高的困境。”朱克力表示,核心矛盾集中在三者适配失衡:词元生产依赖算力、电力、硬件等刚性投入,固定成本占比高,规模扩张很难摊薄单位成本,算力负载越高反而会推高能耗与运维成本;定价模式相对粗放,大多采用简单按量计费,没有结合使用场景、服务质量、响应等级做差异化划分,高价值场景难以形成合理溢价,无法覆盖高端算力投入成本;盈利模式过于单一,大多依靠售卖词元流量变现,缺少增值服务支撑,长期研发与基建投入和短期收益形成明显错配。

想要跑通词元相关商业模式,实现可持续发展,朱克力表示关键要做好三重优化。他认为,生产端聚焦能效提升,通过算法优化、智能算力调度、算力集群协同降低单位词元能耗与生产成本,逐步实现规模效应;定价端建立精细化分层体系,区分通用需求与高端专业需求,按服务时效、模型精度、场景专属化实现按质定价,化解价值与成本不匹配问题;经营端跳出单纯售卖词元的思维,延伸配套服务、行业解决方案、定制化模型优化等增值业态,提升词元复用效率,让规模增长逐步转化为经营效益,实现可持续发展。

优先筑牢发展根基

在词元经济产业链加速成型、商业实践遍地开花的同时,市场也出现了对概念炒作、标准不一、合规风险等问题的担忧。如何确保这一新兴经济形态在高速发展中不脱轨、不失序,实现健康可持续增长,成为业界关注的焦点。

“词元产业热度持续攀升,但概念炒作、合规隐患、行业乱象等问题相伴而生,想要行稳致远,必须优先筑牢发展根基。”在朱克力看来,首先要统一行业计量与标准体系,明确词元统计口径、质量分级、核算规则,杜绝市场中虚假计量、虚标词元价值等乱象,为交易流通、行业监管、成本核算提供统一依据。

同时,朱克力认为,要严守数据安全与权益合规底线,词元流转贯穿数据采集、模型训练、内容生成全流程,需要厘清数据使用边界、个人信息保护要求,做好内容溯源与版权管控,防范信息泄露、数据滥用、侵权盗版等风险。还要明晰产权与权责界定,理顺词元生成内容、训练数据的权益归属,明确平台、用户、合作方的权责划分,化解经营中的各类纠纷隐患。

此外,朱克力指出,要完善适配性监管治理体系,平衡创新发展与风险防控,坚决打击借词元概念进行的炒作、非法交易、非法集资等行为,同时适配产业发展特点建立包容审慎的柔性监管机制,避免过度监管抑制创新。兼顾产业商业化落地需求与长期合规发展底线,引导行业回归技术创新与价值创造本质,让词元经济在规范中创新、在发展中提质,真正赋能智能经济高质量发展。

记者手记

让AI从“技术特权”变为“普惠工具”

□ 杜 壮

当AI从实验室的炫技走向千行百业的赋能,一个根本性问题浮出水面:如何让它像水电一样可计量、可获取、可负担?答案正藏于两个看似技术性的概念之中——词元(Token)与算力超市。二者协同,正为AI大规模应用铺设底层基石,激活新质生产力的关键引擎。

长期以来,AI服务的商业化面临价值难以度量、资源难以触达的问题。企业调用AI,可能按次数、按时长或按模块付费,标准混乱、成本模糊。这就像用电没有“度”、用水没有“吨”,市场无法形成透明高效的交易。词元的实质,正是为AI的“智力产出”建立统一的度量单位。它将AI处理的信息,无论是文本、代码还是指令,都可以拆解为标准化的最小计价单元。如同电力以“度”、流量以“GB”计量,词元让AI服务的消耗与产出变得清晰可数。

然而,仅有尺子,没有便捷的商店,商品依然无法触达用户。这正是算力超市扮演的角色——它不仅是一个交易场所,更是一套将复杂算力转化为便捷、可靠服务的精细化运营体系。相关资料显示,算力超市核心运作机制可以概括为四个紧密衔接的关键环节。

第一,资源标准化上架。算力超市将异构的底层硬件性能,转化为按词元、算力或时长等标准计费的统一“商品”,按Token数、单精度浮点运算能力、核时计费的算力规格计算,消除技术门槛。

第二,智能搜索与比价。平台通过动态比价、智能推荐,帮助用户,尤其是中小企业,在众多选择中高效匹配成本与性能的最优解,有效破解了市场信息不对称,让决策回归效率与效益。

第三,开箱即用服务。平台预集成主流模型与工具环境,实现“下单即用”,降低开发与运维的启动门槛,让用户能将精力聚焦于业务创新本身,而非底层技术适配。

第四,普惠支付与结算。创新性地对接“算力券”等补贴政策,并提供预付、后付、包年包月等支付模式。平台同时承担服务质量监控与公正结算的职责,切实降低了企业的试错成本与资金压力,真正做到了“让企业敢用、好用”。

通过以上四个环节的系统性设计,算力超市得以将庞大、分散的算力资源整合为即取即用、按需付费的服务套餐,并通过智能调度,高效匹配供需。例如,将西部绿电产出的充裕算力,精准输送给东部密集的应用需求。这极大降低了使用门槛,使中小企业乃至个人开发者能以极低成本共享顶尖的AI能力,同时优化了全国范围的资源配置效率。

实际上,二者的关系是相互依存的。词元是算力超市的“交易货币”,使复杂的算力转化为易懂的“服务用量”。算力超市是词元流通的“核心平台”,其普及极大拓展了词元的应用场景与规模。

发展词元与算力超市,根本目的在于让AI从“技术特权”变为“普惠工具”。它解决的正是发展新质生产力如何普及与如何增效的核心问题。当智能的价值有了公认的尺度,当算力的获取变得像网购一样方便,创新的潜力将被极大释放。这背后,是一场深刻的生产关系调整,也就是从占有资源转向获取服务,从集中控制转向开放协同。

未来,词元与算力超市的成熟,将像电网与电表推动电气化一样,成为社会全面智能化的基石。这意味着中小微企业无需自建“发电厂”就能用上智能服务。一条更宽广、更包容的智能之路,正在我们面前铺开。

新闻链接

关于Token的知识清单

Token究竟是什么?

@知乎答主笙囧同学

你可以把词元(Token)理解成乐高积木。语言是用积木拼出来的东西。模型处理语言,不是看你拼好的成品,而是把它拆回一块一块的积木,然后处理这些积木。

积木的大小不是固定的。常见的组合,积木块大。生僻的组合,积木块小,甚至拆到一块砖头一块砖头的粒度。你说一句话,模型把它拆成积木进行处理,然后一块一块地把回答拼出来,再还原成你能看懂的文字。整个过程,积木就是Token。

我国日均Token消耗量已突破140万亿 这是一个什么量级的概念?

@知乎答主平凡

不同的AI任务在每次调用时消耗的Token数量差异显著,可以大致分为几个级别。微型任务(Micro)每次约消耗200个Token,典型应用包括意图识别、情感分析或关键信息抽取。小型任务(Small)约消耗750个Token,适用于一句话问答、短句翻译或标题改写。标准任务(Standard)消耗量在1500个Token左右,例如进行普通问答或段落润色。

对于更复杂的增强型任务(Enhanced),每次消耗约4000个Token,可用于撰写邮件或小方案、总结5页文档,或进行附带示例的代码讲解。处理长文档或会议纪要(Long-form/Meeting)则需约1.15万Token,能完成10~20页的综述或整理1小时的会议内容。

而涉及超长上下文(Very Long Context)的任务消耗剧增至约8.5万Token,典型场景是检索增强生成(RAG)或对长代码库进行问答。最消耗资源的当属智能体工作流,每次调用可高达60万Token,用于驱动多步骤工具链或执行“写作—评审—改写”这类复杂循环任务。

不同的用法消耗的Token数量天差地别,很有可能10%的人消耗了90%的Token,特别是智能体类型的工作,因为它的每次消耗大多是以完成任务为主要目的,所以用起Token会非常快。随着大模型的发展,Token价格会越来越便宜。

Token和我们普通人有什么关系?

@《中国战略新兴产业》杂志记者 卜文娟

Token就像我们生活里的“智能电表”,它默默计量着我们使用AI的每一分“脑力”。每次你让AI帮忙写总结、做翻译、甚至规划旅行,都在消耗Token,这正让它像水和电一样,成为我们触手可及的基础资源。正因为有了这个清晰的计量单位,企业才能推出灵活的服务,让AI从昂贵的技术变成人人用得起的工具。更深刻的是,Token消耗的暴增意味着AI正从新奇玩具变成每个人的“能力放大器”——它让我们用更少的时间消化信息、用更低的门槛完成专业任务,就像给普通人的双手和大脑装上了高效引擎。所以,Token不只是一个技术参数,它实实在在地记录并推动着每个人更高效、更富有创造力的新生活。

(本报记者杜壮编辑整理,设计崔一)

智源FlagOS完成DeepSeekV4八款芯片适配

本报讯 记者王健生报道 近日,DeepSeek发布DeepSeek-V4-Pro 1.6T旗舰模型(1.86万亿参数)及DeepSeek-V4-Flash 284B高效模型(2840亿)。由智源研究院牵头研发的众智FlagOS第一时间对两个“巨无霸”模型进行全量适配,已经完成DeepSeek-V4-Flash在8款以上AI芯片上的全量适配与推理部署,包括海光、沐曦、华为昇腾、摩尔线程(FP8)、昆仑芯、平头哥真武、天数、英伟达(FP8)等芯片。FlagOS同时正在推进DeepSeek-V4-Pro模型在多个芯片的迁移适配,后续即将开源。

DeepSeek-V4-Flash是深度求索推出的V4系列两大模型之一,采用混合专家(MoE)架构,总参数量284B,激活参数仅13B,支持100万Token上下文长度。

围绕DeepSeek-V4-Flash多芯适配,此次FlagOS系统软件技术栈突破了三大关键技术:FlagGems全算子替代(实现多芯片统一适配)、为o-group采用独立张量并行策略解锁更多低显存场景,以及“FP4+FP8混合精度”的原生权重到FP8/BF16的精度路径转换。这三项关键技术,使DeepSeekV4能够在当前各种厂商的主流AI芯片上稳定运行,而非仅限于支持FP4和大显存的少数高端AI加速卡。