• 专题:双清论坛“大模型时代数智商务的理论与方法” •
DOI: 10.3724/BNSFC-2025-0016

面向数智商务的大模型垂域化关键技术研究*

高超越[1] 刘和福[1**] 刘建伟[1] 姜广鑫[2] 姜贺敏[1] 秦 娟[1] 夏 昊[2]

1. 中国科学技术大学 管理学院,合肥 230026
2. 哈尔滨工业大学 经济与管理学院,哈尔滨 150001

[摘 要] 随着全球数智商务的快速发展,大模型技术作为核心驱动力,正在重塑产业链与价值链。本文系统综述了面向数智商务的大模型垂域化关键技术研究进展,围绕算力调度、算法优化、数据治理、架构适配和应用构建与运维管理五个维度展开分析。研究发现,大模型垂域化面临分布式算力调度瓶颈、轻量化与性能平衡、数据治理体系不完善、基础模型架构适配性不足以及全生命周期管理缺失等核心挑战。针对这些挑战,本文梳理了当前的研究进展,包括基于算力网络的动态调度策略、参数高效微调与轻量化算法、多模态数据治理框架、模块化与增量式架构优化,以及覆盖全生命周期的智能化运维体系。此外,本文还探讨了未来研究方向,如跨区域算力协同调度、任务特化算法设计、动态数据权属机制、弹性架构优化以及可信人机协同等。基于技术与管理深度融合的视角,为大模型垂域化进一步推动数智商务和产业创新提供了理论参考。

[关键词] 数智商务;大模型;垂域化;算力调度;算法优化;数据治理;架构升级;人机协同

伴随全球数智商务的迅猛发展,以大模型为典型代表的数智技术,正逐渐成为重塑全球产业链与价值链的核心要素。这一发展态势与党的第二十届中央委员会第三次全体会议所提出的“以国家标准提升引领传统产业优化升级,支持企业运用数据智能技术、绿色技术改造提升传统产业”①目标高度契合。2025年政府工作报告进一步将“人工智能+”行动列为重点任务,从而明确了大模型在产业融合中的战略地位②。作为数智化的基石,大模型技术已彰显出卓越的通用性,而数智商务的快速推进也催生出丰富多样且复杂的大模型垂直领域应用[1]。例如,科大讯飞医疗、政务和法律等垂直大模型已逐步落地:其中“智医助理”覆盖300多种医学场景,在约6万家基层医疗机构应用,累计提供超8.77亿次AI辅诊建议;政务大模型服务4 500多项标准化事项,显著提升审核与登记效率;法律领域与近百家司法单位合作搭建AI专业辅助平台,实现行业智能化升级③。在多模态大模型领域,智象未来作为国内多模态大模型的领军企业,自主研发产品已实现全球范围技术领先和商业化落地,面向影视、广告、设计、文旅等行业,服务超过2 000万用户及数万企业,赋能内容创作、营销和城市特色产业,成为中国多模态AI商业生态的代表④。在出行领域,携程“携程问道”垂直大模型基于筛选200亿高质量旅游数据,专注于行业场景,为用户提供智能、精准的旅游推荐及预订服务,是国内旅游行业数智化升级的重要突破⑤。这些案例充分证明,大模型技术正以前所未有的深度和广度融入数智商务的各个层面,以其卓越的通用性与各

收稿日期:2025-07-27; 修回日期:2025-10-16
* 本文根据国家自然科学基金委员会第406期“双清论坛”讨论的内容整理。
** 通信作者,Email:liuhf@ustc.edu.cn
本文受到国家自然科学基金项目(72121001,72332007)的资助。
① https://fdi.mofcom.gov.cn/selfBuild/1729671575172/content.html?id=9914.
② https://www.gov.cn/yaowen/liebiao/202503/content_7013163.htm.
③ http://dz.jjckb.cn/www/pages/webpage2009/html/2025-02/27/content_104973.htm.
④ https://finance.sina.com.cn/tech/roll/2025-07-29/doc-inficfwe7908218.shtml.
⑤ http://dz.jjckb.cn/www/pages/webpage2009/html/2023-07/27/content_94173.htm.

引用格式: 高超越,刘和福,刘建伟,等. 面向数智商务的大模型垂域化关键技术研究. 中国科学基金,2025,39(5):761−772.
Gao CY,Liu HF,Liu JW,et al. Research on key technologies for domain-specific large language models in digital intelligence business. Bulletin of National Natural Science Foundation of China,2025,39(5):761−772.(in Chinese)

行业日益增长的复杂需求相结合,正在催生一个丰富、多样且充满活力的垂直领域应用生态。在此情形下,深入探究大模型垂域化关键技术的管理与应用,成为达成技术与产业精准对接的关键切入点,对推动我国智能商务发展具有深远意义[2,3]。

在当前大模型垂域化关键技术的管理与应用过程中,业界与学界共同面临五项核心挑战。这些问题不仅是学术前沿的研究难题,也是实际产业落地过程中亟需解决的关键性障碍[4,5]。其一,分布式算力调度瓶颈尚未得到解决。在算力受限的情况下,异构资源的协调调度以及训练任务的智能分配成为大模型垂域化的关键难题。为此,需要构建高效的资源感知与监控体系,开发兼顾垂域性能与能耗的调度算法,以实现算力资源的最优配置和高效利用。其二,轻量化模型与性能之间难以实现平衡。若要在有限的有效算力下实现高质量推理,就必须突破模型压缩与推理精度之间的权衡。知识蒸馏、参数剪枝和量化压缩等方法亟待在垂域泛化能力、平台适应性和资源约束之间达成平衡。其三,数据治理体系尚不完善。对于来源多样、结构复杂的数据,从采集、处理到治理的一体化治理框架,包括数据标准、质量评估和合规机制等,仍存在欠缺。其四,基础模型架构与垂域场景的适配性不足。通用模型难以精准响应场景需求,需要建立“预训练—微调—部署”的动态闭环机制,提升领域知识注入能力和场景感知能力,打造具备自适应演化能力的协同模型架构。其五,垂域化应用的全生命周期管理体系仍有待构建。大模型垂域化尚未构建具备可解释性和行为安全边界的全链条管理机制,难以保障其在复杂商务流程中的稳定运行和持续优化。

针对上述挑战,本文聚焦于数智商务的典型场景,致力于系统梳理大模型垂域化的关键技术体系。该体系以算力调度为基础资源支撑,以算法设计与数据治理为核心要素,以架构适配为系统层面的整合保障,并以智能体部署为价值实现的最终途径。本文将依次对这些层面的研究进展进行梳理,并剖析其在应用过程中的管理瓶颈与潜在优势,进而提出面向效率提升、模态融合、系统协同与场景落地的未来研究重点。本论文的目的在于为构建具备自主可控、可持续演进特性的大模型垂域技术的管理与应用研究提供方向性的指引。

1 大模型垂域化关键技术研究进展

针对面向数智商务领域的大模型垂域化关键技术,本部分将分别对分布式计算资源的智能调度策略展开探讨,剖析在资源受限环境下高效算法的优化方法,探究多模态数据开发与治理的技术路径,阐述基础大模型与垂直应用场景之间的协同适配机制,进而研讨商务智能体全生命周期的管理框架。通过对这五个关键维度的研究进展进行系统性梳理,本部分旨在全面总结大模型垂域化技术的管理与应用研究现状,为未来研究方向的研讨奠定基础。

相关文献梳理的整体逻辑架构图如图1所示,以资源、要素、架构、应用四大层次结构,系统梳理了大模型在垂直行业落地的技术体系和资源流动路径。首先以算力调度为基础,聚焦于算法优化和数据治理,强调架构弹性与多模态适配,并最终落脚于应用场景的构建和运维。层与层之间通过反馈和支撑机制实现上下贯通、动态协同,数据与算力整体贯穿、要素和架构相互驱动,突出体现了大模型数智商务应用多维联动、持续演进和价值导向的系统性创新逻辑,为产业级智能化升级提供了扎实的基础和框架。

1.1 大模型垂域化的算力调度与分配

算力资源分为三类:以CPU为主的通用算力、以GPU和AI芯片为主的智能算力和超级计算机提供的超算算力[6]。由于大模型在通用领域广泛应用,我国智能算力需求迅速增长,预计2026年将达到1 271.4 EFLOPS,其中大模型垂域算力占比将达50%以上①。由此,计算能力和资源有限的单一算力中心难以满足大模型垂域化的算力需求。基于算力网络的算力调度与分配成为当前的主流发展方向。然而,算力调度与资源分配正面临多任务、高并发、异构资源并存的复杂局面。传统静态算力分配机制难以满足模型训练与推理对实时性、灵活性和能效的高要求,算力调度正演化为集技术优化与管理决策于一体的复杂系统工程。当前研究对任务调度、网络资源分配、数据布局优化与跨系统协同调度等进行了探索。其中,任务调度被认为可以决定资源利用率与系统响应速度。Chowdaiah等[7]通过能耗感知的RWTS(Resource-efficient Workload Task Scheduling)模型,计算了执行物联网任务所需的资源数量,从而实现低资源占用下的高性能。网络资源调度则是通过优化网络性能与服务质量应对带宽、缓存等资源有限的挑战。Blöcher[8]以最小成本最大流算法为基础,构建HIRE(Holistic Resource Scheduling)资源调度器,统一协调服务器与网络计算资源。另外,数据布局合理性被认为能决定存储与访问效率。Jin等[9]通过数据“搭载”来

① http://www.xinhuanet.com/tech/20230111/ea44384fdad446578a57cbdcc6994840/c.html.

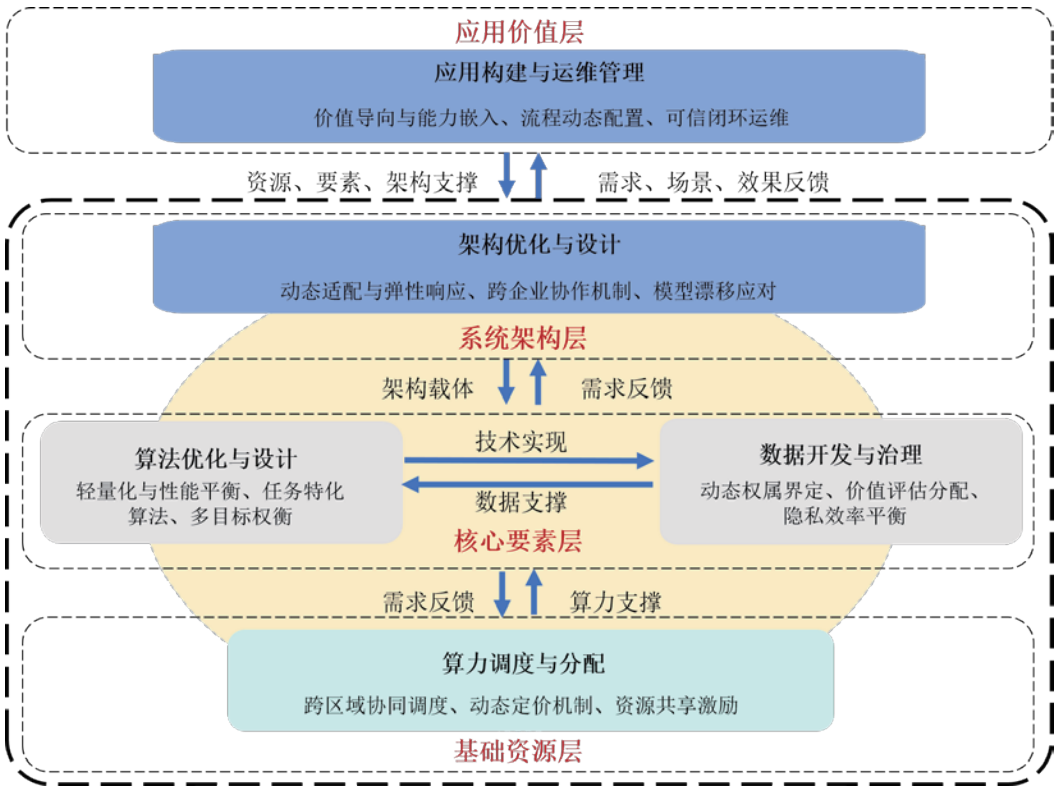

图1　大模型垂域化关键技术研究进展逻辑架构图

Fig.1　Logical Architecture of Key Technologies for Vertical Localization of Large Models

优化卸载路径,缩短任务响应时间。而在多资源、多实体场景下,多源协同调度正成为研究重点。Yang等[10]设计的SplitDL算法就结合了VCG(Vickrey-Clark-Groves)机制和物联设备偏好,实现资源与性能的最优匹配。

管理学中有关生产调度、资源协调与动态分配的研究对算力调度与分配的研究具有借鉴意义。如基于拍卖谈判机制的分布式多项目调度[11]通过多Agent系统协调全局资源冲突,为算力网络的实时调度提供了参考。类似地,云制造中的分布式鲁棒优化[12]通过事件关联模糊集处理不确定性,为算力调度中的动态决策能提供方法论支持。李飞飞和徐哲[13]提出的两阶段资源分配机制可应用于算力动态调度场景,通过预分配和再分配策略优化算力资源利用率。王敏和刘国山[14]的任务组合优化方法则为分布式计算任务的高效调度提供了组合决策的新思路。

1.2　大模型垂域化的算法优化与设计

大模型垂域化常受限于计算、数据及业务资源,并对算法调优提出更高要求。具体而言,其需对基座模型进行按需微调,实现高效推理。参数高效微调(Parameter-Efficient Fine-Tuning, PEFT)通过冻结原始模型,仅引入小规模可训练模块来实现高效适配。主流方法包括Adapter、Prefix-tuning等可加性微调,以及LoRA(Low-Rank Adaptation)等重参数化策略,后者通过低秩矩阵分解,在不增加推理成本的前提下保持精度。近年来,Parallel Adapter、AdapterFusion、Prefix-Propagation、AutoLoRA等方法进一步提升了多任务适应能力[15]。同时,结合4-bit量化的PEQA(Parameter-Efficient and Quantization-aware Adaptation)在算力受限场景下表现良好,体现了PEFT与模型压缩的协同优化潜力[16]。推理优化方面,重点在结构稀疏化和注意力机制重构。动态输入剪枝(Dynamic Input Pruning)[17]与无训练动态稀疏(Dynamic Sparse No Training, DSNT)等方法,可根据输入敏感度动态稀疏计算,显著降低算力需求,无需重新训练[18]。还有注意力机制优化,其中的窗口化(Sliding Window Attention)[19]和稀疏注意力(Progressive Sparse Attention)[20]机制通过限制Token间计算范围,缓解了长文本处理中的内存压力。而LLM-KICK(Knowledge-Intensive Compressed LLM Benchmark)[21]和LLMC(LLM Compression Toolkit)[22]等则为量化模型的部署提供了性能基线。

轻量化是降低大模型资源消耗,提升部署灵活性的另一关键路径,主要包括神经架构搜索(Neural Architecture Search, NAS),混合专家(Mixture of Experts, MoE)系统,以及知识蒸馏和迁移学习等。神经架构搜索(Neural Architecture Search, NAS)通过自

动化结构设计优化模型体积与性能，如引入加权主成分分析(W-PCA)的零样本NAS，可在无需完整训练的情况下高效评估架构[23]；LoNAS结合LoRA，探索弹性 LoRA 搜索空间，进一步降低全规模NAS的内存和计算消耗[24]。混合专家系统(Mixture of Experts, MoE)则通过稀疏激活机制，在保证模型整体能力的同时，实现每次推理仅调用部分专家，显著提升单位算力效率[25]。如DeepSeek-V3虽然全量模型参数量为6 710亿，但每次使用时实际仅激活370亿参数。MoE-CAP(Cost, Accuracy, and Performance)为专家网络提供了成本、准确性和性能的权衡评估框架[26]，而Ling-Coder-Lite等模型也验证了MoE在垂直任务中的实用性[27]。为解决MoE部署中显存与通信消耗问题，MoLE通过查找表方式预加载专家参数[28]，Speculative MoE则优化并行策略，有效提升推理效率，拓展其在资源受限下的可用性[29]。在知识迁移方面，知识蒸馏技术通过教师-学生模型架构将通用知识压缩迁移至小型模型。特别是在特定垂域场景中，领域自适应蒸馏(如Distilling Domain Knowledge, DDK)[30]与ANON方法可动态调整训练样本或集成Adapter技术，显著降低标注依赖[31]。迁移学习方面，多阶段渐进式策略在任务相关性强的垂直应用中表现突出[32]，而知识图谱辅助迁移可强化结构化领域信息的建模能力[33]。

部分大模型垂域部署还面临额外挑战。如AI客服对延迟响应极为敏感，需要保障推理实时性[34]；运行于移动终端、边缘设备的大模型在算力、存储、散热等方面受到更严苛的约束，需进行高度轻量化与定制化[35]。同时，大模型垂域应用还需遵守业务所在地数据隐私法规(如欧盟《通用数据保护条例》、中国《数据安全法》《个人信息保护法》等)。为此，联邦学习、差分隐私等隐私计算技术被引入大模型优化流程中以保护用户隐私和商业机密[36]；FedSpine框架将PEFT和结构化剪枝应用于联邦学习，实现多边缘设备协同高效微调[37]；HeLoRA进一步结合联邦学习与LoRA，支持异构设备间的秩值优化[38]。面对资源、性能、安全等多目标并存的优化需求，运筹学中的多目标优化(Multi-objective Optimization, MOO)理论提供了系统解法。如Chang[39]提出MOO框架支持在资源约束下的任务性能权衡，Xia等[40]则利用MOO对多个LoRA适配器进行量化配置管理，降低推理成本。还有学者通过MOO嵌入NAS流程实现结构搜索的帕累托最优解[41]，或基于多目标对齐增强模型行为的人类价值对齐与业务契合度[42]。此外，在运营层面，学者们发现引入强化学习与随机最优控制策略，可应对动态用户需求与系统波动，通过内存感知的动态批处理、智能任务分配等策略提升整体吞吐与SLA(Service Level Agreement)保障能力[40,43]。

1.3 大模型垂域化的数据开发与治理

大模型垂域化过程中，数据采集、加工、评估与管理构成技术实现与价值释放的基础。大模型已广泛应用于数据清洗、建模与语义提取等流程，助力医疗、金融、教育等实现智能化升级[44]，但同时其对数据质量、权属明晰与治理机制也有更高要求。首先，数据采集与清洗是构建垂域模型的起点，涉及了技术优化到权属界定等问题。传统基于规则与模板的数据清洗难以处理复杂语义与语用信息，大模型驱动的框架如IterClean显著提升了清洗效果[45]。同时，可解释性也被证明在复杂商务数据清洗中具有重要作用[46,47]。其次，高质量标注与数据增强是提升垂域模型贡献度量化与利益分配的关键支撑。医疗和金融领域垂类大模型均依赖长期积累的专业标注数据实现了大模型垂域化的成功[48,49]。但在垂域中，标注任务需专业知识支持，难以规模化推广。半监督、弱监督以及“小模型初筛+大模型精标”的协同机制正成为主流路径。新近研究通过模拟人类判断分布(HJD)实现低资源环境下的高效标注，从而降低人工依赖[50,51]。因此，建立基于稀缺性、标注质量与模型提升贡献的动态收益分配机制，正成为推动数据资源价值共创与共享的主流。最后，数据质量评估与治理体系从合规审查到全域治理的转化，是大模型可持续运行的重要保障。从缺失值、异常值检测到数据漂移监控，相关研究构建了多维度质量框架[52,53]。同时，自动化治理系统如SPADE与SEED已实现对数据质量报告生成、清洗、标注、解析等任务的覆盖[54,55]。

在数智商务场景中，数据隐私保护是大模型垂域化的核心挑战。用户个人信息与交易行为高度敏感，不当处理可能违反《通用数据保护条例》《生成式人工智能服务管理暂行办法》等法规。中国信息通信研究院与中国科学院在《大模型治理蓝皮报告》中明确提出，数据合法性、版权保护、伦理对齐与内容安全是大模型治理的四大核心议题[56]。联邦学习作为一种分布式协同训练范式，正被广泛应用于缓解数据孤岛与隐私泄露风险[57]。如Yang等[58]提出的综合性框架整合了水平、垂直与迁移式联邦学习，以支撑数智商务中的跨组织协同建模；Ali等[59]进一步引入区块链技术，构建异步联邦学习架构，解决传统联邦机制中面临的单点故障、模型篡改与客户端异构参与等信任瓶颈，形成多层级安全保障体系。然而，隐私计算通常伴随较高的算力开销。Mohammadi等[60]通过系统评估指出，在联邦学习实践中，“隐私—效率”间的权衡尚未解决。在大模型垂域化

加速发展的背景下，如何在数据合规保护与高效利用之间找到平衡，成为联邦学习下一阶段研究的关键。

1.4　大模型垂域化的架构优化与设计

通用大模型应用于垂直领域常面临资源消耗大、任务适配性差的问题，推动架构层面的系统优化与设计成为关键突破口。动态稀疏化混合专家架构（MoE++）、模块化架构和增量式领域适应等路径，为大模型在复杂垂直场景中实现高效能、低资源、强适应提供了多维解法[61]。

MoE++架构以按需激活、稀疏计算为核心，通过引入零计算专家、门控残差与负载感知路由，显著提高推理效率与泛化能力[62]。以DeepSeek-MoE-16B为例，其单次推理仅激活约28亿参数，较稠密模型节省75%训练成本[25]。MoE-LLaVA则在3B稀疏模型上实现超越13B稠密模型的性能，展现出在跨模态任务中的高度性价比[63]。

模块化架构则强调系统功能的解耦与组合，支持企业灵活调整模型功能以适应快速演变的业务需求[64,65]。在电商、供应链等领域，模块划分使用户画像、商品匹配、库存优化与物流调度等功能可独立迭代与优化，显著提升系统的可维护性与部署灵活性[66]。然而，模块间协同优化和接口标准化仍是实施难点，特别是在多模块并行部署时，需通过精细化调度机制实现数据流与业务逻辑的有序衔接。高效的动态路由策略与接口语义对齐机制，是提升模块化架构实际效能的关键所在。

增量式领域适应旨在支持模型在动态环境中的持续学习与知识迁移，避免灾难性遗忘[67]。当前学者提到的方法包括对抗式增量适应，通过领域判别器与任务分类器的对抗博弈实现新旧知识协同；记忆增强机制，借助可微分记忆网络或PromptTPP保存并调用历史知识[68]；轻量化参数优化，仅微调特定领域适配器，降低资源负担。增量式领域适应的挑战在于跨领域知识迁移可能引发负迁移，需引入语义对齐机制缓解冲突；同时，增量学习可能导致参数膨胀，从而要求通过动态稀疏化策略实现规模控制。未来，自监督与联邦增量学习将成为关键方向，分别致力于构建领域不变表征与多中心异构数据的高效整合。

1.5　大模型垂域化的应用构建与运维

随着大模型在垂直领域的深度融合，其应用构建与运维成为模型价值持续释放的关键环节。首先，构建层面聚焦于实现从模型开发到业务集成的技术闭环，涵盖知识注入、流程整合、系统适配等关键环节。在实际应用中，大模型正逐步具备将复杂业务流程转化为可学习任务的能力，尤其在采购、建筑运维、供应链管理等领域展现出显著优势。如Allal-Chérif等在采购流程中的智能代理、García Coria等提出的多智能体数字孪生框架，均展示了大模型在流程重构与任务自动化中的价值[69-71]。这一过程离不开语义建模、本体论嵌入与数据治理体系的支撑，它们共同构建了支撑大模型理解业务语义、适应流程结构的底层能力[72,73]。

其次，组织与交互层面的适配机制在实际落地中同样不可忽视。大模型不仅是技术工具，更是嵌入组织运行的智能参与者，在高创造性、决策敏感或情感交互场景中，人机协同的质量直接影响用户体验与业务成效。在此背景下，“信任”成为影响模型接受度和持续使用的关键机制，包括对模型可解释性、稳健性、情绪表达与行为一致性的认知[74-79]。研究表明，用户对于AI的语言风格、拟人化程度、语音特征等具有显著偏好差异，在不同应用场景中需因地制宜设计交互策略。例如，在直播电商、本地化语音助手等场景中，系统需要具备高度的角色适配能力与情绪响应能力，以实现信息精准传达与情感共鸣[80-83]。

在运维层面，大模型需要面向生命周期全流程构建智能化、可持续的管理体系。模型性能会随数据、环境与业务变化而退化，需借助持续监控、自动调优与模型漂移检测等机制维持服务稳定性[84]。此外，合规要求日趋严格，尤其在医疗、金融等高风险领域，系统还需满足GDPR、KYC/AML、HIPAA等监管标准。边缘智能、差分隐私与联邦学习等隐私计算技术为模型合规落地提供了技术路径[85,86]。总体而言，构建一个覆盖“流程建模—能力注入—人机协同—可信运维”的系统化应用管理框架，是推动大模型在数智商务垂直领域实现稳定部署与价值创造的关键所在。

2　面向数智商务的垂域化大模型研究挑战与关键科学问题

随着大模型在数智商务领域的深入应用，其垂域化暴露出的管理问题日益凸显。尽管当前研究在算力调度、算法优化、数据治理、架构适配和运维管理等方面取得了显著进展，但仍存在诸多共性挑战亟待解决。这些问题不仅关乎技术实现，更涉及资源分配、组织协同和价值创造等管理核心。本章着眼于大模型垂域化管理中尚未妥善解决的五大关键科学问题，从算力调度与分配、算法优化与设计、数据开发与治理、架构优化与设计，到应用构建与运维管理，系统分析其技术瓶颈与管理难点，为未来研究提供方向指引。表1为相关核心挑战与关键科学问题的总结。

表1 大模型垂域化核心挑战与关键科学问题

Table 1 Core Challenges and Key Scientific Issues in the Verticalization of Large Models

维度	核心挑战	关键科学问题
算力调度与分配	算力资源供需错配,分布式算力调度瓶颈,异构资源协调调度与智能分配困难	• 如何构建跨区域算力资源的协同调度体系? • 如何设计动态定价机制实现资源的最优时空配置? • 如何通过激励机制设计促进算力资源的共享与协同?
算法优化与设计	轻量化与性能平衡困难,模型压缩与推理精度权衡,高复杂性与异质性使迁移适配困难	• 如何发展具备任务特化能力的轻量化算法体系? • 如何构建兼顾成本效率与业务价值的多维评价框架? • 如何在性能、效率、成本与合规性之间建立权衡关系?
数据开发与治理	数据治理体系不完善,缺乏一体化治理框架,数据标准、质量评估和合规机制欠缺	• 如何构建动态数据权属界定机制? • 如何建立基于任务贡献的数据价值评估与利益分配机制? • 如何实现隐私保护与计算效率之间的动态平衡? • 如何构建跨组织数据协作中的信任机制?
架构优化与设计	基础模型架构与垂域场景适配性不足,通用模型难以精准响应场景需求	• 如何构建动态适配与弹性响应机制? • 如何建立跨企业模型协作的可信运行机制? • 如何解决高频场景中的反馈延迟与模型漂移问题?
应用构建与运维	全生命周期管理体系缺失,缺乏可解释性和行为安全边界的全链条管理机制	• 如何构建“价值导向–能力嵌入”双轮驱动的构建逻辑? • 如何建立面向流程的能力动态配置机制? • 如何构建智能体参与下的流程再造理论与治理机制? • 如何建立“持续学习–可信评估–智能干预”为核心的闭环运维体系?

2.1 面向数智商务场景的大模型垂域化算力调度与分配方法

在垂域化大模型的实际部署过程中,算力资源的供需错配已成为垂域管理所面临的核心科学难题。这不仅涉及分布式商务与通用人工智能领域的资源优化问题,更直接关系到垂域场景下异构模型的结构性资源约束及其协同效益机制。具体科学问题包括:(1)区域协同调度体系的科学挑战:在“东数西算”战略背景下,如何依据垂域模型任务及业务负载,构建跨区域、多中心的算力协同规划与实时弹性调度机制,以突破模型参数规模与区域算力分布不均导致的结构性瓶颈,实现任务—资源动态映射与资源孤岛打通。(2)电力—算力联动的多维定价与资源分配问题:垂域模型计算负载与电力供给之间存在复杂耦合,如何设计动态时空定价模型,联合考量模型负载、服务等级协议、动态能耗、区域供电与成本测度,在保障业务持续性的基础上提升全局配置效率。(3)异构环境下分布式激励机制设计难点:在多主体参与的分布式垂域场景中,如何根据模型异构性和任务特征,设计有效激励约束机制(如激励兼容、协同共享、行为策略选择),充分调动算力资源共享与集体协作行为,从而提升整体系统调度效能。

目前的工程优化主要聚焦于算力和电力的协同调度,但面向垂域化应用的科学研究尚缺乏面向异构模型与异构资源的统一度量标准,难以实现针对模型复杂性和算力灵活性的高精度成本—效益分析。此外,垂域模型调度尚未充分融入组织行为与决策机制,导致算法部署与业务实际情境存在适配障碍。垂域化大模型的算力调度已由单一技术优化问题拓展为涵盖技术参数、经济目标和制度/行为约束的综合科学决策问题,亟需融合运筹优化、博弈论与管理制度设计等跨学科方法。未来研究可以:(1)构建融合技术指标、经济成本和制度摩擦的三维评价体系,并基于异构算力进行统一资源度量与实时监测;(2)引入行为运筹学和多智能体建模理论,考察不同激励模式下参与主体对整体协同效率的影响;(3)推进基于博弈论的多区域垂域模型算力协同优化、结合随机规划方法解决能耗与资源耦合的动态调度难题,以及采用多智能体强化学习算法提升实时调度与资源动态共享水平。

2.2 面向数智商务场景的大模型垂域化算法优化与设计创新

高复杂性与异质性正使大模型算法的迁移适配成为当前学界关注要点。当前多数优化研究聚焦于通用任务(如编程、多模态处理、逻辑推理)或垂域知识(如金融、医疗、法律),但缺乏面向具体管理任务与商务流程的定制算法设计。未来研究应以商务情境为牵引,发展具备任务特化能力的轻量化算法体系,重点探索如何在受限资源约束下完成结构性任务,如用户画像构建、销售预测、流程自动化等,以实现效能—成本的最佳匹配。这不仅是技术优化问题,更涉及“任务—算法—资源”三者之间的战略配置与管理决策。

同时,当前主流的算法评估偏重模型精度与参数压缩率等技术性指标,而忽视模型部署的经济性与业务适配性。未来研究应在管理角度构建兼顾成本效率与业务价值的多维评价框架,涵盖算力资源消耗、部署复杂

度、推理延迟、输出可用性、用户体验与算法公平性等维度,形成支撑组织级大模型部署决策的分析工具。此外,大模型的垂域化部署可视为典型的动态多目标优化过程,其从预训练、微调到部署与推理服务,构成了一个端到端的复杂系统。未来研究应引入管理学中的多目标决策、资源配置与最优控制理论,建立刻画性能、效率、成本与合规性之间权衡关系的数理模型。引入博弈论与系统建模方法,探讨多业务实体间模型能力的协同共享机制。通过管理科学与计算智能的深度融合,为设计高性能、低成本、高可靠性的大模型垂域化算法提供理论基础与实施路径。同时,可结合运筹优化与随机控制方法,根据具体数智商务场景需求,进行全生命周期下的垂域大模型设计与运维管理自适应动态优化。例如,Huang等[87]基于Mistral-7B、DeepSeek-Math-7B、LLaMA-3-8B等开源大模型,定向微调构建出一系列专用于优化建模任务的模型,用来解决业界中的优化建模与求解任务。

2.3　面向数智商务场景的大模型垂域化数据开发与治理框架

在垂域化大模型应用场景中,数据治理已被提升为融合技术实现与制度设计的复合型科学管理问题,其复杂性远超通用数智商务和AI系统。当前垂域模型的数据治理不仅关乎数据清洗、融合与安全等传统工程技术环节,更深刻地嵌入权属动态界定、场景化价值共识与多方协作的管理核心。针对面向多模态、大规模异构和业务敏感数据的模型,构建可信AI系统要求系统性应对数据异构性、治理方式演化、跨组织协作和隐私—效率的微妙平衡[88]。传统“三权分置”等数据制度亟需在跨模态及多主体语境下重构——如引入基于智能合约和区块链的数据确权机制,通过自动登记、实时更新和分润绑定,实现多维场景下的数据权益管理与可信使用。

从管理视角看,数据价值评估与利益分配机制构建已成为推动数据要素市场化的核心议题。与传统“按量定价”模式不同,大模型训练中的数据价值高度依赖场景与模型结构,其边际效用呈现非线性乘数效应。因此,需引入基于任务贡献的定价机制,将可解释性算法用于量化数据在模型性能提升中的边际价值,并基于此设计多方可接受的、激励相容的利益分配模型。该路径不仅增强数据流通的激励约束结构,也有助于构建组织间的数据供给合作网络。与之配套,还应发展基于风险感知的数据隐私保护机制,实现隐私控制与计算效率之间的动态平衡。通过建立情境驱动的策略切换体系,使隐私保护强度随数据敏感度与业务场景灵活调整,从而提升数据在商业流程中的可用性与合规性。

此外,跨组织数据协作中的信任构建问题,已成为数据流通机制能否落地的瓶颈。数智商务典型场景,如供应链协同、客户画像共享等,涉及多方标准、权益、责任与风险的系统整合。尽管区块链增强的联邦学习为跨组织建模提供了路径,但在现实部署中,仍需补足责任划分、风险共担、策略协调等制度配套。未来研究应借助组织协同理论与平台治理机制,构建覆盖数据标准制定、行为边界约束与协同效用分配的治理体系,并以机制设计方法论优化跨域协作中的信任传递链条,推动“数据共建、价值共创、成果共享”的商业生态结构真正落地。

2.4　面向数智商务场景的大模型垂域化架构优化与设计方案

在数智商务场景中,业务动态性强、资源波动显著、跨域协同频繁,推动大模型架构从静态通用向专业化、弹性化转型。“基础大模型+垂直场景”的双耦合架构虽已成为主流,但在实践中常因语义迁移失效、实时响应滞后、协同机制缺位而难以支撑复杂业务需求。例如,金融风控模型迁移至零售领域时偏差率达30%[89],静态规则难以应对电商促销等突发情境。因此,未来应聚焦动态适配与弹性响应机制建设。一方面,构建基于动态本体学习的语义对齐体系,结合联邦学习与MoE++实现跨企业协同与隐私保护;另一方面,发展融合Transformer与深度强化学习的预测-调度联合模型,支持“潮汐式”算力需求下的秒级资源编排,并引入智能合约机制优化资源交付效率,实现资源调度的管理自动化。

在架构治理层面,跨企业模型协作的可信运行机制亟需强化。模块化架构虽具扩展性,但在多主体数据异构与标准不统一背景下易形成“架构孤岛”,限制资源共享与协同推理的规模化落地。尽管联邦学习为数据隔离条件下的协作提供了可行路径,但其高通信成本与缺乏激励约束机制的问题仍未解决[90]。对此,研究需引入区块链技术构建算力与数据资源的可信交易机制,通过智能合约实现资源使用权的可验证切分,并引入基于声誉评估的数据价值度量模型,动态调整算力分配策略,激励多主体协同优化与知识共享。此类制度型架构设计不仅是技术问题的延伸,更是组织管理、激励机制与平台治理理论的交叉融合问题。

此外,当前大模型在商务高频场景中仍面临反馈延迟与模型漂移的瓶颈。传统训练更新机制在面对电商推荐、智能客服等高频任务时存在响应滞后与模型退化问题,尤其在多模态数据融合下,特征对齐困难进一步放大误差累积[91]。未来应构建多模态增量学习架构,结合MoE++的token感知式路由策略,实现异构特征的

有效解耦与动态优化;同时引入事件驱动式实时训练机制,将模型更新周期从小时级压缩至分钟级,以提升系统对业务变化的即时感知与响应能力。综上,面向数智商务的大模型架构优化需从“结构设计—资源调度—协同治理—学习反馈”全链条出发,推动技术体系与管理机制的融合创新,构建一个具备动态适配性、资源弹性与治理可信性的深度应用框架。

2.5 面向数智商务场景的大模型垂域化应用构建与运维管理

在数智商务环境中,大模型垂域化应用构建与运维面临复杂性持续升级的挑战。一方面,模型训练多源于通用语料,难以精准对齐企业战略与业务目标,导致“技术能力强—价值实现弱”的错配问题。大模型虽已嵌入业务流程,但缺乏将模型输出转化为业务价值的有效机制[92]。而缺乏用户体验与交互逻辑的系统性理解,则会限制了模型在业务场景中的高效融入[93]。从管理角度出发,未来需构建“价值导向—能力嵌入”双轮驱动的构建逻辑,明确模型能力与业务目标间的映射机制,通过指标约束、流程嵌套与结果可衡量性实现模型与组织价值的一致性,确保大模型作为工具、代理甚至伙伴,服务于组织的长期战略诉求。

其次,面对流程复杂、响应多样的商务需求,当前大模型仍停留在“通用能力+人工适配”模式,难以满足高度个性化与动态变化的业务场景。用户交互效果高度依赖模型在特定流程节点的语境适配与风格响应[94]。因此,需构建面向流程的能力动态配置机制,探索声明式建模、模块组合与上下文注入等方法,实现模型“按需加载、粒度调用、自动调整”,提升流程敏捷性与应用效率。同时,随着模型从“任务助手”逐步演化为“智能体”,其嵌入组织系统将引发业务流程重构与权责机制重塑。在客户服务、销售、招聘等场景中,模型已具备部分决策能力,需重新审视“角色边界”“责任划分”与“监督机制”。管理研究应重点构建智能体参与下的流程再造理论、风险共担机制与组织协同模型,以适应智能化转型中的治理新范式。

最后,在系统运维层面,大模型面临服务稳定性、模型漂移、合规监管等多重压力,传统以人工巡检与被动响应为主的运维体系难以支撑其全生命周期管理。尽管已有研究探索去中心化训练与持续学习算法[76,79,84,85],但缺乏可扩展的智能化运维框架以支持多维监控、自我诊断、自主修复与实时优化。同时,可信性风险正在上升,涵盖数据污染、算法偏见、非预期行为与伦理违规等多种维度。因此,需从管理技术融合角度,构建以“持续学习—可信评估—智能干预”为核心的闭环运维体系,配套可信模型评估指标、验证工具与监管科技方案,支撑从“可用”向“可靠、安全、可控”的能力跃升。综上,数智商务中的大模型应用构建与运维管理,应构建覆盖能力价值对齐、流程能力适配、智能体协同治理与可信智能运维的系统框架,实现模型效能、组织弹性与技术责任的协同提升。

3 总结与展望

本文系统梳理了面向数智商务领域的大模型垂域化关键技术研究进展,围绕算力调度、算法优化、数据治理、架构适配和应用构建与运维管理五个维度,深入分析了当前的技术现状与管理挑战。研究表明,大模型垂域化需要技术与管理的深度融合,未来研究应重点关注以下方向:(1)算力调度与分配:构建跨区域、多主体的动态协同调度体系,结合博弈论与强化学习优化算力—电力联动机制,提升资源利用效率。(2)算法优化与设计:发展轻量化、任务特化的高效微调与推理技术,建立兼顾性能、成本与业务价值的评估框架,推动算法与场景的精准适配。(3)数据开发与治理:完善动态权属界定与价值分配机制,融合隐私计算与区块链技术,实现数据安全流通与高效利用的平衡。(4)架构优化与设计:探索模块化、增量式与稀疏化架构的创新路径,增强模型在动态环境中的语义对齐与实时响应能力。(5)应用构建与运维管理:构建覆盖全生命周期的智能化运维框架,强化人机协同的可信机制,确保模型在复杂业务流程中的稳定性和合规性。通过上述方向的协同突破,大模型垂域化将进一步提升数智商务的智能化水平,为产业升级与价值创新提供核心驱动力。

参 考 文 献

[1] 江小涓. 数智时代的秩序重构与治理合作:合理合意双重目标. 管理世界,2025,41(5):1—14,58,241.
Jiang XJ. Reconstructing order and collaborating governance in the digital-intelligent era:Pursuing dual objectives of rationality and acceptability. Management World,2025,41(5):1—14,58,241. (in Chinese)

[2] 蔡姝雯,张宣,杨易臻,等. 百模大战 既卷模型更卷应用. 新华日报,2024-09-05(09).

[3] 黄鑫. 发展大模型关键在应用. 经济日报,2024-04-19(06).

[4] 籍欣萌,昝红英,崔婷婷,等. 大模型在垂直领域应用的现状与挑战. 计算机工程与应用,2025,61(12):1—11.
Ji XM,Zan HY,Cui TT,et al. Status and challenges of large language models applications in vertical domains. Computer Engineering and Applications. 2025:1—11. (in Chinese)

[5] 陈浩泷,陈罕之,韩凯峰,等. 垂直领域大模型的定制化:理论基础与关键技术. 数据采集与处理,2024,39(3):524—546.
Chen HL,Chen HZ,Han KF,et al. Domain-specific foundation-model

customization:Theoretical foundation and key technology. Journal of Data Acquisition and Processing. 2024,39(03):524—546. (in Chinese)

[6] 唐卓,蒋冰婷,张嘉鹏,等. 算力网络调度基础理论与关键技术现状及展望. 中国科学基金,2025,39(2):240—249.
Tang Z,Jiang BT,Zhang JP,et al. The foundation theory and key technologies of computing first network scheduling present situation and prospects. Bulletin of National Natural Science Foundation of China. 2025,39(02):240—249. (in Chinese)

[7] Chowdaiah NK,Dammur A. Resource-efficient workload task scheduling for cloud-assisted Internet of Things environment. International Journal of Electrical and Computer Engineering (IJECE),2023,13(5):5898.

[8] Blöcher M,Wang L,Eugster P,et al. Holistic resource scheduling for data center in-network computing. IEEE/ACM Transactions on Networking,2022,30(6):2448—2463.

[9] Jin YB,Qian ZZ,Guo S,et al. run runData:Re-distributing data *via* piggybacking for geo-distributed data analytics over edges. IEEE Transactions on Parallel and Distributed Systems,2022,33(1):40—55.

[10] Yang YT,Wei HY. Edge–IoT computing and networking resource allocation for decomposable deep learning inference. IEEE Internet of Things Journal,2023,10(6):5178—5193.

[11] 有维宝,徐哲,刘东宁. 基于拍卖谈判机制的分布式多技能多项目调度. 运筹与管理,2024,33(1):1—8.
You WB,Xu Z,Liu DN. An auction-based negotiation mechanism to distributed multi-skilled multi-project scheduling problem. Operations Research and Management Science. 2024,33(01):1—8. (in Chinese)

[12] 罗遵昊,余乐安,王杜娟,等. 任务服务时间不确定下基于事件关联的云制造任务调度研究. 系统工程理论与实践,2025,doi:10.12011/SETP2024-1746.
Luo ZH,Yu LA,Wang D,et al. Event-wise cloud manufacturing task scheduling under uncertain task service time. Systems Engineering-Theory & Practice. 2025,doi:10.12011/SETP2024-1746. (in Chinese)

[13] 李飞飞,徐哲. 基于两阶段资源分配协调机制的分布式多项目随机调度. 中国管理科学,2022,30(12):38—51.
Li FF,Xu Z. Distributed multi-project stochastic scheduling with two-stage coordination mechanism of resources allocation. Chinese Journal of Management Science. 2022,30(12):38—51. (in Chinese)

[14] 王敏,刘国山. 一种基于任务组合优化的项目调度机制研究. 管理评论,2025,37(3):228—237.
Wang M,Liu GS. A project scheduling mechanism based on activity combination optimization. Management Review. 2025,37(3):228—237. (in Chinese)

[15] Han ZY,Gao C,Liu JY,et al. Parameter-efficient fine-tuning for large models:A comprehensive survey. (2024-03-21)/[2025-11-01]. https://doi.org/10.48550/arXiv.2403.14608.

[16] Kim J,Lee JH,Kim S,et al. Memory-efficient fine-tuning of compressed large language models via sub-4-bit integer quantization// Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023),New Orleans,LA,USA:2023.

[17] Federici M,Belli D,Baalen MV,et al. Efficient llm inference using dynamic input pruning and cache-aware masking//Proceedings of the 8th Conference on Machine Learning and Systems (MLSys 2025),Santa Clara,CA,USA:2025.

[18] Zhang Y,Zhao L,Lin M,et al. Dynamic sparse no training:Training-free fine-tuning for sparse llms//Proceedings of the 12th International Conference on Learning Representations (ICLR 2024),Vienna,Austria:2024.

[19] Fu Z,Song W,Wang Y,et al. Sliding window attention training for efficient large language models. (2025-02-26)/[2025-10-16]. https://arxiv.org/abs/2502.18845. arXiv:2502.18845.

[20] Zhou Q,Yin P, Zuo P,et al. Progressive sparse attention:Algorithm and system co-design for efficient attention in llm serving. (2025-03-01)/[2025-10-16]. https://arxiv.org/abs/2503.00392. arXiv:2503.00392.

[21] Jaiswal A,Gan Z,Du X,et al. Compressing llms:The truth is rarely pure and never simple// Proceedings of the 12th International Conference on Learning Representations (ICLR 2024),Vienna,Austria:2024.

[22] Gong RH,Yong Y,Gu SQ,et al. LLMC:benchmarking large language model quantization with a versatile compression toolkit. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing:Industry Track. Miami,Florida,US. Stroudsburg,PA,USA:ACL,2024:132—152.

[23] Wang S. W-pca based gradient-free proxy for efficient search of lightweight language models//Proceedings of the 13th International Conference on Learning Representations (ICLR 2025),Singapore:2025.

[24] Munoz J P,Yuan J,Zheng Y,et al. Lonas:Elastic low-rank adapters for efficient large language models//Proceedings of the 2024 Joint International Conference on Computational Linguistics,Language Resources and Evaluation (LREC-COLING 2024),Torino,Italia:2024.

[25] Dai DM,Deng CQ,Zhao CG,et al. DeepSeekMoE:Towards ultimate expert specialization in mixture-of-experts language models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers). Bangkok,Thailand. Stroudsburg,PA,USA:ACL,2024:1280—1297.

[26] Jiang Y,Fu Y,Huang Y,et al. Moe-cap:benchmarking cost,accuracy and performance of sparse mixture-of-experts systems. (2024-12-10)/[2025-10-16]. https://arxiv.org/abs/2412.07067.

[27] Codefuse-Ling-Team. Every sample matters: leveraging mixture-ofexperts and high-quality data for efficient and accurate code llm. (2025-03-22)/[2025-10-16]. https://arxiv.org/abs/2503.17793.

[28] Jie S,Tang Y,Han K,et al. Mixture of lookup experts. (2025-03-20)/[2025-10-16]. https://doi.org/10.48550/arXiv.2503.15798.

[29] Li Y,Zheng P,Chen S,et al.Speculative moe:communication efficient parallel moe inference with speculative token and expert prescheduling. (2025-03-06)/[2025-10-16]. https://arxiv.org/abs/2503.04398.

[30] Saxena P,Janzen S,Maass W. Streamlining LLMs:Adaptive knowledge distillation for tailored language models//Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics:Human Language Technologies (Volume 4:Student Research Workshop). Albuquerque,USA. Stroudsburg,PA,USA:ACL,2025:448—455.

[31] Liu J,Zhang C,Guo J,et al. Ddk:Distilling domain knowledge for efficient large language models//Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS

2024),Vancouver,Canada:2024.

[32] Guan CH,Huang C,Li HL,et al. Multi-stage LLM fine-tuning with a continual learning setting// Findings of the Association for Computational Linguistics:NAACL 2025. Albuquerque,New Mexico. Stroudsburg,PA,USA:ACL,2025:5484—5498.

[33] Deng SM,Ma YB,Zhang NY,et al. Information extraction in low-resource scenarios:Survey and perspective. 2024 IEEE International Conference on Knowledge Graph (ICKG). Abu Dhabi,United Arab Emirates. IEEE,2024:33—49.

[34] Cappelli P,Tambe P S,Yakubovich V. Will large language models really change how work is done?. MIT Sloan Management Review,2024,65(3):48—53.

[35] Kundu A,Lim YCF,Chew A,et al. Efficiently distilling LLMs for edge applications//Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies (Volume 6:Industry Track). Mexico City,Mexico. Stroudsburg,PA,USA:ACL,2024:52—62.

[36] Neel S,Chang P. Privacy issues in large language models:A survey. (2023-12-11)/[2025-07-27]. https://arxiv.org/abs/2312.06717. arXiv:2312.06717.

[37] Yao Z,Xu Y,Xu H,et al. Efficient deployment of large language models on resource-constrained devices. (2025-01-05)/[2025-07-27]. https://arxiv.org/abs/2501.02438. arXiv:2501.02438.

[38] Fan BY,Su X,Tarkoma S,et al. HeLoRA:LoRA-heterogeneous federated fine-tuning for foundation models. ACM Transactions on Internet Technology,2025,25(2):1—22.

[39] Chang TH,Wild SM. Designing a framework for solving multi-objective simulation optimization problems. INFORMS Journal on Computing,2025.

[40] Xia YF,Fu FC,Zhang WT,et al. Efficient multi-task LLM quantization and serving for multiple LoRA adapters. Neural Information Processing Systems.,2024.

[41] Zhao YY,Wang LN,Guo T. Multi-objective neural architecture search by learning search space partitions. Journal of Machine Learning Research,2024,25(177):1–41.

[42] Guo Y,Cui G,Yuan L,et al. Controllable preference optimization:Toward controllable multi-objective alignment// EMNLP 2024,Association for Computational Linguistics:Miami,Florida,USA:2024.

[43] Du M,Yu HT,Kong N. Transfer reinforcement learning for mixed observability Markov decision processes with time-varying interval-valued parameters and its application in pandemic control. INFORMS Journal on Computing,2024,37(2):315—337.

[44] Raza M,Jahangir Z,Riaz MB,et al. Industrial applications of large language models. Scientific Reports,2025,15:13755.

[45] Ni W,Zhang KH,Miao XY,et al. IterClean:An iterative data cleaning framework with large language models. ACM Turing Award Celebration Conference 2024. Changsha China. ACM,2024:100—105.

[46] Kraus M,Tschernutter D,Weinzierl S,et al. Interpretable generalized additive neural networks. European Journal of Operational Research,2024,317(2):303—316.

[47] 浪潮通软. 海岳智能合同管理模块应用实践. (2025-01-10)/[2025-07-27]. https://haiyue.inspur.com/ecloud/gywm/xwdt/2025011411274794523/index.html.

[48] Fleming SL,Lozano A,Haberkorn WJ,et al. Medalign:A clinician-generated dataset for instruction following with electronic medical records//Proceedings of the AAAI Conference on Artificial Intelligence. 2024.

[49] Wu S,Irsoy O,Lu S,et al. Bloomberggpt:A large language model for finance. (2023-03-30)/[2025-10-16]. https://arxiv.org/abs/2303.17564.arXiv:2303.17564.

[50] Chen BD,Wang XP,Peng SY,et al. “Seeing the big through the small”:Can LLMs approximate human judgment distributions on NLI from a few explanations? Findings of the Association for Computational Linguistics:EMNLP 2024. Miami,Florida,USA. Stroudsburg,PA,USA:ACL,2024:14396—14419.

[51] Chen BD,Peng SY,Korhonen A,et al. A rose by any other name:LLM-generated explanations are good proxies for human explanations to collect label distributions on NLI. Findings of the Association for Computational Linguistics:ACL 2025. Vienna,Austria. Stroudsburg,PA,USA:ACL,2025:10777—10802.

[52] Batini C,Cappiello C,Francalanci C,et al. Methodologies for data quality assessment and improvement. ACM Computing Surveys,2009,41(3):1—52.

[53] Cong G,Fan WF,Geerts F,et al. Improving data quality:Consistency and accuracy. Pakistan Journal of Biological Sciences,2007. Cong G,Fan WF,Geerts F,et al. Improving Data Quality:Consistency and Accuracy. In Proceedings of the 33rd International Conference on Very Large Data Bases,University of Vienna,Austria,2007:315—326.

[54] Shankar S,Li HT,Asawa P,et al. Spade:Synthesizing data quality assertions for large language model pipelines.//Proceedings of the VLDB Endowment,2024,17(12):4173—4186.

[55] Chen Z,Cao L,Madden S,et al. Seed:Domain-specific data curation with large language models. (2023-10-01)/[2025-10-16]. https://arxiv.org/abs/2310.00749. arXiv:2310.00749.

[56] 中国信息通信研究院政策与经济研究所,中国科学院计算机技术研究所智能算法安全重点实验室. 大模型治理蓝皮报告——从规则走向实践(2023年). (2023-11-24)/[2025-10-30]. http://www.caict.ac.cn/kxyj/qwfb/ztbg/202311/P020231124526622371194.pdf.

[57] Daly K,Eichner H,Kairouz P,et al. Federated learning in practice:reflections and projections. 2024 IEEE 6th International Conference on Trust,Privacy and Security in Intelligent Systems,and Applications (TPS-ISA). Washington,DC,USA. IEEE,2024:148—156.

[58] Yang Q,Liu Y,Chen TJ,et al. Federated machine learning. ACM Transactions on Intelligent Systems and Technology,2019,10(2):1—19.

[59] Ali W,Kumar R,Zhou XM,et al. Responsible recommendation services with blockchain empowered asynchronous federated learning. ACM Transactions on Intelligent Systems and Technology,2024,15(4):1—24.

[60] Mohammadi M,Al-Fuqaha A,Sorour S,et al. Deep learning for IoT big data and streaming analytics:A survey. IEEE Communications Surveys & Tutorials,2018,20(4):2923—2960.

[61] Jin P,Zhu B,Yuan L,et al. Moe++:Accelerating mixture-of-experts methods with zero-computation experts. (2024-10-09)/[2025-10-16]. https://doi.org/10.48550/arXiv.2410.07348.

[62] Fedus W,Zoph B,Shazeer N. Switch transformers:Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research,2022,23(120):1—39.

[63] Lin B,Tang Z,Ye Y,et al. Moe-llava:Mixture of experts for large vision-language models. (2024-01-29)/[2025-10-16]. https://arxiv.org/abs/2401.15947.

[64] Ostapenko O,Su Z,Ponti EM,et al. Towards modular llms by building and reusing a library of loras.(2024-05-18)/[2025-10-26]. https://doi.org/10.48550/arXiv.2405.11157.

[65] Ye Q,Xu H,Xu G,et al. Mplug-owl:modularization empowers large language models with multimodality. (2023-04-27)/[2025-10-26]. https://arxiv.org/abs/2304.14178.

[66] Gomez-Uribe CA,Hunt N. The netflix recommender system. ACM Transactions on Management Information Systems,2016,6(4):1—19.

[67] Jovanovic M,Voss P. Towards incremental learning in large language models:A critical review. Expert Systems,2025,42(10):e70127.

[68] Xue S,Wang Y,Chu Z,et al. Prompt-augmented temporal point process for streaming event sequence. Advances in Neural Information Processing Systems,2023,36:18885—18905.

[69] Allal-Chérif O,Simón-Moya V,Ballester ACC. Intelligent purchasing: How artificial intelligence can redefine the purchasing function. Journal of Business Research,2021,124:69—76.

[70] García Coria JA,Castellanos-Garzón JA,Corchado JM. Intelligent business processes composition based on multi-agent systems. Expert Systems with Applications,2014,41(4):1189—1205.

[71] Yoon S,Song J,Li JT. Ontology-enabled AI agent-driven intelligent digital twins for building operations and maintenance. Journal of Building Engineering,2025,108:112802.

[72] Bag S,Gupta S,Kumar A,et al. An integrated artificial intelligence framework for knowledge creation and B2B marketing rational decision making for improving firm performance. Industrial Marketing Management,2021,92:178—189.

[73] Lalicic L,Weismayer C. Consumers' reasons and perceived value co-creation of using artificial intelligence-enabled travel service agents. Journal of Business Research,2021,129:891—901.

[74] Luo XM,Qin MS,Fang Z,et al. Artificial intelligence coaches for sales agents:Caveats and solutions. Journal of Marketing,2021,85(2):14—32.

[75] Jia N,Luo XM,Fang Z,et al. When and how artificial intelligence augments employee creativity. Academy of Management Journal, 2024,67(1):5—32.

[76] Emaminejad N,Akhavian R. Trustworthy AI and robotics:Implications for the AEC industry. Automation in Construction,2022,139:104298.

[77] Moussawi S,Koufaris M,Benbunan-Fich R. How perceptions of intelligence and anthropomorphism affect adoption of personal intelligent agents. Electronic Markets,2021,31(2):343—364.

[78] Seaborn K,Miyake NP,Pennefather P,et al. Voice in human–agent interaction. ACM Computing Surveys,2022,54(4):1—43.

[79] Han E,Yin DZ,Zhang H. Bots with feelings:Should AI agents express positive emotion in customer service. Information Systems Research, 2022,34(3):1296—1311.

[80] Whang JB,Song JH,Lee JH,et al. Interacting with Chatbots:Message type and consumers' control. Journal of Business Research,2022,153: 309—318.

[81] Garvey A M,Kim T,Duhachek A. Bad news? Send an ai. Good news? Send a human. Journal of Marketing,2022,87(1):10—25.

[82] Bharadwaj N,Ballings M,Naik PA,et al. A new livestream retail analytics framework to assess the sales impact of emotional displays. Journal of Marketing,2022,86(1):27—47.

[83] Rohit K,Shankar A,Katiyar G,et al. Consumer engagement in chatbots and voicebots. A multiple-experiment approach in online retailing context. Journal of Retailing and Consumer Services,2024, 78:103728.

[84] Helo P,Hao Y. Artificial intelligence in operations management and supply chain management:An exploratory case study. Production Planning & Control,2022,33(16):1573—1590.

[85] Chen SZ,Yu DX,Zou YF,et al. Decentralized wireless federated learning with differential privacy. IEEE Transactions on Industrial Informatics,2022,18(9):6273—6282.

[86] Cao LB. Decentralized AI:Edge intelligence and smart blockchain, metaverse,Web3,and DeSci. IEEE Intelligent Systems,2022,37(3):6—19.

[87] Huang C,Tang Z,Hu S,et al. Orlm:A customizable framework in training large models for automated optimization modeling. Operations Research, 2025.

[88] Janssen M,Brous P,Estevez E,et al. Data governance:Organizing data for trustworthy artificial intelligence. Government Information Quarterly,2020,37(3):101493.

[89] Brynjolfsson E,Mcafee A. The business of artificial intelligence. Harvard Business Review,2017,7(1):1—2.

[90] Kairouz P,McMahan HB,Avent B,et al. Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 2021,14(1—2):1—210.

[91] Feng YF,You HX,Zhang ZZ,et al. Hypergraph neural networks. Proceedings of the AAAI Conference on Artificial Intelligence,2019, 33(1):3558—3565.

[92] Allal-Chérif O,Yela Aránega A,Castaño Sánchez R. Intelligent recruitment:How to identify,select,and retain talents from around the world using artificial intelligence. Technological Forecasting and Social Change,2021,169:120822.

[93] Elshan E,Zierau N,Engel C,et al. Understanding the design elements affecting user acceptance of intelligent agents:Past,present and future. Information Systems Frontiers,2022,24(3):699—730.

[94] Choi S,Zhou JR. Inducing consumers'self-disclosure through the fit between Chatbot's interaction styles and regulatory focus. Journal of Business Research,2023,166:114127.

Research on Key Technologies for Domain-specific Large Language Models in Digital Intelligence Business

Chaoyue Gao[1] Hefu Liu[1*] Jianwei Liu[1] Guangxin Jiang[2] Hemin Jiang[1] Juan Qin[1] Hao Xia[2]

1. *School of Management, University of Science and Technology of China, Hefei 230026, China*
2. *School of Management, Harbin Institute of Technology, Harbin 150001, China*

Abstract With the rapid development of global digital and intelligent commerce, large language model has become a core driving force, reshaping industrial chains and value chains. This paper systematically reviews research progress on key technologies for domain-specific adaptation of large models in digital and intelligent commerce, analyzing five dimensions: computing power scheduling, algorithm optimization, data governance, architecture adaptation, and application development and operation management. The study finds that domain-specific adaptation of large models faces core challenges including: bottlenecks in distributed computing power scheduling, balancing lightweight design with performance, incomplete data governance systems, insufficient adaptation of foundational model architectures, and lack of full lifecycle management. To address these challenges, the paper organizes current research advances, including: dynamic scheduling strategies based on computing power networks, parameter-efficient fine-tuning and lightweight algorithms, multimodal data governance frameworks, modular and incremental architecture optimization, and intelligent operation systems covering the full lifecycle. Additionally, the paper explores future research directions such as: cross-regional collaborative computing power scheduling, task-specialized algorithm design, dynamic data ownership mechanisms, elastic architecture optimization, and trustworthy human-AI collaboration. From the perspective of integration between technology and management, it provides theoretical references for further advancing domain-specific adaptation of large models to promote digital and intelligent commerce and industrial innovation.

Keywords digital and intelligent commerce; large language models; domain-specific adaptation; computing power scheduling; algorithm optimization; data governance; architecture enhancement; human-AI collaboration

刘和福 中国科学技术大学管理学院教授、博士生导师，主要研究领域包括数智化管理、IT价值创造、数字化商业模式等。主持国家自然科学基金重点项目、青年科学基金项目(B类)及科技部重大专项课题等科研项目。研究成果发表在*MIS Quarterly*、*Journal of Operations Management*、*Production and Operations Management*等学术期刊上。曾获教育部自然科学奖一等奖、爱思唯尔(2020—2024)中国高被引学者等荣誉。

高超越 中国科学技术大学管理学院特任副教授，主要研究方向为数字经济与金融科技，具体包括大模型与区块链技术应用及影响、社交媒体与金融市场、量化交易等。研究成果发表在*Journal of Operations Management*、*Production and Operations Management*、《管理世界》等国内外学术期刊上。曾多次获得国际顶级会议最优论文奖和最优论文提名。

(责任编辑 张 强)

* Corresponding Author, Email: liuhf@ustc.edu.cn