上海證券報
◎2026年4月21日 星期二 ◎责编 阎基旺◎美编 蔡剑平

AI生成图

新闻小贴士

词元(token)是人工智能时代智能设备中信息存储、处理和交换的具有一定语义的基本符号单元，特别是在大模型中作为模型处理和交换信息的最小单位。

裂变 词元超千倍

从“烧钱”到“烧出价值”，产业“大考”已然开始

“今年3月，我国日均词元（token）调用量突破140万亿。”国家统计局副局长毛盛勇日前表示，我国人工智能商业化、规模化利用取得阶段性突破。

从2024年初到2026年3月，词元在两年间实现超1400倍的裂变式增长，创下数字经济发展史上的增长奇迹。上海证券报记者调研获悉，此次爆发，是政策、技术、场景等多重力量的共振，更是智能体普及带来的产业范式重构。为解决高速增长带来的“成长之痛”，产业竞争全面转向拼效率、拼价值、拼付费的价值深水区，一场从“烧钱”到“烧出价值”的产业“大考”已开始。

◎记者 邹俭朴

多重驱动因素叠加引爆指数级消耗

3月底，无问芯穹词元调用量较年初增长超10倍，部分顶尖模型甚至迎来百倍增长；MiniMax云端AI助手Max-Claw上线120小时内四次紧急扩容；智谱GLM Coding Plan编程套餐推出半年内，其词元调用量大涨15倍……企业的数据直观印证了词元调用的爆发态势。业内人士认为，这背后是政策、技术、智能体、场景四大引擎的合力驱动。

《“人工智能+”行动意见》全面布局产业融合，2026年政府工作报告提出“打造智能经济新形态”，东数西算工程全面推进……从算力建设、场景落地到标准制定，政策全链条赋能，让词元从技术概念走向产业刚需。

在技术层面，成本骤降与国产算力崛起打破了应用门槛。国产模型实现“高性能，低成本”双重突破，API调用价格大幅下探，极大降低了中小企业与开发者的接入成本。MiniMax副总裁严奕骏说：“国产模型在实现同等能力的前提下，性价比优势凸显，不仅激活了国内市场，更打开了全球词元消费市场，成为词元量增长的核心推力。”智谱CEO张鹏表示，模型推理侧的极致工程优化，能大幅降低词元单位成本，为词元规模化普及扫清成本障碍。

在应用层面，智能体成为引爆词元消耗的核心引擎。AI主战场已全面转向智能体（Agent），从“百模大战”进入“智能体大战”阶段。以OpenClaw（龙虾）为代表的智能体应用，直接引爆高质量词元消耗潮。严奕骏打了个比方：“智能体相当于电脑里的‘隐形员工’，在处理复杂任务时会自我规划、持续思考、多次调用工具，后台自主运行带来的词元消耗量呈指数级放大。”

在趋势层面，算力服务形态正在发生根本性转变。“所谓词元经济，本质就是算力即服务的智能经济新形态，核心是普惠、高效、可规模化。”摩尔线程相关技术负责人表示，“人与AI交互、AI与AI协作，都以词元为核心媒介。随着AI智能体迈入应用元年，推理算力需求增速已远超训练，成为词元消耗的主要增量来源。”

免费依赖、算力饥渴与泡沫隐忧

词元调用量的突飞猛进，让AI产业迎来空前繁荣，但粗放式发展带来的深层问题也日益突出，这场由技术与资本共同催化的增长盛宴，多重难题亟待破解。

在经营层面，商业模式脆弱，财务可持续性面临严峻挑战。

当前产业最突出的困境，是“规模增长”与“商业健康度”的割裂，最直接的表现是“免费依赖症”难解。复旦大学中国研究院副研究员刘典表示：“当前，超95%的词元消耗来自免费补贴用户，真正具备商业价值的付费调用占比很低，行业陷入‘调用量虚高，营收能力薄弱’的困境。”大量企业依靠免费额度、补贴政策拉动词元量，数据看似亮眼，却无法形成可持续商业闭环，一旦补贴退坡，增长便面临断崖式风险。

运营效率低下、成本控制失序的问题同样突出。并行科技董事长陈健援引第三方数据称，当前行业算力平均利用率仅30%至60%，大量GPU处于闲置或半闲置状态。按此利用率核算，算力租赁成本要高于词元的总收入。这表明，当前商业模式的脆弱性与成本倒挂风险。

在供需层面，多重供给瓶颈突出，结构性矛盾待解。

张鹏直言，2026年以来词元市场持续供不应求，算力供给已成制约高质量词元产出的核心瓶颈。无问芯穹CEO夏立雪分析称：“传统云计算是‘为人设计’，任务以分钟、小时计；而智能体是毫秒级高并发，长程任务可持续运行数小时、生成百万级词元。传统云的响应速度、并发支持等都难以支撑，底层算力基建亟待重构。”

在业内人士看来，算力资源的低效消耗加剧了供需矛盾，其本质是“资源太散，调度太粗，电价太贵，计量太乱”。陈健算了一笔账：按目前利用率核算，电力成本占词元总成本的七成，西部绿电价格便宜，但一根专线月费要十几万元，数据来回传一趟，省的电费全交网费了。他还表示：“算力太‘散’，比如英伟达、华为、海光等十几类芯片互不兼容，各地算力中心成了一个个‘孤岛’。我们投入大量研发精力，就是在做算力界的‘同声传译’。”此外，跨省调度还面临多重博弈，进一步加剧供需错配。

与此同时，概念炒作升温，价值虚增与投资风险暗流涌动。

词元经济的火热，吸引了大量资本和创业者涌入。“概念股不等于受益股。”刘典表示，部分上市公司仅凭“token”概念蹭热点，实际业务与词元产业链无实质关联。有业内人士称，当前的炒作，“是一场典型的利用技术信息差进行的金融收割”。

香港科技大学计算机科学及工程学系副教授王帅说：“在做模型审计研究时观察到一个现象，调用量在涨，但有效产出未必同步在涨。这也是词元产业价值错配的特殊之处——不是没人用，而是用的过程中，价值在悄悄漏掉。”

然而，词元并非虚无缥缈的营销符号。沐曦股份相关负责人说：“词元正从AI计算的最小计量单位，转向AI消费的‘财务指标’。算力规模与成本，直接决定了企业的词元预算、词元支出等核心经营指标。”这意味着，词元正在进入企业的财务报表，成为实在的成本与支出项。当前，部分模型厂商仍依靠“烧钱”补贴换取市场份额和用户增长，如果词元的“量价齐升”逻辑因竞争加剧或成本失控而无法持续，企业将面临巨大的盈利压力。

“量”“质”齐升解锁未来新可能

面对成长阵痛，行业正朝四大核心方向发力，推动词元产业链实现从“量的爆发”到“质的提升”。

方向一：从“堆砌算力”到“优化每瓦词元产出”。

随着单纯扩大算力规模的边际效益急剧递减，未来行业的竞争不再是模型参数、词元总量的比拼，而是单位词元产出效率与算力性价比的较量。

沐曦股份相关负责人介绍：公司一方面通过规模化落地降本增效；另一方面构建全栈自研生态，以软硬协同提升词元产出效率。夏立雪表示，无问芯穹通过上下文缓存优化，将词元储存复用效率提升2至3倍，大幅降低重复计算带来的算力损耗。中科曙光则通过新技术，同卡数吞吐性能提升4倍以上。

硬件创新是提升“单兵作战能力”，智能调度则是优化“兵团作战效率”的关键。并行科技董事长陈健将算力调度定位为词元产业链的“效率引擎”和“价值转换器”。并行科技正通过全域异构调度与算电协同，将分散的算力变成标准化的词元服务。

陈健则举例称：“我们服务一个大模型集群，通过池化调度，千卡级别的需求直接压缩到两百卡就能承载，让词元真正成为了‘普惠水电’。”

中科曙光相关负责人注意到一个现象：“以前客户租的是GPU卡，现在大家更关注‘每瓦电力能产出多少词元’‘每块钱成本能买到多少智能’。计费体系从‘按卡/按时’转向‘词元级精细化计量’。”他介绍，广州已上线全国首个基于“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，构建按量、按周期等多种灵活计费体系，实现全栈异构算力的统一纳管与资源池化。词元级调度意味着智算中心与模型、应用端的协同可以精细到每一次推理请求，极大地提升了资源匹配效率。

方向二：从通用智能体到行业垂直智能体。

智能体是将词元消耗转化为客户付费意愿和价值增量的关键载体。2026年被产业界普遍视为“智能体规模化应用元年”。从通用智能体到行业垂直智能体，AI正从被动问答转向主动执行，将词元消耗直接转化为企业效率提升与成本下降。

能够解决实际问题的智能体，正驱动词元消耗进入“量价齐升”的新阶段。智谱CEO张鹏披露，3月公司推出了Claw Plan服务，上线仅2天订阅用户突破10万，上线20天订阅用户突破40万，“这验证了智能体长链路任务的巨大商业空间”。

从MiniMax的实践来看，伴随公司M2系列编程模型持续迭代与OpenClaw等智能体应用的爆发，词元调用量呈现指数级增长。M2系列文本模型2月日均词元消耗量较2025年12月增长超6倍，编程场景增长超10倍。

方向三：从“卖硬件、卖机时”到“卖词元服务”。

词元时代，传统的“卖硬件”或“卖机时”模式已无法捕捉产业增长的最大价值，商业模式要与客户的价值创造深度绑定。

以并行科技为例，陈健给出“卖词元服务”的清晰商业化路径：企业已落地按词元计量计费、算力交易撮合、MaaS一体化三条路径。“我们要做AI时代的‘算力运营商’，赚取生态流通的长期价值。”他透露，公司的MaaS平台目前已支持秒级计费、API即开即用，开发者用多少付多少，不再需要“包年包月租机器”。通过算力池化、词元级细粒度调度、算电协同、算力银行和普惠定价五大路径，系统性解决算力低效消耗的痛点，“我们的定位，就是让算力流动起来，让词元生产更便宜、更普惠”。

摩尔线程相关技术负责人也表示，公司坚持训推一体、云边端全场景布局，既要依托智算集群兼顾训练与高性价比推理，又要通过个人算力设备打通全场景算力链路，“让词元经济服务于每一个人和每一家企业”。

方向四：从“技术指标”到“治理对象”。

当词元从技术参数蜕变为经济载体，一个更深层的问题随之浮现：谁来定义词元的质量？行业专家预判，词元质量将成决定产业格局的重要“护城河”。

相关专家表示，规则层面的缺位值得关注——如果缺乏统一标准，词元计费可能带来价格不透明、差异化收费等问题。

“词元的数量不等于能力和价值；如果缺乏统一口径和有效审计，它很容易从技术指标异化为营销指标和收费指标。”王帅表示，当前部分模型服务可能存在减配、‘降智’、能力漂移等问题，从表面指标看服务仍在运行，实际输出质量、稳定性和任务完成度却有所下滑；在聚合平台、分销链条等场景中，存在“货不对版”风险，用户支付的是高水平模型服务，实际获得的却未必是相应能力。

上海交通大学副教授、无问芯穹联合创始人戴国浩认为，不同行业、不同任务，对词元“价值密度”要求不同。由更强模型生成的高质量词元在某些场景中能够产生更高价值，“因此在定价上进行分层，符合基本经济逻辑。”

“基于高阶智能带来的底气，我们的API调用定价在一季度提升83%，即便如此，市场依然呈现出供不应求的情况，调用量增长400%，再次印证了高质量词元是当下的稀缺资源。”张鹏说。

相关专家表示，当前，不同模型对同一文本的词元切分数量存在差异，计量标准尚不统一。谁能在标准与治理的博弈中率先建立清晰的计量口径、审计框架和规则体系，谁就将在这场关乎标准制定权和规则主导权的竞争中，掌握下一轮的先手棋。

记者观察

词元能否真正打通人工智能商业闭环

◎记者 窦世平

词元(token)已成为AI产业商业化的核心计价载体，从算力消耗计量到服务价值结算，它被寄予破解AI“重投入、难变现”困局的厚望。我国日均词元调用量两年间实现千倍增长，云厂商、模型企业纷纷布局词元计费模式……热潮之下，AI商业闭环仍未真正形成，词元能否成为打通这一堵点的关键，有待深入观察。

长期以来，AI产业因缺少统一价值标尺，始终面临定价模糊、成本与收入难以匹配的难题。词元的出现，为这一痛点提供了破解路径，其可计量、可定价、可交易的属性，让AI算力消耗与价值输出得以精准量化，推动行业逐步从“算力即服务”向“词元即服务”转型。

但是，统一的计价单位并不意味着统一的定价能力。随着模型能力趋同，普通模型词元单价持续走低的背后，是模型层向应用层让渡价值的行业现实。部分企业陷入“调用量激增，收入增速滞后”的困境，行业开始思考：词元的定价权究竟掌握在谁手中？业内人士认为，解决复杂推理任务的词元，其价值理应高于普通文本生成的词元，从实践探索来看，头部模型企业已针对代码等高价值场景优化定价。这一趋势表明，词元的价值衡量正从“按量计价”向“按质定价”演进，商业闭环的构建需要在精准计量的基础上，叠加价值分层的能力。

在产业发展中，算力投入与变现能力的失衡，始终影响着闭环构建节奏。全球AI算力投入持续高位增长，北美科技巨头加码数据中心、芯片等基础设施布局，算力资本开支快速攀升，算力通胀推高产业链成本。即便词元实现了精准计价，如何平衡上游高额投入与下游变现能力，维系合理盈利空间，仍是行业必须面对的课题。

值得关注的是，随着上下文窗口突破百万乃至千万级，大量词元被消耗于长文本背景维护、检索增强生成的数据召回以及智能体间的通信协作，这类“开销型词元”消耗了可观算力，却难以向客户单独计费。区分“有效推理词元”与“辅助词元”的价值差异，正在成为产业链上下游博弈的新焦点。商业闭环的真正挑战，不仅在于卖出更多词元，更在于让每一类词元消耗都获得市场认可。

从落地实践来看，AI商业化呈现鲜明结构性特征：在企业级场景中，词元消耗与业务提质增效呈现高度正相关，订阅制、阶梯定价等模式获得市场认可，形成“用量提升—收入增长”的良性循环，成为商业化核心支撑；而在消费端市场，现象级应用仍在培育，AI智能体场景适配能力有待提升，用户付费习惯养成尚需时日，仅靠词元定价调整，难以快速突破消费端变现瓶颈。

从产业逻辑而言，词元是AI商业化的重要工具，为价值核算提供了关键支撑，却无法替代核心价值创造。企业实现商业化突破，既依托词元的精准计量，更源于真实应用价值的供给；部分企业词元消耗规模扩大，但商业化转化效果不佳，也印证了词元应与价值创造协同发力，才能推动闭环落地。

展望未来，词元的角色或将超越单纯的计价单位。在AI智能体加速落地的趋势下，词元正演变为智能体与智能体之间、智能体与工具调用之间的微支付桥梁，一种AI原生经济的底层结算体系呼之欲出。而AI商业闭环的最终形成，仍应在算力投入与变现节奏、工具创新与价值创造、B端实践与C端突破之间不断磨合优化。随着产业实践持续深入，各类要素逐步适配，AI商业化的成熟生态，也将在稳步推进中不断完善。