



人工智能大模型如何实现算力低耗能

□ 宋 洁

筑牢算力底座是释放数据要素价值的关键。近期，随着人工智能大模型的爆发式涌现和快速迭代，算力需求已呈现指数级增长。国际数据公司（IDC）与浪潮信息联合发布的《2025年中国人工智能算力发展评估报告》预测，2023年至2028年，中国智能算力规模将以年复合增长率46.2%的速度扩张，为充分发挥人工智能大模型的生成能力提供了重要的基础能力支撑。

人工智能大模型面临算力高耗能挑战

在智能算力快速增长背后暗藏隐忧，算力基础设施的高能耗与低碳可持续发展目标间的矛盾日益尖锐。根据《算力基础设施高质量发展行动计划》，算力或计算力是数据中心服务器对数据处理并实现结果输出的一种能力，是衡量数据中心计算能力的一个综合指标，包含通用计算能力、超级计算能力和智能计算能力。在作为算力枢纽的数据中心，算力能耗包含直接电力消耗与冷却系统能耗两部分，本质上是“比特到瓦特”的能量转换。而计算模型能力的提升往往依赖于不断增长的算力投入，在以人工智能为代表的许多前沿科技领域，模型的能力获得了优先考虑而能耗则经常被忽视。以大模型训练为例，OpenAI公司训练GPT-4的能耗高达2.4亿度电，凸显了算力扩张与能源约束的尖锐矛盾。当前，人工智能大模型既是算力需求的主要推手，也成为破解能耗困局的关键。通过算法优化、基础设施升级与观念革新，实

现算力低能耗化，已成为推动数字经济可持续发展的必然选择。

从算法到硬件的系统性技术革新

降低模型计算量是实现算力低能耗的核心路径。通过算法创新，可在不牺牲性能的前提下减少算力消耗。例如，模型剪枝算法通过剔除冗余参数，将深层网络转化为精简的“小模型”；模型蒸馏技术则将大模型的知识迁移至轻量级模型，显著降低推理能耗。更具突破性的是动态参数激活机制的应用。以深度求索公司发布的6710亿参数大模型DeepSeek-R1为例，其创新性地采用“按需激活”策略，每次仅调用少量活跃参数，实现了性能与能耗的平衡。此外，低精度计算（如FP16或INT8）的普及，通过减少浮点运算位数，进一步降低了算力需求。这些技术的融合，标志着算法层面的节能革命已从理论走向实践。

算力基础设施的能效效率（即单位算力能耗）是另一关键维度。一方面，硬件设计需兼顾通用性与专用性。针对AI模型的并行计算特性，异构计算芯片（如GPU、NPU）的能效比显著优于传统CPU，而类脑芯片的探索则更具颠覆性。在2025全国类脑智能产业创新发展推进会上展示的脉冲神经网络芯片，通过模拟人脑稀疏激活机制，将能效提升1—2个数量级，为突破“算力—能耗”瓶颈提供了新方向。另一方面，算力集群的集约化布局亦是关键。“东数西算”工程规划布局了八大算力网络国家枢纽

节点，分为东部实时性算力集群与西部非实时性算力保障基地两大类，旨在优化全国算力资源布局，推动绿色集约化发展。例如，我国首个洞库式数据中心，结合山洞山体特性，采用冷通道分离设置等设计手法，能源效率指标达到国际先进水平。人工智能大模型本身可成为算力低能耗化的工具。通过实时分析任务负载，大模型能动态调度算力资源，在低负载时段关闭冗余服务器，在高峰前唤醒备用设备，避免“空转”浪费。例如，我国有的大型调度平台，通过智能算法将通用、智能、超级算力与量子算力协同分配，使算网一体化效率提升20%。在硬件层面，AI与冷却系统的联动也潜力巨大，例如在数据中心通过AI预测服务器温度变化，动态调节冷却强度。AI驱动的能效管理系统正推动算力基础设施向“智能低碳”转型。

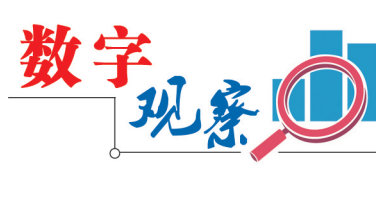
从技术到制度的系统性生态建构

强化标准化的约束与引导。《数据中心算力碳效可信评价技术规范》的发布，首次将“算力碳效”纳入评估体系，通过量化单位算力的碳排放量，为绿色算力提供可衡量的标尺。未来，需进一步完善碳排放双控政策，如明确要求智算中心绿电占比、通过“算力券”补贴激励企业使用低碳算力等。政策层面，需强化对高能耗算力设施的约束。各算力枢纽应该着力探索“绿电聚合供应”模式，强制要求新建数据中心与可再生能源协同布局，

倒逼企业优化用能结构。要让这套政策组合拳为算力低能耗化提供制度保障。

打破技术壁垒畅通产学研全链路。开源社区的崛起为算法优化提供了新路径，如DeepSeek开放其模型蒸馏框架，使中小开发者能低成本构建轻量化模型。产业链协同整合高校、企业的研发资源，在算力产品层面形成“设计—验证—量产”全链条生态，在算力服务层面从“硬件依赖”转向“云端协同”。这种“产学研用”深度融合的模式，正成为算力低能耗化的加速器。形成公众参与的良好氛围。要扭转“以能耗换性能”的发展趋势，必须更大力度地呼吁所有民众对算力能耗的关注，从而让节约用电一样深入人心。通过公开透明的能耗数据报告，企业可以更加直观地感知算力碳足迹，从而优化算力能耗成本，政策制定者也能在成本效益和环境影响之间权衡利弊作出正确决策。

算力低能耗化不仅是技术问题，更是关乎数字文明存续的生态命题。随着先进制造、绿电协同、智能调度等技术的发展成熟，加快实现算力与能源、环境的深度耦合，构建起“技术—经济—生态”共赢的新范式，从而实现人工智能大模型的强大生成能力和算力低能耗“双赢”局面。



作为《“十四五”数字经济发展规划》中的新型基础设施，云计算通过算力资源的高效调度，正加速重构政务治理、跨境金融、智慧医疗等关键领域的数字化转型范式。这种以弹性服务为特征的数字生产力引擎，在驱动千行百业智能化跃迁的同时，也因数据要素的跨境流动与融合共享，衍生出主权边界模糊、隐私泄露加剧、合规成本攀升等系统性风险。工信部等十六部门联合印发的《关于促进数据安全产业发展的指导意见》明确指出“推进新型计算模式和网络架构下数据安全基础理论和技术研究”的目标，强调要推动数据安全产业高质量发展。

云计算环境下的数据安全困局。得益于弹性算力调度机制和资源全局池化架构，云计算正重构数字世界运行规则。在《国务院关于促进云计算创新发展培育信息产业新业态的意见》政策驱动下，企业通过云原生架构实现基础设施成本指数级下降，同时获得分钟级全球业务部署能力。从医疗影像的实时AI辅助诊断，到跨境金融交易的毫秒级风控决策，云端分布式计算正将数据处理效率推向新维度。然而，云环境的开放共享特性在释放生产力的同时，也将安全威胁从单点漏洞演化为系统性风险。多租户资源池中，一个配置失误的存储桶可能引发百万级数据泄露。虚拟化层的攻击面突破会连锁反应至整个资源池。构建数据分类分级制度，以及规范数据本地化存储机制，推动了云数据安全保护体系的建设。然而，传统基于边界的防火墙策略，在动态伸缩的微服务架构前形同虚设。当数据在分布式节点间自由流动时，物理世界的安全边界已被彻底瓦解。

在此背景下，数据加密技术正经历从“配套工具”到“架构基石”的范式跃迁。数据加密的根本价值在于通过数学规则的不可逆性，在数据本体层面建立终极防线。即使攻击者突破网络边界、绕过访问控制，获取的仍是无法解读的密文。在国家网络安全战略层面，数据加密的战略价值集中体现于破解云安全三大核心困局：针对多租户架构下的数据主权侵蚀风险，采用静态加密技术实现存储介质级主权锁定；面对跨境流动中的长臂管辖威胁，依托传输加密技术构建可信通道，阻断境外司法越权访问；破除运行时的侧信道窃取隐患，采用动态加密技术构建可信执行环境，确保核心数据在内存计算全程密态化。

第一重防御：静态数据加密守护安全基线。在云计算重塑全球数字基础设施格局的背景下，静态数据加密已突破传统存储保护范畴，升维为国家数据主权在云端落地的核心抓手。面对云环境多租户资源共享、弹性扩展、跨境流动的架构特性，数据安全法与密码法形成双重制度驱动，将“存储即密文”的防护准则深度植入云原生架构。静态数据加密的本质是将原始明文转化为数学上不可逆的密文，使数据在硬盘、数据库或对象存储中始终保持不可解析状态，从根本上消解多租户环境的数据交叉污染风险。

从算法实现维度，AES-256加密算法与国密SM4算法构成双轨技术体系。AES-256加密算法远超经典计算机算力极限，因而成为全球云平台的默认存储加密标准。国密SM4算法则达到等效128位安全强度，满足我国《信息安全技术网络安全等级保护基本要求》对政务云、金融核心系统的合规强制部署要求。二者虽在全球化兼容与主权可控性上形成技术博弈，但均依托对称加密的单密钥高效特性，完美适配云场景的大数据处理需求。

加密模式的差异决定安全粒度与性能边界。比如，块加密模式将数据切分为固定长度区块独立处理，支持数据库字段级精准防护；透明数据加密深度集成于存储引擎层，在数据写入磁盘时自动完成整页加密，实现用户无感化保护。密钥管理体系需突破传统架构桎梏。根密钥存储于硬件安全模块或云原生密钥管理服务，通过密钥派生函数动态生成唯一数据加密密钥，形成“根密钥物理隔离、数据密钥按需销毁”的零信任链条。这种分层架构完美契合云存储的动态扩展特性。当数据分布式存储在跨地域、跨可用区的对象集群时，每个存储对象对应独立加密密钥，即使单点泄露也无法影响局部数据，系统性风险被数学规则严格约束。

第二重防御：传输数据加密打造可信通道。在云计算架构下，数据跨境流动与多态交互的特性，使得传输加密技术成为国家数字主权的核心防线。云原生环境中的动态网络拓扑与瞬时通信需求，驱动加密协议向高性能、自适应方向演进。传输层安全协议（TLS）1.3通过精简握手流程与强制前向保密机制，在保障政务云跨区数据调度、金融交易毫秒级响应的同时，构筑起抵御密网泄露的数学屏障；服务网格架构深度集成的双向加密认证通信协议（mTLS），通

云计算环境下的数据加密技术

□ 公 备

过身份绑定与动态证书颁发，在万亿级微服务交互中实现“通信即加密”的零信任范式。这种技术演进不仅是对传统网络边界的突破，更是国家网络安全战略在云时代的具象化落地。比如，当数据在“东数西算”工程的东西部枢纽间流动时，加密技术通过算法自主性与协议可控性，确保国家关键数据资源的传输主权始终握于掌心。

面对量子计算对传统密码体系的颠覆性威胁，传输加密技术的前瞻布局已上升至国家战略安全高度。基于格密码的NTRU算法与量子密钥分发技术形成双重防御纵深：前者通过多项式环结构优化，在云原生集群中实现抗量子密钥协商的工程化部署；后者依托量子不可克隆定理，在核心算力枢纽间构建物理定律担保的密钥分发网络。这种算法革新与物理免疫相融合的架构，既满足跨境金融、智慧城市等高敏感场景的实时通信需求，又为数字基础设施预埋了抵御未来十年量子攻击的防御基因。作为全球少数实现抗量子加密全栈技术落地的国家，中国正通过自主可控的密码技术标准，重塑云计算时代的国际安全规则。当加密协议从技术工具升维为数字疆域的主权标识，这不仅是网络安全能力的突破，更是国家在全球数字经济博弈中掌握规则定义权的关键落子。

第三重防御：动态数据加密重构信任边界。动态数据加密技术正在重塑国家数据主权的防御体系。在云计算环境中，数据处理环节的明文暴露风险已成为国家安全的关键威胁。从政务数据联合分析到跨境金融交易，任何明文计算都可能成为主权防线的缺口。为此，可信执行环境与全同态加密构成双重技术支柱。可信执行环境依托中央处理器（CPU）指令集在内存空间构建硬件级隔离的“安全飞地”，仅允许授权代码访问加密信息，即使云平台被入侵，敏感数据仍受物理级保护。全同态加密基于格密码学实现密文直接运算，使医疗影像分析、工业设计仿真等核心业务无需解密即可完成。这两项技术从硬件和算法维度封堵数据使用环节的泄露风险，让“数据可用不可见”从理论构想变为国家关键信息基础设施的标配能力。

这种密态计算体系从根本上重构了云计算的信任边界。传统模式下，数据安全依赖对云服务商的盲目信任，而动态加密技术将主动权交还数据所有者。比如，当机器学习模型在可信执行环境中加密训练时，云服务商无法窥探原始数据；当跨境金融监管通过全同态密文协同计算时，敏感信息始终处于数学不可逆状态。物理服务器的硬件漏洞、虚拟化层的权限滥用、境外供应链的后门植入，在动态加密体系下均被降维打击。这种以数据本体安全为核心的防护范式，标志着云计算从“基础设施可信”向“数据内生安全”的战略转型，使云端算力真正成为可控的生产力工具而非数据主权让渡的风险源。

数字治理

人工智能大模型需要什么样的数据

□ 吴 洁



近年来，人工智能领域的大语言模型发展迅猛。从GPT-4到Claude，从Kimi到DeepSeek-R1全球大模型百花齐放，技术升级迭代不断加速。一般认为，大模型的进步得益于算力规模和参数堆叠，实际上，决定模型“智能涌现”能力的核心因素是数据的质量。大模型不是靠“吃得多”变聪明，而是靠“吃得有结构有质量”而变得智慧。准确理解“什么样的数据会被人工智能大模型需要”，不仅关乎新质生产力时代关键产业链的升级方向，更关乎我国的国家安全。

大模型为何“偏爱”结构化的数据体系

当前主流大模型多以转换器（Transformer）架构为核心，是一种用于自然语言处理（NLP）和深度学习任务的模型架构，其注意力机制并非基于词语字面含义的理解，而是着重于构建语言单元之间的关联网络。因此，大模型在训练过程中，能够有效掌握和泛化的能力，取决于输入数据是否具备明确的内部逻辑结构。以编程代码、数学题等规模式数据为例，其数据天然带有强逻辑性、严格语法和可预测的函数组织，这使得模型能够学习推理路径与规划策略，形成具备执行力的认知结构。

相较之下，那些内容零散、缺乏上下文、逻辑模糊的非结构化语料，仅能训练模型表面语言生成能力，无法支撑深层理解与可靠输出。这说明，大模型的“理解”行为并非语法的直觉掌握，而是一种依托于“结构识别”的关系构建过程。如果没有明确的结构，模型便无法提取有效的推理路径，最终只能依赖统计模拟，无法进行真正的

知识推理和创新。结构明确、逻辑严密的数据体系，才是推动大模型能力跃升的真正基础。

五类关键数据形态与其对模型能力的支撑

目前，大模型所依赖的关键数据类型主要包括五类，它们对应着模型不同的认知能力。一是规模式数据，如程序代码和数学逻辑题等，此类数据构成了模型进行推理、决策和任务规划的基础，支撑其逻辑严谨性的训练。二是多样化的语料，涵盖口语、方言、网络表达、跨文化文本等。这类语料提升了模型在真实环境下的适应能力，使其具有更强的语言理解广度和多语境迁移能力。三是高质量文本，涵盖新闻报道、学术论文和政务公开报告等，这些数据不仅内容权威、语言严谨，也具备语篇连贯性，有助于提升模型生成内容的准确性与专业可信度。四是对话型数据，如客服对话和问答论坛等，能够训练模型的多轮交互和情感感知能力，强化人机协同效率，尤其在政务服务办事、民生服务等场景具有高度价值。五是跨模态对齐数据，包括图文、音文和视频脚本等，这类数据发展了模型在多模态空间中的表示能力，推动模型实现多模态信息的整合处理，是构建AI辅助教育、智慧医疗、工业自动化等领域智能系统的关键支撑。

这五类数据并非彼此割裂，而是在应用中互相交织，构建出复杂的“数据网络结构”。例如，在智慧教育场景中，图文教材（跨模态）结合问答记录（对话型）与知识点讲解（高质量文本），可实现对学生认知路径的全方位建模，增强模型的适应性和个性化反馈能力。

当前数据生态所面临的挑战及后期应用

尽管近年来训练数据的数量大

幅增长，但在构建高质量、结构优的数据生态方面仍面临挑战甚至会存在意识形态风险。首先，数据的“结构偏斜”问题十分突出，如互联网上代码类和科技类数据占比过高，导致大模型在处理人文史料如历史、艺术类文本时，缺乏足够的训练数据，表现出理解能力上的局限性。其次，偏见残留问题也不容忽视，来源于社交平台等非审核语料的数据，可能含有某种偏见，若未经清洗即用于训练，模型继承这些偏见，易导致模型在公共服务场景中出现不当或失误回应，引发社会信任问题。最后，“低资源领域”数据匮乏。如少数民族语言、特殊行业标准（如基层医疗记录、乡村治理案例等）数据尚未系统化整合，制约了人工智能在基层治理、公共服务等领域的深度应用。为推动构建面向国家发展新质生产力的高质量数据体系，可重点从以下三个方面发力：一是实施认知驱动的数据设计。借鉴儿童语言习得机制，通过“课程学习”方式，分阶段引导模型掌握从基础表述到复杂推理的知识结构。二是加强数据结构标注能力建设。通过加入因果链、时间线、角色关系等标注，帮助模型建立更深层的逻辑网络，提升其事务识别与判断能力。三是探索以人工智能生成合成数据辅助训练的机制。在保障数据真实有效的前提下，利用大模型自身生成结构良好的语料，再由专业人员审核校正，实现“人机共训”，突破高质量数据不足的瓶颈。

以“高质量结构性数据”作为新质生产力时代的新基建支点

大模型不完全是传统意义上靠“堆参数、拼算法”获得突破的技术路径，更是在“高质量结构性数据”之上成长的智能系统。人工智能模型的训练与优化是一个系统性