大语言模型赋能智能网络的应用与挑战

Applications and Challenges of Intelligent Networks Empowered by Large Language Models

牛嘉林/NIU Jialin[1,2]，邢铭哲/XING Mingzhe[1]，
张蕾/ZHANG Lei[1]
（1. 中关村实验室，中国 北京 100194；
2. 北京邮电大学，中国 北京 100876）
(1. Zhongguancun Laboratory, Beijing 100194, China;
2. Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI：10.12142/ZTETJ.202505003
网络出版地址：https://link.cnki.net/urlid/34.1228.TN.20250926.1148.004
网络出版日期：2025-09-26
收稿日期：2025-07-25

摘要：大语言模型（LLM）正逐步融入网络的智能规划建设、智能维护、智能优化与智能网络运营等关键环节，在提升自动化与智能化水平方面展现出显著潜力。基于大模型与智能网络融合的背景，梳理了大模型在智能网络各关键领域的应用路径，总结其在提升决策效率、增强服务适配性、降低运维成本等方面的优势。深入探讨了智能网络环境下大模型面临的解空间组合爆炸与NP难（NP-hard）问题、多维度不确定性、实时性约束、数据异构性、人机协同与成本效益平衡等技术挑战，并归纳了现有应对思路。未来，随着多模态融合、在线学习与人机协同等技术的持续进步，大语言模型有望在推动网络从规则驱动向知识驱动转型的过程中发挥重要作用，为智能网络的发展提供新思路。

关键词：大语言模型；智能网络；人工智能；应用与挑战

Abstract: Large language models (LLMs) are gradually being integrated into key stages of intelligent network development, including network planning and construction, intelligent maintenance, optimization, and operations, demonstrating significant potential in enhancing automation and intelligence. This paper, grounded in the context of the convergence between LLMs and intelligent networks, reviews the application pathways of LLMs across critical areas of intelligent networks. It summarizes their advantages in improving decision-making efficiency, enhancing service adaptability, and reducing operational and maintenance costs. Furthermore, it explores the major technical challenges faced by LLMs in intelligent network environments, including the combinatorial explosion of solution spaces and NP-hard problems, multidimensional uncertainties, real-time constraints, data heterogeneity, human-machine collaboration, and cost-benefit trade-offs, and outlines current strategies for addressing these issues. Looking ahead, with the continued advancement of technologies such as multimodal integration, online learning, and human-machine collaboration, LLMs are expected to play an increasingly important role in facilitating the transition of networks from rule-driven to knowledge-driven paradigms, offering new perspectives for the development of intelligent networks.

Keywords: large language model; intelligent network; artificial intelligence; application and challenge

引用格式：牛嘉林，邢铭哲，张蕾．大语言模型赋能智能网络的应用与挑战 [J]. 中兴通讯技术，2025, 31(5): 11-18. DOI: 10.12142/ZTETJ.202505003

Citation: NIU J L, XING M Z, ZHANG L. Applications and challenges of intelligent networks empowered by large language models [J]. ZTE technology journal, 2025, 31(5): 11-18. DOI: 10.12142/ZTETJ.202505003

人工智能（AI）技术的发展，尤其是以大语言模型（LLM）为代表的模型范式，正在重塑网络智能化的技术范式与演进方向。

1）AI与网络的融合

随着网络规模与业务复杂度的持续提升，AI技术与网络的深度融合已成为推动网络智能化转型的核心路径。AI通过将网络管理问题的输入特征、输出目标与优化策略进行有效映射，显著增强了网络系统的自学习能力与动态自治水平。AI算法在网络中的应用不断深化，覆盖从资源调度、流量预测到安全防御等多个方面。其中，强化学习适用于动态优化任务，而联邦学习等分布式AI框架则在保障数据隐私的前提下，实现了跨站点模型的协同训练与优化[1]。

基金项目：中国工程院战略研究与咨询项目（2023-JB-13）

大模型的迅速发展为AI与网络深度融合带来了新机遇。这类模型具备强大的语义理解、知识抽取和上下文推理能力，已被应用于网络配置自动生成、异常行为检测和安全事件响应等场景，展现出跨层次、跨域调控网络状态的潜力。当前，AI算法的适配性研究逐步深化：监督学习在流量预测、路径选择等有标签任务中表现稳健，而无监督学习通过模式挖掘在恶意行为检测领域实现高精度识别[2]。随着模型能力的持续迭代，AI正驱动网络从“人工可控”向“全域自治”加速演进，成为新一代智能网络的核心引擎。

2）网络智能化需求的提升

现代网络正朝着虚拟化、编程化和弹性化方向快速演进，新型网络架构如软件定义网络（SDN）与网络功能虚拟化（NFV）的推广，使得网络结构更加动态复杂，策略调整的实时性要求显著提升[3]。与此同时，5G与6G网络以更高的速率、更低的时延和更强的连接能力为目标，对网络资源调度、故障恢复、服务保障等方面提出了更高的智能化要求。

在此背景下，传统的人工配置与静态策略已难以应对高速发展的业务需求，AI技术成为满足网络智能化需求的核心驱动力。网络数据的激增，特别是加密流量比例不断上升，进一步加剧了对自动化分析和决策系统的依赖。基于图神经网络的流量识别模型可实现复杂通信模式的分类判断，而大模型驱动的意图识别系统则能够将用户自然语言需求高效转换为可执行的网络策略[4]。此外，边缘计算与AI技术的结合也推动了“算力下沉”，使得边缘节点具备一定的自主决策能力，从而构建更加高效、协同和智能的分布式网络体系[5]。

3）大模型带来的网络变革

大模型正推动网络从规则驱动走向认知驱动、从模块自动化迈向系统智能化。大模型的引入不仅提升了网络系统的性能与效率，更在架构理念、交互方式与运行机制上带来了深刻变革。

（1） 网络性能优化

大模型通过统一感知、理解与决策流程，打破了传统网络中配置、监测与优化的割裂壁垒。模型可自主解析网络状态与业务需求，实现资源按需分配与动态调度，支撑网络从被动调整向前瞻自适应转变。

（2） 网络安全防护

传统安全系统多依赖固定规则与特征匹配，而大模型具备语义理解与上下文建模能力，能够识别未知攻击路径与复杂行为链条。大模型生成式机制也推动了“以攻促防”的新范式，重塑网络安全的响应机制与攻防博弈逻辑。

（3） 智能网络运维

在大模型驱动下，网络运维从脚本执行升级为知识交互。模型可解析自然语言指令，结合上下文生成修复策略，实现从“事后应对”到“实时响应”的跃迁，并通过持续学习与知识图谱支撑经验迁移。

（4） 网络资源调度

相较于传统调度策略依赖静态规则与离线优化，大模型能够实时感知任务优先级与计算资源分布，主动做出跨域协同调度决策，具备“即看即调、即调即优”的动态自演化能力。

（5） 网络多模态融合

大模型打破了网络中结构化与非结构化数据间的壁垒，将日志、指令、拓扑、配置等异构信息转化为统一语义空间。这为网络提供了“理解自己”的能力，实现从数据堆积到知识生成的跃升。

（6） 网络用户体验优化

传统网络服务依赖用户适应系统，而大模型实现了系统适应用户，其对自然语言意图的理解与反馈能力，使网络响应机制转向按语义驱动配置资源，推动了从“功能对接”到“体验协同”的转变。

大模型在性能、安全、运维、调度、多模态融合和用户体验等方面对网络的全面重塑，彰显了其从单一工具向网络认知中枢演进的路径。随着模型在计算效率、跨层协同与实时推理等方面的进一步提升，其在推动网络向更高自治化和智能化阶段迈进中将发挥越来越关键的作用。

1 大模型与网络融合的技术基础与应用

1.1 大模型的发展与优化技术

大模型在智能网络应用中的落地，源于其基础架构与能力体系的演进。早期依赖规则和小规模神经网络的系统，因数据和算力受限而难以满足复杂场景需求。2017年，Transformer将自注意力机制引入序列建模，以并行化和长距离依赖建模能力显著提升了语言理解效果[6]。当前，代表性系统如GPT、Gemini、LLaMA等，依托数十亿至万亿级参数及网页、书籍和代码等跨域语料，实现了语义理解、逻辑推理与跨模态感知，为网络流量预测、智能调度和服务感知等场景提供了坚实架构。

为了高效释放大模型能力并适配网络环境，还需配套完善的训练与优化体系。当前主流做法是先进行自监督预训练以获取通用语义表征，再通过有监督微调增强特定任务性能。为降低大模型在网络部署中的资源消耗，提出了混合精

度训练、模型并行与数据并行等提升算力利用率的方法，以及LoRA、Adapter等参数高效微调技术，显著减少了显存占用和边缘部署成本[7]。此外，诸如对抗训练、指令微调与人类反馈强化学习等技术被用于提升模型在动态网络环境中的稳定性与安全性，确保其在实际运营中保持高可用性和鲁棒性。

1.2 大模型赋能网络的应用

大模型与网络系统的融合正在加速推动网络从数据驱动的“感知型”体系向知识驱动的“认知型”体系转变。在传统网络中，策略配置主要依赖静态规则和有限模型，难以应对动态环境与复杂需求。大模型通过上下文理解与语义建模能力，使网络能够动态感知用户需求、环境变化与服务状态，支撑实时优化与智能决策[8]。代表性通用大模型和网络大模型的发展历程如图1所示。

在实际应用中，大模型已广泛辅助网络完成智能规划建设、智能维护、智能优化与智能网络运营，显著提升了网络的自动化与智能化水平。其上下文建模与语义理解能力，尤其适用于移动通信、边缘计算与物联网（IoT）等高动态环境，能够实现多源异构信息的协同融合与联合优化[9]。同时，大模型在自然语言理解与指令生成方面展现出显著优势，使得用户可通过对话式交互精准表达需求，推动网络控制从传统程序式调用向语义驱动配置转型。随着大模型应用逐步渗透至网络管理核心环节，模型上下文协议（MCP）[10]与智能体间通信协议（A2A）[11]等新兴协议机制也被用于增强上下文共享与多智能体协同推理能力，进一步推动网络智能化水平的提升[12]。当前基于大模型的智能网络管理框架及其应用如图2所示。

1.2.1 智能规划建设

大模型在智能网络布局规划建设的多个核心环节中展现出实际价值，尤其在资源规划、智能选址与自动化验收方面实现了流程的优化与效率提升[13]。

在资源规划方面，运营商已借助大模型分析历史流量日志、用户活跃区域与业务类型，生成动态资源配置策略。通过时间序列建模与场景推理，模型能精准预测不同区域的流

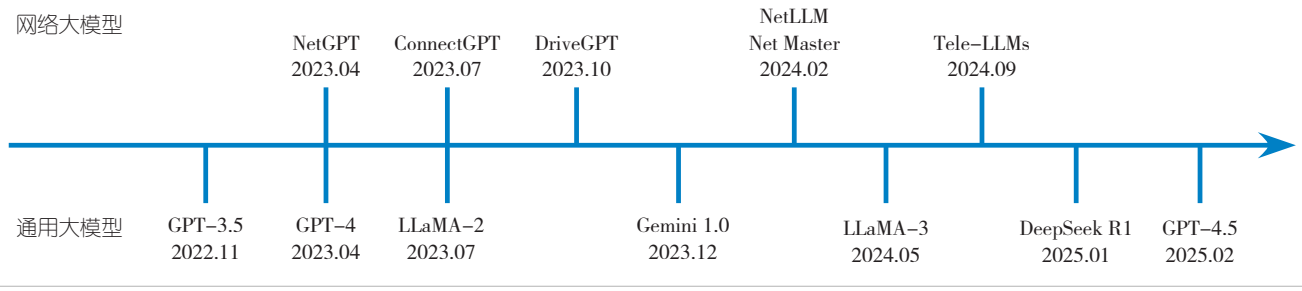


图1　代表性通用大模型和网络大模型的发展历程

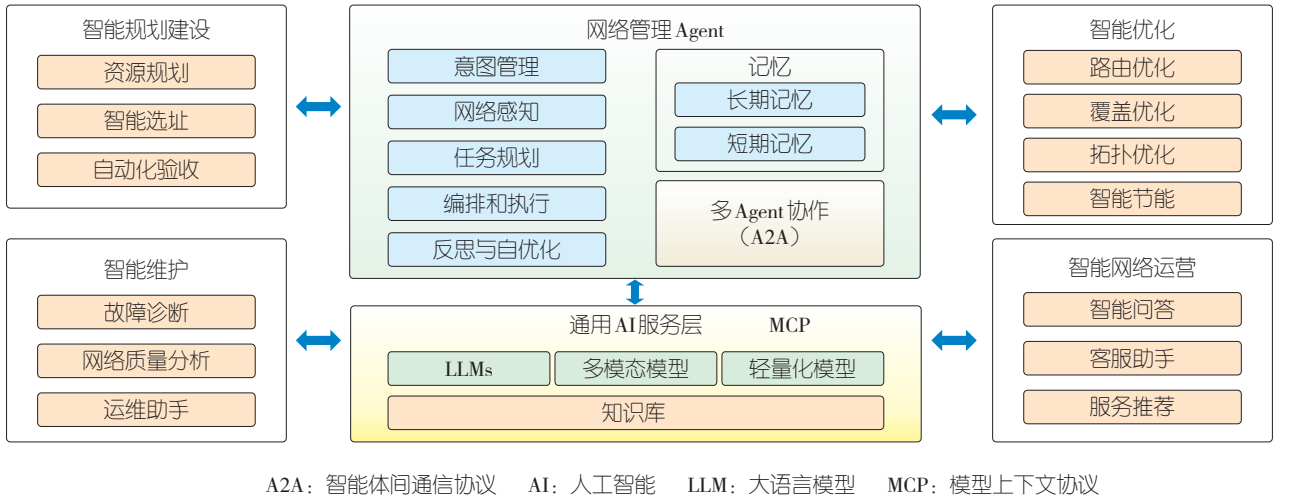


图2　基于大模型的智能网络管理框架及其应用

量峰值，辅助优化基站部署与带宽分配，避免了传统静态规划中常见的资源冗余与服务盲区问题。

在智能选址方面，大模型通过融合地理信息、建筑结构、人口密度、业务需求及部署成本等多源异构数据，自动推荐站点部署方案，减少人工勘测与方案比选的工作量[14]。当前，正在逐步为城区高密集区与乡村低覆盖区提供差异化选址支持，以提升网络部署的成本效益与服务弹性。

自动化验收环节同样得到了大模型的重塑。模型通过自然语言接口解析网络验收标准，并对设备配置、接口参数、运行状态进行自动比对和校验，生成结构化验收报告。相比传统人工检查流程，该方法显著降低了配置错误率，缩短了验收周期，提升了系统上线的一致性与规范性。

1.2.2 智能维护

在网络智能化转型中，大模型已广泛应用于故障诊断、网络质量分析和运维助手等维护核心场景，推动运维流程向自动化、精准化方向演进[15]。

在故障诊断方面，运营商通过接入大模型，对多源告警日志、配置项和历史工单进行语义解析与模式匹配，实现分钟级的故障定位与根因分析。在云网环境中，大模型已被用于对链路震荡、中央处理器（CPU）异常等问题进行快速溯源，并生成修复建议，替代了依赖人工检索的低效流程。

在网络质量分析上，大模型被部署于实时性能监控与异常趋势识别。通过分析丢包率、时延、吞吐等关键指标，模型能够识别视频业务卡顿、物联网（IoT）节点失联等服务劣化现象，并提前发出风险预警。目前已有运营商借助大模型部署了基于性能预测的预防性维护机制，将部分告警量压缩超过50%，显著减轻了一线运维负担[16]。

此外，运维助手类大模型被集成至网络管理平台，服务于日常运维场景。工程师可通过自然语言交互调用配置模板、生成脚本或获取操作指引，简化复杂命令的记忆与执行过程[17]。在网络割接、参数变更等高风险场景中，大模型还可辅助制定操作步骤并生成回滚预案，显著降低人为错误率[18]。同时，通过与工单系统联动，系统可实现从故障感知、策略生成到执行验证的自动闭环处理，加速故障闭环时效。

1.2.3 智能优化

在网络性能优化领域，大模型通过多维度智能决策显著提升网络效率与可靠性，其核心应用涵盖路由优化、覆盖优化、拓扑优化与智能节能等场景[19]。

在路由优化方面，大模型基于实时流量模式与网络拓扑状态，动态调整数据传输路径以降低时延并提升可靠性。在SDN中，模型通过融合流量矩阵与拓扑信息，自主优化路径选择策略，有效缓解网络拥塞并增强吞吐能力。结合强化学习算法，模型可实时感知流量波动并自适应调整路由规则，实现网络性能的动态均衡[20]。

在覆盖优化方面，大模型被应用于基站参数自适应配置与边缘信号质量提升。通过综合分析基站负载、用户密度分布与环境因素，模型自动生成功率调整与天线优化方案[21]。针对移动性强的车联网应用场景，模型基于车对万物通信数据动态调整发射功率，保障关键区域的连续覆盖与服务稳定性。

拓扑优化致力于适应动态业务需求，大模型通过分析节点负载与链路状态，自主优化网络节点布局与资源分配策略[22]。在移动边缘计算场景中，模型基于业务需求预测动态重构拓扑连接，缩短了服务响应时间。在卫星–地面融合网络中，通过轨道与链路状态预测实现全球覆盖的连续性保障。

在智能节能方面，大模型被部署用于基站能耗建模与动态休眠调度。结合流量预测与设备运行监测，模型自动制定节能策略，在保障服务质量的前提下降低网络整体能耗。

1.2.4 智能网络运营

在网络运营层面，大模型的应用已成为连接用户与网络的智能中枢，广泛应用于智能问答、客服助手与服务推荐中，推动了智能化的客户服务与运营管理[23]。

智能问答系统利用自然语言理解技术，快速响应用户关于网络业务和故障排查的咨询。当用户用日常语言描述问题时，系统结合知识库和实时网络数据，提供精准解决方案，自动生成诊断报告和修复建议。

客服助手集成多轮对话引擎与知识图谱，自动处理用户投诉和工单。系统能自动解析投诉并关联网络告警，动态分配优先级，从而提升工单处理效率。开放问答系统的应用使得运营商能够支持多语种服务，减少人工干预并提升服务的全球覆盖性。用户投诉自动分类通过语义分析与模式识别技术，将非结构化投诉准确映射到具体网络问题类别。结合图神经网络和大模型框架，系统精准区分“覆盖问题”“速率问题”等，提升分类准确性。模型还能通过历史投诉数据预测高频故障区域，提前部署维护资源，从而实现主动预防。

在客户留存与服务推荐方面，大模型通过分析用户行为特征、服务使用模式及历史交互记录，构建动态用户画像，基于此生成定制化套餐优化建议与增值服务推荐策略，提升用户满意度与长期留存率。模型进一步结合网络资源供给状态与用户需求趋势，确保推荐方案可行性[24]。

2 大模型赋能网络应用的技术挑战与解决方案

随着大模型在网络中的广泛应用，其模型复杂性与系统融合深度不断提升，也带来多项关键性技术挑战[25]。针对这些挑战，需进一步研究优化相应解决方案[26]。当前大模型赋能智能网络面临的技术挑战与解决方案如图3所示。

2.1 资源调度场景的解空间组合爆炸与NP难

在智能网络管理任务中，虚拟网络功能部署（VNF）、服务功能链设计（SFC）等问题因涉及带宽、计算、存储等多资源联合优化及拓扑约束，已被证明属于NP难（NP-hard）问题[27]。随着网络向分布式云和多接入边缘计算（MEC）扩展，物理节点数量和链路复杂度呈指数级增长，使得传统的穷举搜索和经验式启发式算法在涉及千万级节点的动态编排场景中，难以兼顾毫秒级时延敏感型业务的实时调度，与跨域负载均衡的质量保障，其解空间维度已远超多项式时间求解能力的边界。

针对上述挑战，分布式智能与机器学习深度融合的架构被提出。一方面，云边协同的异构计算架构将整体优化任务分解为多层次子问题。云端负责集中式训练以生成全局策略，边缘节点则应用轻量化模型进行实时推理，从而显著减轻单节点搜索全量解空间的压力[28]。另一方面，混合整数线性规划与启发式松弛技术相结合的方法，在服务功能链嵌入等应用中展现显著优势[29]。通过数学规划与规则引擎的协同，将计算耗时控制在业务可接受范围内，同时保证了解的次优质量。此外，将深度强化学习（DRL）引入动态网络切片部署的研究也取得了突破：DRL智能体通过与网络环境的持续交互，自主适应流量波动和节点故障等不确定性因素，在非稳态条件下展现出较强的策略泛化能力。

现有DRL方案多基于静态环境假设，面对真实网络中的突发流量峰值和拓扑剧变时，仍会遭遇策略震荡和收敛速度下降的问题[30]。因此，如何通过分层强化学习架构实现离线策略预训练与在线微调的深度耦合，以及借助联邦学习机制在多域环境中同步更新模型，平衡解空间探索效率与策略稳定性，是突破NP-hard问题传统求解范式的关键课题。

2.2 网络状态的多维度不确定性

网络系统在实际运行过程中常面临多维度不确定性因素，主要包括用户行为剧烈波动、链路状态不稳定变化以及外部策略的频繁调整[31]。这些因素导致网络状态难以精确感知，策略决策难以稳定执行，影响服务质量保障和资源调度效率。在动态场景下，如移动边缘计算或多租户环境中，不确定性可能引发预测偏差、资源冲突和策略漂移，使得传统依赖静态配置的方案难以满足实时性和可靠性要求。

为应对上述挑战，智能系统需具备动态感知与自适应调整能力。一种有效路径是构建基于概率图模型与在线学习机制的策略框架，通过建模网络状态与环境变量的依赖关系，实时更新参数以修正策略偏差。实践表明，该方法可有效降低流量突变和策略冲突引发的服务中断风险。在此基础上，结合时序建模与隐空间表示技术可进一步增强系统的异常识别能力。长短期记忆网络用于提取流量变化规律，变分自编码器则通过重构机制捕捉潜在异常特征，两者协同可将异常检测和响应延迟控制在毫秒级范围内[32]。同时，为解决跨域数据孤岛与隐私保护问题，联邦学习通过分布式训练方式实现多域信息的安全聚合，在保障数据合规的同时提升预测准

图3 当前大模型赋能智能网络应用的技术挑战与解决方案

确性与系统稳健性。上述技术协同构建了从“动态感知–特征提取–协同决策”的闭环优化链路，为应对多维不确定性提供了系统性解决方案。

2.3 实时场景的时延约束

在工业控制、自动驾驶等对时效性要求极高的应用场景中，网络管理任务需在毫秒级甚至亚毫秒级内完成感知、决策与响应，任何延迟都可能导致系统异常或安全风险。以SDN为例，当网络拓扑发生动态变化时，控制器必须在极短时间内完成规则更新与故障恢复，确保数据流连贯性。传统集中式架构依赖全局同步，受限于信息传递延迟和中央处理瓶颈，难以满足极致实时性的需求[33]。分布式架构虽具备一定的响应优势，但局部感知信息的不完备又容易引发决策偏差，造成“低延迟高误差”或“高精度慢响应”的两难局面。

为有效应对上述挑战，边缘推理与分层协同决策成为关键策略。高实时性任务（如故障检测、负载异常响应）被下沉至边缘节点，借助轻量级神经网络模型（Tiny–YOLO、MobileNet）进行本地快速推理[34]。而复杂的全局优化决策（如多域资源调度）则由云端集中处理，形成云–边–端分层协作体系。同时通过知识蒸馏等技术，将云端大模型压缩迁移至边缘节点，实现不同计算层级间的策略同步与模型适配，保障全局一致性与推理高效性。

此外，为适配边缘设备资源受限的特点，自适应精度调整与渐进式推理机制被引入，通过优先输出近似结果并逐步迭代精化，进一步降低决策延迟[35]。在实际应用中，这一分层推理与动态协同框架显著降低了关键路径延迟，支撑了工业网络、车联网等领域对高可靠、低时延智能控制的需求。

2.4 异构数据的融合难问题

在智能网络管理过程中，系统需依赖大规模、多源数据支持决策制定。这些数据涵盖设备日志、性能指标、流量报文等多个维度，存在显著的结构、粒度与更新频率差异，形成高度异构的信息环境[36]。同时，由于设备故障、链路中断等因素，部分数据存在缺失、损坏或延迟上报现象，进一步削弱了数据完整性与时效性。尤其在5G及未来网络架构中，实时流量激增带来的高数据速度，加剧了数据处理的复杂性，对系统的吞吐能力与数据质量保障提出更高要求。传统集中式数据处理方案受限于全局同步延迟与边缘节点资源受限，难以在保证实时响应的同时兼顾特征提取精度，易导致模型训练偏差和推理决策滞后。

为应对上述挑战，可通过分层数据处理与自适应特征抽象的协同机制加以优化。一方面，可在边缘节点引入轻量化预处理机制，对原始数据进行采样、冗余字段过滤及时间窗聚合，显著降低上行数据量与通信开销，缓解边缘计算资源压力。云端则集中处理高价值异构数据，通过跨域日志关联、异常模式挖掘等方式提升分析深度与准确率。另一方面，利用机器学习模型对多源数据特征进行动态抽象与自适应建模，可根据实时环境变化调整规则阈值，避免传统静态规则引擎在动态场景下失效的问题[37]。此外，基于Transformer的统一编码架构可有效整合异构输入，结合掩码自动编码器对缺失特征进行恢复，提高数据一致性与特征完整性[38]。为进一步提升实时处理能力，可在数据采集层引入滑动窗口机制与流式标准化处理技术，稳定特征分布，保障模型输入质量。

2.5 运维管理下的人机协同障碍

在智能网络管理体系中，实现自动化流程与人工操作的高效融合，是突破大规模智能部署瓶颈的关键。虽然大模型在故障诊断、资源调度、性能优化等任务中展现出卓越的自动推理与决策能力，但在复杂场景下，如策略冲突仲裁、未知故障应对等，人类专家仍不可或缺。当前人机融合机制主要面临3个方面的挑战：一是语义鸿沟，即人工指令与机器执行语义之间存在映射偏差，易导致意图理解失真；二是信任壁垒，由于AI模型多为黑盒结构，专家难以快速理解决策依据，降低了系统可控性与可信度；三是应急断层，一旦AI系统推理失效，人工接管往往因上下文同步不足而响应滞后，易导致故障进一步扩散。

为解决上述问题，当前研究提出了双向可解释、动态可干预的人机协同框架。一方面，采用SHAP（SHapley Additive exPlanations）、反事实推理等可解释性技术，自动生成决策因果链，并通过可视化方式直观展现模型推理过程，帮助专家快速理解系统行为[39]。另一方面，通过模仿学习采集专家操作序列，并结合知识图谱进行领域知识建模，将人类经验转化为模型可用的策略约束，提升系统的策略合规性与决策稳健性[40]。此外，为提升人机协同的韧性，设计基于置信度的分级接管机制：当模型输出的置信水平低于预设阈值或推理后果超出安全边界时，系统自动推送决策上下文摘要，触发人工干预流程。该机制不仅缩短了故障响应时间，也降低了人工介入频率，在大规模网络运维与智能控制系统中初步验证了其实用性与有效性。

2.6 边缘部署的成本效益平衡

在边缘计算、IoT终端等资源受限的环境中，部署大模

型面临算力–精度–时延的“不可能三角”约束。由于大模型通常具备极高的参数规模和复杂的计算图，其推理延迟和内存占用常常超出边缘设备的硬件承载能力，导致实时性能下降、功耗飙升，甚至系统失效。此外，跨多节点协同推理时，频繁的模型同步操作带来了巨大的通信带宽压力和能耗开销，进一步制约了智能网络在大规模场景下的可扩展性[41]。

为实现部署成本与性能效益的平衡，当前主要采用模型轻量化、分布式推理与硬件–软件协同优化等多维技术路径。模型轻量化方面，剪枝、量化、知识蒸馏等方法被广泛应用，通过移除冗余参数、降低计算精度或转移知识表征，在尽可能保留准确率的前提下，压缩模型体积与推理复杂度[42]。分布式推理系统通过动态负载感知机制，合理划分推理子任务，使边缘节点根据实时资源状况自适应选择推理路径，减少无效计算与通信冗余，从而提升系统整体吞吐率[43]。硬件–软件协同优化则依托定制化加速器与高效推理引擎，通过算子融合、内存共享等手段进一步压缩延迟与功耗，确保智能网络在低资源环境下仍具备稳定可靠的运行能力。这些技术体系的协同应用，已在工业边缘、智能交通、智能制造等场景中展现出显著的成本效益提升潜力。

3 结束语

大模型作为新一代智能引擎，已在网络领域的多个核心环节实现初步落地。其在网络规划、运维管理、性能优化及用户服务中的应用，展现出良好的可扩展性与智能化潜力。随着模型训练技术与算力平台的持续演进，大模型将进一步推动网络从规则驱动向知识驱动转型。

在未来网络建设进程中，大模型赋能的智能网络将逐步成为主流技术路径之一。通过融合多模态数据、在线学习、人机协同等技术手段，网络运营将朝着更自动化、更个性化的方向发展。随着大模型在通信行业生态中的不断融合与优化，其产业应用前景将更加广阔，将为运营商和终端用户带来更高效、便捷的服务体验。

参考文献

[1] WU J, FANG X. Collaborative optimization of wireless communication and computing resource allocation based on multi-agent federated weighting deep reinforcement learning [EB/OL]. (2024-04-02)[2025-08-10]. https://arxiv.org/abs/2404.01638

[2] TANG F X, MAO B M, KATO N, et al. Comprehensive survey on machine learning in vehicular network: technology, applications and challenges [J]. IEEE communications surveys & tutorials, 2021, 23 (3): 2027-2057. DOI: 10.1109/COMST.2021.3089688

[3] 崔勇, 张蕾, 马川. 面向多目标的一体化融合网络体系结构 [J]. 电子学报, 2023, 51(9): 2277-2288

[4] ZHOU H, HU C M, YUAN Y, et al. Large language model (LLM) for telecommunications: a comprehensive survey on principles, key techniques, and opportunities [J]. IEEE communications surveys & tutorials, 2025, 27(3): 1955-2005. DOI: 10.1109/COMST.2024.3465447

[5] CHEN Y, LI R, ZHAO Z, et al. NetGPT: a native-AI network architecture beyond provisioning personalized generative services [EB/OL]. (2023-07-12) [2025-08-10]. https://arxiv. org/abs/2307.06148

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [EB/OL]. [2025-08-10]. https://proceedings. neurips. cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[7] WU D, WANG X D, QIAO Y Q, et al. NetLLM: adapting large language models for networking [C]//Proceedings of the ACM SIGCOMM 2024 Conference. ACM, 2024: 661-678. DOI: 10.1145/3651890.3672268

[8] LEE W, PARK J. LLM-empowered resource allocation in wireless communications systems [EB/OL]. (2024-08-06) [2025-08-10]. https://arxiv.org/abs/2408.02944

[9] ZHOU H, HU C, YUAN D, et al. Large language model (LLM) -enabled in-context learning for wireless network optimization: a case study of power control [EB/OL]. (2024-08-01) [2025-08-10]. https://arxiv.org/abs/2408.00214

[10] VENTUREBEAT. Anthropic releases model context protocol to standardize AI-data integration [EB/OL]. [2025-08-10]. https://venturebeat. com/data-infrastructure/anthropic-releases-model-context-protocol-to-standardize-ai-data-integration/

[11] Google. Announcing the agent2agent protocol (A2A) [EB/OL]. [2024-05-15]. https://developers. googleblog. com/en/a2a-a-new-era-of-agent-interoperability/

[12] ZHAO X, WANG M, CECCARELLI D, et al. AI based network management agent (NMA): concepts and architecture [EB/OL]. [2025-04-16]. https://datatracker. ietf. org/doc/draft-zhao-nmop-network-management-agent/

[13] BOATENG G O, SAMI H, ALAGHA A, et al. A survey on large language models for communication, network, and service management: application insights, challenges, and future directions [EB/OL]. (2024-12-16) [2025-08-10]. https://arxiv. org/abs/2412.19823

[14] LI Z, XU J, WANG S, et al. StreetviewLLM: extracting geographic information using a chain-of-thought multimodal large language model [EB/OL]. (2024-11-19) [2025-08-10]. https://arxiv. org/abs/2411.14476

[15] YAO K, CHEN D, JEONG J, et al. Use cases and practices for intent-based networking [EB/OL]. [2025-04-16]. https://datatracker.ietf.org/doc/draft-irtf-nmrg-ibn-usecases/.

[16] YU Z Y, MA M H, ZHANG C Y, et al. MonitorAssistant: simplifying cloud service monitoring via large language models [C]// Proceedings of Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. ACM, 2024: 38-49. DOI: 10.1145/3663529.3663826

[17] LIN L, JIN Y, ZHOU Y, et al. MAO: a framework for process model generation with multi-agent orchestration [EB/OL]. (2024-08-04) [2025-08-10]. https://arxiv.org/abs/2408.01916

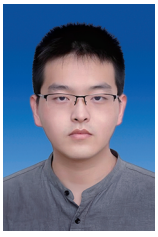
[18] GOEL D, HUSAIN F, SINGH A, et al. X-lifecycle learning for cloud incident management using LLMs [C]//Proceedings of Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering. ACM, 2024: 417-428. DOI: 10.1145/3663529.3663861

[19] HUANG Y D, DU H Y, ZHANG X Y, et al. Large language models for networking: applications, enabling techniques, and challenges [J]. IEEE network, 2025, 39(1): 235-242. DOI: 10.1109/MNET.2024.3435752

[20] SONG Y, QIAN X S, ZHANG N, et al. QoS routing optimization based on deep reinforcement learning in SDN [J]. Computers, materials & continua, 2024, 79(2): 3007–3021. DOI: 10.32604/cmc.2024.051217
[21] QUAN H Y, NI W L, ZHANG T, et al. Large language model agents for radio map generation and wireless network planning [J]. IEEE networking letters, 2025, 7(3): 1. DOI: 10.1109/LNET.2025.3539829
[22] YE M, HUANG L Q, WANG X L, et al. A new intelligent cross–domain routing method in SDN based on a proposed multiagent reinforcement learning algorithm [J]. International journal of intelligent computing and cybernetics, 2024, 17(2): 330–362. DOI: 10.1108/ijicc–09–2023–0269
[23] XU Z T, CRUZ M J, GUEVARA M, et al. Retrieval–augmented generation with knowledge graphs for customer service question answering [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2024: 2905–2909. DOI: 10.1145/3626772.3661370
[24] LU H, CHAI Z, ZHENG Y, et al. Large memory network for recommendation [EB/OL]. (2025–02–08) [2025–08–10]. https://arxiv.org/abs/2502.05558
[25] Datatracker. Research challenges in coupling artificial intelligence and network management [EB/OL]. [2025–04–16]. https://datatracker.ietf.org/doc/draft–irtf–nmrg–ai–challenges/
[26] Datatracker. Considerations of network/system for AI services [EB/OL]. [2025–04–16]. https://datatracker. ietf. org/doc/draft–hong–nmrg–ai–deploy/
[27] ATTAOUI W, SABIR E, ELBIAZE H, et al. VNF and CNF placement in 5G: recent advances and future trends [J]. IEEE transactions on network and service management, 2023, 20(4): 4698–4733
[28] GOLKARIFARD M, CHIASSERINI C F, MALANDRINO F, et al. Dynamic VNF placement, resource allocation and traffic routing in 5G [J]. Computer networks, 2021, 188: 107830. DOI: 10.1016/j.comnet.2021.107830
[29] BEHRAVESH R, HARUTYUNYAN D, CORONADO E, et al. Time–sensitive mobile user association and SFC placement in MEC–enabled 5G networks [J]. IEEE transactions on network and service management, 2021, 18(3): 3006–3020. DOI: 10.1109/TNSM.2021.3078814
[30] YAN Z X, GE J G, WU Y L, et al. Automatic virtual network embedding: a deep reinforcement learning approach with graph convolutional networks [J]. IEEE journal on selected areas in communications, 2020, 38(6): 1040–1057. DOI: 10.1109/JSAC.2020.2986662
[31] Datatracker. Artificial intelligence framework for network management [EB/OL]. [2025–04–16]. https://datatracker. ietf. org/doc/draft–pedro–nmrg–ai–framework/
[32] GAO Z, ZHANG Z, ZHANG Y, et al. Online client scheduling and resource allocation for efficient federated edge learning [EB/OL]. (2024–09–29)[2025–08–10]. https://arxiv.org/abs/2410.10833
[33] LÓPEZ J, LABONNE M, POLETTI C, et al. Priority flow admission and routing in SDN: exact and heuristic approaches [C]// Proceedings of IEEE 19th International Symposium on Network Computing and Applications (NCA). IEEE, 2020: 1–10. DOI: 10.1109/NCA51143.2020.9306725
[34] HENG L, YIN G F, ZHAO X F. Energy aware cloud–edge service placement approaches in the Internet of Things communications [J]. International journal of communication systems, 2022, 35: e4899. DOI: 10.1002/dac.4899
[35] ZHAN H, ZHANG X, TAN H, et al. PICE: a semantic–driven progressive inference system for LLM serving in cloud–edge networks [EB/OL]. (2025–01–16) [2025–08–10]. https://arxiv. org/abs/2501.09367
[36] SUN P, ZHAO B, LI X. Research on multi–source heterogeneous data fusion method of substation based on cloud edge collaboration and AI technology [J]. Discover applied sciences, 2025, 7(4): 262. DOI: 10.1007/s42452–025–06725–8
[37] JIN B W, ZHANG Y, ZHU Q, et al. Heterformer: transformer–based deep node representation learning on heterogeneous text–rich networks [C]//Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. ACM, 2023: 1020–1031. DOI: 10.1145/3580305.3599376
[38] JEONG J, KU T Y, PARK W K. Denoising masked autoencoder–based missing imputation within constrained environments for electric load data [J]. Energies, 2023, 16(24): 7933. DOI: 10.3390/en16247933
[39] GILPIN L H, BAU D, YUAN B Z, et al. Explaining explanations: an overview of interpretability of machine learning [C]//Proceedings of IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018: 80–89. DOI: 10.1109/DSAA.2018.00018
[40] CHRISTIANO P, LEIKE J, Brown T, et al. Deep reinforcement learning from human preferences [EB/OL]. (2017–06–12) [2025–08–10]. https://arxiv.org/abs/1706.03741
[41] SUN Q, YIN Z, LI X, et al. Corex: pushing the boundaries of complex reasoning through multi–model collaboration [EB/OL]. (2023–09–30)[2025–08–10]. https://arxiv.org/abs/2310.00280
[42] MIRZADEH S I, FARAJTABAR M, LI A, et al. Improved knowledge distillation via teacher assistant [J]. Proceedings of the AAAI conference on artificial intelligence, 2020, 34(4): 5191–5198. DOI: 10.1609/aaai.v34i04.5963
[43] ZHOU H, EROL–KANTARCI M, POOR H V. Knowledge transfer and reuse: a case study of ai–enabled resource management in RAN slicing [J]. IEEE wireless communications, 2023, 30(5): 160–169. DOI: 10.1109/MWC.004.2200025

作 者 简 介

牛嘉林，中关村实验室与北京邮电大学联培在读博士研究生；主要研究方向为大语言模型在网络中的应用、基于知识图谱的安全威胁建模与推理。



邢铭哲，中关村实验室助理研究员；主要研究方向为AI智能体和AI赋能网络；发表论文10余篇。

张蕾，中关村实验室副研究员；主要研究方向为网络安全和网络系统；发表论文10余篇。