

AI时代的Token经济:商业价值、企业责任与治理挑战(笔谈)

张省,赵占恒,杨若愚,樊浩,郑立

编者按:Token(词元)是大模型处理信息的最小单元,具备可计量、可定价、可交易三大核心属性,是AI服务计费、结算的基础单位。以Token为核心计价单位的新型数字经济体系,正深刻改变人工智能产业的商业模式、竞争格局和资源配置方式。本期策划的笔谈“AI时代的token经济:商业价值、企业责任与治理挑战”,围绕Token的定价、核算、评估、配置与治理展开讨论,旨在构建一套较为成熟的智能经济学框架,推动人工智能从实验室能力走向商业世界中的稳定生产力。五篇稿件从不同角度阐释人工智能核心资源——Token的生产、分配和消费的经济原理,论证其商业模式的可行性和参与治理的复杂性。《AI时代的“新石油”:Token的价值究竟从何而来?》一文,从信息熵、算力价值、预测能力三个维度揭示了人工智能时代Token的价值来源。《算力公司数字责任的生成逻辑:Token的视角》则从Token的主体——算力公司的角度,指出其应履行更好的数字责任,包括价值计量责任中的透明化与公允性、绿色低碳与价值共创责任的内在关联、智能经济红利的公平普惠、构筑安全合规责任与全链路风险防控体系。《智能经济的“新度量衡”:Token(词元)的治理需求识别与规则体系构建》在论证中国在全球数字治理领域角色全面升级的基础上,系统梳理了智能经济下Token的治理需求,并据此搭建了适配性的Token治理规则体系。《无声计量:Token如何塑造智能社会的治理秩序?》一文认为Token正在成为智能社会当中一种隐性的治理规则,其通过成本约束影响治理精度的选择,通过容量限制影响信息传递的完整性,通过配额差异影响智能服务获取的公平性,这要求政策制定者要关注Token资源的分配公平。《人工智能时代的中美竞合态势与全球治理方略——基于算力与Token(词元)的分析》从全球视野分析了Token领域中美竞争与博弈的核心焦点,并衍生出一系列涉及国际安全的重要议题,提出构建公平合理、安全有序全球治理体系的建议。

关键词:Token经济;人工智能;算法;数字责任;智能经济

中图分类号:F49;D61 **文献标识码:**A **DOI:**10.12186/2026.05.001

文章编号:2096-9864(2026)05-0001-20

AI时代的“新石油”:Token的价值究竟从何而来?

张省

郑州轻工业大学 经济与管理学院,河南 郑州 450001

随着人工智能全面步入商业化落地阶段,Token已跨越单纯的技术要素范畴,跃升为连接模型底层能力、算力资源消耗和商业结算的核心技术中介。国家数据局将Token正式译为“词元”,这不仅是语言层面的规范命名,更是我国构建自主AI经济体系的战略举措。可见,Token的价值源于其作为AI时代基础计算单元的技术不可替代性,叠加其在商业化、低碳化、制度化和本土化进程中所承载的经济与战略意义。它不仅是基本的计费单位,更是重构智能经济底层逻辑的关键基础设施。

一、信息熵的结晶:Token的价值来源于人类文明信息经验的高维压缩

如果Token仅仅是一个技术切分单位,或者一串高维空间中的数学向量,它还不值得被称作“新石油”。真正使Token拥有深层价值的,是它背后承载的东西——数千年来人类文明所积累的信息、经验和智慧。同理,一个Token所具有的价值,不只是来自训练时所消耗的电力和算力,更来自它所具有的隐性脑力劳动。例如,当模型开展对万有引力定律这个Token序列的读取时,它事实上是在获取牛顿当年闭关思考的成果;当它对相对论相关文本进行处理的时候,它是在间接利用爱因斯坦十年

心血构建起的理论框架。这些知识不是毫无根据产生的,而是人类最为杰出的头脑用一生或者几代人的时间换来的,每一个优质Token的背后,都有着一位或多位不露面的思想劳动者。

并非所有Token都“生而平等”。当前市场上亦存在着大量的垃圾Token,其所具有的信息熵很低,逻辑密度几乎为零,训练进去不仅无益,反而会造成模型的能力下降。反过来看,优质Token的信息密度高,逻辑也足够严谨,正好把一种有效的认知基础提供给大模型,让大模型能逐步学到推理、论证、批判,从而创造性生成一些更高级的能力。

科技巨头肯花大价钱来获取这些数据,甚至肯付出高额的版权费用,这绝对不是碰巧的事情。拿OpenAI训练GPT-4用到的语料库来举例,仅版权授权支出就达到数千万美元,主要原因有以下几个方面:一是这些平台产出的文本被Token化处理之后,可维持住很高的信息熵与逻辑密度;二是它们拥有稀缺的认知价值以及训练价值。大模型训练过程可类比于系统化人才培养,训练数据的质量与知识结构,直接决定了模型最终的认知水平与能力边界。

Token的第一重价值是历史所赋予的,也是人类智慧在数字世界里的一个投影。每一个优质Token都能看成一个微型的知识载体,它

收稿日期:2026-04-20

基金项目:河南省高校哲学社会科学应用研究重大项目(2026-YYZD-11)

作者简介:张省(1981—),男,河南省桐柏县人,郑州轻工业大学教授,博士,主要研究方向:数字治理。

能把学者的理论思辨、工程师的技术调试、创作者的思想沉淀这类个体化的认知劳动,被封装成一种可计算、可调用、可重组的标准化数字单元。当企业为 Token 付费的时候,表面上是在支付算力所产生的成本,实际上则是在租赁人类几千年来所积累下来的那份认知劳动成果,这也是 Token 被冠上“新石油”这个名称背后的深层原因。石油是生物质经过亿万年地质作用的压缩才形成的,而 Token 则是千百年来人类智能用一种数字形式做的压缩,二者都属于时间沉淀下来的稀缺性资源,只是在演化周期和存在形态上有些差别。正因为这样,强化知识产权保护、做好数据溯源、构建公平有序的 Token 经济体系,就不只是法律或伦理问题,更是 AI 产业能否持续拿到高质量原料的战略性问题。

二、算力即价值:Token 的价值来源于支撑其生成的 AI 基础设施的成本映射与商业化度量

Token 已成为现代 AI 经济中衡量算力消耗和商业规模的核心计量单位,并被各大模型服务商统一用于定价。一个看似没什么重量的 Token,背后却有一整套重工业量级的工程在作支撑。厂商为了能产生 Token,得预先支出数亿美元去搭建大规模的 GPU 集群,拿 GPT-4 来说,它在训练阶段投入了大约 25 000 张 A100、H100 这类高端算力芯片,并让这些芯片连续不间断地跑上好几个月,光是电力成本就达到数千万美元。到了模型推理的场景,用户每一次发出的交互请求,都得调度大规模的算力集群去做协同运算,并在毫秒级的时间之内完成数万亿次的浮点计算;而 GPU 在高速运行的过程当中会产生大量的热量,需要足够庞大

的冷却系统把这些热量给带走,大型数据中心一年消耗的水量跟一座小型城镇的用水量差不多。这就是“算力即电力”这句话所代表的真实含义,也就是说,每一个 Token 的生成,都会燃烧掉一些实打实的能源。

Token 作为度量衡的经济学逻辑,也就能变得清楚起来了:算力与电力本身所包含的价值不太容易被直接感知、度量或者交易,而 Token 正好就发挥了像“金本位”时期黄金那样的作用,它能把摸不着的那些计算资源变成一种可以被准确切分、标价和售卖的标准化商品。企业购买 Token 的行为,从本质上看就是在购买 GPU 集群的运算时间、电厂发出来的电的规模、制冷设备的制冷规模,还有背后大量工程师所做的那些维护工作的作业量。Token 的价值构成有着很明确的成本导向性,其中涵盖 GPU 设备的折旧、电力的消耗、水资源的消耗、工程师的薪酬等多个方面,每一项的价值构成都对应着实实在在的产业成本投入。

从更宏观的视角去看,Token 的定价机制揭示了 AI 基础设施的商业化逻辑。早期互联网是用字节来做计量单位的,但字节只是表示信息量,跟计算复杂度没有联系。后来云计算按 CPU 小时和 API 调用次数来收费,但是这些计费单元要么比较粗放,要么跟业务价值有偏差。而 Token 同时体现计算量以及业务价值,其价格由算力成本、算法效率以及市场供需决定。当前,Token 的单价正呈现出下降的趋势,推理专用硬件的迭代和模型架构的优化不断压低边际成本,这个过程会让 Token 逐步从“奢侈品”转变成“日用品”,这就有点像电力从爱迪生时代的稀缺资源,逐步演变成当今社会生产生活中随手就能拿到的公共基础资源。

然而,低价的 Token 并不总是好事。当有

的厂家以远低于成本的价格卖出 Token 时,其背后或是为了抢占市场做战略性补贴,或是利用了低精度算力或者降级模型。真正可长久维持的 Token 定价,应覆盖物理世界的实际损耗,包括电费、水费、硬件折旧损耗、人力成本等。企业客户去选择供应商时,不能只着眼于单价,还要关注算力质量以及基础设施的可靠性,因为算力不足或者模型发生降级所产生的“幻觉 Token”,不但没有价值,反而可能造成不可挽回的损失。例如,合同审查场景中出现的错误 Token,可能导致企业面临巨额经济赔偿。Token 的第二重价值是把数十亿美元的芯片资金投入、数亿千瓦时的电力消耗、数千名工程师的日夜苦干浓缩起来,归结为每百万 Token 几美元的价格标记,这是 AI 时代最具客观性与诚实性的商业度量语言。

三、超越统计的涌现:Token 的价值来源于预测能力跃迁后的涌现效应

真正让 Token 被视为“新石油”的原因,在于它的第三个也是最厉害的一个特性,那就是 Token 既是一种能促使智慧涌现出来的活跃催化剂,也是一座迈向通用人工智能的桥梁。要想把这一重价值给理解清楚,就得回到大语言模型的那个基本运行原理上去,也就是预测下一个词元。当模型看到“今天天气很”这几个字的时候,它会先把所有可能出现的 Token 的概率分布给计算出来,再从里面选出最合适的一个,比如“热”或者“冷”,然后把选出来的这个词加到上文里头去,之后再继续去预测下一个词,就这样反复做下去,最终就能生成一个完整的回复。在整个生成的过程当中,模型手里并没有一个预先规划好的宏大蓝图,它只是在每一个局部的步骤里面,做一次看起来有点像

机械操作的文字接龙而已。

不过,正是这种看起来挺简单的局部计算,在跨过了规模上的那个临界点之后,会产生一种质的飞跃,也就是人工智能领域里常说的那种涌现效应。当模型的参数量只有几百万的时候,它只能去模仿一些简单的短语搭配,输出的内容会带有逻辑碎片化的问题,语义连贯性也不太好。而当模型的参数量增加到百亿甚至千亿的规模,训练数据也跟着增加到了数万亿 Token 的时候,模型就开始学会去做类比推理了,比如拿“苹果是水果,那么菠菜是什么”这个问题去问它,它能正确回答出“蔬菜”,它甚至能学会根据自然语言描述去编程并写出可以运行的代码。这些能力并没有被人做过显式的编程,也没有在训练数据里以一模一样的形式出现过,它们是从海量 Token 的预测实践当中自发地生成、并靠着内生演化一步步长出来的。

OpenAI 首席科学家 Ilya Sutskever 有一句著名论断:预测即压缩,压缩即理解。这句话揭示了 Token 预测和智能之间的深层联系。为了能把众多的人类文本给压缩起来,模型会被迫去学习数据背后的那些规律和结构,并由此逐步推动理解能力的涌现。从这里能看出,Token 预测可以说是解锁人工智能高阶智能的一个核心关键。思维链技术的出现,则对 Token 的生产力价值作了进一步的说明,它的核心思想是与其让模型直接输出最终的答案,不如先引导模型把中间的推理步骤给输出出来,然后再去推导出结论。举个例子,碰到一道复杂的数学题时,直接让模型输出答案的话,准确率也许只有 30%;但如果让模型先把推理过程输出出来,准确率就会大幅上涨到 80%以上。这就说明每一个处在中间环节的推理 Token,都在为最终的那个正确答案做着铺垫。所以说,Token

不只是一种用来承载输出结果的东西,它还承担了模型在做逻辑推演以及自主思考时的那个运行角色。

这一发现具有深远的生产力含义。在传统软件开发的过程当中,程序员得一行一行地把逻辑和推理写成代码;而到了大模型时代,用户只要把问题给表述出来,模型就会利用生成的推理 Token 去自动搭建解决问题的路线,并能直接替代掉很多传统的脑力活儿。从法律合同条款的审查、医疗影像报告的判读,再到金融研报的自动产出,每一个场景的背后,都有成千上万个推理 Token 在无声地做着工作。再进一步去看,Token 经由涌现效应所释放出来的那种价值,还会带有一种更宏观的战略指向。也就是说,它是一条通往通用人工智能(AGI)的关键纽带。在目前最前沿的那些研究探索当中,拿 OpenAI 的 o1 系列模型举例来说,这个模型的创新之处主要就体现在它会在推理阶段为“思考”这个过程分配更多的 Token,模型会被允许去生成大量的内部推理 Token,并借此进行多次的校验与回溯,从而修正思路,然后再给出最终的答案。可见,随着计算量的增加,Token 的作用从单纯记录转向参与模型的推理与复杂问题的解决。

四、总结与展望

Token 是 AI 时代的基础单位,也是一座能把人类智慧、工业算力还有智能涌现给连接起来的桥梁。它把文明所积累的信息转化成可计算单元,同时把算力、能源、资本做统一的计价,把物理世界的付出变为可流通的经济资产。借助预测以及涌现机制,它还可让模型从被动模

仿转为主动思考,产生训练数据没涵盖的新能力。但 Token 化也有代价:文本的机械拆分,会破坏语言的韵律、情感和文化方面的语境,尤其是低资源的相关语言;高维映射容易造成数据偏见的放大;大量的算力消耗会造成能源与环境的负担;Token 只是统计方面的单位,模型在逻辑推理以及因果理解上的能力仍有限度,所谓的“智能涌现”一般是偶然的模式,并非真正的推理。即便如此,Token 还是为语言数字化打下了基础,释放了 AI 的能力,推动知识流通和价值实现。从更大视野看,它不仅是 AI 时代的“新石油”,更是数字经济和认知变革的核心动力,其战略意义远超表面价值。

Token 作为 AI 时代其中的一个核心要素,虽说在当下可展现出一定价值,但也面临着诸多挑战。只有妥善应对这些挑战,才有助于它获得进一步的发展。未来 AI 产业以及经济方面的发展,势必会围绕怎样高效获取 Token、怎样优化使用 Token、怎样创造性产生 Token 这几方面展开。一方面,要提升获取效率,探索更合理的切分与编码方式,减少信息损失;另一方面,要优化使用方式,改进算法与模型架构,降低能耗,提高可靠性。同时还要推动创造性生成 Token 的研究,为 AI 开辟更多可能。同时,也要正视 Token 化过程中造成的损耗、偏见和高能耗等弊端。一是要加大对训练数据审核和管理的力度,以便降低数据偏差给模型造成的影响;二是要加强对 AI 领域绿色能源应用的研究,降低环境所产生的成本。唯如此,Token 的价值和潜力才能于智能持续涌现的进程中获得更健康也更持久的增长,并推动 AI 技术朝着更成熟、更可持续的方向发展。

算力公司数字责任的生成逻辑:Token 的视角

赵占恒

郑州轻工业大学 经济与管理学院,河南 郑州 450001

2026 年 3 月,国家数据局公布,我国日均 Token(词元)调用量突破 140 万亿,相比 2024 年初增长超千倍。Token 调用量的爆发式增长正在改变数字经济的底层运行规则,它不再只是算法背后的技术指标,而是贯穿算力生产、模型服务和商业结算的主要核算单位。对算力公司来说,Token 的核心作用是把设备损耗、电力成本和人工研发投入整合为可以定价、交易的价值标准,让算力产能第一次实现可量化、可交易,算力公司也因此从普通的技术服务商变成了智能经济秩序的塑造者。随着 Token 经济的快速发展,定价标准混乱、能源消耗转嫁、低质量 Token 泛滥及数据安全存在漏洞等问题不断显现,影响了市场公平和行业信任。作为掌握 Token 资源定价、分配与质量管控权限的主体,算力企业的数字责任是一个随着 AI 大模型和云计算普及而迅速凸显的议题,已超越普通商业经营范畴,深入到技术伦理、数据主权、能源可持续性核心生存问题。

一、定价权的归属:价值计量责任中的透明化与公允性

算力公司的第一重数字责任,源于 Token 的定价权应当如何配置。Token 定价机制形成在极为复杂的成本体系之上,这个体系对终端用户来说非常不透明。如果定价权力得不到有效的监督,信息不对称就会引发经济利益大受损失,而且会滋生价格偏见和产品质量含混不清

的情况,极大地破坏 Token 市场整体的信任基础。价值计量责任包含相互依存的两个方面:其一,采用透明度举措,保证消费者可以全面知晓有关费用的一切情况,捍卫消费者的知情权;其二,按照公正的原则,使得定价机制不偏袒某些利益集团,进而做到市场竞争有序,协同共进。

透明责任源于信息分配不均衡,Token 价格形成受硬件投资规模与能耗量的明显影响。例如,GPT-4(生成式预训练 Transformer 模型第四代)开发期间,训练阶段用到约 25 000 个高性能处理器,耗时数月,伴随而来的还有巨额电费,合计达数千万美元,整个研发项目总投资估计在 1 亿~2 亿美元之间^[1]。Token 整体运作成本包含芯片磨损、电力供应、水资源消耗,以及工程师和标注人员薪资等各种支出。服务提供商掌握着详细的成本信息,但是用户却不知道单次调用具体消耗了多少资源,这就造成了一种单向的信息不对称状况。从执行困难方面来看,透明化遭遇的主要障碍表现在算力公司没有自发公布信息的积极性上,行业内还没有形成一种被广泛接受的披露标准和审查体系,这又增加了透明化进程在实际操作中的难度。

信息不对称造成定价权力过度集中。控制核心计算能力基础设施的主要平台,凭借对芯片架构设计、模型训练效果以及分发网络的完全掌握,事实上握有 Token 定价机制的重要权力。通过细致划分定价层级,这些平台可以针对具备不同支付能力的用户制定专属的价格策

收稿日期:2026-04-23

基金项目:河南省软科学研究项目(252400410125)

作者简介:赵占恒(1982—),男,河南省汝州市人,郑州轻工业大学副教授,博士,主要研究方向:公司治理。

略,进而最大程度地获取高于一般水平的利润。更为严峻的难题在于算力质量存在隐蔽性特征,部分算力公司以远低于成本的价格出售Token时,消费者很难区分这种低价到底是由于算力效能大幅增强所致,还是靠牺牲算力精准度甚至模型改良幅度而达成的一种虚假现象。

Token定价机制既要守住透明度和公平性这些核心原则,还要融入一套清晰划分责任的体系。算力公司需清楚地公开自己的定价依据,细致讲解不同种类Token的定价规则、各种模型版本之间的差异,以及额外收取费用的地方,这样才能保证全过程非常透明。Token兑换比率的调节,也要有严格的程序设置,防止服务商擅自改变比率从而侵害用户权益。对已购买Token的定价保持不变,仅对新购入的Token进行调整。充分披露透明度并形成公正定价机制,这属于价值量化责任的根基,一旦缺失便会严重妨碍后续有效的监督及必要调整地开展。

二、双重约束:绿色低碳与价值共创责任的内在关联

算力公司的第二重数字责任,源于Token生产的可持续性,这重责任由绿色低碳与价值共创两个维度构成:前者要求降低Token生产的环境代价,后者要求维护高质量Token的生态供给。该责任的履行难度在于,能耗约束和质量治理都会增加短期成本,而算力公司倾向于优先把规模做大,这种短期盈利与长期可持续发展之间的冲突仅靠企业自觉是难以克服的。

绿色低碳责任源于Token的生产与能源消耗紧密相关。Token的生成机制就是把能源转变成计算能力,每次执行模型训练都会产生很大的电能消耗。算力企业为了稳固市场份额和维持自身业务增长,会持续扩容算力基础设施,使得技术迭代产生的单位能效提升能够被不断增长的算力需求抵消,最终推高行业整体的能源消耗规模。黄仁勋在GTC 2026(英伟达GPU技术大会2026)提出的Token工厂经济学,将

数据中心类比为生产Token的工厂,以每瓦性能作为核心竞争力标尺,揭示了Token工厂与工业时代的工厂一样都在消耗实在的能源,排放实在的碳。算力企业若想应对成本难题,便迫切要把能源效率纳入自身的责任范畴之内,大力推动可再生能源的应用,还要展开全方位的碳足迹考量,覆盖Token的全部生产阶段,以此精准计算出每单位算力产生的碳排放量,从而达到资源利用最大化、环境影响最小化的目标。

价值共创的责任源自Token品质的异质性以及优质Token的稀有性特点。在社交媒体平台上流传的低质量内容以及毫无意义的生成文本,它们的信息熵比较小,无法很好地推动模型性能得到改善。为了保证数据品质,科技企业一般会花大量版权费来取得优质数据集。OpenAI是美国处于领先地位的人工智能研究组织,在训练GPT-4模型时所支付的版权费用很高,这么做就是为了保证所使用的素材具有很高的逻辑连贯性和认知价值。如果训练数据包含大量低质量的标记单元,那么大型模型就必然掉进“差的输入带来差的输出”这个恶性循环当中。随着训练数据量的快速增长,很多研究机构早已预见,到2028年,全球优质的文本资源也许会出现极度匮乏的情况。在此情形之下,算力公司的身份正在慢慢改变,不但要努力营造Token的良好生态体系,促使高质量语料库得以流传和分享,而且还要优化数据训练、选取的过程,这样才能保证所用的数据集既准确又有效。

绿色低碳责任重点在于批判那种漠视环境成本的规模化扩张模式,即过度追求产量增长而忽视能源消耗的真实代价。相反,价值共创责任更多着眼于信息层面的问题,也就是批判那些在创建大规模数据集时没有有效地控制低质量Token的大量涌现,进而造成信息价值被淡化和浪费的行为。这两种批评表明要达成可持续发展,算力公司的规模扩张策略既要提高能源利用率,也要提高信息质量。这不是要阻

止增长,而是要提倡一种反省性的扩张方式,就是要将反思性思维融入 Token 生产过程中,使其目标从仅仅谋求数量增多转向兼顾质量和效率。这有助于实现更高效、更可持续的技术提升和资源调配。

三、从垄断到普惠:公平普惠的社会责任

算力公司的第三重数字责任,源于 Token 经济向垄断收敛的趋势与中小主体对算力可及性的需求之间的张力。这重责任的核心问题是 Token 经济创造的红利能不能超越少数头部企业的边界,被更广泛的社会群体所共享。履行这重责任的难度在于 Token 生产的规模报酬递增效应本身是市场机制的结果,算力公司从中获利属于正常的商业行为。然而,要求企业在追求效率的同时兼顾公平,仅靠市场自发调节难以实现。

Token 生产的规模报酬递增效应使市场结构天然向寡头垄断收敛。数据中心规模越大、芯片数量越多,单位 Token 的平均成本就越低。这种成本递减规律源于三个层面的叠加:硬件研发成本在大规模产量上的摊销、大型数据中心在能源利用效率上的优势,以及海量用户调用产生的反馈数据对模型的持续优化。Token 生产具有突出的自然垄断特征,一些大型算力公司在算力服务市场上占据优势,它们的平均成本远小于分散竞争情况下的水平。实际上,已有的垄断形式有着明显的层级化组织框架,拿美国 GPU(图形处理器)芯片设计行业的龙头企业英伟达来说,它凭借自身更新的 GPU 架构及其相关 CUDA(统一计算设备架构)生态系统,有效地引领了芯片技术的发展走向,并使其不断向应用场景拓展。而且,在算力租赁市场,主要平台凭借先进的调度算法,做到了 70%~80% 的资源利用率,这个成绩远远高于业内一般只能达到的 40%~50% 的利用率水准。于是,Token 经济领域的市场结构表现出高度稳

定性,即由少数寡头把控的独特态势。如果这种垄断状况得不到约束,很可能引发部分优势企业掌控 Token 的供应和定价权,致使整个 Token 经济体系的开放性和包容性明显下降。

如果这种垄断局面不断向下延伸,将极大地限制中小企业接触 Token 机制的可能性。很多人工智能初创企业偏好租赁计算能力而不是直接购买硬件芯片,这样做可以减少创建自身系统时的高额初期投入以及不断更新技术带来的风险。算力租赁模式给用户带来很大的灵活性和可扩展性,但是它的经济成本很高。例如,租用一块英伟达 H100 显卡,每月要花 55 000~66 000 元人民币,如果需要的数量是 64~128 张,那么一年的租赁费用就可能达到 42 000 000~100 000 000 元人民币^[2]。因此,初创企业承担着较高的投入成本,其主要目标就是有效地应对技术飞速更新换代的需求,并保障稳定的现金流,从而保持业务长远的竞争力和可持续性。如果算力公司没有履行好公平普惠的责任,Token 资产便会一直朝着大企业聚集,最后造成 Token 经济体系变成被少数几个主体控制的封闭市场。

算力公司应从三个方面积极履行公平普惠的责任义务:其一,采用算力池化策略,此机制可达成资源的动态分配与共享,而且给出灵活又价廉的套餐选择,以此适应不同规模企业的计算需求;其二,创建“东数西算”的国家级算力基础设施,战略性地改善 Token 产生的地理分布状况;其三,全面利用西部地区丰富而又成本较低的绿色能源资源,把它们变成具有很高市场价值的 Token,以大幅缩减区域之间的数字差距。据统计,中国电力成本比美国低约 50%,西部地区的成本优势更加明显^[3]。2025 年,全国电力总需求量增至超 10 万亿千瓦时^[4],这就为电力与算力的高效协同奠定了坚实基础,也使反垄断政策的执行变得关键。公平普惠并非慈善之举,而是算力公司在其 Token 经济体系中应担负的关键职责,目的在于防止此领

域形成被极少数领头者控制的市场态势。

四、底线的构筑:安全合规责任与全链路风险防控

算力公司的第四重数字责任,源于Token全链路中潜藏的两重风险——数据风险与内容风险。这重责任要回应的是算力公司如何确保Token在生成、传输与消费的全过程中不侵害用户权益和不威胁公共安全。

Token生产流程附带的用户数据隐患构成了AI产业典型的数据风险。用户在使用AI服务过程中录入的提示词中包含的个人身份资料、商业机密及各类敏感数据会在模型服务端进行不规范的存储与调用操作,从而引发隐私侵权问题。2026年3月,由于安全方面的担忧,OpenClaw(开源智能体)被官方告知不得部署在政府机构和企业的设备上,这表明令牌传递系统本身就存在一定的安全风险。如果数据风险得不到有效管控,那么用户隐私的泄露和信任的流失将直接动摇整个Token经济的基础。

内容风险重点在于考量Token产生阶段的准确度与可控度两个方面。大模型生成Token的时候,实际上是凭借概率预测来判定后续序列元素出现的概率大小,并非依靠对具体事实的认知或者逻辑推理,一旦遭遇算力不足或者模型性能下降造成生成虚假Token的情况,在诸如合同审查、医疗判断这些非常关键的应用场景当中,这种失误可能带来的后果十分严重,超出普通商业纠纷的层级。

算力公司应加强安全保障举措,这样才能较好地承担起安全合规的相关责任。其数据经营策略应严格执行最小必要性原则,从而系统引导Token生成流程中涉及的数据采集、处理以及保存步骤,并采用数据加密手段,形成详细的访问权限控制架构。同时,运用差分隐私法来全面保障用户的个人信息安全。要想让用户数据在模型训练期间既得到有效利用又合乎伦理标准,就要创建专门的授权系统,清楚地划定

数据使用的范围及其前提,以此做到全方位捍卫和保护用户隐私。

五、总结与展望

数字责任得到落实时,算力公司不再只是消极的数据处理者,还是积极投身内容保护、安全捍卫以及伦理准则的创立者与执行者。算力公司履行数字责任可通过两条战略途径:一是加强责任内化机制,采用透明定价法,加强能耗监管,巩固质量保障体系并推动安全合规运作,全方位融入自身日常经营过程;二是推动制度向外拓展,对Token价值考量体系中的定价模型革新、数据所有权划分、能耗信息公示以及算法审查等重要方面,形成完备的规则架构,使技术发展和社会责任相互协调一致。算力公司若把自己的责任、承诺变成可量化、可追踪、可经营的制度措施,并依托Token机制来实施,就会从数字经济发展的基础设施——如水、电、气般不可或缺——升级为引领数字文明创建的关键力量,发挥类似“碳基土壤”的作用,从而推动整个数字生态系统兴盛起来。

参考文献:

- [1] Ludvigsen K G A. The carbon footprint of GPT-4 [EB/OL]. (2023-07-18) [2026-04-10]. <https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae/>.
- [2] 胡琳. H100 租金半年涨三成,算力租赁概念持续走强 [EB/OL]. (2026-04-22) [2026-04-23]. <https://www.sfccn.com/2026/4-22/2NMDE1MjBfMjEzMzQ2NQ.html>.
- [3] 李雪萍. 美媒:AI竞赛下半场拼电力,中国恰恰拥有决定性优势 [EB/OL]. (2025-11-04) [2026-04-10]. <https://hkcn.hk/docDetail.jsp?id=101154459&channel=4657>.
- [4] 2025 能源成绩单 用电量数据反映我国产业布局调整趋势 [EB/OL]. (2026-03-06) [2026-04-10]. <https://cec.org.cn/detail/index.html?3-354491>.

智能经济的“新度量衡”： Token(词元)的治理需求识别与规则体系构建

杨若愚

中国民航大学 经济与管理学院,天津 300300

当前,人工智能(Artificial Intelligence, AI)技术持续迭代升级,不断重塑社会经济原有结构,催生全新生产要素与价值创造方式。步入智能经济阶段,Token(词元)已经被定位为大模型处理信息的“新度量衡”,深度参与数字底层秩序搭建,在全球数字治理当中发挥关键作用^[1]。2026年3月,中国国家数据局局长刘烈宏指出,词元不仅是智能时代的价值锚点,更是联结技术供给与商业需求的计量尺度^[2]。随着AI服务计费方式的不断调整,其逐步脱离传统授权与账号收费模式,转为依照词元消耗核算成本。这意味着Token逐步具备基础资源计量属性,对标水电、网络流量等通用统计口径,其产业价值与经济属性持续释放。与之相伴,治理层面的挑战逐步显现。2026年3月,世界数据组织在北京正式成立,这是全球首个国际性

数据治理专业组织,彰显了我国对全球数字经济发展的重视。从数据大国到治理强国的关键跨越,表明了我国将在全球数字治理领域发挥重要作用。基于这一背景,本研究拟系统梳理智能经济下Token的利益相关者及其治理需求,搭建适配性的Token治理规则体系,以期为相关治理工作提供理论与实践支撑。

一、Token的利益相关者

Token治理涉及多元主体,各方的角色定位、利益诉求与责任边界各不相同。从治理视角出发,可将其归为监管供给层、技术供给层、应用需求层与社会监督层四大类别(见表1)。

首先,从监管供给角度来看,政府监管部门承担着核心管理职能,主要是制定法律法规与监管政策,维持市场秩序,依托政策法规引导技

表1 Token治理的利益相关者分析

类别	主体	核心角色	主要利益或关注点
监管供给层	政府监管部门	制定法律法规与监管政策,保障监管标准化	数据安全与隐私、市场秩序、国家安全
	行业协会或组织	制定行业标准,倡导自律	技术规范、行业声誉、标准竞争力
技术供给层	AI开发者	提供模型与算力服务	创新动力、知识产权保护、商业利润
	服务集成商	提供API、开发套件、定制化解决方案	服务质量、客户满意度、运营效率
	数据提供者	提供训练和应用所需数据	数据权益、隐私授权、收益分配
应用需求层	计算资源提供者	提供算力、存储、电力等基础资源	利润、可持续发展、负责任使用资源
	企业用户	使用AI服务提升生产力	成本效益、安全合规、数据控制
	个人消费者	作为智能产品和服务的最终用户	隐私保护、安全可信、合理定价
社会监督层	公众与社会组织	社会公共利益监督者	公平性、透明度、社会伦理

收稿日期:2026-03-24

基金项目:2025年天津市哲学社会科学规划智库项目(TJYT25-07)

作者简介:杨若愚(1993—),女,河南省周口市人,中国民航大学副教授,博士,主要研究方向:科技创新与人工智能。

术良性发展,守住公共利益与国家安全底线。在国内治理格局中,国家数据局、网信办、工业和信息化部多部门联动协作,共同参与AI与数据领域治理工作,搭建多层级的跨部门监管框架;行业协会及标准化组织可补足监管短板,依靠自律约束与技术规范完善行业生态,稳步提升整体口碑与标准竞争力。其次,在技术供给环节,Token的技术产出与功能供给,主要依靠开发团队和运营平台。技术安全与内容品质属于首要考量因素,商业化推进过程中的现实难题同样无法忽视。AI开发者和服务集成商等主体持续输出开发工具与定制化服务内容。各方需要合理调配现有资源,调整定价方式,规划基础设施与算力排布。此外,数据提供者与计算资源提供者,主要负责训练数据、标注服务以及提供算力、存储、电力等基础资源,是Token生态中的基础要素支撑方。Token的产生高度依赖于数据输入与算力消耗,因此这两类主体的利益诉求,如数据权益、隐私授权、收益分配,以及利润回报、可持续发展与资源负责任使用,构成治理设计的关键考量。再次,聚焦终端使用场景,不同群体构成Token服务的主要受众。企业用户侧重综合效益、合规标准和数据掌控范围;个人消费者更重视隐私安全、服务体验与收费标准。使用者对Token的应用诉求也会影响整体治理规则的优化方向。最后,公众与社会组织作为公共利益监督者,构成Token治理的社会监督层级。他们的核心关切集中于治理过程的公平性、透明度与社会伦理合规性,并通过舆论监督与公共参与对治理措施施加影响。

可见,不同利益相关者在关注点与治理需求维度上存在显著差异,政府掌握规则制定与执行权,而行业与企业拥有技术话语权,公众与社会组织则主要通过舆论与社会选择施加间接

影响。所以深入理解各方立场与互动关系,是制定包容、有效且可执行治理制度的必要前提。

二、Token 面临的风险

各类参与主体共同组成完整的Token生态,他们普遍追求技术升级带来的商业收益。角色分工的天然区别,会让各方暴露在不对等的风险环境里。本研究将围绕安全防护、金融稳定、伦理规范及环境这四类关键风险,分层展开具体探讨。

其一,信息安全风险是Token生态面临的最为直接的技术性威胁。Token常充当身份凭证与安全核验工具,其传输和储存环节一旦出现泄露或篡改,会引发账号盗用、隐私外泄等问题,严重时还会造成系统入侵。不法人员可能会利用脚本攻击截取有效Token,顺势盗用用户身份权限。其篡改权限相关字段后,还能避开常规认证流程,从而私自盗取平台内部敏感数据。这类隐患不只损害个人与企业的信息资产,还会顺着技术链路不断扩散,形成大范围安全隐患,进一步影响整体网络环境稳定。

其二,金融秩序风险源于Token概念的泛化与错用。Token本质上只是统计算力消耗的技术载体,本身具备量化统计与流转交易的属性。这一特点容易被刻意包装成带有升值空间的金融投资品类。多地监管部门虽早已警示相关乱象,但各类假借Token投资制造的诈骗行为仍持续不断。部分不法人员虚假宣传、刻意混淆AI词元和数字货币的概念并大肆炒作,干扰金融市场原有运行节奏,损害普通投资者切身权益。由于现阶段Token交易缺少完善监管约束,人为操控价格、恶意操盘等违规行为很难得到及时约束与管控。

其三,社会伦理风险涉及Token经济在竞

争公平性与数字包容性方面的深层挑战。如果Token分配规则缺少公开透明的约束,大型平台就会依靠技术优势与规则主导权筑起市场壁垒。这将导致中小创新企业准入空间被压缩,进而加剧行业垄断,扩大数字差距。在监管约束缺位的环境里,各类违规刷量与算力造假行为不断发生,直接冲击正常市场竞争秩序。放眼全球,Token交易串联着跨境数据及跨境金融,其外部风险具有扩散性和现实危害性。

其四,资源环境外部性是Token经济扩张中不可回避的物理约束。算力、电力是Token产出的核心消耗条件。频繁调用相关功能,会让数据中心保持高负荷运转,形成高额能源损耗。缺少绿色算力政策约束导致的Token经济扩张,会加重碳排放,进而加大环境承载压力。Token批量产出还要依托海量数据支撑,若肆意抬高采集、存储相关成本,就会产生隐私泄露与数据安全隐患。因此,有必要将上述外部环境因素纳入治理框架,通过推广绿色计算、强化数据治理等制度手段实现外部成本的内部化。

Token生态的风险体系呈现出技术、经济、社会与环境的三维嵌套特征^[3]。信息安全风险构成最底层的技术底座,金融秩序风险与社会伦理风险反映市场运行与制度设计层面的张力,而资源环境外部性则将治理视野延伸至Token经济的物理边界与代际公平。四类风险彼此关联,依托数据、算力、资本等多种核心要素形成联动传导关系。金融投机容易造成算力资源无序消耗,加重碳排放带来的环境压力。信息安全漏洞若被用于市场操控,会同时破坏金融秩序,削弱公众信任基础。因此,面向Token生态的治理设计必须具备系统性视角,既要分类施策以精准回应各类风险的特殊性,又需统筹兼顾以化解风险之间的关联放大效应。

三、Token的治理规则体系

本研究将Token生态中的多元利益主体、多维风险类型纳入统一的治理需求框架,这些治理需求并非孤立存在,而是通过利益博弈、风险传导与资源约束等机制相互关联。因此,相应的规则体系设计必须超越单一维度的修补式调整,转向多层次、全周期的系统性构建,具体应包括治理原则、制度设计、合规机制、激励与制裁以及跨域协调五个层面。

1. 治理原则是规则体系的价值基石,为制度设计提供方向性指引

结合现实风险与多方主体诉求,Token的规范治理应遵循四项核心准则,分别是安全韧性、公平正义、可持续发展、人本理念。治理全过程应优先保障生态信息安全与网络运行稳定,提前规避技术漏洞、恶意攻击诱发的各类系统风险,把安全防护放在首要位置;平衡各方主体发展机会,杜绝头部平台借标准优势划定行业壁垒,约束资本投机与人为控盘行为,维护良性竞争环境;正视能源与环境层面的衍生问题,普及低碳算力应用,维持Token经济平稳运转,杜绝短期逐利造成的过度排放与资源消耗;重视普通用户的隐私与数据防护,让技术落地始终贴合公共利益,切实发挥民生价值。

2. 制度设计可围绕法规约束、技术标准、行业自律三条路径推进

守住治理底线、划定运行规则、灵活补足监管空白,Token治理的制度设计应围绕这三方面内容展开,并将其整合以形成完整的运行体系。法律条文是Token治理的核心依托,依托现有《生成式人工智能服务管理暂行办法》《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》等法律法规规章,清晰界定

Token的法律属性与监管边界,补充专项法规约束Token生产、计量、交易全流程,为实际治理提供稳定法律依据。各类技术标准保障制度应落地执行,统一Token格式、算力核算与安全要求,让产业各环节依照统一准则运行。行业自律用于缓和恶性竞争,改善整体经营环境,稳步提升服务水平。除此之外,还应结合政府指引、行业统筹、企业落实与群众监督、多方协同的共治模式,适配当下需求,也有利于完善整体治理格局。

3. 合规机制是连接制度设计与治理实效的关键环节,涵盖企业内部自查与外部穿透式监管两个维度

企业需结合各类Token应用场景,搭建常态化风险评估流程,优化数据管理与权限管控模式,参考大模型备案管理思路,定期上报调用规模与安全事件。外部监管端应强化全链路动态监测,覆盖Token发行流转的全部环节;依靠专项检查与第三方审计,定向整治投机炒作等违规行为,压缩灰色操作空间;参考金融行业风控思路,采用反洗钱核查逻辑,追踪排查异常资金流向,切断灰色资金流通路径。同时,应考虑行业发展压力,采用差异化认证管理:正向鼓励合规经营项目,给予准入及政策优待;对不合规经营主体区分问题程度,落实分层处罚与法律约束。

4. 规则体系的有效运行不仅需要强制性约束,还需要合理的激励政策与清晰的责任追究机制

政府相关部门应灵活运用各类政策手段,平衡行业合规发展速度与技术创新节奏。对于合规服务质量达标的企业,政府应给予资金扶

持、税收利好与准入简化等激励;对于低碳生产模式的落地项目,政府也应依托专项补助与绿色金融政策,使其获得实际帮扶。对于各类金融投机行为,可按情节轻重划分监管处罚与刑事追责层级;对盗用数据、诈骗洗钱等违规主体,同步落实行业禁入、信用惩戒等多项管控措施。另外,应主动对接境外治理规则,联手压制跨境类Token违法活动。总之,应以奖惩并行的管理模式,推动整个Token产业生态走向规范透明,以维持长期稳定发展。

5. Token治理具有显著的多领域交叉属性,涉及人工智能、网络安全、金融监管、能源环境等多个领域,单一部门的治理能力难以覆盖其全部需求

由于Token面临的各类风险相互串联渗透,突破行业与领域边界,分散的监管模式存在天然短板,容易出现监管空白和处置延迟。对内,我国亟待搭建稳定的跨部门协作体系,以顶层统筹方式建立常态化沟通机制;统筹多类职能部门职能,协同履职,打通信息壁垒,合理完成风险研判与突发情况处置。对外,应保持开放协作姿态,深度参与全球治理交流;主动对接国际通用标准,统一Token计量与安全规范,减少地域监管差异造成的经营负担;压缩跨境灰色操作空间,逐步达成全球统一治理共识。

参考文献:

- [1] 王佳宜,吴勇.智能体Token(词元)的法律定性与规范治理[J].重庆邮电大学学报(社会科学版),2026,38(4):108-118.
- [2] 孙庭阳,姚坤.词元(Token)和词元经济[J].中国经济周刊,2026(6):108-109.
- [3] 薛澜.AI治理,需要多元工具协同应用[J].中国经济周刊,2025(16):101-103.

无声计量:Token 如何塑造智能社会的治理秩序?

樊浩

河南师范大学 政治与公共管理学院,河南 新乡 453007

Token 是大语言模型在处理文本时的最小计量单位。在技术层面,一个常见的英文单词需要消耗 1~2 个 Token,而单个汉字通常需要消耗 1.5~2.5 个 Token。在经济层面,它是大语言模型服务定价的基本单元。而在应用层面,它又决定了我们每次和 AI 对话要花多少成本、能说多少内容。在 2026 年初,主流大语言模型的上下文窗口已扩展到百万 Token 级别,单次交互成本持续下降。但是,Token 的底层逻辑从未发生改变,每次交互都需要消耗它,窗口总是存在硬性上限,免费服务与付费服务之间仍然存在差异。

这些技术设定看上去是中立的,却在公共部门与公众日益依赖 AI 的宏观背景下,产生了超出技术范畴的制度性后果。本文认为:Token 正在成为智能社会当中一种隐性的治理规则,它不像法律条文那样明确了权利和义务,也不像行政命令那样让人必须服从,而是通过成本约束、容量限制与配额分配三个维度,隐蔽地限定谁能用得起智能服务、谁的信息能被完整传递、谁又能够平等参与数字化治理进程。

需要指出的是,将 Token 纳入治理分析的视野并非牵强附会。政治学长期关注的核心问题之一就是稀缺资源的分配逻辑如何塑造权力结构与行为模式。例如,土地、财政资金、行政编制和信息通道等资源的计量与分配方式,无一不深刻影响治理秩序。而 Token 在智能社会

中的地位,类似于工业社会中的能源与货币,是驱动智能服务的一项基础资源,它的计量与分配不可避免地产生分配效应与秩序后果。

当然,Token 在运用过程中产生的治理效应并不是显性的。它和传统治理变量的关键区别在于技术的隐蔽性,Token 的切分规则由算法自动执行,上下文窗口容量是由模型架构预先设定的,配额消耗与计费,都在后台悄然进行。用户在使用 AI 服务时,既看不到 Token 的切割边界,也无法感知消耗配额的实时进度,只有在额度耗尽或者对话被截断时才会体会技术的刚性约束。这种“无声”的特征使 Token 产生的治理效应极易被忽视,它不像一项新政策那样引起公众热议,也不像一次机构改革那样进入公众视野,却在每一次人机交互过程中,从治理成本、信息完整性和参与门槛这三个维度影响着公共治理的微观秩序。

一、Token 与治理成本:算力约束下的资源配置

任何一项治理活动都有成本。传统公共治理成本主要有人力、时间和财政投入。而到了智能社会,Token 正成为一项新增的、不可忽视的重要成本。这种转变随即产生了两个问题:一是 Token 成本影响了治理精度的选择;二是 Token 费用分担塑造政府部门的行为选择。

首先看治理精度选择问题。假设一个地方

收稿日期:2026-04-22

基金项目:国家资助博士后研究人员计划项目(GZC20252487);中国博士后科学基金会第 78 批面上资助项目(2025M783784)

作者简介:樊浩(1994—),男,安徽省铜陵市人,河南师范大学讲师,博士,硕士生导师,主要研究方向:智能政务理论与实践。

政府用大模型来分析市民政务服务热线产生的海量诉求文本,以实现自动分类、情感识别与优先级排序的目标。这在技术上是完全可行的,但是 Token 消耗构成现实约束,每条诉求均需输入模型,每次分析都需消耗 Token,而部门 Token 预算却是有上限的。于是,治理主体需在精度与成本间进行审慎权衡。可能因 Token 不足而无法实现全量数据处理,抽样分析成为比较现实的次优选择;长文本可能因超文窗口被截断,后半部分信息被忽略;精细多轮交互分析因成本高也会让位于粗略的单轮分类。

在技术视域下,这些权衡是效率优化,而从治理角度来看,则是治理中的信息损失和精度下降,换言之,Token 成本转化为治理效能的损失函数。传统治理也面临资源约束,如人力不足可加班、借调或外包,财政紧张可调剂或追加预算等。而 Token 约束是硬性技术上限,上下文窗口不可逾越,额度用尽服务就会中断。这种刚性约束要求治理主体具备新型应对能力,如精准判断信息优先级、动态监控 Token 消耗、实时调整对话策略。这些能力目前尚未纳入公共部门的能力建设当中,但随着 AI 应用深化,它会越来越重要。

其次是费用分担所产生的激励问题。消耗 Token 过程中产生的费用如何分担,是典型的治理难题。费用分担目前主要有三种模式:第一种是由信息化主管部门集中购买大模型 API 额度,让各部门按需调用;第二种是部门自筹,由各部门从自身预算列支 Token 费用;第三种是项目外包,将 Token 成本嵌入信息系统开发合同,由承建企业来承担。上述三种费用分担模式在激励结构上存在着显著的差异:在统一采购模式下,政府部门无需承担直接的成本压力,但可能过度使用或浪费,产生“公地悲剧”;在部门自筹模式下,经费充裕的部门可充分使用,而经费紧张部门会因成本顾虑减少使用或选择低配服务,造成服务能力上的落差;在项目

外包模式下,Token 成本隐藏,透明度低,存在着监管困难的问题。

问题的实质在于,Token 是一种无形的有价资源,其预算编制、执行监控与绩效评估的制度框架还没建立起来。尽管公共财政管理已经形成了较为成熟的制度体系,但是 Token 资源分配还处于制度真空之中。当公共服务靠 Token 消耗来实现时,这种制度滞后必须得到重视。另外,Token 费用分担机制也涉及央地关系与跨区域协调。经济发达地区能承担高预算,提供优质服务;而欠发达地区因预算约束在智能化治理竞争中持续落后。因此,Token 成本可能加剧数智治理能力的区域不平衡,成为数智时代的新型财政转移支付难题。

二、Token 与信息完整性:容量限制下的传递损耗

公共治理有效性极度依赖信息的准确传递。无论是民众向政府表达诉求,还是政府部门之间的沟通,信息的完整性与准确度都会直接影响决策的质量和执行效果。而上下文窗口这一技术设定对信息传递施加了容量限制,产生的治理后果也值得认真对待。

在智能政务系统中,字数限制并不新鲜。政务网站的留言框一般都设有字数上限,公文格式也有篇幅限制,政策文件起草会考虑审阅者的时间和精力。这些限制主要是为了提高处理效率、降低信息过载风险。Token 窗口在形式上类似于这些字数限制,但是有两点关键的差异。

其一,Token 窗口存在硬性的容量上限。尽管一部分高级模型能够实现长记忆缓存,但在底层逻辑上,超长对话仍面临信息压缩与细节丢失风险。这意味着,较早输入的精确原文,一旦超出窗口,在当前轮次的注意力计算中就会被边缘化或近似重构,用户无法一次性完整地呈现所有信息,必须在交互中持续关注哪些

信息仍在窗口内、哪些已经面临丢失风险。

其二,Token 窗口对不同语言的容纳能力也存在差异。由于分词算法不同,同样的意思表达,中文所需的 Token 数量通常多于英文。也就是说,在相同上下文窗口下,中文承载的有效内容相对会更少。例如,一位市民向政务 AI 咨询复杂事项时,他可能需要先陈述自己的具体身份,再说清楚具体问题。如果 Token 窗口不能容纳完整陈述,市民就需要在“说清背景”与“说清问题”之间作出取舍。无论选择哪种,输入系统的信息都将是不完整的。

Token 的容量限制还可能带来一种系统性的信息压缩倾向。当用户意识到 Token 有限时,会自发精简表达,删去细节信息,保留核心内容。这种行为在个体层面是效率优化,但在系统层面可能导致重要信息的结构性缺失。治理实践也表明,关键信息往往隐藏在细节当中,如果信息不能被完整传递,那么决策者看到的将是一个被简化甚至扭曲的问题。在传统治理中,科层制就具有信息逐级衰减特征,每一层级都会对信息进行筛选、归纳与重新表述,这一过程既能够大幅提升效率,同时也内含着信息失真的风险。

因此,Token 窗口可能形成一种新的衰减机制,与科层组织既有的信息衰减相叠加。更值得关注的是,这种衰减具有不同于组织衰减的特征,它不涉及人的主观判断与取舍,纯粹由技术容量阈值触发,因此更加机械和不可协商。当某个信息片段因超出窗口而被删减时,没有任何人为此承担责任,也没有任何申诉或补救渠道。与科层制中靠人判断形成的过滤不同,这种“非人化”的信息过滤机制,既可能减少主观偏见,也可能丧失对信息重要性的灵活判断。这个问题在中文语境下尤为突出。当全球主流大模型的窗口设计以英文为基准时,中文用户事实上承担了一种“容量税”,为表达相同内容会消耗更多 Token,或在相同 Token 配额下传递

更少的信息。这不仅是技术上的差异,更是一种治理层面的非对称约束,不经意间削弱非英语用户在智能治理基础设施中的表达效能。

三、Token 与参与门槛:配额差异下的服务分层

公共治理具有内在的公平价值倾向,这要求公共服务对全体民众平等开放。而 Token 配额的存在,无论是免费版还是付费版,亦或不同用户之间的额度分配,都可能在不经意间形成新的参与门槛与智能服务分层。

当前大模型服务普遍使用分层定价的策略。免费用户获得有限的 Token 配额、小上下文窗口和低服务优先级,付费用户则享有更大容量、长窗口和更快的响应速度。因为算力资源需成本支撑,这种差异化定价策略符合商业逻辑,能够为不同支付能力的人提供多样化选择。但是,当 AI 服务嵌入公共治理场景时,商业定价分层可能转化为公共服务的使用能力分层。例如,地方政务服务平台的基础服务调用免费版接口,但是 Token 配额有限;增值服务或企业办事通道调用付费版接口,Token 却配额充裕。这就会导致有支付能力者可享受更完整的智能服务,而支付能力有限群体只能获得有限服务,人们的诉求表达就受到了 Token 制约。尽管民众还可以通过其他渠道表达诉求,但是,当 AI 成为主流甚至唯一高效通道时,Token 配额差异便构成了服务获取的门槛,这类似于数字鸿沟。从治理公平看,这种分层又极为隐蔽,用户不清楚 Token 配额及与付费版的具体差异,难以在实际使用中判断服务是否发生降级,信息不透明使公众难以就 Token 的配额公平性形成有效利益表达与政策诉求。

就算 Token 配额是形式平等的,如每人每天分配相同的免费额度,实际服务质量也会因为使用能力不同而产生极大分化。那些善于设计提

示词、熟悉上下文窗口特性、了解分词原理的用户能用更小额度获取精准回答、避免信息截断以及采用更高效表达方式,这些能力可以称为“Token 素养”,即在 Token 额度限制下,实现信息获取效率的最大化。“Token 素养”与社会经济地位、受教育程度、数字技能水平存在较强的关联性,导致即便 Token 配额平等,实际服务获取质量仍会因使用能力而分化。高 Token 素养者能在有限配额内完成复杂的信息交互,低 Token 素养者可能只是简单查询一下就用完了额度。这种基于能力的服务分层更为隐蔽,难以用政策矫正,最终导致作为技术计量单位的 Token 成了加剧社会不平等的微观生产机制。这要求政策制定者不仅要关注 Token 资源的分配公平,更要推进 Token 使用能力的大众化培育和普及。

四、结语:对“无声计量”保持制度主义警觉

Token 不是政治学概念,也不在公共治理的分析框架之内。但是,当政府和民众之间的交互越来越多地由 Token 作为中介时,必须关

注这一技术计量的治理后果。这些后果总体上表现为 Token 通过成本约束影响治理精度的选择,通过容量限制影响信息传递的完整性,通过配额差异影响智能服务获取的公平性。这些影响是 Token 作为一种计量标准的客观技术属性在公共治理场景中产生的制度性后果。

社会科学研究者关注 Token,不在于改进分词算法,而是建立一种制度主义警觉,对任何一种计量标准背后的利益分配与影响效应保持敏感。货币如何计量价值、选票如何计量偏好、指标又如何计量绩效,这些都是政治学领域的经典议题。Token 如何计量信息交互,也应当被纳入同一学术脉络之中。面向未来,Token 治理效应将随着 AI 技术的普及而持续放大。如何设计更公平的 Token 分配机制? 如何避免 Token 容量限制成为信息传递的新瓶颈? 如何缩小 Token 素养差异带来的服务获取分化? 解决这些问题不仅需要技术进步,更需要治理智慧。如何让这种计量服务于而非侵蚀公共治理的价值,需要我们对这种“无声计量”进行理性认知和主动回应。

人工智能时代的中美竞合态势与全球治理方略 ——基于算力与 Token(词元)的分析

郑立

云南大学 政府管理学院,云南 昆明 650091

人类社会已迈入数智时代,人工智能技术作为第四次工业革命的标志,成为重塑全球科技格局与国际安全的重要力量。算力与 Token(词元)是人工智能运行的物理根基与信息处

理基础单位。通过串联起技术、数据与市场,两者成为数字时代新的关键资本,是企业逐利和国家竞争的核心对象。从目前全球 AI 技术的发展态势来看,中美两国是全球 AI 领域的两大

收稿日期:2026-04-26

基金项目:国家社科基金项目(24BGJ018)

作者简介:郑立(1984—),男,陕西省西安市人,云南大学副教授,博士,博士生导师,主要研究方向:国际政治与计算政治学。

关键领导力量。两国在算力与 Token 领域的竞争与合作,不仅深刻影响着两国自身的科技发展,更深刻形塑国际政治格局的走向与人类命运共同体的构建。

究其本质,中美在算力与 Token 领域的竞争是“技术-市场”主导权与规则制定权的复合博弈。在算力领域,美国凭借长期的技术积累、全球人才吸引力与相关产业优势,已经占据相当稳固的领先地位。从高端算力芯片方面来看,英伟达(GPU)和谷歌(TPU ASIC)的产品合计占据全球数据中心 AI 加速器市场规模的八成以上。英伟达目前依托 4 nm 先进制程,在单卡算力、显存容量与互联带宽上领跑。在软件生态方面,CUDA(Computer Unified Device Architecture)平台以及 PyTorch/TensorFlow 框架形成了强大的技术壁垒,支撑着超大规模算力集群的高效运行,使得美国在前沿大模型训练、AGI(通用人工智能)研发方面保持两年左右的领先优势。更为关键的是,美国通过商务部工业与安全局(BIS)的芯片出口管制、实体清单等手段,将算力技术作为地缘竞争泛安全化的工具。这种“小院高墙”式的精准封锁虽无法从根本上遏制中国人工智能技术的自主发展,但无疑在一定程度上拖延了中国获取最尖端算力的技术迭代速度^[1]。

与美国依赖少数企业主导的技术突破路径不同,我国在算力领域走的是国家统筹、规模化部署与区域协同的发展道路。首先,依托华为昇腾、寒武纪等国内企业的技术突破,我国已构建起自主研发主导的算力队伍,尤其在推理算力等领域正在加速追赶,且现有技术能够满足国内绝大多数 AI 应用场景的需求。此外,随着“东数西算”与“算电协同”工程的深入推进,算力资源与西部绿电资源的高效匹配将大幅降低 Token 处理的能耗与成本,我国单 Token 推理成本维持在美国的 1/5~1/10。低成本为大范围

和规模化应用提供了比较优势。同时,我国庞大的互联网用户基数与全产业链应用场景,使得人工智能的落地能力与 Token 调用总量位居全球前列。随着人工智能在政务、工业、教育、医疗等众多领域的深度嵌入与应用,算力在我国的应用价值会得到全面提升。

在 Token 领域,中美竞争与博弈的核心主要聚焦于质量、成本与全球市场覆盖三个方面。美国依托 GPT-4、Claude 3.5 等高端闭源模型,在 Token 质量上占据明显优势。其长上下文处理能力、多模态融合水平与逻辑推理能力,能够较好地满足高端企业和尖端科研领域的复杂需求。同时,美国厂商掌握着 Token 的定价主导权,通过较高的 API 标准锁定高价值客户,并结合成熟的 API 生态与数据闭环,形成一个用户越多、数据越大、模型越强进而吸引越多用户(消耗更多 Token)的正向反馈循环。这种闭源的 AI 生态实际上为其构筑起坚固的技术与市场壁垒。相比之下,我国 Token 产业的优势集中在低成本与规模化上。根据全球最大 AI 模型 API 聚合平台 OpenRouter 的最新数据,2026 年 3 月 16—22 日,全球 AI 大模型总 Token 调用量为 20.4 万亿,仅中国就达 7.359 万亿,占全球的三成以上。

我国日均 Token 调用量已突破万亿级,呈现爆发式增长态势,能够有效覆盖中小企业、个人用户与新兴市场^[2]。此外,我国的多语言 Token 优化、垂直行业场景适配能力也在不断提升。然而,当前推动 Token 服务出海和挑战美国的定价垄断地位仍面临下述多方面制约。其一,我国在 Token 计量与计费上缺乏统一标准,导致产业链各环节成本核算口径不一。其二,我国 Token 出海遭遇美国的市场屏蔽与其他西方国家的严苛合规限制。其三,我国的高端 Token 不足,在多模态融合、长上下文推理等方面与美国闭源模型存在一定差距,难以满足

科研与金融等高端市场需求。此外,特朗普政府的人工智能战略将持续加大我国 AI 发展与出海风险。特朗普上任首日即撤销拜登时期的行政令,并通过 2025 年的新行政令和《美国 AI 行动计划》重新界定 AI 治理的合法性基础。即以维护美国技术主导地位为核心,减少联邦监管约束,以盟友协调和出口管制取代普遍性的多边机制,并借助国际标准制定平台对冲中国影响。随后,美国退出 UNESCO、拒绝签署巴黎 AI 峰会相关声明,进一步表明其以主权优先和市场竞争为导向的 AI 治理战略^[3]。即便面临美国技术右翼的巨大转向,我国人工智能仍然不断取得重大突破。当前,DeepSeek V4 的发布将极大地纾解我国在高端 Token 不足方面的困局。更为重要的是,V4 的发布将为国产大模型和 AI 产业提供一条不完全依赖英伟达 CUDA 体系的自主发展道路,对美国的 AI 生态壁垒形成冲击。

在当前局势下,算力与 Token 领域的竞争确实是中美之间竞争的主线,不过在不涉及国家安全的非敏感领域,两国的合作空间仍然宽广。中美在人工智能领域的合作既是两国产业均衡发展的现实需求,也是推动全球 AI 技术普惠发展的责任担当。在开源生态领域,中美企业与科研机构在开源模型、基础算法、推理优化等方面的共享合作,既能够降低全球 AI 技术的研发门槛、推动技术快速迭代,又有助于 AI 技术的普惠包容发展。像国内的 Qwen、DeepSeek 与美国的 Llama 系列,彼此之间有借鉴也有协同,共同拉低了全球 AI 研发的门槛。中小企业甚至其他全球南方国家的科研团队不用从头开始学习 AI 技术,直接借鉴和学习这些开源成果便可进行迭代。在标准共建领域,中美作为全球 AI 产业的两大核心力量,如能致力于算力能效、Token 计量与数据安全等国际标准的协调制定,能够极大避免全球 AI 治理体系的碎片

化。此举同时也能为形成统一和规范的全球 AI 产业秩序奠定重要基础。此外,在不涉及国家安全的非敏感科研领域,比如两国在气候、医疗和基础科学等领域关于算力与 Token 的联合研发活动,能够通过整合双方资源破解全球性难题。具体而言,我国在算力基建、电力资源、中低端芯片与服务器方面具有比较优势。而美方在高端芯片设计、软件工具与前沿研发方向上引领潮流。从合作潜力与效能来看,中美在算力与 Token 领域完全形成了短期虽然有限但是长远极为适配的互补双赢格局。

基于上述中美在算力与 Token 领域的竞争态势与合作潜力,连带衍生出一系列涉及国际安全的重要议题^[4]。在 Token 领域,安全风险主要集中在数据主权、技术滥用与伦理冲突三个层面。首先,Token 处理过程中伴随海量用户数据与敏感信息的跨境流动,美国的云法案(CLOUD Act)与我国的数据出境安全评估制度形成双向约束,使得数据跨境流动面临合规困境,也存在敏感数据被强制调取、泄露的风险,冲击全球数据安全治理体系^[5]。其次,大模型生成的 Token 可能被用于深度伪造、网络攻击等违法违规行为,甚至被应用于军事领域,推动 AI 军备竞赛。目前,中美在此方面尚缺乏统一的管控规则,技术滥用风险较高。此外,Token 训练数据中的文化、价值观差异,导致模型输出存在算法偏见,容易引发国际舆论摩擦,加剧中美及各国间的地缘分歧^[6]。

面对中美在算力与 Token 领域的竞争态势、合作空间及衍生的国际安全风险,亟须构建公平合理、安全有序的全球治理体系。从全球治理来看,中美两国作为全球 AI 领域的领导者,应承担起大国责任,摒弃零和博弈思维,推动构建良性互动的竞争合作关系^[7]。在数智时代,中美战略竞争的主战场不同于美苏冷战时期。国家资源竞争的重点从地理空间、自然资

源转移到人工智能及其相应技术的全方位扩展与迭代发展。目前,人工智能技术加速与市场、军事以及国家的国际地位深度绑定,大国战略竞争势必聚焦于人工智能技术。阎学通^[8]指出,由于数字经济领域的竞争具有垄断性和跨越性特性,中美竞争成败的决定性因素不再是意识形态与制度,而是国家与政府能否持续推进技术创新的领导力。基于上述讨论,作为世界大国的中美应建立算力与 Token 领域的安全对话机制,划定敏感领域红线,避免技术脱钩与全面对立;共同推动多边算力与 Token 治理规则的制定,强化国际合作,防范技术滥用与地缘风险。此外,中美应联合向发展中国家提供低成本的算力与 Token 服务,缩小全球数字鸿沟,推动 AI 技术普惠发展。同时,呼吁各国加强协同配合,构建多元共治的全球 AI 治理体系,兼顾各国的发展利益与安全需求,推动 AI 技术朝着有利于人类共同利益和构建人类命运共同体的方向发展^[9]。

参考文献:

- [1] 申青青,孙成昊.“硅幕”已至:美国对华人工智能出口管制与中国应对[J]. 国际经济合作, 2026,42(1):1-13,90.
- [2] 日均 Token 调用量达 140 万亿[EB/OL]. (2026-

03-26)[2026-04-23]. <https://www.news.cn/tech/20260326/8026bd54df3f489a8c79d443-8cac-96b3/c.html>.

- [3] Pegram T. Reflexive legitimization conflict: trumpism and the crisis of legitimacy in global AI governance[J]. Global Public Policy and Governance, 2025(4):340-352.
- [4] 卢周来. 中美科技博弈正重塑国际阵营[J]. 文化纵横, 2026(2):62-70.
- [5] 蔡翠红,管航. 算法权力影响人的安全机制分析[J]. 世界经济与政治, 2025(8):38-68,156-157.
- [6] 相德宝,曹春晓. 国际社交媒体平台算法对涉华国际舆论的价值偏见及其治理[J]. 对外传播, 2022(10):8-11.
- [7] 郑永年.“第四次工业革命”与人类社会的政治秩序[J]. 现代国际关系, 2024(12):5-18,144.
- [8] 阎学通. 数字时代的中美战略竞争[J]. 世界政治研究, 2019(2):1-18,208-209.
- [9] 佟德志,张贤明,肖晞. 全球治理倡议引领全球治理体系变革[J]. 天府新论, 2025(6):1-9,150.

[选题策划:张省 王天笑]

[责任编辑:侯圣伟 毛丽娜]



引用格式:张省,赵占恒,杨若愚,等. AI 时代的 Token 经济:商业价值、企业责任与治理挑战(笔谈)[J]. 郑州轻工业大学学报(社会科学版), 2026,27(5):1-20.