



Center for Education,
Innovation and
Sustainable Development

上海科技大学教育、创新和可持续发展研究中心

Strategy Note | 战略研判系列

总 No. 15 期/2026 年 9 月

词元经济：AI 计算基元与资源分配的演进及重构

执行摘要

随着大语言模型（LLM）及其衍生的多智能体（Multi-Agent）系统向高度复杂化与规模化演进，人工智能基础设施的核心运转逻辑正发生基础性变迁。在这一结构性范式转换中，“词元”（Token）已跨越自然语言处理（NLP）领域的数据预处理概念边界，下沉为驱动生成式 AI 运行与底层算力商业化结算的基础经济基元。词元的指数级消耗规模，不仅引发了庞大的物理算力开支，更在系统架构设计、智能体通信协议、云端基础设施定价等各个维度引发了错综复杂的供需博弈。

本报告对词元经济及其衍生讨论进行系统的理论溯源与实证解构。有别于算法工程优化视角，本报告引入“计算-经济双重视角”，将 Token 定义为人工智能时代的生产要素、交换媒介以及计价单位。研究从微观的分词算法（Tokenization）底层出发，剖析其对模型显存利用率及算力消耗的边际效应；继而将宏观维度的物理基建约束（如能源供给、电网排队、芯片制造周期）纳入统一的分析框架。物理世界固有的供给限度与资源刚性构成了 Token 经济的硬底座，决定了宏观算力成本并非随摩尔定律呈线性下降趋势。与此同时，随着国内算力金融化前沿实践

（以广州海珠区“Token 贷”为代表）的落地，Token 正在脱离纯粹的技术与计费场域，演变为信贷核定与企业核心资产定价的新型锚点。这一宏观衍生趋势要求产业界在构建多智能体生态与进行资源配置时，须同时建立具备严密审计标准与可追溯性的底层可信数据体系。

第一章 词元：智能时代的核心经济基元与认知具象化

1.1 词元的概念剥离与历史演进

在自然语言处理的早期阶段，词元被定义为将连续文本转换为离散符号的语言学产物。彼时，其主要功能是为浅层神经网络提供离散的输入向量。随着 Transformer 架构的确立及自回归生成模型（Autoregressive Generative Models）的爆发式增长，模型机制发生根本转变：系统不再直接处理自然语言，而是摄取并生成代表词元的固定长度整数标识符。

由于大型语言模型的上下文窗口容量与注意力机制（Self-Attention）的计算复杂度直接与输入输出的词元序列长度挂钩，词元在物理执行层面实质上控制了显存消耗量、计算延迟（Latency）、网络带宽占用以及加速器（GPU/TPU）算力资源的分配。进入 2026 年的技术周期，大模型推理服务已高度商业化。输入词元消耗着稀缺的上下文窗口并占据键值（KV）缓存空间，输出词元则引发高昂的顺序解码计算负荷。在此背景下，词元的定义跨越了计算科学的边界，演变为衡量 AI 系统生产力与资源消耗的基础度量衡，成为现代人工智能基础设施运行的底层硬约束。

1.2 认知的“可读性”与隐性劳动的词元化账单

探究 Token 的经济学本质，需将其置于更宏观的社会组织学与政治经济学框架内。政治学家 James C. Scott 在探讨国家视野与社会管理时，提出了极具解释力的“可读性”（Legibility）理论。该理论指出，国家机器为了实施有效的征税、征兵与社会控制，必须将混沌、复杂、非标准化的自然与社会生态，通过度量衡统一、土地确权与人口普查等手段，简化为可被系统记录、比较与控制的标准表格与图表。在这一过程中，根植于地方的隐性经验知识（Mētis）被转化为自上而下的标准化技术知识（Technē）。

在当前的生成式人工智能时代，**Token 化（Tokenization）**过程对人类的符号与认知活动执行了极其相似的社会学功能。语言、代码、图像、指令序列乃至企业内部高度复杂的隐性工作流，一旦被转化为**Token**，平台便能对其进行精准的计数、定价、路由、限额与合规审计。非标准化的社会活动与认知劳动通过分词器映射为离散的**Token**序列，最终形成可供云服务商进行路由与结算的机器账单。过去深藏于人脑与传统科层组织内部的模糊认知活动，首次被具象化为机器账单上高度可度量、可分割的字符单位。

这种标准化的转换绝非中立的数据预处理，而是对人类符号活动的重构。**Token**化的深层逻辑直接决定了智力资源的分配权力：谁掌握了分词器的训练语料、上下文窗口的设定标准、模型路由的权限，谁就提前规定了机器能够“看见”什么、忽略什么，以及哪些认知活动能够以最低的成本被处理。这一机制实现了对复杂物理世界与人类思维的降维，使得不可量化的认知劳动能够无缝接入现代计算计价体系，从而在经济学意义上成为可供大规模交易与分配的标准化商品。

1.3 智力租差：通用智能的泛滥与稀缺性转移

当隐性的认知活动被拆解为廉价且高度标准化的**Token**在云端流通时，微观个体与企业的产出逻辑随之发生剧变。算力基础设施的大规模建设与算法压缩技术的进步，共同驱动了单**Token**生成边际成本的快速下行。然而，其宏观结果并非简单的全社会效率无差异提升，而是引发了深刻的结构性“智力租差”（**Intellectual Rent Gap**）现象。

通用生成能力的边际成本趋近于电力、数据中心折旧与芯片损耗的物理底线，这意味着常规的文本撰写、代码生成、逻辑推演与初步数据分析不再具备传统意义上的稀缺性。普通的**API**调用者或轻量级应用开发者在购买低价**Token**后，往往只能获取短暂的效率平权。在充分竞争的市场中，他们被迫将这种效率提升以更低的服务报价让渡给下游消费者，同时将主要利润以“**Token**过路费”的形式持续向上游掌握核心算

力、闭源前沿模型与底层数据基础设施的寡头厂商交租。

在 AI 能力帕累托前沿不断外移的产业背景下，真正的稀缺价值发生了向两端的转移。**Token** 能够完美搬运语言的结构与逻辑形态，但其物理机制决定了它无法背负语言在现实社会中产生的后果。系统可以消耗数百万 **Token** 生成一份精准的医疗诊断方案或复杂的并购法律合同，但 **Token** 本身不承担误诊的医疗事故责任，也不具备契约背后的商业信任与社会信用属性。因此，在词元经济学框架中，**Token** 仅仅决定了机器重构世界的颗粒度与计算尺度。人类的剩余价值与经济系统中的超额利润，被极限挤压至决定“什么必须被 **Token** 化”的目标设定环节，以及“什么必须留在账单之外”的逻辑判断、结果验证与最终责任承担环节。

第二章 词元的微观生成机制与效率边界

在评估宏观算力调度与推理层面的经济模型之前，必须严格审视词元的微观生成机制。词元化机制将连续的自然语言与代码向离散的机器表征进行映射。分词算法的信息压缩效率直接决定了模型架构的静态显存利用率及全生命周期的算力开销总额。

2.1 分词算法的演进路径与压缩逻辑

大语言模型的底层张量运算只能处理固定词表范围内的整数代码。分词算法的核心目标是在极度有限的词表容量约束下，实现对近乎无限的自然语言及代码输入的最优无损或低损压缩。

当前主流分词范式在理论基础与工程实现上存在严格的技术分歧。字节对编码（BPE, Byte-Pair Encoding）作为生成式大模型领域的绝对主流（如 GPT-4、Llama 3 等均采用此架构），其本质是基于纯频率的贪心压缩算法。它初始化于基础字节级别，持续扫描预训练语料库并合并出现频率最高的相邻字符序列对，直至达到预设的词表规模上限。这一逻辑将高频信息编码为短代码（单一词元），而将低频信息强制拆解为长代码（碎片化子词）。尽管该机制在统计学上实现了特定语种的最优压缩，但也为后续跨语种处理的经济不平等埋下了底层算法逻辑的伏笔。

2.2 词表容量与内存带宽的零和博弈

词表大小（Vocabulary Size）是大模型架构设计与微调环节中最关键的超参数之一，它直接锚定并决定了模型的“词元效率”。从 GPT-2 时代受制于显存限制的 50,257 规模，跃升至 Llama 3 的 128,256 及以上，扩大词表能够显著提高文本的压缩率，即利用更少的词元标识符来表达同等信息熵密度的文本内容。

由于 Transformer 架构的核心机制——自注意力（Self-Attention）的

计算复杂度随序列长度呈平方级 ($O(N^2)$) 增长，缩短总体序列长度能够以指数级幅度削减计算周期 (FLOPs)，并大幅减少键值缓存 (KV Cache) 的动态内存开销。然而，这种词表扩张伴随着严格的边际成本递增效应。庞大的词表需要维持体积成正比扩张的嵌入矩阵 (Embedding Tables) 与输出层投影矩阵。这不仅大幅推高了模型的静态显存占用率，挤压了推理时极其宝贵的高带宽内存 (HBM) 空间，还会导致长尾低频词元在预训练期间无法获取充足的梯度更新，进而损害模型在稀有术语上的泛化表现。因此，词表的设计本质上是一场关于硬件资源的零和博弈，必须在减少动态计算开销与控制静态显存占用之间寻找精确的帕累托最优平衡点。

2.3 跨语言语境下的结构性“词元税”

词元经济学微观领域最为核心的不平等现象，是多语言处理过程中的结构性“词元税” (Token Tax)。分词器的信息压缩效率高度依赖于其初始训练语料的统计分布特征。由于绝大多数通用基础大模型的初始分词器是在以英文占绝对主导的语料库上训练的，英文单词通常能够被近乎完美地映射为单个或极少数词元，均值维持在 1.3 词元/单词的极高效率区间。

相比之下，对于形态复杂的屈折语，以及缺乏空格分界的亚洲语言 (中文、日文、韩文等)，同样语义厚度的文本在进入分词器时，会被拆解为大量无语义关联的碎化子词甚至基础字节。这种低效的映射逻辑在商业应用层引发了双重经济惩罚：一方面，在按调用量计费的商业 API 市场中，非英文业务开发者为获取或输入等量的信息，必须向云服务商支付数倍的商业费用；另一方面，碎片化且冗长的词元序列迅速挤占了昂贵的上下文窗口 (Context Window)，变相削弱了模型在处理跨语言长文本复杂推理任务时的有效认知容量。尽管企业可通过训练定制化分词器来提升垂直领域的压缩效率，但由此引发的重新对齐底层模型权重及巨大的重训练开销，使得企业往往陷入两难抉择。

第三章 词元的硬件级配置与推理经济学模型

当大语言模型从实验室训练阶段步入工业界规模化部署，其全生命周期的算力支出结构发生了逆转。全球范围内累计的推理成本（Inference Costs）正在迅猛攀升，其总绝对量级在诸多商业应用场景下已大幅超越了初始的预训练沉没成本。构建严密的推理经济学模型，寻找算力利用率与生成速度的物理理论极限，已成为 AI 工程基建的核心课题。

3.1 推理流程的底层解构与双重硬件瓶颈

计算阶段	核心机制与运算特征	复杂度	硬件受限类型与表现
预填充阶段 (Prefill Stage)	并行处理所有输入提示词元。调用矩阵乘法单元计算注意力层中的查询(Q)、键(K)和值(V)投影，存入显存构建 KV 缓存。	$O(N^2)$ (高度并行的 GEMM 运算)	算力受限 (Compute Bound): 系统运行高度依赖 GPU 算术逻辑单元的绝对浮点运算能力，算力利用率极高。
解码阶段 (Decode Stage)	自回归顺序生成。每生成一个新词元，需重新从 HBM 读取预填充及历史生成的所有 KV 缓存进行矩阵乘向量运算。	$O(N)$ (密集的 GEMV 内存数据搬运)	内存带宽受限 (Memory Bandwidth Bound): 算术逻辑单元大面积闲置，计算速度被 HBM 的物理读写带宽锁定。

大语言模型的推理流程在底层硬件执行机制上，被切分为两个具备异质化特征的计算阶段。这种计算范式的严重割裂，直接决定了算力优

化的天花板与成本下降的极限所在。

3.2 房顶线模型与帕累托前沿的数学推演

在底层硬件内核编程（CUDA）中，工程师依赖“房顶线模型”（Roofline Model）来衡量算法在算术时间与内存读写时间取极小值时的理论表现极限。映射至宏观的推理经济学中，云服务商必须在单次词元生成成本与串行生成速度之间寻找经济学意义上的最优配置边界，即帕累托前沿（Pareto Frontier）。

理论分析框架揭示了主导这一边界的核心物理标度律：密集型 Transformer 架构的词元串行生成速度，与模型总参数量的平方根成反比，同时与提供推理服务的 GPU 高带宽内存（HBM）带宽的立方根成正比。这意味着，在不改变底层硅基材料物理特性的前提下，单纯缩减模型体积对生成速度的提升呈现极其陡峭的边际效用递减。

在并行化调度策略上，增加批处理规模（Batch Size）可显著摊薄庞大模型权重在显存与计算单元之间的传输时间成本，拉升 GPU 资源利用率并压制单次生成的边际成本。但这一策略存在绝对的物理拐点。当并发请求的吞吐量达到显存带宽与网络通信的饱和临界点后，盲目扩大批处理规模将不再带来任何单位成本收益，反而会导致系统的首词元延迟（Time to First Token, TTFT）发生灾难性恶化，严重损害前端业务交互的实时性体验。此外，为追求极致单任务生成速度而采用张量并行（Tensor Parallelism）跨多 GPU 分布推理时，受制于物理光速与总线协议的限制，网络延迟（Network Latency）成为了不可逾越的硬约束。

3.3 投机解码与工程优化的成本折让效应

面对解码阶段难以通过堆叠算力跨越的内存带宽物理瓶颈，算法层面的投机解码（Speculative Decoding）被证明是当前最有效的经济学对抗手段。该技术通过部署一个参数量极小、计算成本低廉的“草稿模型”（Draft Model）快速推测出后续若干词元的序列，再交由庞大的主模型

进行一次性的并行评估与验证。

从经济机制底层审视，由于现代高算力 GPU 在解码单一词元时算术单元大量闲置，主模型对多个草稿词元的并行验证，实质上是在不增加任何额外内存读取次数与时间开销的前提下，充分压榨了冗余的算术计算能力。研究表明，在维持大模型原有生成质量和固定总成本的前提下，投机解码可实现内存读取时间开销的多重分摊。结合低比特量化（将内存占用降至基线的四分之一）以及持续批处理调度（Continuous Batching），现代数据中心在面对特定推理工作负载时，其有效推理成本最高可实现 16 倍的数量级缩减，彻底重构了 AI 产品商业落地的利润率模型边界。

第四章 物理基础设施资本支出与双时钟错配机制

算力租赁定价体系的底层，并非能够随着软件复制而趋近于零边际成本的虚拟资源，而是由沉淀着巨量资本、受制于地理与能源约束的物理基建构成。

4.1 黄仁勋的“五层蛋糕”与成本颗粒度拆解

业界长期存在将 Token 等同于纯粹软件代码复制或单一电力转售的认知谬误。基于英伟达 CEO 黄仁勋在世界经济论坛等场合提出的“五层蛋糕”理论，一枚 Token 的完全生产成本（Fully Burdened Cost）需要自下而上进行极细颗粒度的物理拆解：

1. **能源层（约占完全成本 10%-20%）**：GPU 矩阵运算、冷却系统与网络传输依赖庞大的电力支撑。按美国工业电价测算，一座 500 兆瓦的数据中心年电费高达数亿美元。但一旦计入重资产资本折旧，电力仅占单位 Token 完全成本的底层一环，并非决定性因素。

2. **芯片层（约占完全成本 40%-55%）**：构成供给的绝对咽喉。模型参数必须驻留于芯片封装内的高带宽内存（HBM）中，其容量直接决定推理并发上限。而全球 HBM 产能高度集中于 SK 海力士、三星与美光三家寡头手中，这种寡头垄断叠加 AI 需求爆发，使其成为推高总拥有成本（TCO）的最核心驱动力。

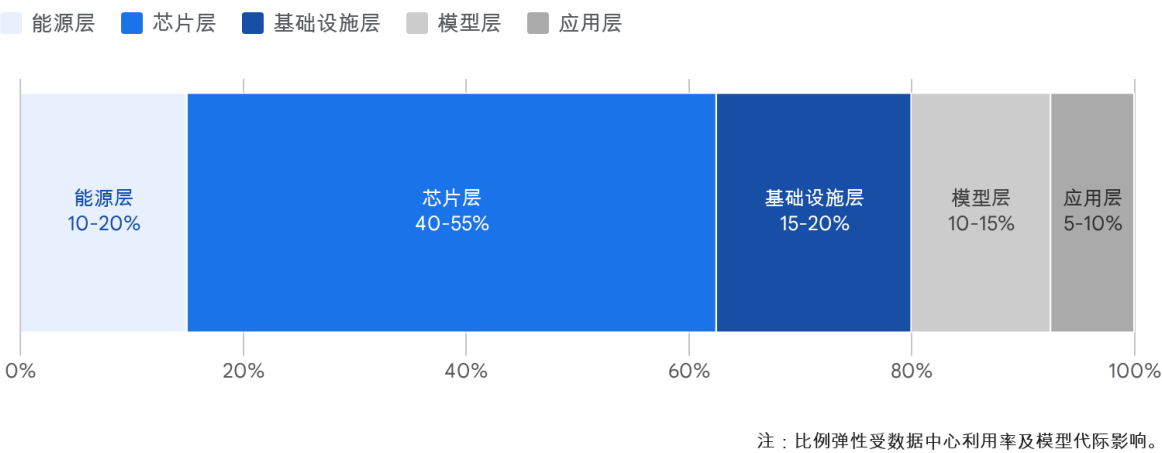
3. **基础设施层（约占完全成本 15%-20%）**：涵盖机房配电单元（PDU）、液冷系统、光模块与 InfiniBand 交换网络。其任务是将离散的计算芯片组织为协同运转的 AI 工厂。GPU 利用率的微小波动都会导致该层折旧分摊发生剧烈变化。

4. **模型层（约占完全成本 10%-15%）**：包含高昂的预训练沉没成本摊销与推理框架优化（如混合专家 MoE 架构、模型蒸馏）。这是工

程魔法最密集且边际成本下降最快的软性环节。

5. 应用层（约占完全成本 5%-10%）： 涉及接口路由、系统集成与内容合规安全审核，是将原始模型能力转化为可售卖服务不可或缺的封装成本。

Token生产总成本拆解 (基于大规模商用推理口径)



在大规模商用推理场景下，芯片硬件与基础设施的资本折旧占据了Token完全成本的主体地位，能源费用并非最核心约束。

4.2 物理与算法的结构性错配：“双时钟”张力

理解五层结构后，必须直面算力供给体系中最最为剧烈的结构性错配——即物理世界刚性供给的“慢时钟”与软件算法优化的“快时钟”之间的同场运转张力。

算法侧的快时钟： 技术迭代以日、周和月为单位高速推进。斯坦福 AI 指数报告等数据表明，通过注意力机制优化、推测性解码与蒸馏

技术，同等能力水平的推理成本在三年内可录得超 99% 的下降幅度。这种无须替换物理硬件即可生效的软件级优化，给予了市场单 Token 成本将无限趋近于零的线性下降错觉。

物理侧的慢时钟： Token 的生产受制于严格的物理约束与宏观工业建设周期。

1. 电网接入排队与输电网扩容： 国际能源署数据显示，全球主要数据中心枢纽（如阿姆斯特丹、都柏林）的电网新接入排队时间平均已超过 7 至 10 年。截至 2025 年底，美国电网互联排队规模高达 2060 吉瓦。输电网的系统级扩容受制于立法与环境审批，遵循以 10-15 年为单位的规定周期，纯靠资本突击无法压缩这一物理时间差。

2. 重型电气设备交付周期： 核心卡脖子组件如标准电力变压器，其交付周期已从 2023 年的 140 周延长至 2026 年的逾 160 周。一个投资 20 亿美元的算力园区，极易因区区数千万美元的电气设备缺货而陷入长达数年的停滞空置。

3. 能耗双控与行政审批： 在中国一线城市（如北京），能耗指标双控政策与极低的 PUE（电能利用效率）上限要求叠加，构成了复合型的行政与物理约束。这一约束直接催生了“东数西算”国家级算力网络工程，推动新建超大型数据中心向内蒙古、贵州等西部枢纽转移，以空间换取能源与审批时间的宽裕。

两套时钟的摩擦，彻底破除了“单 Token 成本线性下降”的宏观迷思。服务商的市场报价可以因某项算法突破在短时间内跳水；但物理容量见顶、电网系统重建以及重资产机房的阶梯式资本跃升，将立刻且周期性地拉高算力的全盘拥有成本。

4.3 云端 GPU 租赁的市场分层与隐性摩擦

硬件制造成本向词元计价传导的核心枢纽在于云端算力租赁市场。

云端费率不等同于裸机租赁价格，其背后隐藏着庞大的运营摩擦开销（Operational Overhead）。随着单张算力卡的热设计功耗（TDP）突破千瓦大关，昂贵的液冷系统折旧、数据中心配电单元的专项改造以及全天候网络运维团队的支出，使得物理 GPU 的真实运行成本需在裸卡基准上大幅上浮。这种庞大且不可避免的隐性摩擦，迫使绝大多数缺乏极大规模经济效应的科技企业，放弃自建算力池战略，转而全面拥抱按 API 调用次数和 Token 生成量结算的 Serverless（无服务器）按需付费模式。

第五章 词元计价体系与 API 经济学

在将底层计算严格的物理约束转化为上层灵活的软件路由逻辑后，词元正式成为生成式人工智能生态商业化兑现的标准计价通货。云服务提供商在平衡物理边际成本与抢占开发者心智的双重压力下，正推动一场深刻的定价体系重构。

5.1 推理价格的超通缩与市场梯队结构

回顾计算机工业发展史，大语言模型推理计算市场的降价轨迹打破了传统的摩尔定律迭代周期律。以处理每一百万个词元为基准，在商业化起步阶段调用成本高达数十美元；而历经工程加速与开源模型的冲击，获取同等计算能力的成本呈每年数量级下降的“超级通缩”现象。

截至最近，API 服务市场已形成极为清晰的三层定价梯队：

1. **极速入门与长尾场景模型：** 主打高吞吐量与极限低单价（如下探至\$0.06-\$0.075/百万词元），以极低的算力花销支撑起标准化的通用应用基础。

2. **生产主力平衡模型：** 承载企业级核心业务逻辑。值得注意的是，DeepSeek R1 等高推理智商、低架构成本的模型，以每百万输入 1 元人民币（缓存命中）、输出 16 元人民币的极低单价，对西方同类竞品构成了降维成本打击，颠覆了“越强越贵”的旧有定价逻辑，极大加剧了云端算力市场的竞价压力。

3. **前沿认知与长逻辑链模型：** 用于解决高度复杂的代理任务协调及多步逻辑推演。依靠对海量上下文的把控与极低幻觉率，维持着高昂的护城河溢价。

5.2 输入/输出定价机制的系统不对称性

贯穿所有梯队定价模型的一个核心财务特征，是极为显著的“输入/

输出（I/O）定价倒挂”现象。服务商普遍对输出词元施加了数倍于输入词元的计费单价。这一商业计费策略精确地映射了前文所述的推理双阶段物理机制：输入词元享有矩阵乘并行的算力红利，边际能耗极低；而输出词元深陷自回归串行解码的内存带宽泥沼。这种算力利用效率的断崖式跌落，必然且合理地反映在终端 API 的高昂定价上。

5.3 提示词缓存（Prompt Caching）的系统级重构

在工程架构层面优化词元消费账单，提示词缓存已成为效能杠杆最高的经济调节阀。它将首次计算生成的注意力层中间状态（即 KV 缓存）安全地持久化存储于推理服务器集群中。当携带相同长文本前缀的新请求抵达时，系统直接挂载复用缓存矩阵状态，绕开繁重的投影矩阵相乘，瞬间转入快速解码流。

这种机制对经济模型造成了三重深远影响：首先，针对大规模语料的交互场景，首词元延迟（TTFT）可缩减高达 80%以上；其次，云供应商针对“缓存读命中”提供高达 90%的基础输入价格抵扣（如 DeepSeek 的缓存命中单价降至 1 元/百万词元）；第三，在开发重构维度，它倒逼软件架构师重新编排系统指令、工具集及高频参考文档的输入语序，将其前置固化以最大化缓存命中收益。

第六章 复杂系统中的多智能体词元经济学

当单次交互（Prompt-Response）的简单工作流向具备自主规划、工具调用、长期记忆反思功能的多智能体系统（MAS）演进，系统运行的复杂性发生了质的跃迁。在多智能体生态中，词元的生成从单线程的线性累加蜕变为高频循环的内部交锋，直接映射至组织经济学的深层分析框架。

6.1 微观层组织：单体智能体的新古典要素替代

在独立的单体智能体内部结构中，任务执行过程本质上是微观经济学中面临严格预算约束的“新古典企业生产最优化问题”。每个智能体被视为一个理性的微观经济实体，其最终目标是在延迟容忍度与既定可用词元总预算的双重约束下，最大化产出质量。

系统设计演变为不同生产要素间边际替代率（MRTS）的权衡：面对复杂任务，是将有限的词元配额投入到提供海量上下文以拓宽输入带宽？还是将配额授权给隐性思维链模块（如 Reasoning Models 逐步列出思路和过程），容忍其产生大量暂存于黑盒中并不直接输出的过渡推理词元？这种架构决策反映了对直接数据喂养资本与深度自回归推理资本之间效用转换的判定。

6.2 中观层组织：多智能体协作网络中的交易成本

一旦系统扩展至涵盖审查者、编码者、规划者等多重专业角色的多智能体协作网络，科斯“交易成本理论”便构成了其底层运转机理。

在这个互联网络中，系统摒弃了单体模型的独断专行，代之以反复的辩论、共识对齐以及多轮结果交叉审查。然而，智能体间的所有交互沟通状态确认与记忆同步，本质上都是信息交互的阻力成本。在执行复杂长流程链条任务时，网络常常被海量不直接接触达最终结果的冗余交际词元淹没，恰如现代科层制企业中繁冗的跨部门沟通摩擦。架构师必须

研发动态缩放的协同协议，通过设计精简的共享内存架构与高效的任务授权握手信号，最大限度地压缩组织层面的内部交易费用。

6.3 宏观生态层面：拥堵外部性与防线建设

放眼云端宏观生态，海量自治智能体同时连入共享 API 算力集群，无休止地争夺物理 GPU 资源，引发了严重的网络“拥堵外部性”。缺乏调度的狂热探索性请求极易导致处理队列满载，使得高价值核心业务遭受延迟折磨。此时必须引入动态定价机制进行稀缺产能调度。

在系统安全维度，防御对抗性威胁及提示词劫持不再仅是外部合规指标，而是保卫财务底线的刚性要求。恶意攻击可操纵智能体触发失控的调用死循环，这种高频无效推理运算会在极短时间内耗尽企业设置的词元信用上限。构建内生的安全过滤节点与稳健的模型对齐护栏，实质上是一种确保存储与计算资源不被非法掏空的经济安全保险行为。

第七章 宏观生态、金融化衍生与产权边界

随着人工智能应用深度接管企业核心业务流，算力 Token 已突破了纯技术的工程物理与测度学边界，正式向资产定价、信贷风控与金融衍生领域延展。国内前沿的产业实践表明，词元正在被赋予类似常规核心资产的金融核定属性，并在产业链上下游催生出新型的融通机制。

7.1 算力金融化的前沿实践：广州海珠“Token 贷”运作机制

2026 年 8 月 14 日，广东省首个词元经济专项金融产品“中银·算力 Token 贷”在广州市海珠区正式发布。这一由地方政府、中国银行等金融机构及核心 AI 企业联合推动的产业动作，标志着算力金融化正式从底层的“硬件租赁”跨入前端的“应用流转”层。

该系列产品精准聚焦算力供给、应用与服务三大核心场景，推出了“算力 Token 供给贷”、“算力 Token 应用贷”及“算力 Token 服务贷”三个细分子产品，实现了对大中小微各类 AI 企业的全覆盖。其最核心的突破在于产品设计逻辑：打破了以往依赖固定资产评估的传统框架，转而将企业的“Token 产出与消耗量、算力服务合约价值、算力业务产生的应收账款及 Token 佣金结算量”作为核心授信依据，为企业提供最长三年、最高达 3000 万元人民币的授信总量。

信贷核定标准的底层转移：按算力合同和 Token 流转数据来核定额度，分别对应并升级了传统订单融资和经营贷的审核思路。对于大量轻资产、高研发投入的人工智能企业而言，这种核定标准从常规的厂房、服务器硬件向无形的“算力消耗与服务价值”转移，实质上承认了 Token 流转数据作为 AI 企业核心生产力及盈利能力指标类资产属性。中期来看，以 Token 流转为纽带，将催生出锚定人工智能核心大企业与上下游海量应用服务商企业的复杂链式金融产品。

此外，与金融创新并行的地方政策同样印证了 Token 的资产化地

位。海珠区同步落地的“词元八条”政策，针对词元生产端（按模型推理优化投入最高补贴 50%，单企每年最高 150 万）、流通平台端及应用端（最高 500 万扶持），提供了全链条的资金支持，从政策层面为 Token 经济的合法性与商业变现提供了兜底与催化。

7.2 融资瓶颈与底层基建：数据真实性与存证审计平台

尽管“Token 贷”为人工智能企业开辟了全新的信贷输血通道，但将高度虚拟化的算力消耗直接与银行信贷挂钩，面临着极为复杂的贷前审核与贷后风控考验。

业内专家强调，算力融资与算力信贷大规模铺开的关键瓶颈，在于底层数据的真实性与可获取性。Token 数据天生具有脆弱性：一方面，Token 消耗量仅代表平台前端的业务调用流量，它并不直接等同于后端持续的盈利与回款能力；另一方面，AI 业务的 Token 消耗常带有显著的脉冲式特征，项目周期结束即骤降，极易引发贷后误判。更为致命的是信贷操纵风险——如果银行仅将 Token 消耗绝对值作为“放贷 KPI”的目标函数，而不去穿透验证其真实的商业逻辑，企业极易通过短期刷单、编写循环调用脚本或进行无效的测试调用来虚增 Token 消耗数据，进而套取信贷资金。

因此，算力信贷走向成熟的刚性前提，是跳出单一金融机构的数据孤岛，在产业层面建立一套统一、透明、可审计的 Token 数据存证标准。国家级“可信数据空间”概念的引入成为破局关键。依托合规的密码学与可信计算技术，将多模态的复杂合同文本结构化，将 Token 的产出、调用明细、服务合同链条及回款流水进行全流程的上链固化与合规审计。只有当数据流的不可篡改性商业合同的真实性被硬性绑定，构筑起防范流量欺诈的风控中枢，算力金融化的闭环才算真正打通。

7.3 边界勘定：LLM 计算词元与 Crypto 投机属性的本质隔离

在确立词元作为 AI 产业计量基础与新兴信贷锚点的地位后，学术

界与政策制定者须对行业内因英文术语高度重叠（Token Economics / Tokenomics）而引发的概念混淆，进行边界勘定。尽管国内开启了合规的 Token 金融化信贷尝试，但生成式人工智能体系下的计算词元，与 Web3 及区块链分布式账本环境中的加密代币（Crypto Token），在底层经济学内核与法律属性上存在着不可逾越的本质对立。

其一，稀缺性来源的错位。 加密代币通过数学协议制造了人工层面的绝对供给短缺（如比特币硬编码的 2100 万枚发行上限）；而大模型计算词元的虚拟总量在理论上趋于无限，它唯一的供需天花板，受制于真实物理世界中台积电制程工艺的晶圆良率、全球高带宽内存（HBM）的硅片产出总和，以及区域数据中心互联电网的承载力极值。**其二，资产属性与价值锚点的断裂。** 密码学资产具有极强的二级市场自由流转交易属性，其市值完全依靠网络参与者的群体共识预期以及资产流动性溢价支撑，带有明确的金融投机特征。相反，AI 计算词元是纯粹且即时消亡的计算负担度量单位，它不可转移至第三方交易平台流通，也不构成对 AI 开发企业的任何股权或债权主张。其终端定价，精确地锚定于云服务器硬件折旧损耗、液冷设备运维支出、晶体管能耗转换率等真实可观测的工业制造成本。

唯有认清计算词元作为物理算力映射与智能测度衡的本质属性，剥离其与投机金融资产的关联，方能在当前这场席卷全球的 AI 生产力要素重构浪潮中，合法、合规且可持续地驾驭算力经济的演进。

联系我们: CEISD@Shanghaitech.edu.cn