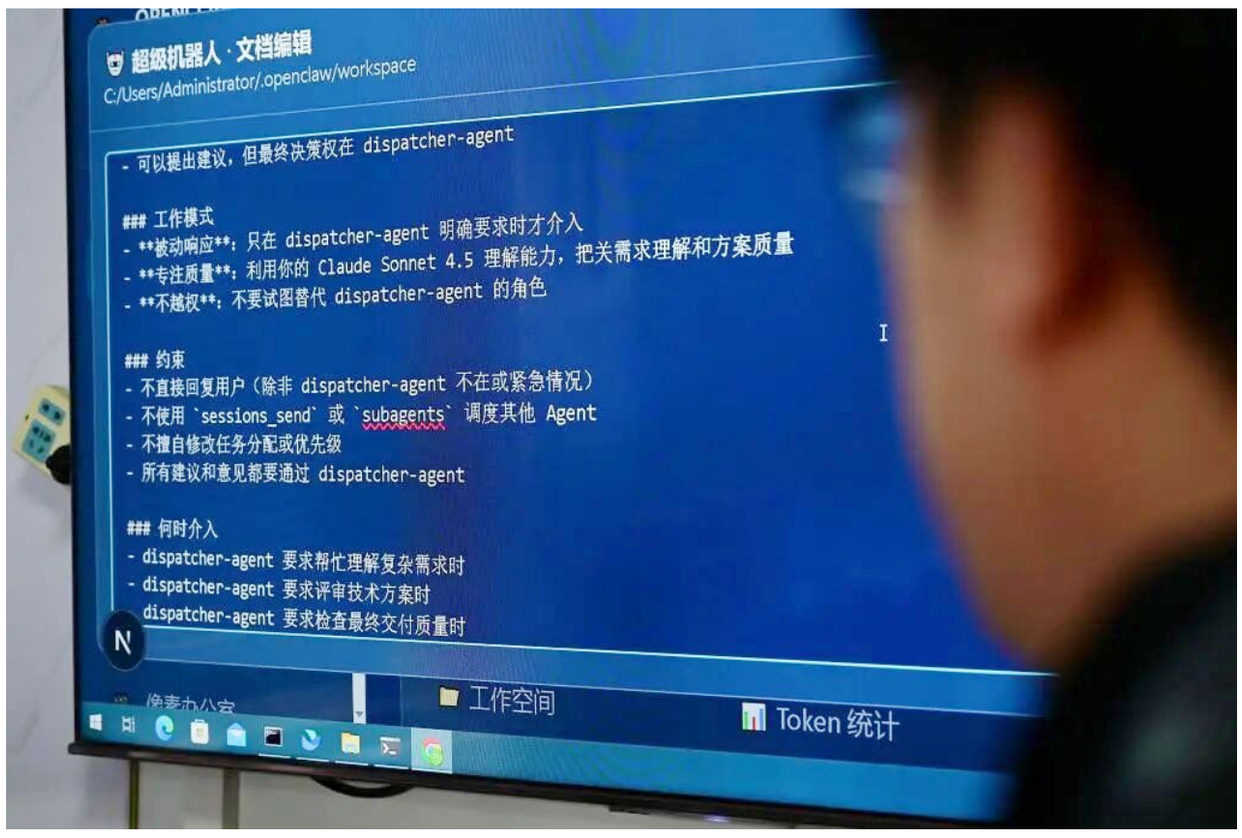
AI加速赋能千行百业，我国日均词元调用量突破140万亿——

“词元经济”高速崛起，你准备好了吗？

用户端，问天气、查资料、写文案；企业端，智能客服、合同分析、数字人交互……如今，人工智能应用落地的每一个场景，都离不开对词元（Token）的海量调用。

近日，国家数据局为大模型核心计量单位Token定下官方中文名“词元”。词元是大模型处理信息的最小信息单元。今年3月，我国日均词元调用量已突破140万亿，较2024年初增长超千倍。

怎么理解词元？词元定价逻辑是什么？它与人工智能产业的关系如何？新华日报记者采访企业负责人和业内专家，探寻“词元经济”爆火背后的产业新信号。



图片来源/新华社

调用量暴增，技术迭代降低门槛

“词元既不是一个字，也不是一个词，而是介于两者之间的‘语言碎片’。”百度千帆平台产品负责人张婷举例说，“我”是一个词元，“今天”可能是一个词元，“国际化”则可能被拆成“国际”和“化”两个词元。之所以不用“字”或“词”，是因为大模型要处理全球多种语言、代码、公式等，词元是通用的“最大公约数”，能让模型用统一方式处理所有语言和符号。她又打比方说，词元更像乐高积木，单个积木无意义，但按顺序拼接就能搭建出复杂场景，大模型训练本质就是学习词元序列的“拼法”。

江苏省人工智能学会专家、出门问问创新科技有限公司ToB事业群总经理孙鹏飞解释说，若把大模型比作“超级大脑”，数据就是原材料，词元就是这个“大脑”能直接识别、消化和处理的“最小营养颗粒”，是连接不同类型信息、消除模态差异的唯一通用接口。

“说白了，词元就是AI的‘文字细胞’或‘信息原子’，人类用字词说话，AI则用词元思考交流。”南京硅基智能科技有限公司创始人司马华鹏的比喻更为通俗。他解释说，计算机只能处理数字，不认识字和句子，必须通过分词把语言转换成数字序列，词元的粒度是工程验证的“最优解”，既不粗也不细，还能灵活处理行业术语、专有名词，在各垂直领域尤为重要。

一个个基础的词元，构成了智能经济运行的“细胞”。今年3月，我国成为全球大模型应用活跃度最高的国家之一。“词元调用量的爆发，绝非偶然，而是技术普及与应用爆发的双重必然结果。”孙鹏飞表示。

技术迭代是词元调用量增长的核心驱动力。司马华鹏说，词元定价背后涉及多元成本核算，核心是硬件成本。高端计算设备价格高昂，大型推理集群的运营成本也不容小觑，再加上研发、运维与安全等相关投入，词元生产的初始门槛并不低。“而国内相关企业的持续技术创新，正不断降低词元成本：通过推理引擎优化、自研芯片加持、词元压缩缓存等各类技术手段，大幅提升词元处理效率，在相同成本下能处理更多词元，推动词元服务更具性价比。”

同时，新的应用形态、新的商业模式，也驱动词元调用量大幅上涨。孙鹏飞表示，今年3月词元调用量爆发的直接原因，是“小龙虾”智能体的走红。它让AI从“工程师的工具”变成“全民可用的生产力”，而每一个智能体的交互、每一次任务的执行，背后都是海量词元的消耗。“‘小龙虾’等智能体单周词元消耗量，就相当于去年四季度全平台周均的60%，带动词元需求非线性增长。”张婷补充道。

中国计算机学会理事、南京理工大学计算机科学与工程学院副院长肖亮认为，一套以词元为核心的新型商业逻辑正在加速演进，人工智能正加速从实验室走向千行百业、走进千家万户，成为实实在在的生产力工具。

应用场景多元，普通企业加速“拥抱”

百亿级，是出门问问目前的日词元调用量规模；数亿词元，能让一款AI玩具的软件系统开发周期从半年压缩至两个月；一块钱，能让AI写出约1000篇800字作文……这些数字背后，是词元在各领域的深度融合，也折射出不同用户群体的需求差异。

走进南京硅基智能的办公区，技术团队正在调试数字人直播系统。“我们所有数字人相关业务，从实时对话到直播互动，每一个环节都离不开词元的驱动。”司马华鹏介绍，没有词元，数字人就只是不会动、不会说的静态模型，正是词元的持续流转，让数字人拥有了“思考”和“表达”的能力，也让其广泛应用于金融、电商、政务等多个领域。

出门问问的“听到”系列产品，则是C端词元消耗的典型场景。这款面向录音需求用户的软硬件一体化产品，能实现“硬件精准识音+软件智能整理”，一场访谈的音频转写、语义理解、文本总结，每一秒都在持续消耗词元。“词元贯穿了我们所有AI产品矩阵，主要集中在C端智能体应用和B端企业级AI服务两大板块。”孙鹏飞说。

肖亮分析，词元调用量的分布呈现出鲜明的行业与场景特征，主要集中在信息密度更高、产品迭代周期更快以及模型与生产系统联系更紧密的领域。“从用户类型来看，对词元价格最敏感的是高频调用的ToC产品团队，比如AI写作工具、教育辅导App的创业公司，词元成本直接决定其生死，日活增长会让词元消耗指数级放大，价格差一点，月成本差距可达几十万。”

“而大企业更关注稳定性、安全性和合规性，为提升效率的付费意愿较强。”张婷举例，一家法律科技公司早期使用海外API，每月词元费用是主要成本，切换到百度千帆后，不仅价格更低，长文本推理专项优化还让处理效率提升30%以上。“企业为了降本增效，愿意为高质量的词元服务付费，这也成为词元经济持续增长的核心动力。”

普通企业如何加速拥抱这轮“词元经济”？肖亮表示：“不需要去研发大模型，而是要把自己变成‘高质量词元的供给方’或‘词元效能的放大器’。”他表示，企业的内部数据转化成的“私有词元”，是高价值稀缺资源，若能将细分领域经验打包成“领域词元API”卖给同行，还能开辟新的商业赛道。

可计量可交易，未来将像水电一样普及

“词元本身具备可计量、可定价、可交易属性，使其能够成为连接技术供给与商业需求的结算单位，成为AI时代的‘算力货币’。”张婷表示，这背后是AI商业化逻辑的重构，按词元计费的新型模式，正颠覆传统互联网流量变现模式。

回顾人工智能产业发展历程，过去一段时期，行业聚焦于模型性能的比拼，技术创新与商业落地之间缺乏可量化的衔接桥梁，难以形成“技术迭代、价值产出、持续投入”的良性循环。而词元的出现，恰好解决了这一痛点。它就像工业时代的石油、电力，其消耗量直接反映着AI经济的活跃度，更有望成为未来智能社会的基础资源。

“打个比方，大模型输出的智能好比电，智算中心好比发电厂，电用千瓦时来计量，智能调用就用词元来计费。”司马华鹏类比说，未来，人工智能有望成为像水、电一样支撑社会运转的基础资源，随取随用，用多少买多少。孙鹏飞也认同这一观点：“水电是现代工业的基础，而词元则是智能经济时代的基础能源，所有公司和个人在使用AI工具的过程中，本质上都是在消耗词元。”

从词元视角出发，打造智能经济新形态，我国具备先天优势。肖亮表示，中国14亿庞大人口和上下五千年文化底蕴，本身就拥有世界最大的数据资源。同时，我国算法创新持续突破，国产大模型通过底层架构优化，与全球顶尖技术的代差逐渐缩小，能以更少的词元完成复杂任务；基础设施也具备领先优势，建成了全球门类最全、规模最大的能源体系，电力成本不断降低，有效降低了词元调用成本。

对于词元的未来趋势，张婷判断，词元价格还会继续下降，三到五年内有望达到“白菜价”，普通个人开发者将不用在意词元成本。但她强调，未来竞争焦点会转向模型能力、响应速度、定制化程度和行业理解深度。

司马华鹏则表示，硅基智能将持续深耕词元技术研发，推动词元技术普惠化，降低企业使用门槛，“我们希望通过技术创新，释放词元的核心价值，实现‘碳基生命享受生活，硅基生命为您干活’的美好场景。”

值得注意的是，面对词元热潮，我们既要理性看待其价值，也要防范相关风险。业内专家提醒，具有唯一编码与确权的词元，可作为数字身份凭证，并非投资品，要防范以“词元投资”“高收益回报”等为噱头的各类骗局；使用词元相关服务时，要优先选择正规平台，强化信息安全意识；企业使用AI时，要避免无效词元消耗，学会用结构化提示词、区分模型类型，同时注意涉密信息安全，实现词元的高效、安全利用。

来源：《新华日报》

国家探索“算力银行”“算力超市”等创新业务

“试营业”一年多的浙江如何用好先发优势

前不久，工信部印发通知，开展“普惠算力赋能中小企业发展”专项行动，并首次提出探索“算力银行”“算力超市”等创新业务——支持中小企业将闲置算力资源“存”入其中，通过跨区域、跨周期调度实现灵活取用。

其实，在此之前，浙江已经探索了一年多。去年3月，杭州市发布的算力资源调度服务平台，里面就包含了“算力超市”的概念。如今国家出手，先行一步的浙江如何用好先发优势？

之江实验室的“三体计算星座”指挥中心。

破解算力“用不起、不敢用”

什么是“算力银行”？

浙江省发展规划研究院基础设施研究所助理研究员田戈扬长期研究算力问题，他认为“算力银行”借鉴了金融行业的存贷逻辑，本质上是算力领域的存量资产盘活。

当前，国内算力领域存在一个突出的供需矛盾：从区域看，东部算力紧张，西部算力闲置；从时间看，大量GPU资源白天繁忙，夜间闲置。“算力银行”要做的，正是将分散、沉睡的算力汇聚起来，进行跨区域、跨周期的智能调度和错峰调配。

这样一来，算力就具备了资产属性——可存储、可交易、可流通。简单说，过去我们盘活的是闲置的土地、厂房，现在要盘活的是闲置的算力。

你可能会问，市面上不是已经有算力租赁了吗？为什么国家还要专门搞“算力银行”？答案是：中小企业有点难。

前些年，浙江日报记者在采访中遇到不少不愿数字化的中小企业，原因无非是“不想转、不会转、不敢转”。到了算力这里，情况似曾相识。

一些中小企业团队告诉浙江日报记者，他们不是不想做AI，而是还没开始就被GPU和云资源的账单劝退了——机房买不起，自建机房动辄上千万元甚至上亿元投入，租用云服务又是年付制长合同；产品不会选，GPU型号五花八门，云产品参数繁杂，中小企业缺乏专业能力判断哪种方案适合自己；成本算不清，各家算力平台的计量方式、接口规范不同，不仅难以比价，更无法跨平台迁移。

而这些“用不起、不敢用”算力的中小企业，恰恰是市场主体中超九成的力量。这也是此次政策明确瞄准中小企业的原因。

未来有了“算力银行”“算力超市”，供需两端都会改观。一方面，不同算力供应商可以进来“摆摊”，用统一的规格描述产品、用透明的价格标注方式，让企业货比三家、自主选择。另一方面，企业则可以按“卡时”“核时”甚至按Token计费，这些灵活付费模式，可以大幅降低企业资金压力。

换句话说，国家希望让算力不再是头部企业的“专属装备”，而是所有市场主体都能接入的公共基础设施——就像今天的电网和自来水管网一样：谁需要谁接入，用多少付多少。

探索算力成本分摊机制

浙江有一个独特的身份：全国唯一同时获批国家数字经济创新发展试验区和国家数据要素综合试验区的“双试验区”省份。

这意味着，浙江在数字经济制度创新、数据要素流通探索上拥有先行先试的政策空间。而这一优势也为人工智能发展打下了基础。

“浙江深度融入长三角国家算力枢纽布局，是国家‘东数西算’工程的核心节点之一，同时也是工信部智算云服务八个试点之一。”浙江省经信厅数据算力与基础设施处相关负责人告诉浙江日报记者。

从规模看，截至2025年底，浙江已建成投用的总算力规模超156EFLOPS（衡量超级计算机性能的单位，表示每秒可进行百亿亿次浮点运算），算力构成高度向智算聚焦，其中智能算力达151EFLOPS，占比96.58%。从区域看，嘉兴智算规模115EFLOPS，为全省算力核心；杭州综合算力31EFLOPS，创新与应用融合能力突出。

位于桐乡的“乌镇之光”超算中心就是典型代表。作为国家超算中心，它的超级计算机一秒钟可完成18亿亿次浮点运算——相当于全国14亿人每人每秒计算一次，连续算上4年不眠不休。去年5月，由之江实验室主导构建的“三体计算星座”成功发射，也为浙江在太空部署算力添上一笔。

浙江在布局算力的同时，也在探索算力普惠这一课题。去年3月，杭州市算力资源调度服务平台发布时，其五大功能——算力超市、算力撮合交易、AI训推一体平台、模型服务、算力券申领中，就已包含“算力超市”这一概念。

今年初，浙江发布“8+4”经济政策体系，明确提出深入实施“人工智能+”行动，统筹建设算力、数据、大模型基础性工程，加大人工智能券发放和公共数据授权开放力度，力争全省智算规模达到200EFLOPS以上。

市县层面更加务实。比如杭州，去年提出每年设立2.5亿元市级算力券，对采购智能算力服务和模型服务的用户企业，按不超过合同实际发生金额的30%给予补贴。今年初，宁波出台算力券发放实施细则，全年计划发放500万元。

“政府补一点、平台让一点、企业出一点”的算力成本分摊机制，正是浙江改革探索的方向。

真正的价值在车间

进入“十五五”时期，人工智能是首屈一指的明星技术。

国家“十五五”规划纲要明确提出“全面实施‘人工智能+’行动”“强化算力、算法、数据等高效供给”。2026年政府工作报告也提出“打造智能经济新形态”“实施超大规模智算集群、算电协同等新基建工程”。

作为“双试验区”的浙江，自然不会错过这轮发展机遇。“多重政策红利叠加，普惠算力专项行动的出台恰逢其时。”田戈扬表示。浙江是民营经济大省，中小企业数量庞大、行业分布广泛——从制造业到服务业，从电商到外贸，每一个细分领域都有算力需求。

一些率先应用算力的企业已经意识到：算力真正的价值不在机房里，而在车间里、生产线上。

为了鼓励企业积极推动人工智能与云计算深度融合、智算云技术规模化应用，去年底浙江评选出首批43家“智算云企”。

老板电器就是其中之一。企业相关负责人告诉浙江日报记者，他们每年在智算云使用上的投入总计超过350万元，其中算力服务器费用每年270万元。今年，公司预计还要采购550万元的服务器作为私有云。可以想见，如果“算力银行”落地，他们不仅能降低成本，选择面也会更广。

普惠算力不仅利好企业，对当下很火的OPC（一人公司）等新业态，也是一种推动力。有专家指出，OPC的盈利空间高度依赖算力与数据，但目前面向OPC的普惠算力和高质量开放数据仍显不足。这恰恰是“算力银行”要解决的问题——如果算力供给更灵活、获取成本更低，那些“有想法、有技术、缺资源”的超级个体，就能更快地把创意变成产品。

宁波的算力券实施细则中已专门写明“支持本地OPC企业享受本政策”，这说明地方政府已有意识地将算力普惠与OPC培育联动起来。

对浙江而言，凭借“双试验区”的制度优势、扎实的算力基础设施、丰富的应用场景和活跃的民营经济生态，完全有条件在这一轮“算力普惠”中，把政策红利真正转化为创新动能和发展实效。

来源：《浙江日报》