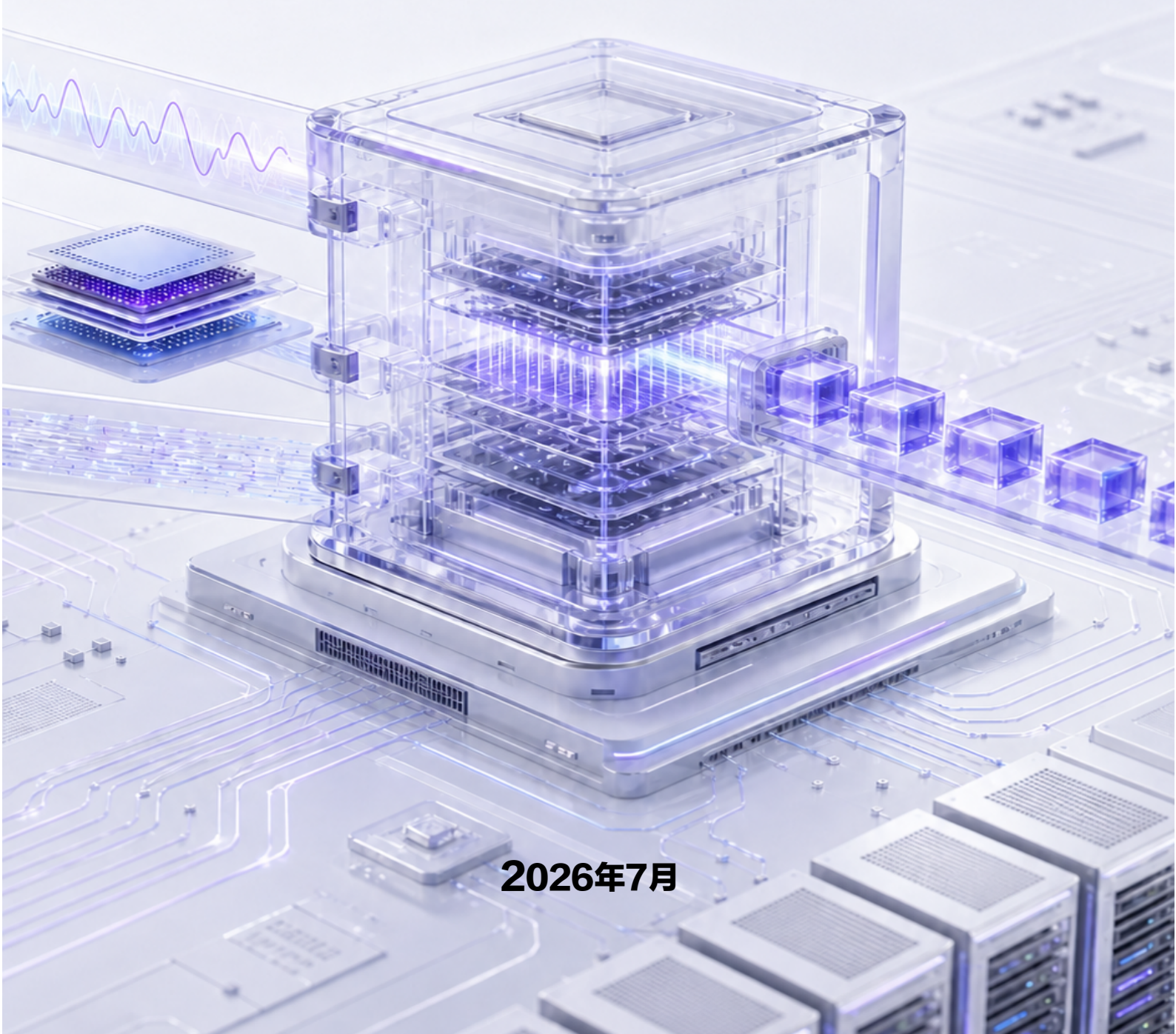




智慧城市行业分析
www.smartcity.team

中国词元工厂发展全景洞察与深度 分析报告（2026）



2026年7月

目录

摘要	4
1. 研究背景与核心概念界定	4
1.1. 词元工厂的兴起背景与产业定位	4
1.2. 核心定义、功能边界与价值主张	5
1.3. 报告研究范围、方法与结构说明	7
2. 中国词元工厂发展的政策环境分析	8
2.1. 国家部委层面政策导向与监管框架	8
2.2. 数据要素与人工智能专项政策影响	9
2.3. 行业协会与地方政府的配套举措	10
2.4. 官方媒体舆论导向与行业规范动态	12
3. 词元工厂的核心构成要素与技术路径	13
3.1. 基础设施层：算力供给、存储与网络	13
3.2. 模型优化层：推理加速与成本控制关键技术	15
3.3. 平台服务层：调度、部署、监控与管理	16
3.4. 面向AI智能体的原生服务能力构建	17
4. 中国词元工厂典型案例与实践分析	19
4.1. 按运营主体分类：国有企业、民营企业与电信运营商	19
4.2. 按服务领域分类：政务、工业、金融、医疗等垂直行业	21
4.3. 按业务模式分类：通用基础设施、行业专用、AI智能体服务	23
4.4. 代表性案例深度剖析	25

5. 国外词元工厂发展实践与典型案例	30
5.1. 北美地区：技术创新与商业模式前沿	30
5.2. 欧洲地区：合规导向与行业应用	31
5.3. 亚洲（除中国外）及其他地区：差异化发展路径	33
5.4. 国外代表性案例及其特点分析	34
6. 业务模式、差异化能力与综合比较分析	39
6.1. 主流业务模式解构：IaaS、PaaS、MaaS、BaaS	39
6.2. 核心竞争力与差异化能力维度分析	40
6.3. 中外词元工厂发展路径与能力对比	42
6.4. 市场竞争格局与关键成功因素	44
7. 发展趋势、挑战与战略建议	46
7.1. 技术融合、行业渗透与生态演进趋势	46
7.2. 面临的主要挑战：技术、成本、安全与合规	47
7.3. 对市场参与者的战略发展建议	49
7.4. 对监管与政策制定的相关建议	50
参考文献	52



摘要

中国词元工厂正成为人工智能时代的核心基础设施，通过将电力与数据转化为可计量、可交易的词元，推动AI服务从算力租赁向“卖智能”模式演进。在政策强力驱动下，国有资本、民营企业与电信运营商协同布局，形成以安全可信、普惠服务和工程化效率为特色的多元化生态。技术上，行业聚焦异构算力调度、推理加速与存储优化，实现单位能耗下更高词元产出。相较于国外，中国依托规模化市场、国产化算力适配与低成本优势，在普惠化和工程落地层面表现突出，但原创模型与基础框架仍存差距。未来竞争将围绕成本控制、安全合规、行业深耕与智能体原生服务能力展开，构建开放生态与标准化体系是产业可持续发展的关键。

1. 研究背景与核心概念界定

1.1. 词元工厂的兴起背景与产业定位

人工智能产业正经历一场从模型训练为中心到以推理应用为核心的深刻范式转移^[1]。这一转变的直接驱动力是AI智能体（Agent）在全球范围内的加速渗透与爆发式增长^[2]。随着OpenClaw等开源智能体产品的问世，市场直观感受到AI自主完成复杂任务的能力，但随之而来的是对支撑智能体多步推理与长程任务的海量词元（Token）需求的急剧膨胀^[3]。行业正从“AI原生”迈向“Agent原生”的新阶段，AI系统的核心使用者正从人转变为智能体，这使得“词元”开始替代“浮点运算次数（Flops）”成为衡量AI能力供给与消耗的新度量衡^[2]。传统的云原生架构已难以适配智能体时代的发展需求，迫切需要构建以“词元”为中心的AI基础设施，即词元工厂，从交互范式到整个基础设施架构进行重新设计与优化^[2]。

市场数据直观印证了这一变革的规模与速度。截至2026年3月，我国日均词元调用量已突破140万亿，相比2024年初的约1000亿规模，两年间增长了超过1400倍^[4]。这一爆发式增长不仅解决了AI服务难以量化结算的痛点，更将推动数字经济从“卖连接、卖带宽”彻底迈向“卖算力、卖智力”的新阶段^[5]。OpenRouter数据显示，中国模型的词元消耗量已占全球市场的54.1%，全面超越美国模型的41.5%，在编程场景中的占比更是高达67.6%^[4]。词元需求的指数级增长与算力供给的线性增长之间的矛盾日益突出，用户普遍感觉AI服务昂贵，其本质是供不应求^[6]。因此，如何高效、稳定、低

成本地规模化生产词元，成为产业发展的核心瓶颈与关键机遇。

在此背景下，英伟达首席执行官黄仁勋于2026年3月的GTC大会上首次提出了“词元工厂”的概念^[2]。他指出，数据中心如今已经成为全天候运转的“词元工厂”，输入电力和数据，输出词元^[2]。黄仁勋同时提出了衡量其经济效益的核心公式：收入 = 每瓦词元数 × 可用千兆瓦数，将词元生产的效率与规模直接与商业收入挂钩^[2]。这一概念的提出，为AI基础设施的发展指明了新的方向。紧随其后，产业界加速布局，商汤科技、百度智能云、中国移动等企业纷纷提出并实践各自的词元工厂战略^[2]。中国庆阳、韶关等多个算力枢纽节点以及百度、阿里巴巴、电信运营商等多方力量均在积极建设词元工厂^[2]。运营商正推动算力租赁从传统的固定月租模式转向按实际词元用量计费，并已推出面向个人用户和企业的词元套餐^[2]。词元工厂的兴起，标志着AI产业进入了以智能产出效率为核心竞争力的工业化生产新阶段。

从产业定位来看，词元工厂的核心角色类似于“水电煤通讯”等传统基础设施供应商，旨在为用户提供稳定、优质、可计量的词元服务^[2]。其本质是将底层复杂的芯片、模型和调度能力，封装成标准化的词元服务，把算力从企业的成本中心转变为可精准管理和调度的生产原料^[2]。在AI出海的新分工中，词元工厂位于产业链的中层：底层是算力基础设施，中层是词元工厂提供标准化的词元产品，上层是行业智能体和AI应用提供面向终端用户的品牌和服务^[4]。因此，词元工厂是连接底层算力资源与上层AI应用的关键枢纽，是AI能力实现规模化、普惠化交付的核心载体。

1.2. 核心定义、功能边界与价值主张

“词元工厂”又称Token工厂，其核心定义是指人工智能时代将电力与数据转化为词元（Token）的算力数据中心或基础设施^[2]。词元（Token）是AI大模型处理文本、图像、语音等信息的最小基本单元，一个智能的基本单元，具备可计量、可定价、可交易的特性^[2]。2026年3月，国家数据局正式将Token的中文译名确定为“词元”，为其在国内的规范化发展奠定了基础^[2]。词元工厂的运作模式可以形象地理解为：以算力硬件（如GPU/加速芯片）与高质量数据为投入，历经AI模型的训练与部署加工，进入实时词元推理生产环节，再进行持续的监控与评测实现反馈优化，最终输出可供API调用的词元，用于聊天回答、代码生成、图片生成等各类AI应用^[7]。整个流程如精密工厂般高效运转，确保词元产出率持续提升^[7]。

词元工厂的功能边界远超传统的算力数据中心或云服务。传统算力中心像一个“五

金超市”，摆满了服务器、显卡、存储等“原材料”，用户需要自行采购并组装、调试、运维^[8]。而词元工厂则直接交付“标准件”——用户只需通过API接口，就能按需调用问答、摘要、推理、生成等AI能力，不再需要自建昂贵的AI算力集群，也无需管理复杂的硬件设施，只需按实际消耗的词元数量付费即可^[8]。具体而言，词元工厂的核心功能在于完成“资源池化、平台化调度、模型服务化和运营产品化”的全链条闭环^[8]。这包括：将底层的智算服务器、存储、网络、安全等资源纳入统一资源池；通过云管平台实现资源的纳管、监控告警和智能调度编排；在智算服务平台上承载模型服务、租户权限、服务目录、调用计量和运行监控；最终将这些能力封装成企业可以直接使用的、标准化的词元服务产品^[8]。

词元工厂的价值主张主要体现在以下几个方面：

1. 降低使用门槛与总拥有成本（TCO）：词元工厂彻底颠覆了传统的硬件售卖和机柜/带宽计费模式^[7]。客户无需承担沉重的硬件采购与运维成本，即可按需获取稳定、高效的AI能力，技术门槛大幅降低^[7]。按词元消耗、订阅或API调用的灵活计费模式，让成本透明可控，综合使用成本较传统模式显著降低^[9]。例如，贵州移动的词元工厂通过智能路由和叠加贵州低温节能优势，使得Token综合调用成本较行业平均水平降低20%以上^[8]。
2. 提供稳定、可靠、可预期的智能供给：词元工厂通过工厂级的操作系统和高可靠的物理平台，确保词元生产的稳定性和服务等级协议（SLA）保障^[10]。它将模型厂商的“造脑”能力转化为可规模化、具备SLA保障、可计量的智能服务，与模型厂商共同构成从模型能力到产业应用的完整闭环^[10]。这种稳定供给对于企业将AI深度嵌入业务流程至关重要。
3. 实现AI能力的普惠化与标准化：词元工厂旨在将算力生产工厂化，并通过“卖智能”的创新模式，让算力像水电一样触手可及，成为赋能千行百业的普惠生产力^[7]。通过标准化封装和统一API，词元工厂使得AI能力能够以标准化产品的方式快速交付，大幅缩短了业务落地周期^[11]。个人用户甚至可以像购买手机流量包一样，以9.9元的价格购买1000万词元，使得AI服务真正走进全民日常^[5]。
4. 优化产业效率与资源配置：词元工厂的核心任务不仅是产出词元，更是让每一块芯片的每一秒算力都最高效地转化为对应场景下的高价值词元^[6]。它通过全栈软件优化技术、智能调度和异构算力统一管理，提升单位算力的有效词元产出数量，减少产能损失和算力闲置，从而提升整个AI产业的资源利用效率和投入产出比^[12]。

1.3. 报告研究范围、方法与结构说明

本报告旨在对国内词元工厂进行一场深度、系统性的分析，全面审视其在人工智能新阶段中的发展全貌。报告的研究范围聚焦于中国境内的词元工厂建设与发展实践，涵盖其政策环境、核心构成要素、典型案例、业务模式及差异化能力，同时包含与国外同类实践的对比分析，以提供更广阔的行业视野。研究将基于广泛而深入的信息采集，确保覆盖关键参与者（包括国有企业、民营企业、电信运营商等）、政策动向、技术路径与商业模式创新，力求体现扎实的行业洞察与专业深度。

在研究方法上，本报告主要采用以下三种方法：

1. 文献与政策分析法：系统收集、梳理和分析国家部委、地方政府及行业协会发布的与人工智能、数据要素、算力基础设施相关的政策文件、规划纲要及行业标准，以厘清词元工厂发展的政策导向与监管框架。
2. 案例研究法：深度剖析国内不同运营主体（如国资、民企、运营商）、不同服务领域（如政务、工业、金融、医疗）及不同业务模式下的代表性词元工厂案例，通过对其技术架构、商业模式、市场策略和落地成效的细致分析，提炼共性规律与差异化特点。
3. 比较分析法：在纵向维度上，比较词元工厂与传统算力服务模式的演进与差异；在横向维度上，对比国内外词元工厂在技术路径、商业模式、市场生态等方面的异同，从而更准确地定位中国词元工厂的发展阶段与核心竞争力。

基于以上研究范围与方法，本报告的结构安排如下：紧随本章对研究背景与核心概念的界定，第二章将深入分析中国词元工厂发展的政策环境，从国家部委、专项政策、行业配套及舆论导向等多层面展开。第三章将解构词元工厂的核心构成要素与技术路径，分层剖析其基础设施、模型优化、平台服务及智能体原生服务能力。第四章将系统梳理并深度剖析中国词元工厂的典型案例，按不同分类维度进行归纳与解读。第五章将视野扩展至全球，分析北美、欧洲、亚洲等其他地区的词元工厂发展实践与典型案例。第六章将在前文基础上，对词元工厂的主流业务模式进行解构，并从多个维度分析其核心竞争力与差异化能力，进行中外综合比较。最后，第七章将总结词元工厂的发展趋势与面临的主要挑战，并为市场参与者、监管与政策制定者提出具有前瞻性和可操作性的战略建议。通过这一完整的结构，本报告致力于为行业研究者与高层决策者提供一份内容详实、论证充分、见解深刻的参考指南。

2. 中国词元工厂发展的政策环境分析

2.1. 国家部委层面政策导向与监管框架

中国词元工厂的兴起与发展，与国家级部委层面积极、前瞻性的政策引导和监管框架的构建密不可分。国家数据局、工业和信息化部、国家发展和改革委员会等核心部委通过发布专项政策、行动方案和指导意见，为词元工厂这一新型产业基础设施的规范化、规模化发展奠定了坚实的制度基础，明确了其在国家人工智能与数据要素战略中的核心地位。

国家数据局作为统筹数据要素发展的核心机构，其政策导向对词元工厂的定位与发展路径具有决定性影响。2026年6月，国家数据局正式印发《关于推进行业高质量数据集建设行动的实施方案》，该文件是国家层面部署数据赋能人工智能发展的专项政策^[13]。该方案在“价值释放行动”板块中，首次在国家顶层设计中明确纳入了词元相关的制度安排，提出“探索词元交易等新型交易模式，构建以词元为基础，可量化、可定价的数据价值体系”^[13]。这一表述标志着“词元”概念获得了官方的命名与正式认可，为数据价值的标准化计量与流通提供了关键的政策依据。方案同时提出要“打造集‘数据集生产加工和流通利用、支撑模型训练应用’于一体的数据赋能工场”，这预示着一类新型产业服务主体的出现，即负责将原始数据加工为标准“词元”并确保其安全高效流通的运营者，与“词元工厂”的核心功能高度契合^[13]。此外，方案强调要“完善人工智能训练阶段数据使用规则”，试图为训练数据的合法来源与使用划定边界，这直接回应了AIGC时代的数据版权与合规性质疑，为词元工厂处理数据提供了法律遵循^[13]。

工业和信息化部则从促进平台经济与中小企业协同发展的角度，为词元工厂的普惠服务能力提出了明确要求。2026年，工业和信息化部等七部门联合印发《促进平台经济大中小企业协同发展行动方案（2026—2028年）》，其中明确提出“提升平台企业词元（Token）普惠服务能力，面向中小企业共性需求优化智能体服务，降低中小企业获取与应用门槛”^[14]。这一政策导向直接推动了大型科技企业和平台型企业将其强大的算力与模型能力，通过“词元”这一标准化单元，以普惠化的方式向广大中小企业开放，从而催生了面向B端市场的词元工厂服务模式。同时，方案要求“推动算力资源开放，开展算力并网池化及互联工作”，并“推进全国一体化算力监测调度服务平台建设”，这为词元工厂实现跨区域、跨主体的算力协同与高效调度提供了基础设施层面的

政策支持^[14]。

北京市经济和信息化局作为地方经济主管部门，其政策实践体现了国家导向在区域层面的具体落实与创新。该局于2026年6月先后发布了两项与词元工厂密切相关的政策。在《北京市关于支持工业企业提质增效若干措施》中，明确“支持企业采购并消费词元（Token）开展人工智能应用，对符合条件的企业，根据其实际使用Token情况给予资金支持”^[15]。这一措施通过财政补贴直接激励企业消费词元，从需求侧激活了市场，为词元工厂的商业化运营创造了有利条件。同时，该局还印发了《支持人工智能OPC创新发展行动方案（试行）》，其中提出支持OPC（一人公司）成长社区“自建词元（Token）服务平台或链接智算服务机构”，并为入驻企业发放包含“Token券（模型券）”在内的“全栈资源包”^[16]。这项政策精准地瞄准了人工智能领域的微创新主体，通过构建“社区+词元服务”的生态，为词元工厂开辟了服务于海量、小微、个性化需求的新市场，推动了产业生态的繁荣。

综上所述，国家部委层面的政策已从概念认可、基础设施互联、服务能力要求、市场激励与生态培育等多个维度，构建了一个相对完整且相互支撑的监管与引导框架。这一框架的核心在于，将“词元”确立为人工智能时代可标准化计量和交易的价值载体，并鼓励建设能够高效生产、调度和交付此类载体的新型基础设施——词元工厂，从而打通从数据要素到智能应用的价值转化通道。

2.2. 数据要素与人工智能专项政策影响

词元工厂的蓬勃发展，深度受益于国家关于数据要素市场化与人工智能产业化的专项政策。这些政策不仅为词元工厂提供了丰富的“原料”（数据）和明确的“产品”方向（智能服务），更通过构建“数据飞轮”等创新机制，将词元工厂置于价值创造的核心环节，使其从简单的算力提供者升级为驱动产业智能化的引擎。

国家数据局发布的《关于推进行业高质量数据集建设行动的实施方案》是理解政策影响的关键文本。该方案系统性地描绘了从数据供给到价值释放的完整闭环，而词元工厂正是实现这一闭环的核心载体。方案提出的“数据飞轮”概念，即“场景牵引数据、数据驱动模型、模型赋能应用、应用创造价值”的良性循环，深刻揭示了未来数据生态的运行逻辑^[13]。在这一逻辑下，词元工厂的角色是“数据飞轮”的运营者和加速器。它负责将政策鼓励建设的“行业高质量数据集”加工为标准化的、可交易的“词元”，并通过高效的调度网络，将这些词元输送给千行百业的模型与应用，最终产生的价值增

量又能通过交易机制反馈给数据提供方，从而激励更多高质量数据的持续供给^[13]。这种动态的、价值共创的商业模式，彻底改变了传统一次性数据售卖或项目制服务的静态模式。

该方案在供给侧极具前瞻性地列出了近20个重点与创新领域，要求加快建设行业高质量数据集，如智能驾驶、智慧能源等^[13]。这不仅仅是简单的领域列举，更是为“词元化”交易指明了最具潜力的垂直赛道。这些领域数据价值密度高、应用场景明确，是生成标准化、高价值数据词元产品的最佳试验田，从而为面向特定行业的垂直型词元工厂提供了清晰的市场指引。

在数据治理与合规层面，专项政策为词元工厂的合规运营扫清了障碍。方案明确要“研究合成数据等新情况新问题”并“完善人工智能训练阶段数据使用规则”，直接回应了AIGC时代最紧迫的数据版权与合规性质疑^[13]。同时，方案要求“落实数据持有权、使用权、经营权三权分置制度”，并“完善数据授权使用机制和收益分配规则”^[13]。这些制度安排为词元工厂合法获取、加工和使用多源数据，并清晰界定各方权益与收益提供了法律基础。例如，振途信息科技有限公司在云南肉牛产业的实践中，正是通过AI大模型完成数据分层词元拆解，并凭借数据产品知识产权登记证书实现合规存证，最终拿下全国首单牧业数据知识产权质押融资^[5]，这生动体现了数据要素政策与词元经济实践的紧密结合。

人工智能专项政策则从需求侧和应用侧为词元工厂创造了巨大的市场空间。各地出台的“人工智能+”政策，均将支持企业调用词元、使用算力作为核心激励措施。例如，贵州省发布的20条专项扶持政策中，明确“支持企业词元调用”，对符合条件的词元调用给予补贴，每家企业最高可获得百亿级词元且不超过50万元的词元补贴^[17]。同时，政策“支持企业购买贵州算力”，对购买智能算力服务的需求主体，按服务金额的30%给予激励^[17]。这类“补需方”的政策，显著降低了企业使用AI服务的门槛，直接拉动了词元工厂产出服务的需求，为词元工厂的规模化、商业化运营提供了稳定的市场保障。

2.3. 行业协会与地方政府的配套举措

在国家顶层设计的指引下，各级地方政府和行业协会积极发挥主动性与创造性，出台了一系列具有地方特色和行业针对性的配套政策与举措。这些举措不仅将国家战略在地方层面做实、做细，更通过差异化探索和先行先试，丰富了词元工厂的发展模式与应

用场景，形成了“中央统筹、地方创新、百花齐放”的良好局面。

省级政府层面，广东省的举措具有标杆意义。广东省政务服务和数据管理局于2026年6月印发了《广东省释放数据要素价值 促进词元经济高质量发展若干措施》，提出了系统性的六大措施。该文件目标明确，提出“到2027年底...广东成为全国词元供给高地、智能应用标杆和词元出海枢纽”^{[18][20]}。其措施直接指向词元工厂的核心环节：一是“夯实基础设施建设，打造协同高效的‘词元工厂’”，强调形成“场景—数据—模型”协同发展的“数据飞轮”应用闭环^[18]；二是“建立健全词元交易体系，促进词元合规高效流通”，旨在建立标准化交易规则和建设省级词元交易专区^[18]；三是“畅通国际数据跨境通道，探索词元出海新模式”，依托粤港澳大湾区的政策优势，打造词元出海枢纽^[18]。广东的政策体现了将数据要素优势、产业场景优势转化为词元经济胜势的清晰思路。

其他省份也结合自身资源禀赋，推出了特色化的发展路径。陕西省通过省数据和政务服务局统筹协调，推动陕西丝路数据交易中心与浪潮云签署协议，共建“丝路词元工厂”^[21]。该工厂旨在统筹智算调度、数据治理、模型训练全链条能力，并先行探索词元标准化定价、交易结算机制，其突出特点是服务于“丝绸之路经济带数字化转型”，具有鲜明的区域战略导向^[21]。云南省则依托独特的“绿电”与“跨境”优势，推动运营商构建差异化算力网络。云南电信通过“算电协同”，将绿电算力接入集团“息壤”算力网，并将成本优势体现在普惠定价上，个人词元套餐低至9.9元/月^[5]。云南移动则依托昆明国际数据交易平台与智算中心，打造了千卡级智算集群和“Token工厂”平台^[5]。这些案例表明，地方政府正将词元工厂建设与本地能源、区位、产业等优势相结合，打造差异化竞争力。

在城市与区县层面，政策更加聚焦于生态培育与应用落地。武汉市硚口区发布了全市首个针对Token经济发展的人工智能政策——《硚口区促进Token经济和人工智能OPC发展三年行动方案及若干措施》^[22]。该方案以“Token普惠共营”为核心引擎，创新推出十条措施，包括对OPC企业算力服务费用按50%补贴、每年免费提供50P以上算力、推行数据券机制等^[22]。同时，该区推动“星火众达Token工厂”和“CSDN龙虾工厂”落地，并集中签约了首批16个特色OPC社区，形成了“政策+工厂+社区”的微生态闭环^[22]。北京市则通过设立“OPC开办服务专区”、发放“Token券”、开发“OPC贷”专属信贷产品等方式，全力支持人工智能OPC创新发展，为词元工厂培育了海量的潜在小微客户群^[16]。

行业协会与产业界通过座谈会、联盟等形式，积极反馈行业诉求，参与政策与标准

共建。例如，广东省政务服务和数据管理局组织召开词元经济座谈会，华为、腾讯云、云天励飞、广州数据交易所等企业代表参会并提出建议^[23]。华为建议“依托公共数据平台搭建省级数据智能底座，建立词元质量评估标准”；腾讯云建议“广东率先制定词元定义、计量、结算、质量四类地方标准”；云天励飞建议将“‘本土推理芯片+词元工厂’一体化建设纳入省级产业统筹规划”^[23]。这些来自市场一线的声音，正在被吸纳进地方政府的政策制定中，推动形成更贴近产业实际、更有利于创新发展的政策环境。

2.4. 官方媒体舆论导向与行业规范动态

官方媒体的报道与舆论导向，以及行业内部在标准、规范上的动态，共同塑造了词元工厂发展的社会认知与市场环境。这些动态在普及概念、树立标杆、引导预期、凝聚共识方面发挥着不可替代的作用，为词元工厂从技术概念走向产业实践营造了积极的氛围。

官方媒体通过广泛报道词元经济的实践案例，有效普及了概念并凸显了其战略价值。央广网、地方政务门户网站等权威媒体对“丝路词元工厂”、“昆明词元经济”等案例进行了详细报道^{[21][5]}。这些报道不仅介绍了业务模式，更强调了其战略意义。例如，报道指出“丝路词元工厂”对“推动数字技术深度嵌入实体经济各领域，以普惠智能服务助力丝绸之路经济带数字化转型”具有重要意义^[21]；将昆明运营商的词元套餐描述为让“AI像水电一样走进全民日常”，突出了其普惠性^[5]。这种舆论导向将词元工厂从企业商业行为，提升至服务国家战略和民生普惠的层面，为其发展赢得了更广泛的社会认同与政策支持。

行业分析报告与专家解读则从专业角度深化了对词元工厂内涵与趋势的认知。深度研究报告明确指出，词元工厂是“人工智能时代将电力与数据转化为词元（Token）的算力数据中心或基础设施”，其本质是“将底层复杂的芯片、模型和调度能力，封装成标准化的Token服务”^[24]。报告还梳理了企业主导的词元工厂多元化格局，包括互联网科技巨头、电信运营商、专业算力服务公司等^[24]。专家观点则被媒体引用以提升分析的权威性。数据交易网创始人张瑶指出，“词元交易的推进，是我国在全球AI产业角逐中抢占主动权的关键举措...其交易规则与定价体系的话语权，直接决定一国在全球AI赛道上的位次”^[25]。中国人民大学国际关系学院教授王义桅认为，词元出海的底气源于“整套基础设施体系的系统性优势”^[25]。这些深度分析有助于市场参与者、投资

者和决策者更准确地把握词元工厂的产业定位与价值。

在行业规范与标准建设方面，虽然统一的国标尚未出台，但地方和行业已开始积极探索。前文提到的企业座谈会上，腾讯云、华为等公司提出的制定词元定义、计量、结算、质量等地方标准的建议，反映了行业对标准化建设的迫切需求^[23]。同时，政府主导的词元交易平台的出现，本身就是规范化、规模化流通的实质性进展。例如，内蒙古词元交易平台作为自治区首个专业化、规范化词元交易平台，采用政府引导、国企主导、市场运营模式，并在揭牌仪式上完成了首批合规词元交易签约，实现了词元交易“零的突破”^[24]。这类平台的建立和运营，为词元作为新型生产要素的规范化流通提供了重要的基础设施和规则试验场。

官方媒体与行业动态也密切关注并探讨发展中的挑战，引导行业健康前行。报道中不仅展示成绩，也客观反映问题。例如，提及词元出海面临“跨境金融配套不足问题”，需要金融机构搭建合规高效的跨境资金通道^[25]；指出行业存在“词元定价机制不完善、流通标准缺失”等瓶颈^[23]；分析认为通往万亿级市场仍需跨越标准化、安全信任与生态共识三道关键门槛^[13]。这种全面、客观的舆论环境，有助于产业理性看待机遇与挑战，避免盲目跟风，推动词元工厂行业在解决实际问题中实现扎实、可持续发展。

3. 词元工厂的核心构成要素与技术路径

词元工厂作为AI时代的新型基础设施，其核心价值在于将异构算力、复杂模型与工程化能力整合为稳定、高效、可计量的Token产出服务。其技术架构并非单一产品或平台的简单堆叠，而是一个分层解耦、协同优化的系统工程。本章将深入剖析词元工厂从底层基础设施到上层原生服务的完整技术栈，揭示其实现“高效、低成本、规模化”Token生产的关键技术路径。

3.1. 基础设施层：算力供给、存储与网络

基础设施层是词元工厂的物理基石，其核心任务是为大规模、高并发的AI推理任务提供稳定、高性能且成本可控的算力、存储与网络资源。这一层的优化直接决定了Token生产的规模上限与基础成本。

算力供给的异构化与池化是当前词元工厂的显著特征。随着国产AI芯片的成熟与多样化，单一依赖某类芯片的架构已无法满足性价比最优的需求。领先的词元工厂普遍构建了兼容多元芯片的异构算力池。例如，是石科技的“拓元”国产Token优化工厂全面兼容昇腾、昆仑芯、天数智芯、太初、瀚博半导体、摩尔线程、沐曦、燧原等超过10种国产算力芯片，并适配超过20个主流模型，实现了国产异构算力池的统一调度与推理加速^[26]。商汤大装置则通过自建超大规模AIDC智算中心，具备同时改造硬件组合、网络架构和软件系统的能力，尝试不同国产芯片与高端芯片的异构组合，以寻找系统最优解。这种异构池化策略，使得企业客户无需关心底层具体芯片型号，算力资源能被统一管理、优化调度，最终产出更多可用Token^[26]。

存储架构的创新是突破“显存墙”瓶颈的关键。传统依赖GPU高带宽内存（HBM）存储KV Cache的方式，在面对百万级长上下文和超高并发推理时，成本高昂且容量受限。业界正通过“以存换算”的架构创新，构建分层的异构存储体系。核心思路是将KV Cache根据访问频率智能调度到不同层级的存储介质中。例如，将活跃数据保留在GPU显存，将不活跃的、较早生成的KV Cache块卸载到主机内存（DRAM），甚至将最冷端数据卸载至本地NVMe SSD固态硬盘或远端分布式存储系统^[28]。焱融科技的YRCache推理存储系统即构建了GPU显存、主机内存、本地NVMe SSD及高性能分布式存储的多级缓存架构，旨在显著扩展KV缓存空间，打破显存限制。华为OceanDisk 1800智能盘框则以DPU为核心，构建原生KV语义的存储加速架构，通过DPU硬件卸载存储管理服务，并与NPU/GPU通过参数网络直连，将访问路径压缩，实现大容量KV Cache池的高效供给。芯展速的AI90方案则构建了HBM+DRAM+SSD三级存储架构，通过将KV Cache卸载至大容量SSD，实测显存占用率降低39%。这些方案共同推动了从“计算驱动”向“存储驱动推理”的新范式转变。

高速互联网络是保障异构资源协同工作的血脉。无论是跨多GPU卡、跨多服务器节点，还是跨地域数据中心的算力调度，低延迟、高带宽的网络都是性能保障的前提。超聚变的TokenBox™企业级Token生产平台与博通联合打造了全球首款PCIe Fabric Gen6产品TokenFabric™，支持从4卡到128卡的全互联扩展，旨在彻底绕开传统CPU中转带来的延迟损耗^[32]。阡视科技的wylon超节点系统除了实现所有AI芯片无延迟互通，还新增纵向存储贯通架构，打通不同层级存储硬件。在跨地域层面，无问芯穹联合中国电信完成了国内首例超4000公里距离的大模型跨域混训，验证了超远距离跨域集群协同训练的可行性^[34]。上海AI实验室DeepLink团队开源的DLSlime高效通信库，兼容RDMA、NVShmem、NVLink、HCCS等多种高速通信链路，其通信性能在部分场景带宽利用率超过97%，为大模型异构算力混合推理提供了核心通信支撑。这些网络

技术的突破，使得算力资源能够像电网一样被灵活调度与聚合。

3.2. 模型优化层：推理加速与成本控制关键技术

在稳固的基础设施之上，词元工厂通过一系列模型层的优化技术，将原始算力高效转化为Token产出，其核心目标是提升吞吐量（Throughput）、降低延迟（Latency）并优化单位Token的综合成本。

推理阶段解耦与异构混合推理是近年来最重要的架构创新之一。传统单体推理架构中，计算密集的Prefill（预填充）阶段和访存密集的Decode（解码）阶段共享同一计算资源，相互干扰。PD分离架构将这两个阶段拆分到不同的计算节点或集群。商汤大装置、无问芯穹、DeepLink等厂商进一步提出了异构混合推理方案，即利用不同芯片的特性差异，将Prefill任务调度到高算力芯片集群，将Decode任务调度到高带宽/大显存芯片集群^{[27][34][35]}。例如，无问芯穹首创的跨集群异构PD分离架构（Prefill-RelayDecode-MainDecode, PDD），用高性价比的广域网以太网连接各地“偏科”的同构数据中心，让算力强的集群做Prefill，显存带宽高的集群做Decode，实测在实现首Token延迟降低51.5%的同时，单Token成本可降低37.5%^[34]。这种“让擅长游泳的运动员只游泳”的策略，实现了系统整体效率的显著提升。

KV Cache的精细化管理与压缩是降低显存压力和推理成本的核心技术。随着上下文增长，KV Cache规模指数级扩张。优化技术包括：1) 动态管理：如受虚拟内存分页启发的PagedAttention技术，将KV Cache划分为块进行管理，实现近乎零浪费的碎片化内存利用^[28]。2) 选择性缓存与逐出：通过计算注意力得分识别并保留重要的KV，逐出冗余部分^[28]。3) 共享与压缩：在批处理中共享多个请求的相同提示词前缀KV Cache；或采用压缩技术，如是石科技的“结果感知驱动KV Cache压缩技术”，可以在不牺牲模型效果的前提下显著压缩KV Cache、降低显存压力^[26]。4) 全模态Token压缩：针对文本、视频、语音等不同输入，实现基于模态自适应的免训练Token压缩方法，保留关键Token，减少无效计算^[26]。新华三基于高性能AI服务器平台，融合合作伙伴方案打造的KV Cache加速实践，在DeepSeek-V3-671B模型下，使首Token延迟（TTFT）下降70%，每个Token生成的平均延迟（TPOT）下降30%^[28]。

长上下文与持续推理优化是满足智能体等复杂场景需求的关键。企业知识库、智能助手、复杂分析任务要求模型具备超长记忆和稳定执行能力。是石科技通过混合位置索引合成的长上下文偏好训练方法，以及记忆引导的重读机制，在远小于同类方法的训练

数据量下实现了长上下文处理能力的显著提升^[26]。针对智能体长程任务，其“智能体长程任务记忆机制”帮助模型实现更低训练成本、更高上下文理解能力和更强复杂推理能力，确保长程Agent任务的稳定执行^[26]。此外，“基于元奖励的可扩展奖励建模方法”使得系统能够理解企业标准，在复杂业务场景中持续输出稳定、可靠的结果，从“需要不断人工训练的AI”变为“能够持续自我优化的AI系统”^[26]。

模型量化与高效推理框架是提升效率的通用手段。将模型权重从FP16/BF16精度量化至INT8/FP8甚至FP4，能大幅减少显存占用和计算量，提升吞吐。浪潮信息联合腾讯的金融行业方案显示，通过模型量化技术，FP8精度可降低50%的机器开销，FP4精度可降低75%的机器开销。同时，自研或深度优化推理框架（如vLLM, SGLang, TensorRT-LLM）以及与之配套的算子库、编译优化技术，对于释放硬件算力至关重要。例如，腾讯云TokenHub通过全异步推理流水线、自研FlexKV框架及训推一体调度，实现推理效率提升100%、缓存命中率提升至85%以上、综合降本40%^[37]。SmartX AI基础设施也融合vLLM、SGLang等推理引擎能力，面向长上下文和多轮对话场景提供高吞吐、低延迟的推理服务^[38]。

3.3. 平台服务层：调度、部署、监控与管理

平台服务层是词元工厂的“操作系统”，它将底层异构资源和上层优化技术封装成稳定、易用、可观测的服务，其核心能力包括智能调度、统一部署、全链路监控与精细化治理。

智能调度与资源编排是平台层的核心大脑。它需要根据实时负载、任务特性（如Prefill/Decode）、成本约束和服务等级协议（SLA），动态地将计算任务分配给最合适的算力资源。贵州移动的“词元工厂”实践路径清晰地展示了这一点：底层将智算服务器、存储、网络、安全等资源纳入统一资源池；通过云管平台实现资源纳管、监控告警和调度编排^[8]。派欧云（PPIO）的智能模型网关具备动态路由策略，将简单查询导向轻量模型，复杂推理分配至高性能模型，配合上下文压缩、计算结果复用等技术，实现质量不变前提下成本降低50%-60%^[3]。联想问天的万全异构智算平台V5.0则通过集群训推加速与芯模编译优化两大核心技术，适配多元国产与通用算力芯片，实现从百卡至万卡全场景的稳定支撑与智能调度。

统一模型服务与部署管理降低了模型使用的复杂性。平台需要统一接入和管理各类公有、私有及开源模型，提供标准化的API。SmartX AI基础设施的模型服务层统一

管理各类模型，支持模型快速上线并提供稳定可靠的模型服务，通过内置模型网关统一内外部模型调用入口，实现用量统计、配额管理、访问控制和访问日志等精细化管理^[38]。中科闻歌的海汇TokSea平台整合了自研模型与DeepSeek等主流通用模型，形成“通用能力+科学智能+决策智能”的企业AI能力组合，企业通过一套API和账号体系即可使用多类模型。超聚变TokenBox™引入的ModelEver模型永新能力，提供模型预验证、预集成工作，用户可第一时间获取精选模型及其配套工具镜像，实现新模型的可视化部署、评测和平滑升级^[32]。

全链路可观测与治理是保障服务稳定性和合规性的基石。可观测性需要覆盖从硬件资源到模型调用的每一个环节。SmartX AI基础设施的可观测治理覆盖GPU资源消耗、推理实例性能、调用日志和Token用量等指标^[38]。腾讯云Agent Runtime实现100% Session全留痕，支持任务追踪、性能分析与优化^[41]。治理则包括成本控制、安全审计和合规性管理。盛邦安全的AI安全网关（RayToken）提供从部门/用户/模型/应用多维度用量统计与分析，彻底告别“AI算力黑盒”，同时构建覆盖输入、内容、行为全链路的AI安全屏障，包括数据防泄漏、提示注入防护、内容合规检测等。腾讯云TokenHub提供从预算管控、权限隔离到合规审计的全链路可视化看板，解决企业深入应用大模型时的治理难题^[37]。

企业级集成与API网关是连接企业现有IT系统的桥梁。词元工厂需要提供安全、高效的API网关，处理认证、鉴权、限流、计量等通用功能。阿里云IDaaS EIAM的Agent ID Guard通过SDK支持Agent以机器或用户身份安全获取下游服务（如大模型API、第三方SaaS）的API Key与OAuth Token，实现凭证的集中托管与动态下发，避免凭证散落与泄露风险^[43]。亚信安全发布的智能体信任框架ATF，围绕“身份可信、意图对齐、生成有界、行为可控、链路可审、责权可溯”六大原则，为智能体建立运行治理平面，确保任何智能体行为都在“有身份、有边界、有记录、有责任”的前提下进行。

3.4. 面向AI智能体的原生服务能力构建

随着AI应用从简单的对话交互向能够自主执行复杂任务的智能体（Agent）演进，词元工厂的基础设施也需进行相应演进，构建原生于智能体工作负载的服务能力，以支撑其“跑得快、跑得省、跑得放心”。

智能体专属运行时与沙箱环境是保障智能体隔离、安全与弹性的基础。智能体任务

具有长周期、多步骤、工具调用频繁等特点，传统应用运行时难以满足。腾讯云推出的Agent Runtime是面向AI智能体场景的原生基础设施，其核心模块“Agent沙箱”采用“一沙箱、一任务”模式，提供会话独占的运行环境，支持毫秒级启动（60ms）、极致弹性伸缩（1分钟可拉起16万沙箱）与大规模并发（100万级）^[41]。派欧云开发的Agent沙箱基于microVM技术，支持毫秒级冷启动，每个任务在独立虚拟机中运行，防止数据泄露和资源争抢，其独创的Auto Pause/Resume机制使空闲任务自动暂停，恢复时瞬间唤醒，较传统方案成本降低超90%^[3]。这些沙箱环境为智能体提供了安全、隔离且资源高效的执行场所。

持久化记忆与状态管理是智能体完成长程任务的关键。智能体需要在多轮交互和长时间执行中保持上下文和状态。腾讯云Agent Runtime的“Agent存储”模块提供快照/回退、任务自动打快照、一键恢复能力，支持任务暂停时不计费，唤醒即用，显著降低Agent任务闲置成本^[41]。这解决了智能体任务中断后需重新执行、成本高昂的痛点。蚂蚁数科的商业智能体超级工厂平台Agentar 2.0，通过预置近200个岗位级数字专家模板和数百个可订阅的Skills级工具，将企业经验沉淀为可复用的智能体能力，本质上也是一种组织级的记忆与知识管理^[45]。

工具调用与外部系统集成能力是智能体延伸其行动范围的核心。智能体需要能够安全、可控地调用外部API、操作软件、访问数据库。派欧云Agent Harness层内置Browser Use模块允许智能体自主操控浏览器，Computer Use功能实现桌面环境的GUI自动化操作，并通过MCP Server协议无缝接入飞书、数据库等外部系统^[3]。明心数智的Ming Harness平台深度适配跨境行业高并发、多地区合规、长链路系统集成特性，内置跨境财税、物流追踪、多平台API对接等垂直领域专属代码模板。阿里云IDaaS EIAM的凭证托管与动态下发机制，则为智能体安全调用下游服务提供了基础设施级的支持^[43]。

多智能体协作与编排平台是应对复杂业务场景的必然要求。当单个智能体无法完成任务时，需要多个智能体分工协作。蚂蚁数科主张通过区块链技术为每个Agent赋予可信身份，将决策、协作及资金流转等全流程记录上链，促进多智能体的可信协同^[45]。迈富时的智能体中台AI-Agentforce具备多智能体协同（MAS）能力，支持多个智能体分工协作完成复杂业务链路，并通过全生命周期管理实现智能体的搭建、部署、监控、归因和成本核算。字节跳动的扣子（Coze）3.0版本也支持多人多Agent协作，提供了智能体团队化运作的平台基础^[48]。

低代码/零代码的智能体开发与部署是降低使用门槛、加速AI应用落地的关键。

通过可视化编排，让业务人员也能参与智能体创建。Dify平台允许用户通过拖拽式界面配置 workflow、智能体人设与回复逻辑，无需编写复杂代码即可构建AI应用，并与EdgeOne Pages等部署平台结合，实现从开发到全球部署的快速闭环。扣子（Coze）平台同样通过低代码或零代码的可视化方式，降低AI应用开发门槛，让用户快速创建、调试并部署AI聊天机器人、智能体及自动化 workflow^[48]。这些平台使词元工厂产出的Token能力能够被更广泛、更便捷地消费，真正赋能千行百业。

4. 中国词元工厂典型案例与实践分析

4.1. 按运营主体分类：国有企业、民营企业与电信运营商

中国词元工厂的产业生态呈现出多元化主体共同参与的格局，主要可分为国有企业、民营企业以及电信运营商三大类别。不同运营主体凭借其独特的资源禀赋、战略定位和商业模式，在市场中形成了差异化的竞争态势。

国有企业在词元工厂建设中扮演着“国家队”和“压舱石”的角色，其核心优势在于安全可信、资源整合与政策引导。北京数据集团旗下京算公司发布的“京算Token工厂”是北京市首家国资词元工厂，其定位是构建安全合规的算力新底座，以满足国央企及大型企业对“稳定、安全、可信”的底线要求^[11]。该工厂具备七大核心能力，包括PD分离智能调度架构、全栈推理加速框架、异构算力统一调度、全面适配国产大模型生态、精准支撑多元落地场景、弹性计费按需取用以及标准化统一接口^[11]。与传统的“卖硬件”模式不同，京算Token工厂交付的是标准化的Token和API成品能力，客户按实际消耗付费，平台承担底层硬件和集群运维风险，旨在降低企业AI使用门槛和综合成本^[11]。另一典型案例是协鑫能源科技股份有限公司在宁夏中卫落地的规模化Token工厂，该项目分三期建设，旨在填补西北地区规模化Token标准化供给的空白^[50]。项目选择中卫的核心考量是其显著的清洁电力优势，中卫市新能源装机规模达1812万千瓦，年发电量超160亿千瓦时，双方将共同建设新能源算电一体化项目，致力于打造全国算电协同支撑Token算力的示范工程^[50]。国有企业的布局往往与区域发展战略紧密结合，例如中卫项目的落地被视为巩固全国绿色算力标杆优势、培育新质生产力的关键布局。

民营企业是词元工厂技术创新和商业模式探索最为活跃的力量，覆盖了从基础设

施、优化平台到行业解决方案的全链条。软通动力旗下的“北京壹号词元工厂”是其“词元工厂计划”的首个标杆示范项目，一期规划每日Token产能达1.4万亿^[10]。软通动力认为，Token工厂的核心是面向Token时代的工厂级操作系统，其业务逻辑在于由模型厂商负责“造脑”，Token工厂则负责把模型能力转化为可规模化、具备SLA保障、可计量的智能服务^[10]。在技术层面，该公司从芯片算力适配、分布式通信到调度策略进行端到端调优，并已适配主流大模型，建立了面向真实场景的服务质量评测标准^[10]。是石科技发布的国产Token优化工厂“拓元”（Vectron）则代表了技术优化服务商的路径^[52]。其核心逻辑是“一套系统完成AI推理全链路优化”，在任务自适应、算力库、模型与推理框架、异构集群调度四个层面进行深度优化，旨在同等算力投入下产出更多有效Token，同等大模型部署下实现更快的推理响应^[52]。“拓元”已兼容10余款国产算力硬件，可适配20余个主流模型，每日Token吞吐量达到千亿级，服务客户涵盖互联网企业、头部大模型厂商以及高端制造、航空航天、生物制药、新能源等行业^[52]。硅基流动作为第三方MaaS平台的代表，其定位是Token工厂的最早践行者，致力于提升Token生产效率并通过规模化生产实现技术普惠^[53]。该公司通过自研推理引擎、异构算力纳管与调度平台，以及超160款模型的快速适配能力，构建了世界领先的全自研Token生产线^[53]。根据第三方数据，硅基流动在国内公有云MaaS市场份额位列前四，是其中唯一的创业公司；在OpenRouter平台上，其日Token消耗量曾连续数周位居第一^[53]。此外，超聚变发布的TokenBox™企业Token生产平台，采用软硬一体设计，将超节点能力带入企业业务现场，单机可支持满血版DeepSeek V4 1.6T旗舰大模型，旨在让每家企业都能快速拥有自己的Token Factory^[54]。智微智能联合腾云智算、元川微发布的“Token工厂”，则打造了从算力硬件与高质量数据投入，到AI模型训练部署加工，再到实时Token推理生产的全链条闭环，直接向客户交付规模化、标准化的“智能输出”^[7]。

电信运营商凭借其覆盖全国的骨干网络、庞大的用户基础、云网融合能力以及丰富的存量数据中心资源，正成为词元经济生态中至关重要的组织者和服务提供商。中国移动、中国电信、中国联通均已明确提出从传统“流量经营”向“Token经营”转型的战略。运营商入局Token服务，被行业视为Token工厂商业化的重要标志，意味着AI产业链将从“算力租赁”逐步演进为“Token生产—Token运营—Token消费”的新商业模式^[56]。

中国电信构建了以“星辰TokenHub”运营服务平台为核心的一体化Token经营体系^[57]。该平台具备多模型聚合与智能路由能力，为智能体开发直供最具性价比的Token，并纳管自研与生态智能体，实现端到端交付^[57]。在落地层面，中国电信推出

了面向个人和开发者的商用Token套餐。例如，云南电信的个人套餐低至9.9元/月，包含1000万词元，企业套餐39.9元起^[5]；上海电信则推出按量订购服务，1元对应25万额度点（以Kimi-K2.5模型为例）^[57]。中国电信还启动了大规模的“Token工厂”生成能力服务集中采购，项目整体预估规模达164.51亿元（不含税），显示出其将Token经营从概念推向实质落地的决心^[58]。

中国移动将Token定义为连接算力、模型、应用与用户的“通用货币”，致力于构建贯通“能力底座、运营平台、应用入口”的Token生态运营体系^[57]。其省公司积极先行试点，例如云南移动推出AI算力服务平台“Token工厂”，依托千卡级智算集群、绿电液冷算力支撑与全栈AI生态优势，搭建了“规模化生产、普惠化定价、轻量化使用”的全链路Token供给体系^[5]。北京移动发布了“一站式Token运营服务平台”，整合多元算力和模型资源，用户一次接入即可调用平台上的全部模型资源，并通过自研推理加速引擎实现成本压降约30%^[59]。中国移动浙江公司联合中国移动云公司打造的“浙南Token工厂”，则重点构建“算力底座、模型超市、运营服务”三大能力，依托中国移动MoMA模型服务平台聚合300多款主流大模型，为企业提供稳定的Token调用服务和按量实时计费^[60]。

中国联通通过“联通星罗”Token服务平台，依托算网一体与九位一体全栈智算底座，为全域客户提供一站式Token服务^[61]。其创新构建了“Agent+Token+AI云”一体化服务范式，提供“算力+数据+模型+Agent”全栈定制化服务^[61]。联通数智的元景MaaS一站式Token运营平台，则打造了算网模协同、模算融合、模型调优和多模智选四大能力，旨在强化高性价比Token供给，并向上搭建完整的Token运营管理和用户服务能力^[62]。

4.2. 按服务领域分类：政务、工业、金融、医疗等垂直行业

词元工厂的价值最终通过赋能千行百业得以体现。不同垂直行业因其业务特性、数据形态、安全要求和智能化场景的差异，对词元工厂的服务提出了个性化需求，也催生了众多聚焦特定领域的行业级解决方案和实践。

政务领域是词元工厂落地的重要场景，核心诉求是安全可控、高效普惠与治理能力提升。云天励飞在广东湛江打造的全国首个面向政务的“国模+国芯Token工厂”是典型代表^[63]。该工厂基于自主研发的GPNPU推理芯片和“算力积木”创新架构，支持DeepSeek等千亿级大模型的部署，并已赋能政务、制造、应急、医疗、海洋、教育等

多元场景^[1]。其目标是构建自主可控的技术底座，为区域人工智能产业发展提供算力支撑^[1]。在政务应用层面，云天励飞在深圳龙岗区打造了自进化城市治理智能体“龙管家”，推动城市治理从人力密集、经验驱动向知识驱动、AI研判全面跃升^[1]。云南电信的Token服务则在政务民生场景落地了标杆案例，如全省“12345热线”全面接入AI能力^[5]。深圳市福田区推出的“福小i”智能服务体系，通过AI赋能实现了政务服务从“人找服务”到“服务找人”的转变，例如“AI视频办”累计办件超3万单，实现了跨城通办^[63]。

工业与制造业是Token经济与实体经济深度融合的主战场，需求聚焦于降本增效、质量控制与产业链协同。宁波作为制造业城市，其人工智能超算中心正从“建起来”向“用起来、转起来、融进去”的公共能力进阶，累计服务超110家企事业单位，覆盖智能制造、医疗健康等多个领域^[64]。该市通过城市级公共算力底座，旨在把AI能力做成像水、电、网络一样可调用的基础能力，实现稳定、低成本、高效率的Token供给^[64]。迅策科技与TCL旗下格创东智的战略合作，旨在共同探索工业智能制造领域的AI数据基础设施与“Token工厂”模式^[65]。双方将围绕泛半导体与新能源行业，落地工厂全局能耗优化、碳足迹追踪、设备预测性维护、多产线智能排产等高价值应用，并将工业核心数据处理、大模型调用等成果纳入按Token计价的结算体系^[65]。温州的中国移动浙南Token工厂，则围绕当地电气、鞋服、泵阀等产业基础，推动AI能力与企业生产、质检、销售、客服等环节结合^[60]。例如，帮助家电企业通过多模态识别客户图片，并联动知识库自动生成报价，提高销售响应效率^[60]。无问芯穹的Agentic Infra产品体系中的“AI生产力商店”，也专注于提供行业解决方案^[66]。无锡高新区与弘信电子、华为昇腾合作打造的Token工厂，明确聚焦“AI+制造”的差异化之路，旨在服务本地乃至长三角的AI企业，并鼓励制造业链主企业、中小企业和“超级个体”（OPC）利用Token工厂的普惠算力进行创新。

金融领域对Token服务的需求突出表现为对安全性、合规性、实时性和精准性的极高要求。虽然参考材料中未提供纯粹的金融领域词元工厂独立案例，但多家综合性词元工厂平台均将金融列为核心服务行业之一。例如，天云数据的TokenFactory与TokenRouter双核架构，特别强调企业级治理、安全策略与审计留痕，其能力天然适配金融行业对成本透明可追溯、风险管控严格的需求。中国电信、中国移动等运营商在推广其Token服务时，也普遍将金融列为重点行业，依托其安全产业链“链长”经验或机密计算能力，满足金融客户数据不出域、合规审计的核心诉求^[62]。

医疗健康领域的词元工厂实践，展现了AI如何通过标准化、可计量的Token服务，重塑传统医疗服务模式。武汉兰丁智能医学股份有限公司的实践尤为突出^[69]。该公司

在武汉市“汉数通”平台上完成了全国首张AI病理数据交易发票的开具，交易产品为30例妇女宫颈癌AI早筛报告，消耗3600万个Token^[69]。兰丁已拥有全球最大的宫颈细胞病理数据库，数据总量达1400TB，其AI大模型可快速完成诊断^[69]。这种模式将AI诊断报告转化为具有医疗服务和数据资产双重价值的AI医疗数据服务产品，实现了医院、数据企业及国有数据流通平台三方按贡献获得价值回报^[69]。云南移动的Token工厂也推出了AI科研平台（YUNNAN AI4S），为生物医药等本土重点产业提供一站式科研服务^[5]。四川电信的“智云四川”Token一体化服务，在医疗领域也已有典型应用落地^[70]。

其他垂直领域同样不乏创新实践。在教育领域，云南移动推出了AI办公助手、AI彩云智能体、企业数字员工以及心理健康教室等自研智能应用^[5]。在文旅领域，扎根云南的振途信息科技有限公司利用词元经济发展模式，打造“找数据”智能体，在肉牛产业、水务、医疗、文旅领域都接连落地了标杆案例^[5]。例如，楚雄110家酒店的异构数据，仅需十分钟即可完成清洗治理并生成完整数据产品报告^[5]。在科研领域，中科闻歌的海汇TokSea平台整合了其自研的磐石ScienceOne科学基础大模型，服务于科研与产业场景^[71]。范式智能则因其在工业AI领域的可靠性和落地实力，入选了“全球工业人工智能旗舰案例”，获评首批全球可信工业AI解决方案。

4.3. 按业务模式分类：通用基础设施、行业专用、AI智能体服务

从业务模式看，中国词元工厂的发展路径呈现出从提供通用基础设施，到深耕行业专用解决方案，再到原生支持AI智能体服务的演进趋势。

通用基础设施模式是词元工厂的基石，其核心是规模化、高效率、低成本地生产标准化Token，类似于云计算中的IaaS（基础设施即服务）。这类工厂不强调对特定行业知识的深度整合，而是专注于优化从算力到Token的转化效率。软通动力的“北京壹号词元工厂”、协鑫能源在中卫的规模化Token工厂、趋境科技的ATaaS（AI Token as a Service）平台均属此类^{[10][50][12]}。它们的共同特点是追求极致的Token产能（如日产能万亿级）和单位成本优化^{[10][12]}。硅基流动的公有云MaaS平台也是通用基础设施模式的代表，它通过统一的平台接口，向开发者和企业提供来自多种模型的标准化Token调用服务^[53]。电信运营商推出的各类Token套餐和运营平台，本质上也属于通用基础设施服务，它们将复杂的算力、模型资源封装成易于消费的Token产品，提供给广泛的

个人、开发者及企业用户^{[5][56]}。

行业专用模式（或称“垂类行业Token工厂”）则在通用基础设施之上，注入了特定行业的专业知识、工作流程和高质量数据，从而生产出能解决具体行业问题、具有更高价值的“专业Token”。长江证券的研报指出，垂类行业Token的定价可能从“成本定价”演变为“价值定价”，即按照其所能解决的任务价值进行定价，在专业性强、数据有壁垒的行业中具备高定价天花板^[56]。迅策科技与格创东智合作探索的“制造业Token工厂”、与博泰车联等探索的“智能网联汽车Token工厂”，正是这一模式的典型^[65]。它们将工业制造或汽车研发中的核心数据处理、算法调用成果纳入Token计价体系^[65]。武汉兰丁的AI病理诊断服务，则是医疗领域高度专业化的Token工厂实践，其产出的每一份AI诊断报告都是专业Token的具体体现^[69]。振途科技在云南肉牛产业搭建的“采集—词元拆解—确权入表—质押融资”闭环，也是将行业数据转化为可抵押资本的专业Token生产流程^[5]。九章云极的“AI工厂”战略明确区分了“训练工厂”（生产专业模型）和“Token工厂”（流通专业Token），其目标正是将通用智能“冶炼”为多场景的专业模型，并封装为可精确度量的专业Token进行交付。

AI智能体原生服务模式是随着AI智能体（Agent）的爆发而兴起的新模式。该模式不仅提供Token，更将Token的生产、调度与智能体的规划、执行、工具调用等能力深度集成，提供端到端的智能体服务或支撑平台。中国电信的“星辰TokenHub”运营服务平台强调纳管自研与生态智能体，将智能体与Token一并封装，实现端到端交付^[57]。四川电信推出的TeleAgent超级智能体，融合大模型、工具调用、RAG、多模态感知与自主规划等能力，实现办公任务的一体化执行^[70]。北京移动的“一站式Token运营服务平台”集成了自研的企业级智能体“龙虾” MobileClaw（企业版），用户可快速搭建专属AI数字员工^[59]。超聚变提出的智企进化路径，沿着“WATT-FLOPS-TOKENS-AGENTS-VALUES”价值链，将Token生产与智能体执行紧密连接，其TokenBox™平台旨在支撑智能体的稳定运行^[74]。无问芯穹的Agentic Infra产品体系，则直接以服务下一代Agentic AI在线进化为核心目标，其“算力集散中心”、“Token工厂”、“AI生产力商店”三大布局共同支撑智能体的持续进化^[66]。光创AI为可孚医疗提供的“数据存储+AI应用+企业服务”系统，也集成了AI功能，可视为面向企业智能体应用的基础设施。

4.4. 代表性案例深度剖析

为更深入理解中国词元工厂的多样性与内在逻辑，以下选取几个跨类别、具有鲜明特色的代表性案例进行深度剖析。

案例一：北京数据集团“京算Token工厂”（国有企业/政务领域/通用基础设施+行业专用模式）京算Token工厂的典型性在于其纯正的国资背景和清晰的服务定位^[11]。它诞生于首都数字经济核心区，直接响应了国央企及大型企业在AI时代对“稳定、安全、可信”算力底座的迫切需求^[11]。其七大核心能力设计，如PD分离智能调度解决长文本推理低效、全栈推理加速提升单卡Token产出效率、异构算力统一调度打破硬件生态壁垒等，均直指企业AI推理中的核心痛点^[11]。与大多数民营厂商不同，京算Token工厂的商业模式彻底摒弃了“卖硬件”或“卖机柜”的传统思路，转而交付标准化的Token和API成品能力，按实际消耗计费^[11]。这种模式将底层硬件采购、集群运维的风险和复杂性全部转移至平台方，极大降低了客户，特别是对技术运维能力有限的传统国央企的接入门槛和使用成本^[11]。同时，作为市属算力平台，它联合三大运营商及硅基流动、无问芯穹、是石科技等生态伙伴签署共建协议，旨在筑牢首都自主可控的数字算力底座，体现了国有资本在整合产业生态、引导技术路线方面的独特作用^[11]。

案例二：硅基流动SiliconFlow（民营企业/通用基础设施模式/第三方MaaS平台）硅基流动代表了在巨头林立的AI云服务市场中，以独立第三方身份成功卡位Token工厂赛道的创业公司路径^[53]。其核心优势在于“原子级”的工程优化能力和异构算力纳管调度技术^[53]。通过自研推理引擎融合PD分离、KV缓存管理等领先技术，硅基流动能够在英伟达、昇腾、沐曦、摩尔线程等多元芯片上稳定提供高吞吐、低延迟的商业级Token服务^[53]。这种深度技术优化使其在效率指标上具备了与头部大厂竞争的實力，这在Gartner、IDC等第三方市场报告以及OpenRouter、Dify Marketplace等实际平台数据中得到了印证^[53]。硅基流动的商业模式清晰地地区分了公有云MaaS和私有化MaaS，既满足了广大开发者和中小企业对便捷、低成本Token服务的需求，也满足了国央企、金融机构对数据合规、私有化部署的刚性要求^[53]。它的成功表明，在Token工厂这个确定性极高的AI底座赛道，技术深度、产品体验和生态中立性可以构成强大的竞争壁垒。

案例三：云南移动“Token工厂”（电信运营商/多垂直领域/普惠服务模式）云南移动的实践是电信运营商将Token经营与地方产业特色、资源禀赋相结合的优秀范本^[5]。首先，它充分利用了云南的“绿电”与“跨境”优势，通过“算电协同”将绿电

算力接入集团“息壤”算力网，从而获得成本优势并反馈在定价上，推出低至9.9元/月的个人套餐^[5]。其次，它依托属地化运维团队和省内专网，为用户提供低时延、高稳定性和安全保障的Token服务^[5]。在应用落地方面，云南移动没有停留在简单的模型调用，而是深度切入本地政务（“12345热线”）、文旅、教育、农业等场景，推出了AI办公助手、AI彩云智能体、企业数字员工“好龙虾”、心理健康教室以及面向生物医药和高原特色农业的AI科研平台等一系列自研智能应用^[5]。这种“普惠套餐+属地网络+深度场景应用”的组合打法，生动诠释了运营商如何利用其管道、用户和本地服务优势，将Token服务真正“送达”终端用户，并转化为可感知的生产力提升。

案例四：武汉兰丁AI病理Token服务（民营企业/医疗领域/行业专用模式）兰丁的案例展示了Token经济在高度专业化、强监管的医疗领域落地的完整闭环和价值重构^[69]。其核心是将AI病理诊断这一复杂服务，通过“汉数通”国有数据流通平台，拆解、计量并交易为标准化、可定价的Token产品（AI诊断报告）^[69]。这个过程实现了多重突破：一是价值显性化，国家Token标准使得AI劳动价值有了清晰的量化尺度，3600万Token对应30份报告，使得定价有据可依^[69]；二是流程合规化，通过国有平台交易，确保了数据来源、处理过程、结果交付的合规性与可审计性，为医疗AI服务的规模化推广扫清了制度障碍^[69]；三是生态协同化，医院、数据企业（兰丁）、平台方（武汉数据集团）三方按贡献参与价值分配，形成了可持续的商业生态^[69]；四是普惠可及化，AI检测量无上限，规模越大，单次检测成本越低，使得高质量宫颈癌筛查服务能够惠及更广泛的人群，尤其有利于医疗资源匮乏地区^[69]。兰丁的实践表明，行业专用Token工厂的成功，不仅依赖于强大的AI模型（拥有1400TB数据库和10亿级标注图像），更依赖于与行业监管规则、数据流通机制、商业支付体系的深度融合^[69]。

案例五：无问芯穹Agentic Infra与PDD架构（技术创新型民营企业/AI智能体原生服务模式）无问芯穹的案例凸显了技术驱动型公司在定义Token工厂未来架构方面的前沿探索^[66]。其提出的“前店后厂一中心”（AI生产力商店、Token工厂、算力集散中心）战略布局，系统性地回答了如何为Agentic AI时代构建基础设施的问题^[66]。其中最引人注目的是其创新的PDD（Prefill-Decode-Dispatch）三级分离式推理架构^[66]。该架构针对智能体推理“上下文极长、交互轮次极多、缓存命中率极高”的特征，以及不同芯片在Prefill和Decode阶段性能差异大的现状，通过高性价比的广域网以太网连接各地同构数据中心，让擅长Prefill的集群和擅长Decode的集群协同工作^[66]。其创新的“P-RLD-MD”设计，通过引入RelayDecode先行输出Token以掩盖传输延迟，再无缝切换给MainDecode，在实测中实现了首Token延迟降低51.5%的同时，单Token成本降低37.5%^[66]。无问芯穹宣称，随着该架构的规模化落地，推理成

本仍有10倍的下降空间^[66]。这一案例深刻说明，Token工厂的竞争不仅是规模和价格的竞争，更是底层架构创新和系统工程能力的竞争。谁能更高效地解决异构算力调度、长上下文处理、高并发推理等根本性技术难题，谁就能在未来的Token经济中占据制高点。

案例名称	运营主体	核心领域/模式	关键特点与能力	参考来源
京算Token工厂	北京数据集团 (国有企业)	政务/通用基础设施+行业专用	国资护航，七大核心能力（PD分离调度、全栈推理加速、异构算力统一调度等），按Token消耗交付标准化API，安全可信。	^[11]
硅基流动 SiliconFlow	硅基流动（民营企业）	通用基础设施/第三方MaaS平台	全自研Token生产线，原子级工程优化，异构算力纳管调度，超160款模型快速适配，公有云与私有化部署并举，第三方市场地位领先。	^[53]
云南移动Token工厂	中国移动云南公司（电信运营商）	多垂直领域（政务、文旅、教育、农业）/普惠服务	依托绿电与跨境优势，推出普惠套餐（9.9元/月起），属地化专网保障，深度开发本地场景智能应用（AI办公助手、农业科研平台等）。	^[5]

4.4. 代表性案例深度剖析

武汉兰丁AI病理服务	兰丁智能医学（民营企业）	医疗健康/行业专用模式	基于全球最大宫颈细胞病理数据库，通过国有“汉数通”平台完成AI病理数据Token化交易，实现价值显性化、流程合规化与生态协同化。	[69]
无问芯穹Agentic Infra	无问芯穹（民营企业）	AI智能体原生服务/技术创新	“前店后厂一中心”战略，创新PDD三级分离式推理架构，显著降低智能体推理时延与成本，致力于服务下一代Agentic AI在线进化。	[66]
软通动力北京壹号工厂	软通动力（民营企业）	通用基础设施	日规划产能1.4万亿Token，定位为面向Token时代的工厂级操作系统，端到端调优，适配主流模型，建立真实场景服务质量标准。	[10]

4.4. 代表性案例深度剖析

超聚变 Token-Box™	超聚变（民营企业）	AI智能体原生服务/软硬一体	企业Token生产平台，软硬一体设计，将超节点能力带入企业现场，支持旗舰大模型，打通模型部署、推理加速、智能体执行全链路。	[54]
浙南Token工厂	中国移动浙江公司（电信运营商）	工业制造/通用基础设施	构建“算力底座、模型超市、运营服务”三大能力，聚合300+款模型，为温州电气、鞋服等产业提供按量计费的Token调用服务。	[60]
迅策-格创东智制造业Token工厂	迅策科技 & 格创东智（民营企业）	工业制造/行业专用模式	联合探索工业数据底座与Token工厂模式，将产线数据、算法调用纳入Token计价体系，聚焦泛半导体、新能源行业高价值应用。	[65]

趋境科技ATaaS平台	趋境科技（民营企业）	通用基础设施	清华系团队，聚焦“少模型、深优化”，通过全链路系统工程提升高品质AI Token生产效率与稳定性，已有日万亿级产能项目投产。	[12]
-------------	------------	--------	--	----------------------

5. 国外词元工厂发展实践与典型案例

5.1. 北美地区：技术创新与商业模式前沿

北美地区，特别是美国，是全球词元工厂（Token Factory）技术创新与商业模式探索的前沿阵地。该地区的发展实践呈现出鲜明的技术驱动、资本密集和生态开放特征，其核心参与者不仅包括传统的科技巨头，也涌现出一批专注于AI基础设施与数字资产代币化的新兴企业。

在AI算力基础设施与词元生产领域，北美企业致力于通过底层技术创新来提升词元生产效率并降低成本。欧洲AI新锐Nebius（NASDAQ:NBIS）虽总部位于荷兰，但其技术核心与资本背景深度绑定北美生态，是行业“词元工厂”概念的标杆级产品[\[76\]](#)。Nebius旗下的“词元工厂”平台旨在为企业和开发者提供从模型部署、优化到量产推理的一站式解决方案，核心价值在于高效量产低成本词元[\[76\]](#)。为了强化其技术优势，Nebius于2026年5月宣布以约6.43亿美元现金加股票的组合收购AI推理与模型优化公司Eigen AI[\[76\]](#)。Eigen AI由麻省理工学院（MIT）HAN实验室校友创立，其团队主导开发的AWQ 4位量化、SpAtten稀疏注意力等技术能显著降低显存占用、提升GPU利用率，从而大幅压低单词元生成成本[\[76\]](#)。此次收购旨在吸收Eigen AI在推理性能优化方面的优势，让Nebius的词元工厂平台从通用托管推理平台跃升为全球生产级AI推理的标杆[\[76\]](#)。这一案例深刻反映了北美市场“技术并购驱动能力跃迁”的发展逻辑，即拥有庞大算力底座与资本加持的平台方，通过并购掌握顶尖底层推理优化算法与人才的

团队，快速构建核心竞争力^[76]。

与此同时，专注于数字资产代币化与稳定币支付的基础设施提供商在北美也形成了成熟的商业模式。以Zero Hash为例，该公司是一家总部位于美国的加密货币与区块链基础设施提供商，其定位是一个B2B嵌入式基础设施平台^[77]。它通过API为金融科技公司、银行及经纪商提供加密货币交易、稳定币支付、资产代币化等数字资产服务的合规技术解决方案^[77]。其业务模式的核心在于“嵌入式”与“合规化”：通过现代化的API和SDK，帮助传统金融机构和金融科技公司将加密产品无缝集成到自有应用程序中，同时凭借在美国51个司法辖区拥有的资金汇款牌照以及经美国金融犯罪执法局（FinCEN）注册的货币服务企业身份，解决合规门槛问题^[77]。其合作伙伴包括摩根士丹利（通过E*Trade）、盈透证券、TradeStation等知名金融机构^[77]。另一家美国数字资产基础设施公司I-ON Digital Corp则专注于实体资产代币化、受监管的黄金支持数字工具及数字资产银行服务^[78]。其平台旨在使机构能够在合规的财务框架内对实物及原地资产进行数字化、代币化、管理和分销^[78]。例如，该公司于2026年6月签署协议，收购位于美国西南部的20个BLM金矿开采权的权益，旨在扩大可纳入数字化体系的矿产资产储备，体现了其将实体资产与数字代币深度绑定的业务路径^[78]。

在支付与结算领域，基于稳定币的跨境支付基础设施创新是北美模式的重要延伸。全球最大的稳定币USDT的发行方Tether，通过投资拉丁美洲数字资产平台Parfin，旨在将USDT更深入地推向拉丁美洲的机构市场，并扩大该地区的链上结算能力^[79]。这笔投资凸显了Tether将USDT定位为高价值活动机构结算渠道的努力，应用场景包括跨境支付、现实世界资产（RWA）代币化，以及与贸易融资、商业发票和信用卡应收账款相关的信贷市场^[79]。这显示了北美数字资产基础设施提供商不仅服务于本土市场，更通过资本和技术输出，在全球范围内构建以美元稳定币为核心的支付与结算网络。

5.2. 欧洲地区：合规导向与行业应用

欧洲地区的词元工厂及相关数字资产基础设施发展呈现出强烈的合规导向和深入的行业应用特征。其发展路径深受欧盟统一监管框架的塑造，技术选择与业务模式紧密围绕《加密资产市场法案》（MiCA）、《数字运营弹性法案》（DORA）、《通用数据保护条例》（GDPR）等法规展开，形成了与美国市场不同的“监管制度主导，自上而下式技术演变”的路径^[80]。

在技术路径上，欧盟明确要求金融资产代币化必须可追溯、可审计、可撤销，且数

据必须存储在欧盟境内^[80]。这导致公链无法直接用于监管型资产的发行，因此欧盟选择了许可型分布式账本技术（Permissioned DLT）作为底层架构^[80]。由此形成了以受监管金融机构为主导、多中心联盟的体系，主要玩家包括Euroclear、LuxCSD和法国兴业银行旗下的Forge等^[80]。例如，由欧洲央行与Euroclear共同开发的Euroclear DLT Settlement System，基于R3 Corda企业链，用于主权债与可持续债券的链上发行与结算^[80]；卢森堡中央证券托管机构（LuxCSD）则主导了专为基金代币化设计的平台，使用Hyperledger Besu许可链架构^[80]；法国兴业银行的Forge平台推出了符合MiCA标准的欧元计价稳定币EURCV（Euro CoinVertible），并用于欧洲投资银行（EIB）债券结算^[80]。这些基础设施的核心特征在于“合规驱动”，以审计和监管访问为首要目标，技术角色是金融制度的延伸而非革命^[80]。

在AI词元工厂领域，欧洲企业同样在合规框架下寻求技术创新与市场扩张。前文提及的Nebius是欧洲头部AI云厂商，其发展得到了英伟达20亿美元战略投资的支持，计划到2030年共建超5吉瓦算力中心^[76]。另一典型案例是意大利领先的数字化解决方案服务商Lutech集团与中国AI公司迅策科技的合作^[81]。双方联合开发面向欧洲市场的AI转型解决方案，并共同推动包括AI词元工厂在内的解决方案在欧洲商业化落地^[81]。此次合作的核心逻辑在于Lutech能够同时解决“进得去”和“合规住”两大难题：Lutech拥有覆盖能源、金融、制造等多个行业的欧洲中大型企业客户网络，可为解决方案提供快速渗透的渠道；同时，Lutech贡献其“欧盟监管合规经验”，帮助迅策科技的TokenOS操作系统遵循欧盟数据主权原则与监管要求，大幅缩短产品在欧洲市场的准入周期^[81]。这一合作模式精准卡位了欧盟持续推进的数字主权战略，旨在构建以主权AI与安全数据基础设施为核心的跨境创新生态^[81]。

此外，欧洲在数字资产支付与结算领域也有创新实践。例如，由Circle发行的欧元稳定币EURC，是首个符合MiCA标准的欧元稳定币，可在欧盟合法流通^[82]。EURC采用与欧元1:1挂钩的全额储备模式，部署在以太坊、Solana等多个区块链上，主要应用于跨境支付、欧元计价的加密服务及作为DeFi协议的抵押品^[82]。在欧盟MiCA法规驱动下，EURC市值在2025年从7000万欧元增长至超过3亿欧元^[82]。另一个案例是，列支敦士登电信推出了由国家控股的合规区块链基础设施LTIN，面向企业与机构提供服务。这些实践表明，欧洲的词元与数字资产基础设施发展，始终在严格的监管护栏内进行，强调合法性、稳定性和与传统金融体系的融合。

5.3. 亚洲（除中国外）及其他地区：差异化发展路径

亚洲（除中国外）及其他地区的词元工厂发展实践呈现出显著的差异化路径，这些路径深受本地资源禀赋、产业政策、市场需求和地缘政治因素的影响，形成了各具特色的发展模式。

中东地区以阿联酋为代表，展现出雄心勃勃的国家级AI基础设施战略。阿联酋人工智能、数字经济与远程工作应用国务部长奥马尔·阿尔·奥拉马宣布，该国计划通过其庞大的“星际之门”数据中心生产60万亿个AI词元^[84]。阿联酋将专为人工智能设计的数据中心称为“AI工厂”，旨在高效处理海量工作，将原始数据转化为AI“智能”^[84]。这一规划隶属于阿布扎比5吉瓦大型园区规划的一部分，“星际之门”数据中心预计于2026年以200兆瓦的载力投入运营^[84]。阿联酋视AI词元为“智慧原油”，认为货币的未来将指向AI词元，这些新型代币可以转化为提升决策质量、改善生活品质、提高生产力的关键洞察力^[84]。这一战略旨在强化该国在科技领域的国际竞争力，在智能经济时代占据全球核心地位^[84]。此外，在数字金融基础设施领域，华赢集团旗下公司AX Coin在巴林获得了首张Stablecoin发行人牌照，其稳定币可合规锚定巴林第纳尔、美元及其他获批法定货币^[85]。该公司与渣打银行旗下创新投资平台SC Ventures签约，联合孵化AI驱动支付项目AGENPAY，展现了中东地区在合规数字支付领域的积极布局^[85]。

澳大利亚则凭借其丰富的清洁能源资源，吸引了全球资本投入AI算力工厂建设。Firmus Technologies是一家澳大利亚人工智能基础设施公司，初期服务于高性能计算与加密货币领域，后转型专注于设计与建设高效能AI工厂及数据中心^[86]。其旗舰项目“南门项目”计划在澳大利亚塔斯马尼亚州朗塞斯顿打造大型AI基础设施园区，采用液冷等技术以提升能效^[86]。该公司获得了包括黑石集团、Coatue和英伟达在内的巨额投资，2026年2月获得由黑石集团领投的100亿美元债务融资，同年4月完成由Coatue领投、英伟达参与的5.05亿美元C轮融资，估值达55亿美元^[86]。Firmus计划到2028年在澳大利亚建成总容量达1.6吉瓦的数据中心集群，并已签署在墨尔本部署数万块英伟达GPU的大额合同^[86]。另一家上市公司IREN Ltd.（前身为Iris Energy）也从比特币挖矿成功转型为AI数据中心基础设施公司，与英伟达签署了为期五年、价值34亿美元的AI云协议，此前还与微软签订了价值97亿美元的GPU云服务合同。澳大利亚的模式凸显了“绿色算力”的吸引力，即利用低成本的可再生能源为高耗能的AI算力工厂提供动力，从而形成成本优势。

日韩等东亚发达国家的发展路径则与本土产业升级和运营商转型紧密结合。韩国电信运营商KT宣布了一项总额约18万亿韩元（118亿美元）的投资计划，旨在转型为一家AI平台公司^[88]。其中，KT计划投资5万亿韩元建设容量达1吉瓦的AI数据中心，以提供物理AI和自动驾驶所需的超低延迟、实时推理环境^[88]。尤为重要的是，KT明确提出将结合自身的AI数据中心与词元优化引擎，着手建设“词元工厂”，并进军与稳定币相关的业务领域^[88]。这标志着传统电信运营商正利用其网络、数据中心和客户渠道优势，积极向AI词元生产与服务平台转型。日本方面，虽然出现了与政治丑闻相关的“SANAE TOKEN”非法运营事件，暴露了监管漏洞^[89]，但这从反面警示了数字资产合规运营的重要性。主流市场方面，日本金融厅对非法加密货币活动的定性，也反映了其监管体系在应对新兴风险时的存在^[89]。

东南亚、非洲等新兴市场的需求则侧重于利用数字资产和稳定币解决跨境支付成本高昂、金融基础设施不足等痛点。例如，稳定币原生支付公司Daya旨在为非洲跨境企业构建金融操作层，通过单一仪表板整合本地银行、外汇、加密提现与充值及支付处理器等渠道，实现自动路由最优支付路径。非洲支付技术公司Flutterwave与区块链网络Tempo合作，旨在将Tempo作为区块链结算网络整合至其个人汇款应用与企业服务平台中，以解决撒哈拉以南非洲地区平均高达7%的汇款手续费率问题。这些实践表明，在传统金融基础设施薄弱地区，基于区块链和稳定币的数字支付与结算网络能够更快速地填补市场空白，满足企业和个人的实际需求。

5.4. 国外代表性案例及其特点分析

基于前述区域分析，以下对国外词元工厂及相关数字资产基础设施的典型案例进行归纳与特点分析，以清晰呈现其差异化发展路径。

案例名称	所在地区	运营主体	核心业务/定位	关键特点/技术路径
------	------	------	---------	-----------

Nebius AI 词元工厂	欧洲（荷兰）	上市公司 (AI云厂商)	一站式大模型部署推理中枢，专业词元生成平台。	1. 技术并购驱动：收购 Eigen AI 获取顶尖推理优化算法（如 AWQ 量化）。 2. 资本密集：获英伟达 20 亿美元战略投资，共建算力中心。 3. 标杆产品：行业“词元工厂”概念的标杆级产品，主打高性能、低成本词元量产^[76]。
Zero Hash	北美（美国）	私营企业 (B2B 基础设施提供商)	嵌入式加密货币、稳定币支付及资产代币化合规技术解决方案。	1. 合规先行：在美国 51 个司法辖区拥有资金汇款牌照，FinCEN 注册 MSB。 2. 嵌入式 API：通过 API/SDK 帮助传统金融机构无缝集成加密服务。 3. 生态广泛：合作伙伴包括摩根士丹利、盈透证券等主流金融机构^[77]。

阿联酋“星际之门”	中东（阿联酋）	国家主导项目	国家级AI工厂，计划生产60万亿AI词元。	1. 国家战略：作为阿布扎比5吉瓦AI园区核心，提升国家科技竞争力。 2. 宏大目标：将AI词元视为“智慧原油”，定位未来货币。 3. 基础设施驱动：依托大规模数据中心建设，将算力转化为词元产能^[84]。
Firmus Technologies	澳大利亚	私营企业 (AI基础设施公司)	专注于绿色、高效能AI工厂及数据中心的设计与建设。	1. 绿色算力：依托塔斯马尼亚低成本清洁能源，打造能效优势。 2. 资本青睐：获黑石、Coatue、英伟达超百亿美元融资。 3. 项目牵引：“南门项目”计划建设1.6吉瓦数据中心集群，已签大额GPU部署合同^[86]。

KT 词元工厂计划	亚洲（韩国）	电信运营商	电信运营商转型AI平台，建设词元工厂并进军稳定币业务。	1. 运营商转型：结合网络、数据中心优势，向AI词元生产与服务延伸。 2. 端到端布局：投资AI数据中心、海底光缆、词元优化引擎与稳定币业务。 3. 市场拓展：计划将AI转型业务扩展至东南亚市场^[88]。
Euroclear DLT 结算系统	欧洲（比利时）	金融机构联盟 (Euroclear/ 欧洲央行)	基于许可DLT的主权债与可持续债券链上发行与结算平台。	1. 强监管合规：严格遵循MiCA等欧盟法规，采用 R3 Corda 许可链。 2. 机构主导：由央行与中央证券托管机构共同开发运营。 3. 聚焦传统资产：服务于主权债券、可持续债券等传统金融资产的代币化^[80]。

迅策-Lutech 合作	欧洲（意大利）	跨国合作（中国AI公司+欧洲解决方案商）	联合开发面向欧洲市场的AI转型解决方案及词元工厂。	1. 合规渠道双保障：Lutech 提供欧盟合规经验与现有企业客户渠道。 2. 主权 AI 生态：契合欧盟数字主权战略，构建安全数据基础设施。 3. 市场准入捷径：为中国AI技术提供快速进入欧洲市场的合作范式^[81]。
Daya / Flutterwave-Tempo	非洲	金融科技公司	基于稳定币的跨境支付与金融操作层基础设施。	1. 解决痛点：针对非洲高昂汇款手续费、金融覆盖不足的痛点。 2. 技术整合：整合多支付渠道，利用区块链网络实现低成本、快速结算。 3. 普惠金融：服务于个人汇款与企业跨境支付，提升金融包容性^{[90][91]}。

综合分析，国外词元工厂及相关实践呈现出以下鲜明特点：北美以技术创新与资本驱动为核心，通过市场化的并购与融资活动快速构建技术壁垒和生态优势；欧洲则呈

现出强监管与合规导向，技术路径选择服务于监管要求，形成以传统金融机构为主导的联盟化发展模式；中东与澳大利亚充分利用资源禀赋（能源、资本）或国家战略，进行大规模基础设施投入，旨在占据未来AI算力与词元供给的制高点；亚洲新兴工业化国家（如韩国）则依托现有产业巨头（如电信运营商）进行转型，将词元工厂作为其数字化升级的重要组成部分；而在非洲等新兴市场，相关实践更侧重于利用数字资产技术解决具体的金融基础设施痛点，发展路径更为务实和应用导向。这些差异化路径共同构成了全球词元经济基础设施发展的多元图景。

6. 业务模式、差异化能力与综合比较分析

6.1. 主流业务模式解构：IaaS、PaaS、MaaS、BaaS

词元工厂的商业模式正经历从传统的“卖机架、卖算力”向“卖Token、卖智能服务”的深刻演进^[2]。这一演进催生了多种分层、分化的业务模式，其核心在于将底层复杂的芯片、模型和调度能力封装成标准化的、可计量的服务^[2]。当前，主流业务模式可解构为基础设施即服务（IaaS）、平台即服务（PaaS）、模型即服务（MaaS）以及智能体即服务（BaaS）四个层次。

基础设施即服务（IaaS）是词元工厂最基础的商业模式，其核心是提供“裸算力”资源。传统算力租赁模式以“裸金属”服务器租赁时长为基础进行固定月租计费^[92]。然而，随着AI推理需求爆发，单纯的IaaS模式面临资源利用率低、客户使用门槛高、成本不透明等问题。因此，运营商和算力租赁厂商正推动算力租赁从固定月租模式转向按实际Token用量计费^[2]。例如，中国移动、中国电信、中国联通均已推出面向个人和企业的Token套餐，将算力资源封装成可按量付费的标准化产品^[92]。这种转变标志着算力服务从“卖硬件时间”向“卖智能产出”的初步升级，但本质上仍属于资源供给层。

平台即服务（PaaS）模式在IaaS之上，提供了更丰富的软件栈和工具链，旨在降低客户使用AI的技术门槛和运维复杂度。该模式的核心是提供一个统一的调度、部署、监控与管理平台。例如，联通数智依托元景MaaS一站式模型开发与运营服务平台，完整覆盖Token从创造、存放、转运到定价的全流程^[62]。贵州移动的“词元工厂”通过四步路径实现资源池化、平台化调度、模型服务化和运营产品化，将过去的“五金超

市”转变为直接交付“标准件”的平台^[8]。PaaS模式的价值在于，它通过统一的API接口、可视化监控仪表盘和清晰的Token账单，让企业无需自建昂贵的AI算力集群或管理复杂的硬件，只需按实际消耗付费即可调用AI能力^[8]。

模型即服务（MaaS）是当前词元工厂最核心和主流的商业模式。MaaS平台的核心是聚合并管理各类大模型，为用户提供统一的模型调用入口。与单纯提供算力的IaaS和提供工具链的PaaS不同，MaaS直接交付的是经过深度优化后的模型推理能力，其计费单位是Token^[4]。行业主流模式是通过Token按量计费^[4]。例如，硅基流动的公有云MaaS服务了近1000万用户与超1万家企业客户，平台已上架包括DeepSeek、Qwen、GLM、Kimi、MiniMax等在内的超150款模型^[93]。腾讯云的TokenHub平台同样实现了HY（腾讯混元）+开源生态全覆盖，通过“Buddy严选”机制在最强、最高性价比、最快、最长上下文四个维度进行模型筛选^[37]。MaaS模式的优势在于，它屏蔽了底层算力的异构性和模型部署的复杂性，让用户能够根据业务场景灵活选择最合适的模型。

智能体即服务（BaaS）代表了词元工厂商业模式演进的前沿方向。随着AI智能体（Agent）加速渗透，行业正从“AI原生”迈向“Agent原生”新阶段^[2]。BaaS模式不仅提供模型能力，更进一步提供智能体的运行环境、编排工具、技能库和安全保障，直接交付可执行复杂任务的数字员工。例如，腾讯云推出的Agent Runtime是面向AI Agent场景的原生基础设施，核心定位是通过重构云基础设施适配Agent任务负载特征，实现Agent“跑得快、跑得省、跑得放心”^[41]。PPIO完成从传统云服务向Agentic Cloud的战略转型，推出智能模型网关与Agent Harness双层架构体系，其Agent沙箱支持毫秒级冷启动和Auto Pause/Resume机制，显著降低了智能体运行成本^[3]。超聚变提出的“智企”进化路径中，智能体执行是将Token转化为业务价值的关键环节^[74]。BaaS模式将词元工厂的价值链从Token生产进一步延伸至任务执行和价值创造，商业模式也从“按Token量计费”逐步转向“按Token为用户业务创造的实际价值付费”^[6]。

6.2. 核心竞争力与差异化能力维度分析

在激烈的市场竞争中，词元工厂的核心竞争力与差异化能力主要体现在以下几个维度：技术优化与成本控制能力、生态聚合与调度能力、安全合规与可信服务能力、行业场景深耕与价值交付能力。

技术优化与成本控制能力是词元工厂的生存基石。其核心指标是“每瓦Token数”和“Token每焦”，即单位能耗下的Token产出效率^[6]。领先厂商通过全栈技术优化实现极致降本。例如，硅基流动通过对底层算子、框架、模型等层面的极致优化，“榨干”每一份算力潜能，在华为昇腾、沐曦、摩尔线程等国产芯片上实现了业界领先的性能^[93]。是石科技发布的国产Token优化工厂“拓元”，通过结果感知驱动KV Cache压缩、全模态推理的Token压缩等五大技术突破，实现Token成本降低30%-50%^[26]。腾讯云TokenHub通过全异步推理流水线、FlexKV自研框架及训推一体调度，实现推理效率提升100%、综合降本40%^[37]。此外，算电协同成为降低用能成本的关键路径。润建股份的五象云谷Token工厂通过自研RunDo能源中枢和润道星算平台，将能耗管控与算力产出拧成一股绳，并利用西部绿电资源，预计能将Token生产成本降低30%-40%。商汤科技发布的算电协同Agent，使商汤大装置实现单位电量Token产出提升80%^[95]。

生态聚合与调度能力决定了词元工厂的服务广度与灵活性。该能力要求平台能够兼容多元算力芯片、聚合丰富模型生态、并实现高效的智能调度。在算力兼容方面，领先厂商积极构建异构算力统一调度能力。例如，“拓元”全面兼容昇腾、昆仑芯、天数智芯等10余种国产算力芯片，适配20余个主流模型，实现国产异构算力池的统一调度^[26]。无问芯穹的Agentic Infra算力集散中心已接入3.7万P算力，覆盖16款不同芯片^[6]。在模型聚合方面，平台需要快速适配社区最新模型。硅基流动凭借高效的自动化流水线，能快速将社区最新发布的顶级模型适配并上线服务，目前已上架超150款模型^[93]。中国移动的MoMA模型服务平台可聚合300多款主流大模型^[60]。在智能调度方面，平台需根据任务复杂度自动匹配最合适的模型和算力资源，实现成本与体验的最优平衡^[10]。清程极智的AI Ping平台搭载自研毫秒级智能路由技术，依托7×24小时实时性能评测，动态匹配最优模型与算力资源。

安全合规与可信服务能力是赢得政企客户，特别是国央企及大型企业的关键。随着数据主权意识日益高涨，企业对“稳定、安全、可信”的Token服务需求迫在眉睫^[11]。运营商凭借其国资背景和多年网络安全运营经验，在此方面具有天然优势。中国联通作为安全产业链“链长”，能够打通Token服务连接侧可观测、平台侧管控和应用侧治理的全链条能力，把安全能力放进“算力-数据-模型-应用”全流程^[62]。北京数据集团发布的京算Token工厂，依托国资信用和属地化服务，可实现数据不出域，满足等级保护三级合规要求^[9]。技术层面，机密计算等技术的应用提升了服务可信度。北京移动的一站式Token运营服务平台采用机密计算技术，能在算力资源上构建安全可信的机密执行环境，确保大模型在推理过程中保持端到端加密^[59]。腾讯云Agent Runtime

采用零凭证、零信任架构，实现100% Session全留痕^[41]。

行业场景深耕与价值交付能力是词元工厂实现差异化竞争和更高溢价的来源。通用Token的定价更多是“成本定价”，而垂类行业Token未来有望演变为按照其所能解决的任务价值进行“价值定价”^[56]。这要求厂商不仅提供Token，更要深入行业业务流程，打造“平台+应用+Token”的深度服务模式。例如，联通数智重点深耕云犀、工业、医疗、教育等领域，以差异化应用拉动Token规模增长^[62]。中科闻歌发布的海汇TokSea平台，不仅支持主流通用模型，还可依托自研的雅意大模型、磐石ScienceOne和决策机Decitron等专业大模型能力，形成“通用能力+科学智能+决策智能”的企业AI能力组合，服务于科研推理、复杂决策等专业场景^[71]。超聚变围绕AI重构企业生产经营与分析决策，展示以智企ERP为核心的企业软件和应用系统，推动AI从“可用”走向“可运营、可管理、可创造价值”^[74]。这种深度绑定行业Know-how的能力，构成了厂商难以被复制的护城河。

6.3. 中外词元工厂发展路径与能力对比

中外词元工厂的发展因技术基础、市场环境、政策导向和产业生态的不同，呈现出差异化的发展路径和能力特点。总体而言，北美地区引领技术创新与商业模式前沿，中国凭借规模优势、工程化能力和政策驱动实现快速追赶与局部领先，欧洲地区则强调合规导向与行业应用。

北美地区作为人工智能技术的策源地，其词元工厂（或类似基础设施）的发展通常由大型科技公司（如AWS、Google Cloud、Microsoft Azure）和领先的模型厂商（如OpenAI、Anthropic）主导。这些厂商的核心优势在于顶尖的原创模型能力、强大的全球算力网络布局以及成熟的云服务生态。其商业模式往往与自研的闭源大模型深度绑定，通过API提供高质量的Token服务，毛利率已达到40%-50%^[58]。技术创新上，它们在推理引擎、长上下文处理、多模态融合等基础研究领域持续领先。然而，其发展也可能受制于高端芯片供应链（如对英伟达GPU的依赖）和特定地区的政策监管。

中国词元工厂的发展路径呈现出鲜明的中国特色：政策强力驱动、全产业链协同和极致成本优化。国家数据局将推动词元经济发展纳入工作体系，七部门首次将“Token普惠”写进国家级行动方案，为产业发展提供了明确导向。这使得中国形成了国有企业、民营企业、电信运营商、地方政府等多方力量共同参与的繁荣生态。在能力

对比上，中国厂商的突出优势体现在：

- 1. 规模化与工程化能力：中国拥有全球最大的Token供给市场，OpenRouter数据显示，中国模型的Token消耗量已占全球市场的54.1%^[4]。面对海量需求，中国厂商在异构算力调度、推理效率优化、集群运维等工程化方面积累了丰富的经验，例如PPIO的分布式算力网络实现GPU利用率持续高于75%^[3]。
- 2. 国产化生态构建：在供应链安全考量下，中国厂商在国产芯片适配方面进展迅速。华为昇腾等国产算力芯片已经在模型推理服务中规模化使用^[4]。趋境科技、是石科技、硅基流动等企业均推出了全面兼容国产芯片的Token优化方案^{[12][26][93]}。
- 3. 成本优势与普惠化：中国的人才密度、工程师红利以及绿电资源，使得Token服务成本具有显著优势^[4]。运营商推出的个人Token套餐低至9.9元/月，极大地推动了AI能力的普及^[5]。
- 4. 开源模型生态活跃：DeepSeek、阿里Qwen、智谱GLM等中国主流模型厂商几乎都在走开源路线，这为全球各地企业本地部署中国模型提供了可能，创造了“自主可控”的价值^[4]。

然而，与国际顶尖水平相比，中国在原创性基础模型架构、底层AI框架与开发工具的全球影响力方面仍有差距。同时，国内市场竞争异常激烈，价格战初现苗头，可能对产业链的长期创新投入构成挑战^[58]。

欧洲地区的发展路径则紧密围绕数据主权与合规监管展开。欧盟持续推进数字主权战略，要求关键数据处理能力必须在境内完成^[98]。这为能够提供符合GDPR等法规的本地化、合规化Token服务的厂商创造了机会。例如，迅策科技与欧洲Lutech集团合作，Lutech贡献其“欧盟监管合规经验”为迅策的TokenOS产品提供合规赋能，以构建满足欧洲主权AI要求的解决方案^[98]。欧洲词元工厂更侧重于在金融、医疗、工业制造等垂直行业提供高可信、可审计的AI服务，商业模式上可能更倾向于项目制与标准化服务结合。

对比维度	中国词元工厂	北美同类实践	欧洲同类实践
------	--------	--------	--------

驱动核心	政策引导、市场规模、成本优势、供应链安全	技术创新、商业模式、全球生态、顶尖模型	数据主权、合规监管、行业标准
技术路径	异构算力调度、推理极致优化、国产芯片适配、算电协同	自研芯片与框架、基础模型创新、全球算力网络	隐私计算、可信AI、行业模型微调
商业模式	多元化（IaaS/PaaS/MaaS/BaaS）、运营商套餐普惠、垂类价值定价探索	以闭源模型API为核心的MaaS、高毛利率	合规导向的本地化部署、项目制与订阅制结合
生态特点	国有/民营/运营商协同、开源模型生态繁荣、全产业链布局	科技巨头主导、闭源生态强大、全球开发者社区	强调合规与协作、区域性联盟、行业标准组织活跃
主要挑战	原创基础研究待加强、激烈价格战风险、高端算力供给瓶颈	监管压力增大、供应链依赖、技术垄断争议	市场规模相对较小、创新节奏可能较慢、技术依赖性

6.4. 市场竞争格局与关键成功因素

当前，中国词元工厂市场已形成“国家队”、“运营商队”、“互联网云厂商队”和“独立第三方队”多方竞逐的格局，市场竞争从单纯的技术和价格比拼，转向涵盖生态、合规、服务深度的综合能力竞争。

“国家队”以北京数据集团、人民数据、各地数据集团/公司为代表，其核心优势在于国资信用、政策资源和对本地化数据要素的掌控力。它们建设的词元工厂（如京算Token工厂、人民数据·城市词元工厂）主要服务于政务、央国企及重点行业，强调安全可控、数据不出域和长期稳定服务^[11]。其商业模式往往与地方数字经济发展规划紧密结合，承担着夯实数字算力底座、培育本地AI产业生态的战略任务。

“运营商队”（中国移动、中国电信、中国联通）是Token经济中最活跃、最具

网络效应的参与者。它们凭借覆盖全国的骨干网络与算力节点、十亿级用户触达能力以及“云、管、端、边”一体化运营优势，正从“流量经营”全面转向“Token经营”^[57]。运营商不仅推出普惠的个人和企业Token套餐，更致力于构建贯通“能力底座、运营平台、应用入口”的Token生态运营体系^[57]。例如，中国电信打造星辰TokenHub运营服务平台，中国移动构建一站式Token运营服务平台，中国联通推出联通星罗Token服务平台^{[59][61]}。运营商的核心竞争力在于其将通信连接、算力调度、模型服务和智能终端打通的整体交付能力。

“互联网云厂商队”（如腾讯云、百度智能云、阿里云）依托其强大的云计算基础设施、深厚的软件技术积累和丰富的业务场景，快速将原有的MaaS服务升级为词元工厂。例如，百度智能云将“MaaS模型服务”升级为“Token Factory”，腾讯云推出TokenHub和Agent Runtime^{[58][37][41]}。它们的优势在于成熟的云原生技术栈、庞大的开发者生态以及与自身业务（如社交、搜索、电商）结合的垂直场景。其挑战在于如何保持平台的中立性，避免与生态伙伴竞争，并适应政企客户对私有化部署和安全合规的更高要求。

“独立第三方队”包括硅基流动、是石科技、无问芯穹、趋境科技、清程极智等创业公司。它们通常技术背景深厚（多源自清华等顶尖高校）、定位中立（不绑定特定芯片或模型）、专注于某一技术长板（如推理优化、异构调度）。硅基流动作为第三方MaaS平台的代表，其市场影响力已跻身行业第一梯队^[53]。这类厂商的生存关键在于其技术的不可替代性和产品化能力，需要在巨头林立的生态中找到精准的定位，或通过“国家队”、“运营商队”合作来获取市场和场景。

综合来看，词元工厂市场的关键成功因素（KSF）已清晰显现：

1. 极致效率与成本控制能力：在Token单价持续下行的趋势下，能够通过全栈技术优化和算电协同实现更低“每瓦Token成本”的厂商将获得决定性优势^{[6][95]}。
2. 构建开放、繁荣的生态：单一厂商无法通吃所有环节。成功的词元工厂必须是一个“连接器”和“赋能者”，能够兼容多元算力、聚合丰富模型、赋能广大开发者和ISV，形成共赢的生态^[93]。
3. 安全可信与合规运营：这是进入政企市场的入场券。具备国资背景、完备的安全资质、数据本地化处理能力和清晰合规路径的厂商，将在To B和To G市场建立坚固壁垒^{[11][62]}。
4. 深耕行业，实现价值闭环：未来竞争焦点将从“聚焦调用量”向“聚焦应用价

值”深化^[62]。能够深入特定行业，将Token能力与业务流程深度融合，提供“平台+应用+Token”一体化解决方案，并帮助客户量化AI价值的厂商，将获得更高的客户粘性和溢价能力。

5. 把握技术演进趋势：随着智能体（Agent）成为主流交互范式，基础设施必须原生支持Agent的运行、编排与治理^{[41][3]}。同时，对多模态、长上下文、科学智能等前沿方向的支持，也将成为差异化能力的重要组成。

未来，市场竞争格局将进一步分化，可能出现“全能型巨头”与“垂直领域专家”并存的局面。拥有资本、技术、生态和场景综合优势的运营商和头部云厂商有望成为平台型巨头，而具备独特技术专利、深度行业Know-how或特殊资源禀赋的独立厂商，则将在细分赛道建立自己的护城河。

7. 发展趋势、挑战与战略建议

7.1. 技术融合、行业渗透与生态演进趋势

随着人工智能技术加速渗透、AI Agent全面爆发，词元工厂作为连接技术供给与商业需求的核心枢纽，其发展正呈现出技术深度融合、行业广泛渗透与生态协同演进三大趋势^[93]。

在技术融合层面，词元工厂正从单一算力供给向“算力、模型、数据、应用”一体化融合演进。其核心任务已不仅是产出Token，而是通过全栈技术优化，让每一块芯片的每一秒算力都最高效地转化为对应场景下的高价值Token^[6]。这依赖于推理加速引擎的深度优化、异构算力的统一调度与模型生态的快速适配三大核心技术支柱^[93]。例如，硅基流动通过对底层算子、框架、模型等层面的极致优化，在华为昇腾、沐曦、摩尔线程等国产芯片上实现了业界领先的性能，有力证明了国产芯片支撑大规模高强度推理的能力^[93]。同时，技术路径正从单卡性能比拼进入系统级竞争阶段，模型参数突破万亿级、“Token工厂”向吉瓦级演进，降低单位算力成本成为核心命题^[95]。太初元基采用的HCU异构融合计算架构，通过CPU与AI算力核统一高速总线互联，配合共享分级存储架构，提升了整体效率^[95]。网络与存储等基础设施的升级也至关重要，新华三集团指出，800G网络较400G网络可显著提升相同算力下的Token吞吐量，而通过分级缓存可以降低算力成本^[95]。

在行业渗透层面，Token服务正从技术概念转化为企业可感知的生产工具，加速向千行百业渗透^[60]。企业对AI的需求已不再停留于通用问答和文本生成，而是转向更具体的生产经营环节，如电气企业需要能够识别电路图的质检工具，泵阀企业需要快速响应技术参数的报价助手，鞋服企业希望通过模型辅助研判市场趋势^[60]。这种渗透催生了“前店后厂一中心”的产业协同模式，其中“前店”是面向医疗、法律、影视、游戏等垂直行业的解决方案，负责把Token打包“卖出去”，转化为各行业能用、好用的生产力^[6]。硅基流动的“Token工厂”模式已在能源、金融、互联网等多个行业落地生根。在能源与工业领域，为某头部电力央企构建的私有化一站式AI平台，落地了“百人建模训练、千人Agent开发、万人推理应用”的“百千万工程”，在智能办公、风电运维等场景成效显著，故障诊断准确率提升至95%以上^[93]。在金融领域，为头部券商在其自有国产算力集群上部署的高性能推理框架，稳定支撑了智能投顾、合规审查等高并发、低延迟业务^[93]。

在生态演进层面，词元经济正催生以“数据飞轮”为核心的自我强化增长引擎和开放协同的产业新生态^[100]。“数据飞轮”描述了一个“场景牵引数据、数据驱动模型、模型赋能应用、应用创造价值”的良性循环，政策鼓励的不再是静态的数据仓库建设，而是一个动态的、价值不断增殖的活生态系统^[100]。这预示着新型产业服务主体——“数据赋能工场”的出现，它们将是“数据飞轮”的运营者，负责将原始数据加工为标准“词元”，并确保其在生态中安全、可信、高效地流通与增值^[100]。同时，生态呈现开放化趋势，真正的“Token工厂”需满足四大标准：支持多元芯片，具备强大异构算力支持能力；部署丰富模型，为开发者提供广阔选择空间；汇聚充沛算力资源，能灵活调度全球计算力；拥有广泛市场连接，服务海量开发者与企业客户^[93]。无问芯穹联合上海移动正式发布并启动“天问计划生态联盟”，汇聚壁仞、燧原科技、天数智芯、沐曦、海光等国内领先本土GPU芯片厂商，同步联动上海仪电、库帕思、MiniMax等头部大模型研发与行业应用企业，贯通从“硅基算力”到“模型智能”再到“行业落地”的全链路^[6]。

7.2. 面临的主要挑战：技术、成本、安全与合规

尽管词元工厂发展迅猛，但其规模化落地与可持续发展仍面临技术异构、成本高企、安全合规与生态共识等多重挑战。

技术挑战的核心在于高效处理异构算力与复杂工作负载。随着国产化进程加速，

词元工厂需要兼容并优化多种国产芯片，如昇腾、昆仑芯、天数智芯、太初、瀚博半导体、摩尔线程、沐曦、燧原等^[26]。这要求平台具备强大的异构算力池统一调度与推理加速能力，破除跨地域、多芯片算力孤岛^[26]。同时，AI智能体（Agent）场景对基础设施提出了全新要求，其Token消耗量是传统Chatbot的5-30倍，且任务具有长周期、多步骤、高并发的特点^[37]。传统的云原生架构已难以适配智能体发展，需要构建以“词元”（Token）为中心的AI Token Factory，从交互范式到整个基础设施架构进行重新设计与优化^[2]。例如，智能体在执行长程检索任务时，有时会无法判断缺失信息是已经被丢弃还是不存在的，从而产生负面影响^[26]。此外，模型迭代速度极快，企业刚完成上一版模型的部署，新版本又已发布，如何实现模型的平滑、快速更新与部署，是持续的技术挑战^[32]。

成本挑战是制约词元经济普惠化的关键瓶颈。Token需求呈指数级增长，但算力供给仅线性增长，“用户觉得AI服务贵，本质就是供不应求”^[6]。成本构成复杂，包括高昂的硬件采购（尤其是GPU）、持续的电力消耗、模型推理的算力开销以及运维管理成本。推理成本尤其突出，当模型上下文变长时，显存压力急剧攀升^[26]。虽然通过技术优化，如无问芯穹的推理成本已降至三年前的1/10，并相信未来会下降到现在的1/10，但持续降低成本仍是核心任务^[6]。电力成本是另一个重大负担，一个1000P（PFLOPS）标准算力集群年度总耗电量约2000万度^[95]。商业模式的灵活性也面临考验，企业需要按需获取稳定、高效的AI能力，同时要求成本透明可控，这要求词元工厂提供按Token消耗、订阅或API调用等灵活计费模式^[7]。

安全与合规挑战随着数据主权意识高涨而日益严峻。企业，特别是国央企及大型企业，不再仅关注价格与性能，更将“稳定、安全、可信”作为选择Token供应商的底线，倾向于本土化技术体系^[9]。数据安全要求“数据不出域”，满足等级保护三级合规要求，从根本上打消企业对数据安全和合规风险的顾虑^[9]。在数据跨境场景下，挑战更为复杂。例如，TokenOS操作系统要进入欧洲企业，需遵循欧盟数据主权原则与监管要求^[98]。欧盟持续推进数字主权战略，要求关键数据处理能力必须在境内完成^[98]。此外，在人工智能训练阶段数据使用规则、版权作品数据用于模型训练、数据授权使用机制和收益分配规则等方面，仍需完善权责清晰、合规包容的制度环境^[13]。防范数据集建设全流程产生数据偏见与歧视，确保数据集建设成果惠及全民，也是重要的伦理挑战^[13]。

生态与共识挑战是更深层、更长期的障碍。首当其冲的是标准化之难，如何为不同行业、形态各异的数据建立统一的质量、确权、计量与定价标准，是“词元”能否成为通用“货币”的基础^[100]。国家数据局提出的“全国互认”目标道出了其艰巨性^[100]。

最深层的挑战在于培育市场共识，即“培育为数据付费的市场共识”^[13]。这需要打破长期存在的数据本位主义和数据“白嫖”心态，推动需求方，特别是政府部门和国有企业“将数据采买纳入预算编制”，发挥示范作用^[100]。只有当数据像水电煤一样，被普遍认可为需要付费购买的生产要素时，“数据飞轮”才能真正高速运转起来^[100]。此外，在智库等专业领域，还面临专属词元标注体系缺位、“社科+数智”双向复合型人才培养机制尚未成型等挑战^[101]。

7.3. 对市场参与者的战略发展建议

面对趋势与挑战，词元工厂的市场参与者，包括基础设施提供商、模型厂商、行业解决方案商及算力运营方，需采取差异化战略，构建核心竞争力。

对于基础设施与平台提供商（如云厂商、专业AI基础设施公司），核心战略是打造开放、高效、安全的Token生产与调度平台。首先，必须坚持开放生态战略，不绑定单一芯片、不禁锢于特定模型、不局限于自有算力，且不与客户竞争，通过汇聚多元算力与模型生态构建护城河^[93]。其次，需持续投入全栈技术优化，聚焦“token每秒、token每焦”等核心效率指标，通过推理加速、异构调度、KV Cache压缩（如拓元的结果感知驱动KV Cache压缩技术）、动态PD分离、全异步推理流水线等技术，极致提升单位算力的Token产出和能效^{[6][26][95][37]}。第三，构建面向智能体（Agent）时代的原生基础设施能力，例如提供毫秒级启动的Agent沙箱、支持任务暂停不计费的存储、零信任架构的Agent网关以及全链路可观测能力，以适配Agent任务负载特征^[41]。第四，深化与国产芯片厂商的合作，通过端到端调优，持续提升混合集群在真实推理负载下的效率，证明国产算力的规模化应用能力^{[10][93]}。最后，创新商业模式，提供从按量计费、Token套餐、Saving Plan到资源包等多层次、灵活的计费模式，并探索未来按Token创造的实际价值付费的更高阶模式^{[6][8][37]}。

对于模型厂商与AI应用开发者，战略重点在于聚焦模型能力创新与深度行业结合，并善用词元工厂降低运营成本。模型厂商应更专注于“造脑”，即核心算法与模型能力的迭代，将模型能力转化为可规模化、具备SLA保障、可计量的智能服务的工作交由专业的词元工厂完成，形成清晰分工^[10]。开发者应充分利用Token工厂提供的多模型选型、智能路由、成本优化等功能，快速构建和部署AI应用，将宝贵资源聚焦于核心业务逻辑与用户体验优化，无需承担沉重的硬件采购与运维成本^{[7][37]}。特别是对于出海业务，可借助具备全球算力布局和合规能力的Token工厂，无需自建海外基建，即可实

现Token的全球化生产与交付^[4]。

对于算力持有方与运营商（如电信运营商、数据中心），战略核心是从传统的资源租赁商向智能服务运营商转型。运营商应推动算力租赁从固定月租模式转向按实际Token用量计费，并推出面向个人用户和企业的Token套餐^[2]。这需要构建“算力底座、模型超市、运营服务”三大核心能力：夯实算力网络底座，实现本地算力与全国算力资源的联通；聚合主流大模型，打造模型超市降低企业使用门槛；构建面向企业的智能交付与运营体系，提供贴近行业实际的智能服务^[60]。例如，贵州移动通过资源池化、平台化调度、模型服务化和运营产品化四步路径，将算力中心转型为直接交付“标准件”的词元工厂^[8]。此外，积极探索算电协同，基于分时电价、风光发电实时波动灵活迁移算力任务负载，主动错峰使用低价清洁能源，从源头压缩算力用电成本^[95]。

对于行业企业用户，战略建议是积极拥抱Token经济，以业务价值为导向，循序渐进地引入AI能力。企业应首先明确AI赋能业务的具体场景和价值点，例如提升生产效率（如质检、报价）、优化客户体验（如智能客服）、或增强决策能力（如市场趋势研判）^[60]。在实施路径上，初期可通过采购标准化Token服务快速验证业务场景，降低试错成本；随着应用深入，可考虑采用私有化部署或混合云模式，在保障数据安全的前提下，获得更定制化、更稳定的AI能力^[32]。企业需要建立对Token消耗的成本管控与效益评估体系，利用Token工厂提供的可视化看板，精细化管理每一次模型调用，让问题排查、安全审计与资源治理都有据可查^{[38][37]}。同时，注重内部“社科+数智”复合型人才的培养，或与专业服务机构合作，以弥补能力缺口^[101]。

7.4. 对监管与政策制定的相关建议

为引导词元工厂产业健康、有序、创新发展，监管与政策制定者需在以下关键领域着力，构建“政策+标准+人才+安全”四位一体的支撑体系^[18]。

第一，加强顶层设计与战略引导，明确产业发展路径。省级层面可借鉴广东经验，制定促进词元经济高质量发展的若干措施，明确到2027年底，初步构建工作体系，形成全省一盘棋统筹、差异化布局、一体化协同的词元经济发展格局，目标成为全国词元供给高地、智能应用标杆和词元出海枢纽^[18]。国家层面需持续将促进词元经济发展纳入数字经济工作重点，如国家数据局将2026年定为“数据要素价值释放年”，明确探索词元价值体系与交易机制^[99]。政策应鼓励“政产学研用金”多方协同，搭建交流平台，形成共建共享、互利共赢的产业生态，破解“数据孤岛”^[13]。

第二，加快构建词元标准化体系与交易规则，奠定市场化基石。这是释放数据要素价值的关键前提。政策需着力建立健全标准化词元交易规则，建设省级乃至国家级的词元交易专区，培育词元交易服务生态，让词元像普通商品一样合规流通、便捷交易^[18]。核心任务包括：建立覆盖数据采集、清洗、加工、标注、质检、测评、迭代、审计等全生命周期的数据集管理服务标准^[13]；研究并推出词元的质量、确权、计量与定价的全国互认标准^[100]；探索以词元为基础，可量化、可定价的数据价值体系，并创新词元交易等新型交易模式^[13]。北京市经信局支持企业采购并消费词元（Token）开展人工智能应用，并根据实际使用情况给予资金支持的做法，为培育付费市场共识提供了示范^[15]。

第三，夯实基础设施，鼓励技术创新与普惠服务。政策应着力夯实算力设施底座，推动算力资源开放与互联互通，引导平台企业联通分布式算力资源，提升算力资源配置效率^[14]。同时，提升平台企业词元（Token）普惠服务能力，面向中小企业共性需求优化智能体服务，降低中小企业获取与应用门槛^[14]。对于技术创新，应支持企业加大在推理加速、异构调度、算电协同等关键领域的研发投入，并对使用国产算力芯片给予鼓励。例如，可通过建设“国家—省级—行业”三级云边协同的词元算力体系，面向重大战略研究、智库、高校等常态化开放千亿级词元免费调用额度，落实算力减免、专项倾斜等普惠保障机制^[101]。

第四，完善数据要素制度与安全合规框架，保障产业行稳致远。监管需加快完善面向人工智能发展的数据相关制度。这包括落实数据持有权、使用权、经营权三权分置制度；研究合成数据等新情况新问题；完善人工智能训练阶段数据使用规则，推动版权作品数据等有序用于模型训练，完善数据授权使用机制和收益分配规则^[13]。在安全合规方面，必须坚守发展安全底线，强化数据跨境制度规则，在保障安全的前提下优化词元出海服务^[18]。要求企业建立全流程安全治理机制，防范数据投毒与污染、数据泄露等安全风险，守牢数据安全底线^[13]。对于出海企业，监管可提供合规指引，并鼓励其依托本地合作伙伴（如迅策与欧洲Lutech集团的合作）的合规经验，缩短产品在海外市场的准入周期^[98]。

第五，强化复合型人才培养与生态培育，注入长期发展动力。人才是产业发展的根本。政策需强化“社科+数智”或“行业+AI”复合人才培养^{[18][101]}。具体措施可包括：在基础教育阶段前置数智育人环节；推动高校在新文科、工科等专业增设与词元应用、大模型实操相关的交叉实践课程；鼓励企业联合高校共建实验室、开放真实场景进行实训，实施“百企千岗”等行动^{[15][101]}。同时，构建词元企业梯度培育体系，发展词元服务新生态，通过遴选典型案例、举办创新赛事、推动应用场景落地等方式，激发

词元经济发展内生动力^[18]。

参考文献

- [1] 云天励飞陈宁：推理芯片、Token工厂、智能体成AI三大关键 - 科技 - 南方财经网
- [2] 词元工厂（算力数据中心）_百度百科
- [3] 对话PPIO姚欣：Agent时代来临，智能Token工厂如何重塑AI云服务格局？ - 人工智能-ITBear科技资讯
- [4] 从产品出海到Token出海，WAIC上的全球新生意 - 科技 - 南方财经网
- [5] 昆明词元经济乘风起_昆明要闻_昆明市人民政府门户网站
- [6] AI基础设施“未来态”揭开面纱，“Token工厂”究竟卖什么_上观新闻
- [7] WAIC 2026 \{\}| 智微智能Token工厂首发！揭开AI基础设施新篇章 - IT之家
- [8] 调词元 找贵州 \{\}| 综合调用成本较行业平均水平降低20%以上 贵州移动“词元工厂”让Token好用不贵
- [9] 2026全球数字经济大会数据要素发展论坛举办 北京数据集团发布“京算Token工厂”_财经_中国网
- [10] 每天可生产1.4万亿Token，软通动力要做AI工厂 - 科技 - 南方财经网
- [11] 2026全球数字经济大会数据要素发展论坛举办，北京数据集团发布“京算Token工厂”_中国经济网——国家经济门户
- [12] 北京Token工厂，融资10亿-业界动态-ITBear科技资讯
- [13] 国家数据局关于印发《关于推进行业高质量数据集建设行动的实施方案》的通知-国家数据局
- [14] 七部门：提升平台企业词元（Token）普惠服务能力，面向中小企业共性需求优化智能体服务_澎湃财讯_澎湃新闻-The Paper
- [15] 北京市经济和信息化局关于印发《北京市关于支持工业企业提质增效若干措施》的通知_政策文件_首都之窗_北京市人民政府门户网站
- [16] 北京市经济和信息化局关于印发《支持人工智能OPC创新发展行动方案（试行）》

的通知_政策文件_首都之窗_北京市人民政府门户网站

[17] 【贵州：发布20条新政策支持人工智能发展】_信用中国

[18] 抢抓机遇 数据赋能 广东制定促进词元经济发展若干措施_广东省政务服务和数据管理局

[19] 抢抓机遇 数据赋能 广东制定促进词元经济发展若干措施 - 中山市政务服务和数据管理局政务网

[20] 抢抓机遇 数据赋能 广东制定促进词元经济发展若干措施

[21] 陕西打造“丝路词元工厂”_央广网

[22] 聚焦新赛道 武汉硚口区首发Token经济专项政策

[23] 以数据赋能词元经济高质量发展 省政务和数据局召开词元经济座谈会

[24] 【独家深度报告】词元经济发展研究报告：内涵、实践与趋势分析（附下载） - 智慧城市行业分析

[25] 词元出海开辟数字贸易新赛道

[26] 清华系团队发布国产Token优化工厂：兼容10余种国产芯片，日吞吐千亿Token - 智源社区

[27] 国产算力最大的短板，被一座Token工厂补上了？ -36氪

[28] 《数字化领航》AI技术专刊

[29] 推理性能飙升、成本结构重塑：ODCC 携手 NVIDIA、焱融等首发KV Cache评测，焱融 YRCache 交出“提速降本”双优答卷\{\}|界面新闻

[30] Agentic AI时代，如何为百万上下文大模型提供推理加速？ - 华为企业业务

[31] 芯展速AI90方案亮相WAIC 2026：首Token延迟降50倍，显存节省39%-信息流-ITBear科技资讯

[32] 超聚变推出企业级Token生产平台 解锁高效本地AI部署_中华网

[33] WAIC 2026：阡视科技发布专为Token工厂而生的超节点系统_上观新闻

[34] PDD，Token工厂必备的“超级管线”！“前店后厂一中心”布局来了！_上观新闻

[35] 取长补短，解锁推理性能1+1>2，DeepLink首发：生产级国产异构算力混合推理加速方案 - 智源社区

- [36] 全面迈向AI推理时代：金融大模型异构算力纳管与系统级加速实践-腾讯云开发者社区-腾讯云
- [37] TokenHub：面向Agent时代的企业级Token生产与调度平台-腾讯云开发者社区-腾讯云
- [38] SmartX AI 基础设施：让企业在自有云大规模运行生产级智能体Agent
- [39] 算力优化的六种路径：主流AI基础设施厂商技术路线与选型指南\{\}|界面新闻
- [40] 中科闻歌发布海汇TokSea-经济·科技-人民网
- [41] 腾讯云Agent Runtime：Agent原生基础设施技术概要-腾讯云开发者社区-腾讯云
- [42] 接入识别\{\}|安全管控_Token可视\{\}|盛邦安全
- [43] 阿里云IDaaS EIAM的Agent ID Guard通过SDK支持Agent以机器或用户身份安全获取下游服务的API Key与OAuth Token，实现集中托管与动态下发。-应用身份服务 (IDaaS)(IDaaS)-阿里云帮助中心
- [44] 详情-安全产品—亚信安全—护航产业互联 安全数字世界
- [45] 像搭乐高一样建流水线，蚂蚁数科让百万智能体走入真实企业生产_10%公司_澎湃新闻-The Paper
- [46] 从数周到30分钟！多智能体协同重构中国商品出海全链路_上观新闻
- [47] 拆解迈富时“全栈Token工厂”：四层架构如何把算力炼成结果_凤凰网
- [48] 扣子Coze（字节跳动开发的AI智能体搭建与 workflow编排平台）_百度百科
- [49] # Dify × EdgeOne Pages 实战：零代码 30 分钟上线 AI 应用全流程-腾讯云开发者社区-腾讯云
- [50] 中国大型Token工厂落户中卫 - 中国工业新闻网
- [51] Token工厂落地宁夏中卫 助力绿色算力产业升级
- [52] 国产Token优化工厂在2026世界人工智能大会正式发布-经济·科技-人民网
- [53] 硅基流动完成新一轮超20亿元融资，华兴资本担任独家财务顾问_地区财经_财经_中金在线
- [54] 算力战略升维、产品创新、生态共赢，打造每家企业都能拥有的AI生产体系 - 超聚变
- [55] 中国电信副总经理栾晓维：推动从传统流量经营向Token经营转型，目标是让AI

好用、管用、用得起—顶端新闻

[56] 长江证券-软件与服务行业深度报告：AI沉思录（二），Token工厂，从“堆资源”到“榨资源”的产业趋势-260630-研报-行业分析-慧博投研资讯

[57] 从Byte到Token，电信运营商抢抓大模型关键入口 - 21经济网

[58] “Token工厂”开启真实落地：运营商启动百亿级集采 大厂加速卡位

[59] 北京移动推出一站式Token运营服务平台 - 中国商报网

[60] 中国移动浙南Token工厂在温州揭牌

[61] 三大运营商Token转型最新动作_通信世界网

[62] 专访联通数智王恺：以应用锚定价值，联通元景布局Token一站式精细化运营 - 推荐 — C114通信网

[63] 广东省发展和改革委员会 - 人工智能应用典型案例之三：AI+政务服务

[64] 华为中国行2026 \{\}| 一座制造业城市的AI跃迁-腾讯云开发者社区-腾讯云

[65] 迅策与TCL旗下格创东智签署战略合作 推动“制造业Token工厂”落地_证券之星VIP

[66] PDD，Token工厂必备的“超级管线” “前店后厂—中心”布局来了！

[67] 城记 \{\}| 从无锡Token工厂，看东西部“算电协同” - 中国金融信息网

[68] 天云数据首批加入信通院“Token服务能力攀登计划”，硬核产品实力获权威认证

[69] 武汉开出AI数据交易“第一票”！AI病理Token有身价

[70] 四川发布 | 四川电信发布“智云四川”Token一体化服务-四川省政府国有资产监督管理委员会

[71] 中科闻歌海汇token平台发布，从部署模型走向token订阅模式

[72] WAIC 2026收官 | 范式大会亮点集锦，见证AI 2.0从技术突破走向产业实践 – 量子位

[73] 九章云极发布“AI工厂”战略 定义智能规模化新基建-36氪

[74] 超聚变亮相WAIC 2026：共建智能体时代，看见智企进化新路径-中新网

[75] 95后深漂回汉，四人小厂如何拿下上市企业存储订单？ -荆楚网-湖北日报网

[76] 加码Token工厂！欧洲AI新锐6亿美元收购20人公司 剑指全球顶级推理平台

- [77] Zero Hash_百度百科
- [78] I-ON Digital Corp_百度百科
- [79] 一文了解泰达币（Tether）投资Parfin以推动USDT在拉丁美洲的机构采用_资讯_区块链_脚本之家
- [80] 从MiCA到DORA：欧盟如何为RWA搭建制度高速公路凤凰网数字经济_凤凰网
- [81] AI出海新范式！迅策联合欧洲Lutech集团开发场景Token工厂，订单驱动ARR强劲增长_财经频道_证券之星
- [82] EURC_百度百科
- [83] 欧洲列支敦士登主权区块链网络LTIN正式启动-金融界
- [84] 阿联酋“星际之门”：拟建60万亿token工厂，为全球AI提供“智慧原油”
- [85] 华赢集团AXG获巴林首张Stablecoin发行人牌照，全球跨境支付布局再提速-国内-金投财经频道-金投网
- [86] Firmus Technologies_百度百科
- [87] IREN Ltd._百度百科
- [88] 投资118亿美元！韩国运营商KT转型AI平台企业，筹备建设Token工厂
- [89] 高市早苗支持者甩锅日本记者 加密货币丑闻再起波澜_中华网
- [90] 稳定币支付公司 Daya 完成 240 万美元 PreSeed 轮融资，Hivemind ... | Daya, 稳定币 - ChainCatcher
- [91] Flutterwave联手Tempo打造非洲新一代稳定币支付基础设施-移动支付网
- [92] 中信证券：Token工厂及Token运营商推动产业链价值重估
- [93] 硅基流动：打造“Token工厂”，让智能像水电一样普惠 - 新华网客户端
- [94] 推理拐点已至：五象云谷Token工厂，开启算电协同新基建
- [95] 技术升级、算电协同推动算力普惠_财经_中国网
- [96] 清程极智：首创Token“前店后厂” - 中国日报网
- [97] 怡亚通：与「DaoCloud道客」签署战略合作协议 共建“城市Token工厂”
- [98] AI出海新范式！迅策联合欧洲Lutech集团开发场景Token工厂，订单驱动ARR强劲增长-科技频道-和讯网

- [99] “人民数据·城市词元工厂”平台启动建设：让城市像用电一样用上AI-经济·科技-人民网
- [100] 行业动态_福建省数据管理局_福建省发展和改革委员会
- [101] 词元赋能智库建设与人才培养的实践创新-中国社会科学网