

AI 时代词元可信流通

基础设施研究报告

联通数字科技有限公司

中国信息通信研究院

2026 年 9 月

前 言

当前，以人工智能为代表的新一代信息技术迅猛发展，正在深刻改变生产生活方式，重塑全球经济结构和竞争格局。随着大模型技术不断成熟、智能体（AI Agent）加速落地，人工智能正从“工具辅助”迈向“自主智能”新阶段，数字经济加快向智能经济、数智经济演进。词元（Token）作为大模型运算的基本单元，已成为智能时代价值计量、传递与交换的核心载体，词元的规模化生产与可信流通，正在成为驱动人工智能产业高质量发展和数字经济转型升级的关键力量。

党中央、国务院高度重视人工智能健康发展和数据要素市场化配置改革。党的二十大报告明确提出，加快建设网络强国、数字中国，构建新一代信息技术、人工智能等新的增长引擎。《“十四五”数字经济发展规划》深入实施，“人工智能+”行动全面推进，为词元等新型价值载体的培育发展提供了根本遵循。2026年，全国科学技术名词审定委员会正式将“词元”作为Token的标准中文译名向社会发布，国家有关部门相继出台政策文件，明确提出探索词元交易等新型交易模式，词元流通的制度框架和市场环境加快形成。

词元流通在规模快速扩张的同时，仍面临计量口径不一、跨云流转壁垒突出、合规审计体系缺位、跨境规则衔接不畅等突出问题，可信流通保障机制建设滞后于产业发展步伐。加快构建安全、可信、高效的词元流通基础设施，既是释放人工智能价值潜力、培育发展新质生产力的必然要求，也是维护国家数据安全、掌握智能经济发展主动

权的重要保障。

为系统研判词元可信流通的发展态势、关键问题与实现路径，联通数字科技有限公司联合中国信息通信研究院，在深入调研和广泛论证的基础上，编制形成本报告。报告系统梳理了词元可信流通的演进背景与现状特征，提出了词元可信流通基础设施的关键能力要求，总结了政务、医疗、教育、央国企等重点领域的应用实践，并对技术融合与产业生态发展趋势作出展望、提出对策建议，以期为政府部门决策、产业界实践和学术界研究提供参考。

目 录

1. 词元可信流通背景机遇.....	1
1.1. 互联网核心概念持续演进，词元成为智能经济核心计量单元.....	1
1.2. AI Agent 崛起催生数智经济新形态，词元流通成为关键通道.....	4
1.3. 词元流通已成全球产业关注焦点，可信流通保障问题待解决.....	6
2. 词元可信流通现状特征.....	9
2.1. 落地生态逐步聚合，合规审计缺位风险加剧.....	9
2.2. 多云部署成为主流，词元跨云流转壁垒突出.....	12
2.3. 运营商加速转型词元经营，可信流通底座基本成型.....	15
3. 词元可信流通基础设施关键能力.....	17
3.1. 词元可信流转的基础设施要求.....	17
3.2. 词元流转核心技术.....	22
4. 词元可信流通重点应用场景.....	26
4.1. 政务场景词元可信流通实践.....	26
4.2. 医疗场景词元可信流通实践.....	28
4.3. 教育场景词元可信流通实践.....	30
4.4. 央国场景词元可信流通实践.....	32
5. 发展趋势.....	34
5.1. 技术融合趋势.....	34
5.2. 产业生态趋势.....	36
5.3. 未来发展建议	37

图 目 录

图 1 企业多云/混合云部署现状与市场规模	13
图 2 医疗案例架构图	29
图 3 教育案例架构图	32

1. 词元可信流通背景机遇

1.1. 互联网核心概念持续演进，词元成为智能经济核心计量单元

词元（Token）革新信息流通载体，成为价值互联网在 AI 时代的“度量符号”。互联网在过去三十年完成了全球数据的互联互通。这一阶段的互联网可以被称为“信息互联网”，其架构设计围绕信息的复制与传输展开，文本、图片、视频、音频等数字化内容以数据包的形式在网络节点之间自由流动。当互联网的应用从单纯的信息共享转变为可信数据交换与知识构建，“价值互联网”概念应运而生。随着 AI 智能体（Agent）时代的到来，互联网上传输的主要内容发生了根本性变化，从消息变成任务，Agent 接收到任务后，在后台调用模型进行推理、访问数据库、执行代码等操作。词元成为 AI Agent 时代互联网上最主要的价值载体。

词元具备三重特性，推动智能技术商业化发展。词元作为大模型运算的最小单元，填补了 Byte 在价值度量层面的结构性缺位，成为智能经济时代的基础价值度量载体。词元作为大模型对文本、代码、图像、音频、视频等多模态信息进行切分与编码后的最小运算单元，恰好填补了 Byte 在价值度量层面的缺位。2026 年 3 月，全国科学技术名词审定委员会正式将“词元”作为 Token 的标准中文译名向社会发布¹。从技术层面看，词元是大语言模型对输入输出进行分词处理后的基本运算单元，中文文本中的一个汉字、一个标点或一个词语片段，

¹全国科学技术名词审定委员会关于发布试用人工智能领域名词词元（Token）中文名“词元”的公告

皆可被切分为若干词元。模型每一次推理都在消费词元，产出的每一段文本也以词元计量。这一技术属性赋予词元三个特征。**一是，可计量性。**词元将抽象的“智能”转化为可精确计数的最小单元，使 AI 服务具备了类似电力“千瓦时”的标准化量度。**二是，可定价性。**从定价模式看，市场已形成按量付费、词元资源包和订阅制三种主流形态。在按量付费模式下，用户根据实际使用的词元数量或 API 调用次数实时结算费用。在词元资源包模式下，用户预先购买固定数量的词元额度，在实际使用中按消耗量扣减，超出部分再按量付费或另购资源包。在订阅制模式下，用户按月/年支付固定费用，在有效期内获得一定范围的无限或定量使用权限。定价机制的形成标志着词元从技术指标走向经济变量。**三是，可结算性。**词元的流通已不再局限于企业与终端用户之间的 API 调用关系，而是拓展至跨平台分发、智能体间协作、跨境数字贸易等更广泛场景。国家数据局 2026 年 6 月发布的《关于推进行业高质量数据集建设行动的实施方案》明确提出“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”²，标志着词元交易正式进入制度框架。

词元流通逐步从单纯的技术调用演变为价值交换行为，“词元可信流通”的概念由此形成。词元流通，是指词元作为智能服务的价值载体，在模型服务提供方、平台运营方与终端用户之间，以及在不同平台、不同智能体之间，完成从算力生产、调度分发到消费使用、计费结算的价值传递过程。“可信”，则是对这一流通过程的质量约束，

²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

主要包括三方面：一是计量可信，词元数量的统计口径统一、消耗量可独立验证；二是结算可信，费用核算公开透明、可经第三方审计；三是流通可信，词元在跨主体流转中的来源、路径与使用去向全程可溯、合规可查。简言之，词元可信流通就是要让每一份词元的来源、数量、价格与去向都可验证、可审计、可追溯，使其成为智能经济中可信赖的价值计量与交换单元。

从数据确权到词元可信流通，理论框架一脉相承。词元可信流通的理论框架的前身是数据要素的可信流通，数据要素可信流通解决了“数据作为生产要素如何安全、合规地流动”的问题，而词元可信流通则是在此基础上进一步解决“AI 时代，智能服务的价值载体如何被可信计量、可信传递、可信清算”的问题。**首先**，词元可信流通与数据要素流通在概念范畴、核心逻辑与基础设施类型上存在本质差异。“可信流通”与“数据要素流通”是两个须严格区分的概念范畴。数据要素流通的核心逻辑是“确权—定价—交易”，强调数据作为生产要素的产权界定与市场化配置。词元可信流通的核心逻辑则是“生产—调度—消费—结算”的连续闭环，强调智能服务从算力转化到价值交付的全链路连续性与可验证性。**其次**，产业需要的不仅是一套数据产权和合规流通的制度，更需要一套围绕词元的计量、定价、传递和清算的可信机制。这一演进在操作层面可以总结为四个相互衔接的环节。**一是词元计量确权**，将模型推理的每一次词元产出与其来源数据、算力投入和模型版本绑定，形成可追溯的词元“出生证明”。**二是词元定价与分发**，依托模型 API 和 Agent 平台的标准化接口，词元在供需双方之

间按需流转，价格信号实时反映模型能力、算力成本 and 市场需求。三是词元价值交割，用户在实际使用场景中调用词元并将其转化为现实生产力，完成从“技术消耗”到“经济产出”的闭环。四是词元可信清算，通过独立审计、计量校验和合规追溯，确保词元流通全链条的准确性、公平性和合规性。这四个环节共同构成了词元可信流通的完整闭环，将词元经济从分散的 API 调用行为升级为体系化的价值流通网络。

1.2.AI Agent 崛起催生数智经济新形态，词元流通成为关键通道

AI Agent 经过三个阶段演进进入落地变现期，驱动词元流通需求。第一阶段，语言理解与生成。2022 年，ChatGPT 的发布标志着大模型具备了流畅的自然语言交互能力，是解决 Agent 与人类之间“沟通”问题的基础。第二阶段，编程与工具使用。2023 至 2024 年，大模型学会了两项关键能力：代码生成和外部工具调用。OpenAI 的 Function Calling 机制使模型能够操作 API、数据库和文件系统，Claude3.5 和 GPT-4 展示了成熟的编程能力。同期，模型上下文窗口从数万词元扩展至百万词元级别。第三阶段，深度推理与自主规划。2024 年底至 2025 年初，OpenAI o1/o3 系列和 DeepSeek-R1 等推理模型的出现，使模型具备了长链逻辑推理、复杂任务分解、执行计划制定与逐步验证的能力。2025 年是业界公认的“AI Agent 元年”，AI Agent 正式开始产品化落地。2025 年 1 月，OpenAI 推出首个面向消费者的通用浏览器 Agent——Operator。与此同时，Agent 间协作协议也在走向标准化。Anthropic 于 2024 年底推出的 MCP（Model Context

Protocol) 在 2025 年迎来了爆发式增长, 至 2025 年底已有超 1 万个活跃公共 MCP 服务器和 9700 万次月度 SDK 下载³。随着 AI Agent 技术的成熟和相关应用的产业化落地, 词元消耗量也呈指数级攀升。在需求规模层面, 中国日均词元调用量从 2024 年初的 1000 亿飙升至 2025 年底的 100 万亿, 至 2026 年 3 月突破 140 万亿, 两年增长超千倍⁴。摩根大通预测, 2025 年至 2030 年, 中国 AI 推理词元消耗量将从约 10 千万亿增长至约 3900 千万亿, 五年增长约 370 倍⁵。

Agent 对词元流通的驱动力体现在两方面。一是 Agent 的多步推理特性导致词元消耗远超单次问答。以 OpenClaw 为代表的 Agent 执行一次复杂任务需要多轮推理、多工具调用和长链规划, 单次任务 Token 消耗可达简单问答的数十倍甚至数百倍。2026 年 3 月, OpenClaw 贡献了 OpenRouter 平台 20% 的词元消耗量, 仅其单周通过平台流通的词元规模就相当于 2025 年第四季度全平台周均词元流通量的 60%⁶。二是多 Agent 协作体系放大了词元流通的网络效应。当多个专业化 Agent 协同工作时, 词元不仅在用户与 Agent 之间流通, 更在 Agent 与 Agent 之间形成高频交互链路。这种多代理协作架构使词元流通从“单点到单点”升级为“网状互联”, 词元流通的规模和复杂度均呈数量级增长。

词元流通承担了智能经济中“定价—交易—结算”三位一体的通

³Donating the Model Context Protocol and establishing the Agentic AI Foundation 捐赠模型上下文协议并成立智能体人工智能基金会

⁴国家数据局党组书记、局长刘烈宏出席中国发展高层论坛 2026 年年会并演讲 (全文)

⁵China Internet & AI: Pre-Spring Festival Thoughts, From Application Adoption Perspective 《中国互联网与人工智能: 春节节前思考, 从应用落地视角分析》

⁶OpenRouter 官方 2026 年 3 月周度用量统计面板 (公开运营数据)

道功能。经济运行依赖生产者与消费者之间如何就产品价值达成共识并完成支付的底层机制。在工业经济中，货币体系承担这一角色；在信息经济中，流量和数据包构成了主要的计价单元。但在智能经济中，用户消费的不是特定数量的数据包，而是一个由 Agent 自主完成的多步任务，其成本结构由词元消耗直接决定，其价值交付也通过词元调用才得以实现。这一模式下，词元定价、词元分发和词元结算共同构成的词元流通系统成为智能服务商流、信息流和资金流的汇聚点，是智能经济中最基础的商业管道。

1.3.词元流通已成全球产业关注焦点，可信流通保障问题待解决

全球 AI 巨头错位竞争，产出、分发、协议三环结构成型。全球词元流通领域，正围绕词元的规模化产出、词元的跨平台分发以及词元流通的协议标准展开。在词元产出环节，专用推理硬件成为竞争焦点。NVIDIA 于 2026 年推出 NVIDIA Groq3LPX 推理机架，与 Rubin GPU 的并行预填充形成“GPU 预填+LPU 解码”的分工架构，使系统推理吞吐量与功耗比提升 35 倍⁷。微软 Azure 的 Fairwater 数据中心同样瞄准极端词元吞吐能力，通过液冷散热和 NVLink 互联架构达到每秒 110 万词元的处理吞吐⁸。

在词元分发环节，支持多模型统一托管的云平台成为主流形态。

以 Amazon Bedrock 为例，平台同时托管 Claude、GPT、Gemini、Grok

⁷英伟达技术解读文章 “Inside NVIDIA Groq 3 LPX: The Low-Latency Inference Accelerator for the NVIDIA Vera Rubin Platform”

⁸Signal65. Breaking the Million-Token Barrier: The Business Impact of Azure ND GB300 v6 Performance for Enterprise AI[R]. 2025- 11- 04

等多种模型，并在此基础上提供三种多 Agent 编排模式（监督者 - 工作者、顺序链、并行扇出）和 40 余个预建企业工具集成⁹，使 Agent 的每一次任务执行都通过 Bedrock 转化为词元调用。微软 Azure AI Foundry 采取了类似的多模型聚合策略，但更强调词元成本优化，平台同时托管 Anthropic Claude 和 DeepSeekV4，通过模型分层实现“高价模型处理复杂推理、低价模型覆盖高频任务”的词元分配。

在协议标准环节，MCP 和 A2A 构成了词元流通的两层互操作基础。Anthropic 于 2024 年底推出的 MCP（Model Context Protocol）解决了 Agent 调用工具的标准化问题。Google 于 2025 年推出的 A2A（Agent-to-Agent）协议则解决了 Agent 之间的标准化通信。MCP 和 A2A 形成了清晰的协议分工，前者是 Agent 与工具之间的“接口标准”，后者是 Agent 与 Agent 之间的“通信语言”，二者共同构成了词元在异构系统中跨厂商、跨平台流通的协议底座。

我国词元流通生态四层架构初步明确，可信保障机制亟待完善。基础设施层以算力供给和网络调度为核心。三大电信运营商和阿里云、华为云、腾讯云等云厂商构成算力供给的主力。截至 2026 年 7 月，全国智算规模超 245 万 PFlops¹⁰，“东数西算”八大枢纽节点覆盖东中西部 14 个省份，全国一体化算力网已实现对主要枢纽集群和运营商的监测调度。模型层以国产大模型集群为支撑。DeepSeek、通义千问、豆包、文心一言、智谱 GLM、MiniMax、月之暗面（Kimi）等构成国内模型供给的主力阵容，借助混合专家架构（MoE）和推理

⁹Amazon Web Services. Use multi-agent collaboration with Amazon Bedrock Agents

¹⁰国家数据局党组书记、局长刘烈宏 2026 中国国际大数据产业博览会致辞

引擎优化大幅压缩推理成本。应用服务层以 AI 智能体平台和行业应用为核心。在办公智能体赛道，腾讯云 WorkBuddy 打通微信和企业微信生态，支持远程下发并执行任务，同时作为模型聚合平台接入 DeepSeek、混元、GLM、Kimi 等多款模型。阿里云将 QoderWork、悟空、MuleRun 三款 Agent 整合为“千问办公”，实现“沟通即执行”的工作流闭环。字节跳动以豆包 3.45 亿月活的 C 端流量为基本盘¹¹，通过飞书 aily 智能体平台将 Agent 嵌入企业协作流程。交易流通层正在构建词元作为标准化交易品的市场机制。国家数据局提出探索词元交易等新型交易模式，多地已开始试点。

我国词元流通的可信保障机制未能与产业规模同步完善。一是计量可信缺失。不同模型使用独立的分词器算法，同一段输入文本在不同模型下可能被切分为数量相差数倍的词元，导致跨平台的词元消耗量不具备可比性，用户无法验证“被收取的词元数量”是否真实反映实际推理消耗。二是，结算可信缺位。当前词元计费完全由模型提供方或云平台单方统计，缺乏独立、可审计的第三方计量机制，企业级客户的大规模词元采购面临“既是运动员又是裁判员”的信任困境。三是，流通可信滞后。词元在跨平台、跨 Agent 的流通过程中，其来源、质量和使用合规性缺乏可验证的追溯链条。当 Agent A 委托 Agent B 执行任务时，B 消耗的词元是否必要、是否合规、是否有重复计费，在现有技术架构下均无法得到可信回答。词元流通与数据流通一样面临“安全、合规、高效流通”的“不可能三角”，且因词元的实时性、高

¹¹2026 年 3 月 QuestMobile 监测数据

频性和跨系统特征而更加尖锐。若可信问题不解决，词元经济将长期停留在“双边信任”层面，而无法升级为“制度信任”层面的公共基础设施。

2. 词元可信流通现状特征

我国词元流通产业正处于技术验证向规模落地过渡的关键期，生态聚合态势显著，但合规审计体系不健全、跨云流转壁垒突出、跨境合规通道不足等结构性问题凸显，产业发展与制度供给、技术支撑的矛盾逐步显现，对新型数字基础设施建设提出迫切要求。中国信通院数据显示，2025 年我国 Token 云服务市场规模已达 59 亿元，日均 Token 调用量从 2024 年初的约 1000 亿升截至 2026 年 6 月已突破 500 万亿，两年间增长数千倍，标志着 AI 产业已全面进入以推理消费为主导的“Token 经济”时代。

2.1. 落地生态逐步聚合，合规审计缺位风险加剧

2.1.1. 词元流通场景快速聚合，品类与覆盖持续扩大

市场发展层面，词元可信流通的产业生态已初步形成算力词元、数据词元、服务词元三大核心品类，场景覆盖从单点试点向行业级、产业链级聚合演进。据中国信通院《智能算力服务研究报告（2026 年）》数据，2025 年我国智能算力服务市场规模突破 1300 亿元，全国十余家数据交易所开展数据词元试点，上架产品超 2000 项，智能算力服务化已成为产业核心趋势。

市场增长层面，数据要素市场化进程为词元流通提供了底层制度

支撑。国家数据局数据显示，2025 年我国数据交易规模接近 5000 亿元，大数据产业市场规模突破 8500 亿元。截至 2026 年初，全国已建成或筹建 58 家数据交易机构，覆盖国家级、区域级和行业级多层次架构。北京国际大数据交易所截至 2025 年 6 月累计数据交易规模达 2250TB，近三年年均增长超 200%，已与超半数头部 AI 模型企业深度合作，为大模型提供覆盖 32 个行业 475 个数据集。数据交易类型从传统的风控核验、精准营销，向 AI 大模型训练、医疗诊断、科研创新等高价值领域延伸。

2.1.2.词元流通合规体系建设滞后，全链路风险突出

词元可信流通领域的合规审计体系建设明显滞后于产业扩张速度，全链路合规风险已成为制约行业健康发展的突出瓶颈。具体表现为三个层面。**一是专项监管规则与标准缺失。**词元法律属性、业务边界尚未明确，发行、流转、兑换、核销全生命周期缺乏系统性监管规则。不同平台计价口径、计量标准不统一，行业计量标准统一度不足 30%。中国信通院 2026 年 6 月联合十余家企业发布的《基于 Token 的云服务能力要求》标准及评估体系，以及 AIIA 词元服务工作组的成立，标志着行业标准化建设开始加速，但从标准发布到全产业落地仍需较长周期。**二是全链路穿透式审计能力不足。**链上链下数据难以协同映射，跨主体审计数据互不打通。据云安全联盟(CSA)与 Tenable 联合发布的《State of Cloud and AI Security 2025》报告，82%受访企业采用混合云环境，平均管理 2.7 个云环境，由此带来的“割裂式可见性、不一致的身份治理、风险监控盲区”为安全事件提供了可乘之

机。组织在过去 18 个月平均发生 2.17 次云安全事件，其中 59%的企业将“身份安全薄弱与权限风险”列为首要威胁。三是安全合规底线存在隐患。部分平台词元承载的算力、数据资源未经过合规审查，账户与权限管理不完善，存在数据泄露、算力窃取、计量篡改等风险。CSA 报告显示，云安全事件的前四大成因中，三个与身份管理直接相关：配置错误（33%）、权限过度（31%）、访问控制不一致（27%）、身份卫生薄弱（27%）。在 AI 应用层面，55%的企业正在使用 AI，其中 34%已遭遇 AI 相关数据泄漏事件。此外，部分平台发行机制不透明，存在超发滥发的潜在风险。

2.1.3.词元出海市场空间广阔，跨境合规成为首要挑战

数字经济规模化出海，词元跨境流通场景增多。我国数字经济出海向技术密集型升级，算力出海与 AI 大模型出海带动词元跨境流通规模高速增长。出海区域覆盖东南亚、中东、欧洲、拉美等市场。流通模式形成 API 直连、离岸节点、跨境交易三类主流模式，整体仍处于快速发展阶段。市场层面，全球市场已形成三大梯队，分别为以美国、欧盟为代表的成熟市场，法律法规完善，分行业数据确权与跨境流通体系成熟；中国、新加坡、日韩等高速增长区，依托完整制造业基础与跨境金融通道，垂直分层运营体系快速完善；东南亚、中东、拉美等起步市场，以基础消费互联网为主，垂直赛道标准化交易体系尚未建立。中国作为首个将数据作为生产要素的国家，正在体系化推进数据要素市场化配置改革，全国一体化数据市场正在加快构建，为词元跨境流通奠定了制度基础。

全球治理格局碎片化，词元跨境监管难统一。全球数字治理呈现碎片化格局，词元跨境流通面临国内外监管规则双重约束与显著冲突。**国际层面**，欧盟以 GDPR、《人工智能法案》构建严苛合规体系，对跨境数据传输设定充分性认定、标准合同条款（SCC）、约束性企业规则（BCR）等路径，合规要求严格。美国通过出口管制限制算力与 AI 技术跨境，先后将多批中国实体列入“实体清单”，对高端 AI 芯片出口实施许可证审查。新兴市场普遍推行数据本地化政策，要求关键数据在境内存储和处理，各区域规则互不兼容。CSA 报告指出，到 2026 年，作为主权云战略的一部分，35%的组织将寻求提供主权控制的基础设施提供商，反映数据主权意识的全球性上升。**国内层面**，我国已建立安全评估、标准合同、认证机制三大数据出境合规路径，但词元跨境业务属性尚未明确定性，企业难以精准匹配合规路径；规模化词元跨境调用易触发个人信息出境安全评估阈值，合规压力较大。国内外规则在数据可携带权、技术出口管制、数据本地化等维度存在多重矛盾，企业面临合规与业务发展的两难困境。AI 模型训练数据的跨境流通方面，各国对训练数据集的合规要求差异显著，词元承载的训练数据、推理结果在不同法域下的法律属性和合规义务尚未形成共识。

2.2. 多云部署成为主流，词元跨云流转壁垒突出

2.2.1. 多云部署已成企业标配，词元体系呈碎片化分布

多云架构已成为企业数字化转型标配。据 IDC2025 年 7 月发布的《Asia/Pacific Managed Cloud Services Market Forecast》报告，亚太

地区 89%的企业在多个公有云上部署工作负载，72%采用真正的混合云模式，平均管理 2.7 个云环境。据混合云产业推进联盟数据，2025 年我国 88.7%的企业采用多云部署架构，中央企业多云部署比例达 92%。IDC 预测，到 2028 年，超过 90%的新开发应用将具备多云能力，利用平台交付的能力推动创新。

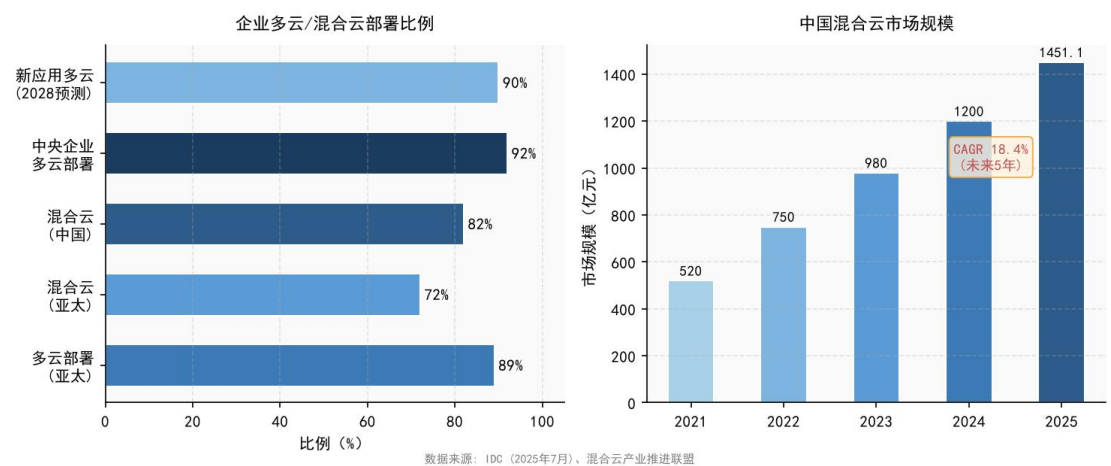


图 1 企业多云/混合云部署现状与市场规模

中国混合云市场规模持续高速增长。据 IDC《中国混合云/私有云市场洞察，2025H2》报告，2025 年中国混合云整体市场规模达 1451.1 亿元，其中 IaaS 软件市场规模 259.4 亿元，PaaS 软件及服务市场规模 536.6 亿元，预计未来五年复合增长率为 18.4%。AI 算力占比显著上升、智能体大规模部署正推动企业级基础设施加速扩容，混合云市场进入深化发展阶段。

词元体系呈现碎片化分布特征。主流云厂商均构建自有词元生态，格式规范、计量规则、加密方式、权限体系各自独立，跨平台互不兼容，形成独立价值孤岛，与算力网络一体化、数据要素全域流通

的发展方向不相匹配。IDC Future Scape 预测，到 2027 年，超过 50% 的企业将现代化改造多达一半的云架构以提升效率和驱动创新；到 2026 年，35% 的组织将寻求提供主权控制的基础设施提供商，60% 的企业将利用专用加速云计算服务优化 AI 应用部署。

2.2.2. 词元流通场景多，技术安全结算标准尚未形成

一是技术标准不互通。各厂商词元格式、接口协议、计量规则均为自有标准，身份信任链无法跨云传递，跨平台价值映射缺乏统一标尺。据 Gartner 2025 年报告，全球 65% 以上的企业因多云技术不兼容，导致数据与服务传输效率降低 38%，该问题在词元流转场景更为突出。IDC 调研显示，44% 的企业将“云技能缺口”列为首要挑战，39% 认为战略不清晰，35% 指出预算不足，31% 表示资源被分流至其他优先事项。

二是安全体系割裂。不同云厂商身份认证、密钥管理、访问控制体系差异显著，跨云流转易出现认证断层、权限失控，存在词元篡改、数据窃取等安全隐患。据云安全联盟调研，多云环境下 59% 的企业将身份安全列为首要风险，三大身份相关成因——权限过度（31%）、访问控制不一致（27%）、身份卫生薄弱（27%）——合计占云安全事件的主体部分。28% 的企业报告云安全团队与 IAM 团队协同不足，21% 的企业在规模化执行最小权限原则方面困难重重。组织在过去 18 个月平均发生 2.17 次云安全事件，仅 8% 被归类为严重事件，但频发的事件表明安全防护体系与云、AI 快速扩张之间存在显著差距。

三是结算机制不通。各云厂商计费与结算体系独立，跨云流量成本高企，企业对账核算成本较单一云环境提升 60%以上，跨云流量成本占云总支出比例平均达 18%。缺乏统一的跨云计量与结算标准，导致词元跨云流转的隐性成本居高不下，严重制约了算力资源的跨域优化配置与按需调度。

2.3. 运营商加速转型词元经营，可信流通底座基本成型

基础电信运营商正从流量时代的“管道提供者”，加速转型为词元生产、调度、分发与运营一体化的综合服务商。依托全国性骨干网络、规模化算力资源池和长期服务政企客户形成的信任基础，运营商正在成为词元可信流通基础设施建设的核心力量。2026 年以来，三大运营商相继将“词元经营”写入战略主线，词元产品化、平台化、集约化布局全面提速，词元开始从模型专用额度转变为类似通话包、流量包的基础资源。

2.3.1. 三大运营商差异化竞速，词元产品规模化落地

词元套餐密集上线，词元成为运营商新增计费单元。2026 年 4 月，中国联通面向个人和团队客户推出了 Token Plan 套餐和各类按量计费资源包产品线，覆盖模型推理调用、模型微调训练和知识库构建等场景。4 月底，推出 AI 算力套餐，包含 Token Plan 三档、Coding Plan 两档和融合套餐，其中融合套餐将手机通信、千兆宽带、AI 算力和 AI 云桌面打包为一个产品。2026 年 5 月，中国电信在世界电信日推出试商用词元套餐，面向个人及家庭客户提供 9.9 元至 49.9 元三档月

费产品（含 1000 万至 8000 万词元），面向开发者及中小微企业提供 39.9 元至 299.9 元三档产品（最高含 1.5 亿词元）；中国移动随即在全国上线词元套餐，最低 5 元月包，并在多地推出“1 元 40 万词元”等普惠产品；2026 年 6 月，三大运营商词元产品正式登陆中国算力平台，天翼云词元 Token Plan、移动云 Coding Plan、联通云 Coding Plan 及 Token Plan 同步上架“算力超市”，面向中小企业提供普惠化词元服务。

经营战略差异化定位初步形成。中国电信提出“以 Token 经营重塑公司业务”，于 2026 年 4 月发布星辰 Token Hub 运营服务平台 1.0，将“天翼 Token 币”定义为 Token 经营流通的统一量纲，构建跨终端、跨业务、跨场景的一体化词元运营服务中枢，并以自研星辰大模型提供稳定可控的自有词元供给。中国移动正式发布词元运营生态体系，通过 MoMA 平台接入超 300 款模型，实现“一次接入、智能优选、普惠可用”，词元集约化运营使单位词元成本压降约 30%、资源占用率降低 50%以上，并打通“Agent 使用词元、词元拉动算力”的服务链路。中国联通提出“Agent+Token+AI 云”算力经营新模式，明确不再单纯做通道经营，而是转向经营智能，对外输出“算力+模型能力+智能应用”一体化服务。

2.3.2.网络、算力、合规三重禀赋构筑可信流通基础能力

运营商参与词元流通具备三重结构性优势。一是网络可信，全国性骨干光网络与海量边缘节点构成词元流式交付的确定性通道，中国电信已部署超 3000 个边缘算力节点，可按需敏捷接入；二是算力可

信，运营商智算集群规模领先且国产化适配程度较高，天然契合“东数西算”一体化算力调度要求；三是合规可信，运营商长期承担关键信息基础设施运营职责，在政务、金融等高安全行业具备显著的安全合规资质与客户信任优势。随着算力、算法、数据矩阵形成，运营商在政务、金融等关键领域的安全合规优势将为其打开新的增长空间。

总体来看，运营商的全面入局推动词元流通从互联网平台单点竞争转向运营商云主导的基础设施级供给，为词元可信流通提供了网络、算力与合规的底座支撑。但也应看到，跨运营商、跨云平台的词元互认、计量口径统一与穿透式审计仍待破题，运营商之间、运营商与互联网云厂商之间的可信互联机制尚未建立。

3. 词元可信流通基础设施关键能力

3.1. 词元可信流转的基础设施要求

词元可信流通基础设施，是指以集约化算力、确定性网络和体系化安全防护为物理底座，融合可信计量、智能调度、隐私计算、全链路审计等关键技术，支撑词元在生产、调度、交付、计量、结算各环节实现来源可溯、计量可信、流转可控、使用合规的新型数字基础设施。其核心功能在于为词元跨主体、跨平台、跨区域流通提供统一的计量基准、可信的执行环境和可审计的流通通道，构成智能经济中价值计量与信任传递的公共底座。

词元可信流转并非单一技术指标的优化，而是“计量、交付、防护”三类能力在算力、网络、安全底座上的综合映射。结合词元流通“请

求接入—语义路由—算力调度—推理生产—流式交付—计量结算”六环节链路，反向推导基础设施的三类客观要求。

3.1.1.算力基础设施要求：规模化、集约化与异构兼容

词元的爆发式增长对算力底座提出规模化要求。截至 2026 年 6 月，我国日均词元调用量已攀升至 500 万亿次以上，2025 年公有云大模型对客侧词元调用量约达 2000 万亿。算力底座须同时满足规模化集约供给、高效集群运行与异构芯片兼容三项技术要求。**一是规模化集约供给能力。**当前主流词元服务平台深耕算力集约化运营，对全域异构算力实现统一纳管、聚合池化与智能调度，显著提升资源利用率、降低资源内耗。在算力布局方面，建立“中心－区域－属地”三级布局实现“算电协同、集约集聚”，通过顶层规划、全域协同，破解“低碳+超大规模”并存的供给难题。建设区域算力节点，实现圈层联动、片区赋能，通过聚合区域算力供给，跨域互通，就近赋能，实现精准匹配 AI 业务诉求。建设属地智算枢纽，满足市域自治、市域强基的需求，统筹全域算力资源，实现集约共建、高效调度。以联通为例的头部云服务商已建成标准机架总规模超 110 万架、落地 7 个百兆瓦级 AIDC 园区、智算规模达 45EFLOPS 的算力底座，全面覆盖“东数西算”国家枢纽节点，为 Token 的稳定、大规模生产提供资源保障。**二是高效集群运行能力。**算力规模须转化为有效词元产能。行业先进水平的 GPU 集群平均利用率训练场景可达 80%以上、推理场景约 50%，训练与推理 GPU 利用率存在显著差距，一方面源于通信模式差异：分布式训练以 AllReduce 同步梯度聚合为主，输入批次规整，负载均

衡度高；另一方面推理业务普遍采用异步动态批处理机制，用户请求的序列长度高度异构，带来显著的负载不均衡。GPU 利用率与词元成本之间存在可量化的因果关系底层 GPU 利用率每提升 10%，上层单词元成本可降低 8%—12%，算力效率直接决定词元的市场竞争力。

三是异构芯片环境下的推理一致性。国产算力多元化格局对词元生产环节提出跨平台一致性要求。2026 年 5 月国家发展改革委明确“指导国产大模型加大力度适配国产算力芯片”，算力底座适配主流国产 AI 芯片已成为生产层的基础技术要求。推理一致性包含两层技术含义，其一，同一模型在不同芯片架构上的输出结果须保持语义等价，避免因数值精度差异导致输出漂移。其二，一套推理引擎须实现跨平台无缝部署，避免为每种芯片维护独立的推理框架分支。

3.1.2.网络基础设施要求：确定性低时延与高可靠传输

词元流式交付特性决定了网络通道的刚性要求。词元以流式方式逐一生成并即时返回，网络时延与抖动直接转化为首词元时延（TTFT）与单词元生成时延的劣化。

一是跨区域确定性低时延。当前主流平台依托全国一体化算力网，以算“织”网、以网促算，构建算力网一体全域纳管与调度能力。依托 RoCE 高速网络与全域高性能算力专网，构建全国一体化算力网，为跨域协同、大规模算力调度提供低时延、高可靠的网络底座。全国一体化算力网要求骨干网络提供确定性传输保障，按照跨省骨干网络传输时延按 95%分位统计，要求京津冀至长三角小于 8ms，京津冀至粤港澳小于 18ms，京津冀至成渝小于 22ms；省际骨干网业务连通可用性可达 99.99%。

二是全域调度与

就近接入能力。词元消费具有显著的地域分布特征，就近接入是降低网络时延的直接手段。要求算力调度平台在全国骨干云池提供中转分发服务，形成覆盖主要城市群的边缘接入网络，实现京津冀、长三角、粤港澳大湾区等重点区域毫秒级超低时延传输，并支持“中训边推”的训推协同架构解决东西部资源错配。这要求当用户请求到达时，调度器首先判断用户地理位置与候选算力节点的网络距离，仅在最近节点负载超过阈值时才触发跨区域调度，且跨区域调度的决策须综合权衡时延增量与目标节点的资源成本节约。**三是大带宽无损传输能力。**跨域协同训练与推理分发要求网络具备 TB 级数据分钟级无损传输能力，为“东数西算”战略下的数据应用提供高速通道。

3.1.3.安全基础设施要求：从访问控制到全链路可证明

AI 环境下词元可信流转的安全基础设施建设，已从单一安全工具堆砌演进为涵盖身份可信、内容可信、路径可信与状态可信四个维度的系统性工程，其成熟度直接决定 AI 系统在规模化应用中的安全边界与信任基线。

一是构建以身份可信为基石的全栈信任根与访问控制体系。基础设施需基于 HSM 或 TPM 建立硬件级身份凭证，为 AI 训练集群、推理节点、模型网关及调用方建立不可伪造的身份标识，确保流转链条每一跳均可验证上一跳身份。部署支持动态策略编排的细粒度访问控制引擎，实现基于属性访问控制与基于风险的自适应授权，确保词元跨租户、跨模型、跨服务流转时每一次访问均经最小权限原则校验，杜绝横向越权与身份冒用。网络层采用 mTLS 与服务网格建立加密通

道与双向身份验证，从源头保障交互主体可信性。

二是部署以内容可信为目标的全生命周期密码学与供应链安全基础设施。该设施需支持国密与国际标准密码算法双轨运行，提供词元生成、传输、存储到销毁的端到端加密，保障公网及跨云流转时的机密性与不可抵赖性。采用密钥封装机制（KEM）与数字签名算法，构建集中化密钥管理平台实现自动轮换与灾难恢复，引入同态加密确保推理过程中词元对非授权实体不可见，同时支持密态合规审计。完善 AI 供应链安全与模型可信验证机制，集成模型签名验证与软件物料清单（SBOM）管理，确保模型文件、推理框架及依赖库均来自可信源，防范词元污染、后门注入及恶意重放攻击，保障输入输出一致性。

三是建立以路径可信为核心的零信任网络与隔离计算环境。通过网络微隔离与东西向流量深度检测阻断异常外泄路径，确保词元只流经预期且经安全审计的通道和节点。结合 API 安全网关实现速率限制、异常模式识别与实时熔断，防止 API 滥用导致词元大规模泄漏。针对模型推理节点与词元处理中间件，基础设施需支持 TEE 硬件隔离实现内存加密与运行态隔离，配套远程证明服务网络，对推理环境启动度量、运行时完整性及软件供应链进行持续验证，防范侧信道攻击与内存转储风险。通过策略即代码引擎将路径策略与业务逻辑解耦，实现安全规则自动化编排，禁止词元流入未授权环境。

四是打造以状态可信为保障的全链路可观测、审计与治理体系。

基础设施须具备端到端追踪能力，通过不可篡改日志与区块链存证记录词元生成来源、流转路径、访问主体及时间戳，形成完整证据链，确保每一次流转均有据可查、不可抵赖。配套部署用户与实体行为分析系统，对访问频率、流向异常等风险行为实时建模告警，支撑事后溯源与合规举证。建立覆盖数据安全与行业监管要求的治理体系，通过 SOC 与红蓝演练持续验证保障效能，形成闭环改进机制，确保流转行为符合策略与合规要求。

3.2. 词元流转核心技术

随着大模型应用全面走向规模化落地，词元服务已成为支撑人工智能产业发展的关键底座，整个行业不再局限于单一算力比拼，逐步形成分工明确、环环相扣的三层架构体系。

一是生产层，作为 AI 产能的核心环节，负责实现词元从无到有，将物理算力资源转化为词元产出能力。竞争焦点集中在集群规模、服务稳定性、能耗控制以及模型适配效率，是整个词元服务体系的基础底座。

二是中转层，定位为资源枢纽与连接器，核心作用是整合市面上分散的算力与模型资源，对外提供统一的词元接入通道。该环节目前以资源整合、流量分发为主，同质化竞争现象较为突出。

三是调度层，是 AI 基础设施的智能核心，依托模型专业评测能力与智能路由算法，实现词元流通全链路异构算力资源的效益最优，竞争核心集中在模型评测精度、路由调度执行效率、跨基础设施

适配能力三个维度，是驱动产业升级的关键支撑。

3.2.1.模型无关性智能调度技术

统一接入与模型无关调度是词元流转的核心技术能力，以模型聚合、智能路由、多级策略自动匹配为核心，旨在为不同场景的词元服务体验提供全方位保障。主要包括三个核心能力，一是**协议兼容与入口收敛能力**。推理网关须兼容 OpenAI API、Anthropic Messages API 等主流协议格式，实现各类大模型“一次对接、全域接入”；以 TPM（每分钟 Token 数）、RPM（每分钟请求数）双维度配额实施入口级流量管控，基于分布式网关集群横向扩展，行业先进平台峰值推理请求吞吐可超过 50 万 QPS、服务可用性达 99.99%。二是**语义画像与智能路由能力**。依托请求语义画像完成意图识别与任务复杂度分级，支持多级路由智能编排，实现智能任务分发与最优专家模型精准匹配；具备实时负载感知能力，对长尾请求智能导流、动态削峰填谷。行业公开基准显示，70B 级模型首 Token 时延 P50 约 180 毫秒、P99 约 450 毫秒，7B 级模型首 Token 时延 P50 约 35 毫秒、P99 约 80 毫秒。三是**缓存亲和与异构感知调度能力**。调度网关与推理后端深度协同，支持 KV Cache 亲和调度实现会话粘性、Prefix Cache 感知实现前缀复用，并根据芯片型号动态加权实现异构算力感知调度；具备秒级容灾切换能力，故障时自动旁路分流。四是**模型聚合与自适应优化能力**。以模型聚合、多级策略自动匹配为核心，提供成本优先、性能优先、均衡优先等调度策略。国内运营商实践中，中国联通构建“多模共生”元景模型家族，自研四大类 20 余项大模型并完成自适应慢思考改造，

打造“联通星罗”先进算力调度平台，与华为联合实现运营商首个商用级大规模专家并行和 PD 分离推理架构，语言模型推理单卡吞吐量提升三倍。中国移动构建“九天”通专大模型矩阵，自研多个行业专用模型，创新 MoMA 聚合服务引擎实现动态路由调度，推理速度提升 42%；中国电信打造“三全”星辰大模型体系，自研语义、推理等系列模型，经多维度优化后单卡吞吐提升 2.5 至 4.3 倍。

3.2.2.全链路可观测与审计技术

全链路可观测与可信审计是词元流转透明化与可归责的核心支撑，以链路追踪、双端计量、行为感知、合规举证为核心，旨在实现词元全生命周期的可视、可管、可溯。主要包括四个核心能力。一是**全链路追踪与归因能力**。基于 OpenTelemetry 等标准化采集框架，实现从芯片到应用的全链路数据采集，每笔词元调用可追溯至具体算力资源、模型实例、调用主体和消费场景，形成完整的调用证据链。二是**双端计量与防篡改存证能力**。网关侧依据协议 Tokenizer 计数，推理引擎侧按模型实际生成量计数，两端定时对账；输入、输出、缓存命中词元分类计价，以哈希链或可信时间戳方式存证，确保不可抵赖、不可修改。三是**异常行为感知与告警能力**。部署用户与实体行为分析系统，对词元访问频率、流向异常、权限提升等风险行为进行实时建模与告警，支撑事后溯源与合规举证。四是**合规审计与闭环改进能力**。建立覆盖数据安全与行业监管要求的审计治理体系，通过安全运营中心与红蓝对抗演练持续验证保障效能，形成策略优化与合规改进的闭环机制。

3.2.3.密码与机密计算保障技术

密码与机密计算是词元作为价值载体的安全根基，以密码防护、机密计算、安全围栏、零信任为核心，旨在构建词元全生命周期的安全闭环，构建“算网安”一体化能力体系。主要包括四个核心能力。

一是商用密码全链路加密能力。按数据全生命周期构建三道防线，传输环节采用 SM2/SM4 或 AES 等国际商用密码算法全链路加密，存储环节以商用密码算法加密并分级管理密钥，使用环节依托密码服务保障数据机密性。

二是 TEE 机密计算隔离能力。通过可信执行环境实现硬件隔离的可信环境推理，使“数据可用不可见”成为可验证的技术事实，确保敏感词元在内存中的加密处理与运行态隔离。

三是模型安全围栏能力。模型侧构建输入过滤与输出拦截机制，对输入词元进行安全检测与风险过滤，对输出词元进行合规审查与内容拦截，防范恶意注入与有害生成。

四是智能体零信任能力。智能体侧实施零信任架构，每个智能体持有可验证身份、按最小权限授权，高风险操作需二次确认，实现代理行为可归责。

3.2.4.跨境合规流通技术

跨境通道是词元流通的跨境延伸，以合规架构、加密传输、网络协同为核心，旨在保障词元服务跨境调用的合法性与安全性。主要包括三个核心能力。

一是“数据不出境”合规架构能力。海外用户算力需求通过合规跨境专线接入境内专区，全部推理运算在境内闭环完成，计算结果经国际海缆回传，原始数据全程不出境，与《促进和规范数据跨境流动规定》及自贸区数据出境负面清单制度衔接。

二是跨境加

密通道能力。跨境专线可采用 SM4/IKEv2 商用密码 IPsec VPN，亦可采用 AES-256/IKEv2 等国际通用商用密码套件加密，确保数据传输安全。**三是国际网络资源协同能力。**依托通达全球的国际海缆和陆缆资源、覆盖全球的境外业务接入点及全球分布式数据中心，为词元跨境流通提供低时延、高可靠的网络通道。跨境调度的核心挑战在于时延。国际链路往返时延通常在 150 至 300 毫秒量级，对交互式词元服务用户体验影响显著。缓解策略包括边缘缓存与预加载，将高频模型参数与提示模板预部署至境外边缘节点，减少跨境传输数据量。合规架构设计须回应国内法与国际法的双重约束，国内数据出境三大合规路径与欧盟 GDPR、美国出口管制等规则之间存在冲突，须预留多法域适配空间。

4. 词元可信流通重点应用场景

本章选取政务、医疗、教育、央国企四大重点场景，按照“应用场景 + 可信流通需求 + 标杆案例 + 行业启示”四层结构展开分析。这四大场景的共同特征是：数据敏感度高、合规要求严、服务连续性要求强，是词元可信流通基础设施价值的集中验证场。

4.1. 政务场景词元可信流通实践

应用场景：数字政府建设正从业务线上化迈向智能化，智能政务热线、政策问答助手、公文辅助写作、城市治理一网统管等应用大规模消耗词元。政务云已形成多云服务商、多云平台厂商并存格局，云资源分布广、技术路线多、业务场景复杂。随着政务热线、城市治理、

应急处突、公安视频分析、科研训练等 AI 应用快速发展，传统政务云仅依靠通用计算资源已难以满足大模型训练、推理、调度和安全审计需求。

可信流通需求：一是安全准入刚性，承载平台须通过中央网信办云计算服务安全评估，模型须完成生成式人工智能备案。二是数据不出域，公民个人信息与政务数据须在政务云域内闭环处理。三是全程可审计，每一次模型调用须可追溯至具体调用主体、业务场景与时间戳。四是高并发弹性保障，政务服务具有显著的潮汐特征，要求词元供给具备秒级弹性调度能力。

解决方案案例：政务云通算扩容智算平台的项目是政务领域推进通算加智算一体化建设的典型案例。平台采用通智一体架构，以 AIDC 五星机房为底座，依托联通云星罗算力平台及私有云平台双引擎构建统一云平台，融合无损网络及高性能 AI 原生存储，实现多元混合异构算力适配和服务编排。项目建成后，省级信创智慧政务云平台已稳定运行多年，承载全省应用系统 200+，提供千余 TB 的存储能力以及万余 vCPU 的计算能力。该案例在词元化设计上具有前瞻性实践，将政务数据目录、模型服务额度、算力配额、调度任务、智能体授权、工单状态和审计日志转化为可验证凭证，通过身份凭证明确谁在调用，通过授权凭证明确为什么调用，通过算力配额词元明确能用多少资源，通过审计凭证明确产生了什么结果并由谁负责。该案例表明，词元可信流通基础设施首先应服务于公共资源的可控调用，形成资源统一供给、应用按需调用、过程全程留痕、风险实时可控的建设范式。

行业启示：词元可信流通基础设施首先应服务于公共资源的可控调用，而非直接交易化。通过身份凭证明确“谁在调用”，通过授权凭证明确“为什么调用”，通过算力配额词元明确“能用多少资源”，通过模型调用词元明确“调用了哪个模型和版本”，通过审计凭证明确“产生了什么结果并由谁负责”。在此基础上，省级政务智算平台可以形成“资源统一供给、应用按需调用、过程全程留痕、风险实时可控、地市可复制推广”的建设范式。

4.2. 医疗场景词元可信流通实践

应用场景：医疗场景下的词元可信流通，本质上是为处方流转、影像互认、急救协同、保险理赔、药械溯源、科研协作、公卫预警、智慧支付等具体业务场景，提供“身份可信、授权可控、数据可溯、传输安全、跨域互认”的基础设施。不同场景对词元的时效性、粒度、合规要求存在显著差异，需要分层分类设计。

可信流通需求：医疗健康数据属于敏感个人信息，受《中华人民共和国个人信息保护法》及健康医疗数据安全专门规定约束。一是数据不出域，患者诊疗记录须采用国密算法加密存储，不同级别医务人员按最小够用原则访问。二是全程留痕与可追溯，辅助诊断类应用的模型输出须支持医疗责任界定所需的追溯与复核。三是公共入口毫秒级响应，挂号、导诊等民生服务入口须在高峰期保持低时延与高可用。四是模型评测准入，模型应用进入医院场景前须通过安全性与准确性评测。

解决方案案例：平台采用国产化算力底座加医疗数据流通加专病模型应用加生态运营四层架构。以 AIDC 五星机房为底座，依托统一云平台，兼容并托管 DeepSeek、Qwen 及专病行业模型，安全层由 AI-Sec 与 AIOP 智算可观测平台纵向贯穿。基地获批医疗卫生行业高质量数据集建设链主单位，围绕先天性心脏病出生缺陷、肾脏病、肺部解剖变异、急性白血病、神经系统重大疾病、儿科疑难重症、致盲性眼病等方向共建 AI-Ready 行业高质量数据集。在词元化设计上，医疗机构通过数据目录词元声明数据资源、敏感级别和授权边界，药企与科研机构通过授权许可词元获得特定场景下的数据调用或计算结果，数据标注过程生成标注任务词元和专家审核凭证，模型训练与推理过程生成算力配额词元和模型调用词元。平台形成高可用、智能化的算力服务闭环，配置 7 人专职团队提供 7*24 小时驻场服务，满足等保 2.0 和医疗数据安全规范要求。项目做到数据不出域、可用不可见、最小必要授权、患者隐私保护、模型评测准入和全链路审计，避免敏感数据滥用、模型误诊风险和算法责任不清。

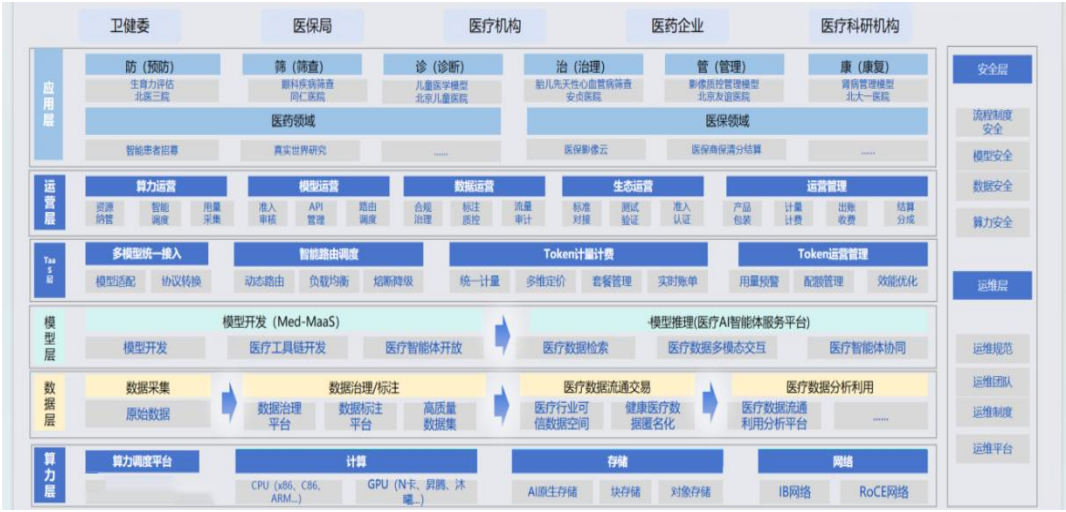


图 2 医疗案例架构图

行业启示：项目把医疗 AI 所需的算力、数据、模型、应用和运营机制放在同一个可信底座上统一建设，既能支撑三甲医院真实项目部署，也能面向药物研发、智能患者招募、医疗 AI 评测和基层诊疗能力提升形成商业闭环。风险控制上，项目做到数据不出域、可用不可见、最小必要授权、患者隐私保护、模型评测准入和全链路审计，避免敏感数据滥用、模型误诊风险和算法责任不清。

4.3. 教育场景词元可信流通实践

应用场景：高校与科研机构作为大模型应用的高活跃群体，其 AI for Science 科研计算、智能教学助手、个性化学习、校园知识库等场景对 Token 可信流通提出了规模化与普惠化的双重需求。教育行业的词元可信流转需解决算力资源分散、使用效率不均衡、资源共享机制不完善等共性痛点，通过统一身份认证体系实现师生、科研团队、校外合作方的身份可信与跨域互认；通过细粒度权限控制，依据科研任务、教学场景、人才培养等不同需求动态分配算力与模型服务访问权限，确保授权可控；通过全链路审计追溯，对科研数据、教学资源及模型训练过程进行留痕，满足学术诚信与数据合规要求；通过加密传输与隐私计算，保障跨学院、跨机构协作中的数据安全，实现原始数据不出域前提下的联合科研与人才培养，最终构建支撑教育数字化转型的词元可信流通基础设施。

可信流通需求：一是成本可控，要求按词元精细计量、按量付费，避免算力闲置浪费。二是数据合规，涉及未成年人个人信息与科研数据的双重保护。三是内容安全，面向师生的模型服务须构建内容安全

围栏。四是资源共享，高校普遍希望以云平台方式共享智算资源，替代分散自建。

解决方案案例：郑州大学 AI 智慧中台项目依托超融合架构和 GPU/4090 等算力资源，联通基于联通云搭建统一 AI 智慧中台。平台以云平台为底座，通过通智一体方式进行整体管理，满足异构算力纳管、模型开发训练服务、AI 资产管理服务、统一运营运维平台的整体要求。在词元化设计上，将校内算力、数据集、模型、标注任务、课程语料和科研服务转化为可管理、可计量、可审计的数字凭证。各学院提交科研任务时，通过算力资源词元明确所需 GPU、CPU、存储和网络资源。科研团队进行模型训练时，通过训练任务词元记录框架、镜像、数据集、参数、日志和结果。教学材料进行 OCR 和结构化转换时，通过语料加工词元记录来源、处理步骤、标注人员、质检结果和使用范围。平台将分散设备和云资源纳入统一入口，形成统一账号、统一权限、统一监控、统一账单和统一运维。通过词元化凭证，学校可掌握每个学院、课题、教师和学生团队的资源申请、训练任务、数据使用、模型产出和教学应用情况，为科研绩效评估、“双一流”学科建设、AI 人才培养和资源预算分配提供依据。科研侧可为多学科团队提供模型开发训练、数据标注、AI 资产管理和统一运营运维能力，降低非计算机学科使用 AI 平台的门槛。

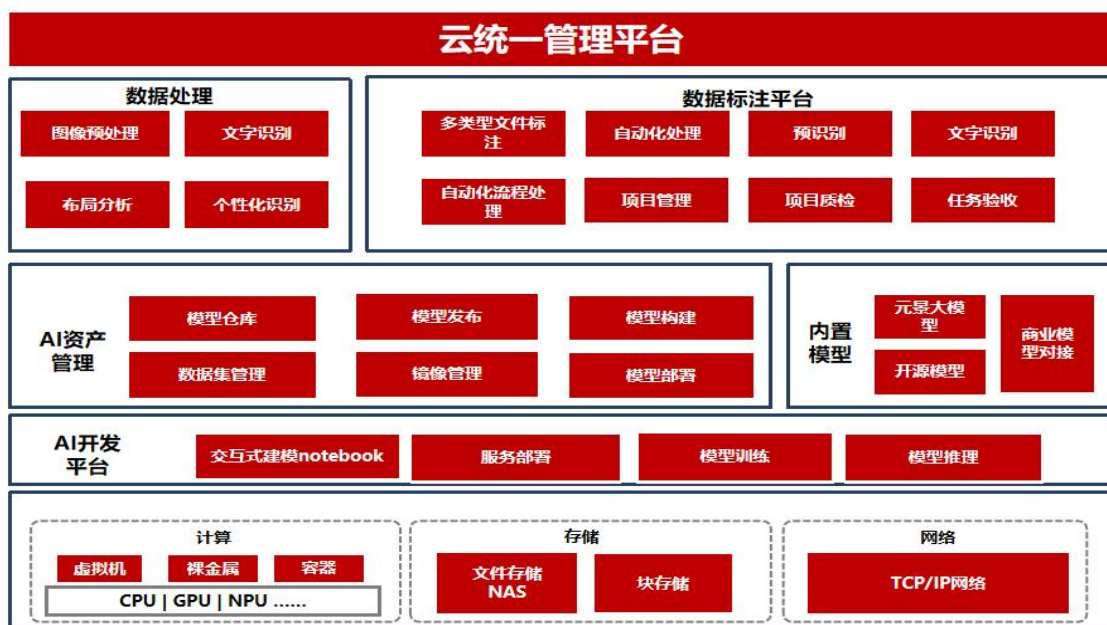


图 3 教育案例架构图

行业启示：该项目的落地价值体现在教学和科研两条线上。**教学侧**，可将课程材料、参考资料和纸质馆藏转化为电子语料，结合知识库和 RAG 技术构建智慧问答模型，服务学生自主学习和教师教学辅助；**科研侧**，为多学科团队提供模型开发训练、数据标注、AI 资产管理和统一运营运维能力，降低非计算机学科使用 AI 平台的门槛。

风险控制上，应重点关注校内数据安全、科研数据授权边界、学生参与标注的质量与责任、模型成果知识产权、跨学院资源公平分配和平台账号权限管理，避免校内数据外泄、资源被少数团队长期占用或模型资产不可追踪。

4.4. 央国场景词元可信流通实践

应用场景：国企作为“人工智能+”行动的主力军，在制造、能源、交通、建筑等领域推动大模型与核心业务融合，其生产调度、设

备运维、供应链管理等场景对词元可信流通提出了高并发、高安全、跨层级协同的严苛要求。央国企词元可信流转需破解集团与子企业间算力分散、效率不均、共享机制不完善等难题，通过统一身份认证实现总部、子企业、生态伙伴的跨域互认。通过细粒度权限管控，依据生产密级与业务条线差异化分配算力与数据访问权限，确保核心数据授权可控。通过全链路审计追溯对模型训练与业务决策留痕，满足国资监管要求。通过隐私计算保障跨地域、跨法人协作的数据安全，实现原始数据不出域前提下的集团级 AI 中台共建，构建支撑央国企数字化转型的词元可信流通基础设施。

可信流通需求：一是底座可控，芯片、模型、密码体系须满足底座信创要求。二是集约管控，集团层面需要统一的词元供给入口、配额管理与成本核算，实现统一接入、分级授权、按量结算。三是行业适配，通用模型须与行业知识、工业数据深度结合。四是长期运维，要求平台方具备央企级的服务保障能力。

解决方案案例：中国电力建设人工智能算力基础设施项目依托云底座、安全网络、运营服务核心能力搭建算力服务体系，围绕集团算力池建设、AI 中台落地、网络安全管控、长效框架运营四大维度开展建设，引入词元可信流通机制完成资源全生命周期溯源管理。在词元化设计上，总部通过集团资源池词元掌握总体算力规模和使用效率。子企业通过项目词元申请 A100、4090、910B 等不同规格算力。AI 中台通过智能体服务词元管理水电工程知识服务、安全培训、智慧投标、科研助手等多元应用。网络和安全侧通过专线接入词元等保

合规凭证和安全审计凭证实现访问范围、时延指标、风险事件和整改闭环可追踪。整体解决方案采用按年预付费模式，覆盖二、三级子企业的算力采购需求。依托订单、账单、服务 SLA 三类词元，结合计费结算、服务质量监控平台，实现多层级采购结算、计费核算、服务质量追踪的全流程管控，打通算力资源采购、交付、运维、结算的闭环管理。该项目提供了通智算力供给加调度平台加安全合规加专线接入的一体化架构，可直接复用于能源、电力、交通、建筑、制造等大型央国企 AI 算力采购项目。

行业启示：该项目提供了“通、智算力供给+调度平台+安全合规+专线接入”的一体化架构，可直接复用于能源、电力、交通、建筑、制造等大型央国企 AI 算力采购项目。对于类似集团客户，词元可信流通基础设施应服务于集团管控和产业协同：一方面打破总部和子企业之间的算力壁垒，避免重复采购和资源闲置；另一方面通过统一凭证体系沉淀智能体、知识库、模型、数据、网络和安全能力，使 AI 中台从单点系统升级为可运营、可计量、可扩容、可审计的集团级数字基础设施。

5. 发展趋势

5.1. 技术融合趋势

一是词元计量与可信结算技术加快融入现有基础设施。针对当前各平台分词口径不一、计费由平台单方统计的信任困境，标准化计量、跨平台词元换算、第三方独立审计与可验证计费等技术将加速融合成

熟，独立于模型提供方的计量结算体系将从“可选能力”升级为词元流通基础设施的“基础配置”，为词元跨平台、规模化流通奠定信任基座。

二是智能体互联网推动词元流通协议标准化。随着 A2A、AP2、x402 等协议成熟，智能体间的服务发现、能力调用与支付结算将走向标准化，词元流通将从平台内循环演进为跨平台开放网络，请求级实时计量与微支付能力将成为词元平台的基础配置。微软预测 2028 年全球 AI 智能体数量达 13 亿个，智能体间流通的词元规模有望超越人机交互。

三是机密计算与全链路审计走向深度融合。TEE 可信执行环境、商用密码算法与 Open Telemetry 标准链路追踪的结合，将使“可用不可见、流通可审计”成为词元流通的默认状态，为高安全行业扫清合规障碍。

四是国产算力适配从市场选择升级为战略导向。国产大模型适配国产芯片已成为国家战略层面的重点要求，多芯片异构调度能力将从“差异化优势”变为“基础准入能力”，词元生产层的国产芯片占比将持续提升。

五是算电协同从工程实践上升为国家新基建。2026 年《政府工作报告》首次将“算电协同”写入国家级新基建工程规划，国家发展改革委、国家能源局、工业和信息化部、国家数据局四部门联合印发《关于促进人工智能与能源双向赋能的行动方案》，统筹优化能源资源与算力布局。词元的生产成本将越来越取决于“电”的成本与“绿电”的比

例，绿电直连、源网荷储、算电协同微电网等模式将成为词元供给成本竞争的关键变量，青海三江源零碳智算园区等实践已先行验证。

5.2. 产业生态趋势

一是词元供给正走向工厂化生产与集约化运营。参照电力工业生产、输配、销售的专业化分工结构，词元产业将形成由词元工厂、算力调度网、智能路由平台构成的专业化分工体系。三大运营商“词元产品”登陆中国算力平台、中国电信星辰 Token Hub、中国联通“三个一”集约化词元运营体系等实践，标志着集约化词元运营已从理念走向落地。

二是行业准入质量评价成为行业生态健康发展的基础保障。随着词元服务市场快速扩张，服务质量参差、计量混乱等问题凸显，产业侧正加速构建全链条质量基准与评估框架。统一质量基线、透明计量规则、可量化服务等级及安全合规要求逐步成为共识，配套评估机制日趋完善。这将抬高市场准入门槛，推动服务商转向质量与可信能力比拼，缺乏评估背书和合规保障的平台将难以立足政企市场并被逐步淘汰，促进行业生态健康发展。

三是跨境规则互认与合规通道建设提速。依托自贸区数据出境负面清单试点和多双边机制，词元跨境流动将形成“负面清单+合规专线+境内闭环推理”的中国方案；信通院牵头的词元跨境流动机制研究与标准编制，将推动跨境词元流通从个案审批走向制度化通道。

四是数据要素市场与词元市场协同促进。《可信数据空间发展行

动计划（2024—2028 年）》提出到 2028 年建成 100 个以上可信数据空间，高质量数据经可信数据空间供给模型训练与推理，词元服务则为数据要素的价值变现提供计量与结算通道，“数据—模型—词元”的价值闭环将逐步形成。

5.3. 未来发展建议

对基础设施建设与运营方，一是坚持“网算安”一体规划，将安全可信能力内生于算力与网络建设，而非事后叠加；二是加快跨云异构调度与模型无关性调度能力建设，在三层架构中技术壁垒最高的调度层确立竞争优势；三是主动参与词元服务标准与评估，通过权威评估认证建立政企市场信任；四是面向出海需求提前布局合规跨境通道与国际版词元产品。

对行业应用方，一是将“可计量、可审计、可追溯”纳入词元服务重点考量标准，要求平台提供链路级审计能力；二是建立集团级词元统一管控与成本核算机制，避免多部门分散采购导致的计量断链与成本失控；三是高安全行业优先选择通过权威评估认证、具备全链路防护能力的服务平台。

对第三方组织，一是加快统一词元计量口径、SLA 规范与审计标准的制定，降低市场摩擦；二是推动词元服务监测平台常态化运行，以公开透明的性能监测引导良性竞争；三是在自贸区等开放载体中先行试点词元跨境流动规则，形成可复制的制度经验；四是研究智能体自主流通场景下的责任界定与治理框架，为智能体经济预留制度空

间。