大模型与网络智能化探讨

Discussion on Large Models and Network Intelligence

李　露，李福昌（中国联通研究院，北京 100048）
Li Lu，Li Fuchang（China Unicom Research Institute，Beijing 100048，China）

摘　要：
深入探讨了大模型在网络智能化中的关键作用与应用模式。大模型凭借其强大的学习能力和泛化性能，能够深度挖掘网络数据中的潜在价值，为网络运维、管理以及性能优化等多方面提供智能决策依据。同时，在架构层面提出了端边云智能协同架构，研究了相关架构特点，为网络智能化演进及支持更多大模型等新AI业务应用提供了参考。

关键词：
大模型；人工智能；端边云协同
doi：10.12045/j.issn.1007-3043.2025.01.001
文章编号：1007-3043（2025）01-0001-05
中图分类号：TN929.5
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
It deeply explores the key role and application modes of large models in network intelligence. Large models, with their powerful learning ability and generalization performance, can deeply mine the latent value in network data, providing intelligent decision-making support for network operation, management, and performance optimization in multiple aspects. Furthermore, it proposes an end-edge-cloud intelligent collaborative architecture at the architecture level, studies the relevant architecture features, and provides reference for the evolution of network intelligence to support new AI services.

Keywords：
Large models；Artificial intelligence；End-edge-cloud collaboration

引用格式：李露，李福昌. 大模型与网络智能化探讨[J]. 邮电设计技术，2025（1）：1-5.

0 引言

大数据、人工智能（artificial intelligence，AI）大模型等IT技术的发展催生了行业领域丰富的新场景和新用例，感知、计算、智能将是5G-A及6G新系统的重要技术组成部分。2023年，ITU发布的《IMT面向2030及未来发展的框架和总体目标建议书》中已明确将通信智能融合写入6G六大应用场景中[1-2]。用户数据的激增、硬件设备的进步以及人工智能模型向大模型的演进引发生成式人工智能（Artificial Intelligence Generated Content，AIGC）技术的浪潮[3]。生成式人工智能（AIGC）及其相关应用成为研究的焦点[4]。国内外领先的科技巨头，如百度、阿里、腾讯、头条、微软和谷歌等，纷纷投入大量资金研发专属GAI模型，旨在为用户提供更全面的数字服务。

大模型的发展也冲击着通信行业及通信网络，2024年，三大运营商纷纷发布自研大模型，中国联通成立AI创新中心，并发布“元景”大模型；中国移动发布“九天”大模型；中国电信发布“星辰”大模型。通信网络为移动互联网产业的发展提供了土壤，ChatGPT

收稿日期：2024-12-16

大模型每半年升级一代，全产业正在加速进入多模态AI，移动AI业务也破土而出。芯模适配加速了移动AI业务的终端、边缘、云端（下称“端边云”）协同，手机/PC全面融合AI，更多AI原生终端也相继出现，高通骁龙8 Gen 3移动平台、联发科旗舰移动处理器天玑9300、苹果A17 Pro芯片已集成智算引擎，支撑GAI端侧模型。苹果已全面开展“AI+”策略，加速智能升级[5]。

在移动AI时代，端边云AI融合对无线网络提出更高要求。移动通信进入AI时代将释放体验红利，经营模式从流量经营走向体验经营。网络需提升性能，从而满足多模态AI交互需求（如满足高低质量时延的AI连接，更高要求的空口时延及上行速率）。同时，AI也为网络注智赋能，基于AI的灵活调度的网络能力，能够使网络效能达到最大化，实现维优无人化或少人化[6]。

本文将从大模型智能化业务需求、大模型在网络中的应用及端边云智能协同架构3个方面探讨通信网络与大模型等AI结合的相关技术及趋势，为网络智能化的演进提供参考。

1 大模型发展引发网络新需求

生成式人工智能正从文字向多模态AI转变，极大改变了交互方式，多模态AI能实现“类真人级”助手全天候实时交互，且这种实时交互已成为行业业务增长驱动力。AI应用正从生成式向执行类AI Agent转变。AI产业正依托大语言模型、多模态大模型，实现从生成式向执行类AI Agent迈进，能完成问答和文图处理等功能。执行类AI Agent可承担长周期且复杂的工作，能直接对现实世界产生实际影响，像订票、回复邮件等具体事务均可处理。AI Agent要达到近真人级的体验，上行需要20~60 Mbit/s速率，同时对AI Agent处理时延也有要求，空口时延需小于50 ms左右。AI Agent的应用需随处可接入的无线网络基础设施（具备高速率、低时延网络特点）提供支撑[7-8]。

多模态AI交互及AI Agent复杂任务执行业务的发展是上行流量增长的主要驱动力。AI产业在功能应用、性能要求以及对流量影响等方面呈现出特定的发展态势，AI聊天助手、AI工业智造、AI自动驾驶车联流量等都展现出较大的发展潜力，各领域都有着不同的流量特点与需求，这些都将影响后续通信产业的布局与技术优化方向[9]。多模型AI业务如图1所示。

图1　多模型AI业务

在当前人工智能的发展进程中，结合场景的多模式并存以及端边云协同已成为明确的发展方向。不同的应用场景有着不同的需求，单一的网络模式很难满足多样化、复杂化的实际情况。端边云协同是提升网络性能、满足不同AI新业务的有效技术方案。通过终端、边缘、云端协同工作，依据具体的场景以及相应的时间节点，合理地分配AI计算的工作负载。例如，在一些对实时性要求较高、数据处理量相对较小的本地场景中，终端可以承担主要的计算任务，快速给出反馈，避免因数据传输等造成的时延问题；而对于那些需要大量计算资源、复杂算法处理的数据密集型任务，则可以将其分配到云端，借助云端强大的算力来高效完成，之后再将结果回传至终端呈现给用户。这样的协同模式一方面可以为用户提供更优质的使用体验，让交互更加流畅、结果更加精准；另一方面，也能实现资源的高效利用，避免终端或云端某一方的资源闲置或被过度使用[10-11]。

2 大模型助力网络智能化

网络智能化已经被探索和实践了很多年，未来会走向高阶智能。未来网络智能化也将从SON、小模型阶段逐步走向大模型阶段，最终发展到网络网元内生智能，每个网元变成智能体。大模型具有强大学习能力、任务分解能力、通用能力及生成能力等，在通信网络的故障诊断、网络优化、智能运维、营销客服等方面具有高度适配性。

a）网络故障诊断。将实时采集的网络数据输入到训练好的大模型中，模型根据学习到的模式和规律对当前网络状态进行评估和判断。当检测到异常数据模式时，模型发出故障预警信号，并初步判断故障类型和可能的故障位置。例如，若模型发现某条链路的流量急剧下降且持续一段时间，同时伴随该链路两端设备端口状态异常的信息，则可能判断为链路故障。大模型能够综合多方面的数据信息进行故障检测，避免了传统单一指标监测方法的局限性。例如，

通过关联分析设备日志、流量数据和性能指标，能够更准确地识别出因软件漏洞导致的设备性能下降或网络拥塞等复杂故障情况。在检测到故障后，大模型可以进一步深入分析故障相关数据，利用其对网络数据的深度理解和推理能力，确定故障的根本原因。它可以通过追溯网络事件的时序关系、分析相关设备的配置信息和运行状态变化，找出引发故障的关键因素[12]。

b）网络性能优化。大模型可以分析网络设备的配置参数与网络性能之间的关系，能够为网络配置优化提供建议。基于大模型可以评估不同配置组合对网络延迟、带宽利用率、可靠性等性能指标的影响，通过模拟和优化算法找到最佳的网络配置方案。在网络环境发生变化（如新增设备、网络拓扑结构调整等）或业务需求变更时，大模型能够快速重新评估网络配置需求，并提供相应的优化策略。这有助于减少人工配置的工作量，降低错误率，提高网络配置的灵活性和适应性。在网络升级或扩容决策中，大模型可以评估不同升级方案对网络性能、成本和业务影响的预测结果，帮助运维人员选择最适合的方案。例如，分析增加网络带宽、升级网络设备或采用新的网络技术等不同方案的优劣，综合考虑投资回报率、业务发展需求等因素，为决策提供科学依据[13]。

c）网络智能运维。当面临网络运维问题或决策场景时，大模型可以充当专家角色，根据其对网络数据的全面分析和理解，将运维工作分解给运维人员，并为运维人员提供智能的运维建议和专业操作指导，提升运维效率，节约人工成本。另一方面，大模型的分析结果和运维建议以直观的可视化方式呈现给运维人员，便于他们快速理解网络状态和问题所在，使运维人员能够全面了解网络运维情况，及时采取相应的措施[14]。

d）营销及客服。在业务营销方面，大模型有助于推动网络个性化服务，根据用户的行为、偏好等数据，大模型能够生成个性化的业务套餐推荐和定制方案，提升用户体验和业务转化率。借助大模型强大的数据分析能力，运营商可以整合客户的基本信息、消费记录、咨询历史、投诉记录等多维度数据，构建更加全面、细致的客户画像。从而深入了解客户的需求、偏好、行为习惯等，为个性化的服务和精准营销提供依据。在客服助手方面，对于一些常见的客户咨询问题，大模型可以直接给出准确的答案，无需人工客服介入。这不仅能够快速响应客户的需求，还可以释放人工客服的时间和精力，使其能够专注于处理更复杂、更重要的问题。大模型可以自动根据客户的问题和需求生成相应的工单，并根据问题的类型、紧急程度等因素，智能地分配给最合适的处理人员或部门。这不仅提高了工单处理的效率，还减少了人工操作可能带来的错误和延误[15]。

3 端边云智能协同架构

端边云协同智能化架构是一种将终端设备（如手机、物联网终端等）、边缘计算节点和云计算中心相结合，实现智能化服务的架构体系。端边云智能协同架构可以为终端用户提供个性化、网络化和包容性的智能服务，让用户能够随时随地享受生成服务。端边云智能协同需要建立由性能驱动的网络资源调度服务。由于大模型等AI新业务服务涉及更频繁的数据收集和处理，需网络持续分配计算资源以提供大模型所需算力，网络架构需要一个独立的计算面，它用于协调计算资源并执行与人工智能相关的功能，作为控制面和用户平面的补充。端边云智能协同架构需设计深度融合的通信和算力融合的资源管理能力，并在这些功能之上设计一个逻辑的AI工作流程来建立生成式服务编排能力。在6G智能内生阶段，各个网元都可被看做具备智能处理能力的智能体，网络通过智能体的协同实现端边云资源的合理调度，更好地服务各项通信业务。如图2所示，端边云协同智能化架构需充分利用端边云三者的优势，实现数据的高效处理和智能决策，以满足不同场景下的应用需求[16-17]。

终端设备未来将演进为端智能体，主要进行数据的采集与初步处理，开展本地实时业务。终端设备是数据的源头，负责收集各种类型的数据，如传感器数据（温度、湿度、光线等）、用户输入（语音、文字、手势等）和设备状态数据。例如，智能手机中的摄像头可以采集图像数据，麦克风可以采集音频数据。同时，终端设备也会对采集的数据进行一些初步的处理，如数据格式的转换、简单的特征提取等。例如，在图像识别应用中，终端设备可以对采集的图像进行初步的滤波、裁剪等操作，减少数据传输量。同时，随着端侧智能的发展，终端设备可以运行一些简单的智能应用，如本地语音助手、本地图像分类等。这些应用可以在没有网络连接或者网络不稳定的情况下，为用户提供基本的服务。例如，手机中的本地语音助手可以

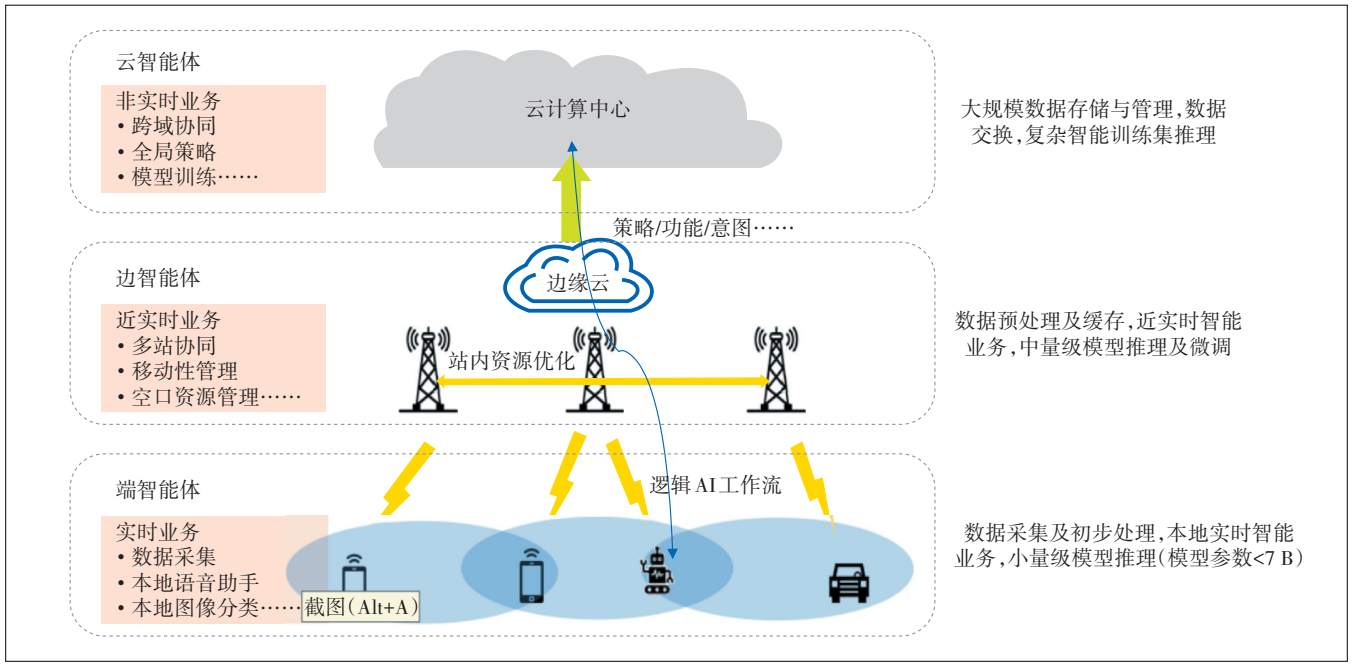

图2　端边云智能协同架构

在离线状态下执行一些简单的语音指令，如打开应用、设置闹钟等。

基站及边缘云设备未来将演进为边智能体，主要进行数据预处理与缓存及近实时业务等。边缘计算节点位于靠近终端设备的位置，如基站、边缘服务器等。它可以对终端设备上传的数据进行进一步的预处理，如数据清洗、数据聚合、数据加密等。例如，在工业物联网中，边缘计算节点可以对多个传感器上传的数据进行清洗和聚合，去除异常数据，并将相关数据组合成有意义的数据集。同时，边缘计算节点还可以缓存一些常用的数据，减少数据传输的延迟和网络带宽的占用。另外，边缘计算节点可以进行实时性要求高的智能处理。例如，在自动驾驶场景中，车辆上的边缘计算设备可以对摄像头和雷达的数据进行实时处理，做出紧急制动、避让等决策，以确保行车安全。

云智能体由云计算中心各类服务器组成，主要进行大规模数据存储与管理、复杂的智能算法运算、非实时业务以及全局智能决策与资源调度。云计算中心拥有巨大的存储容量，可以存储从终端设备和边缘计算节点收集来的海量数据。云计算中心具有强大的计算能力，可以运行复杂的智能算法，如深度学习模型的训练和推理。例如，在自然语言处理领域，云服务器可以利用大规模的文本数据训练语言模型，然后为各种语言处理应用提供服务，如机器翻译、文本生成等。云计算中心可以根据从终端设备和边缘计算节点获取的信息，做出全局的智能决策，并对整个系统的资源进行调度。同时，云中心还可以根据系统的负载情况，调度边缘计算节点和终端设备的资源，以提高系统的整体性能。

端边云协同架构通过一个逻辑 AI 任务流程实现协同工作机制。根据任务的复杂性、实时性要求和资源需求，智能任务会被分配到终端设备、边缘计算节点或云计算中心。简单的任务，如本地设备状态监测、简单的用户交互等可以由终端设备完成；对实时性要求高但计算资源需求相对不大的任务，如实时交通监控、工业设备的实时故障检测等可以由边缘计算节点完成；而复杂的、需要大量数据和计算资源的任务，如人工智能模型的训练、大数据分析等则由云计算中心完成。端边云协同智能化架构会不断根据应用场景的变化、用户需求的变化和系统性能的反馈，对端边云三者的工作进行协同优化。例如，当边缘计算节点的负载过重时，可以将部分任务转移到云计算中心或者终端设备；当云计算中心的存储资源紧张时，可以对数据进行清理或者将部分数据存储在边缘计算节点的缓存中。通过这种协同优化，可以提高整个系统的性能、效率和可靠性。

端边云协同架构需具备融合的网络资源管理能

力。网络需以智能自适应的方式为任务分配最合适的资源。由于AI服务所需资源范围是从云端到终端的分布式计算节点，所以需要新的协议栈来传输AI的应用任务消息并实现模型更新和分发。可以考虑在协议层面引入新的RRC消息或标识，确保网络资源快速分配给实时AI工作流。另外，AI应用涉及大量数据处理及传输，可以通过量化技术改进PDCP压缩算法以更好地识别和压缩AI数据负载，提高数据处理效率。也可以在协议中定义一个动态配置模块，专门针对特定AI任务，可以建立闭环机制进行实时调整。

端边云协同架构需具备数据处理与隐私保护。数据采集处理及微调是生成大语言模型等AI业务的基础。边缘服务器等需具备过滤重复的数据以减轻数据传输的通信负担。另外，网络需引入数据脱敏模块作为关键的数据处理服务，以避免嵌入在数据中的隐私被泄露。同时，可以增加数据策略执行模块来根据监管和非监管规则（如地理限制等）处理数据，以确保数据处理的完整性和合法性。数据策略执行模块可以储存一些数据处理模型库，并通过适当的访问控制向实体网元开放这些能力，以便各个网元更好地利用数据服务进行数据处理。

4 结束语

大模型在网络智能化进程中彰显出了变革性的力量与深远意义。通过对海量数据的深度挖掘与学习，大模型能够精准地理解网络复杂多变的运行状态，实现高效故障诊断与预测、网络资源优化配置等关键功能，极大地提升了网络运维管理的效率与精准度，助力网络服务质量的改善与用户体验的增强。然而，在大模型与网络智能化融合的道路上仍面临诸多挑战，如端边云协同架构、网络上行速率及时延提升、数据隐私与安全问题、模型的可解释性与可靠性等。展望未来，大模型与网络智能化的协同发展将引领网络技术迈向新的高度，催生更多创新应用与服务模式，不仅在通信领域，还将在智能交通、工业互联网、智慧城市等众多领域产生广泛的辐射效应，推动各行业数字化转型加速前行，为构建更加智能、高效、便捷的数字社会注入源源不断的动力与活力，开启网络智能化新时代篇章。

参考文献：

[1] 李露，李福昌，马艳君，等. 6G通感智算一体化无线网络技术研究[J]. 信息通信技术与政策，2023，49(9)：7-12.
[2] 中国联通研究院. 6G通感智算一体化无线网络白皮书[R/OL]. [2024-06-09]. https://13115299.s21i.faiusr.com/61/1/ABUIABA9GAAg7ua4pQYo_IWaoQM.pdf.
[3] KIM J，KIM J，CHOI S. FLAME：free-form languagebased motion synthesis & editing[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2023，37(7)：8255-8263.
[4] WANG W，GHOBADI M，SHAKERI K，et al. How to build low-cost networks for large language models（without sacrificing performance）[EB/OL]. [2024-10-23]. http://arxiv.org/abs/2307.12169.
[5] XU M，DU H，NIYATO D，et al. Unleashing the power of edge-cloud generative AI in mobile networks：A survey of AIGC services[EB/OL]. [2024-10-23]. http://arxiv.org/abs/2303.16129.
[6] WEN WU，MUSHU LI，KAIGE QU，et al. Split Learning Over Wireless Networks：Parallel Design and Resource Management[J]. IEEE Journal on Selected Areas in Communications，2023，41(4)：1051-1066.
[7] 柳宁馨. 2024云栖大会集聚硬科技AI产品迭出、应用场景萌发[N]. 21世纪经济报道，2024-09-20(5).
[8] 谌丽，艾明，孙韶辉. 基于AI内生的无线接入网络架构[J]. 无线电通信技术，2022，48(4)：574-582.
[9] 邓建明，龚循飞，于勤，等. 基于AI大模型的新能源汽车智能座舱多模态交互技术研究综述[J/OL]. [2024-12-14]. https://doi.org/10.19822/j.cnki.1671-6329.20230296.
[10] 孙彦赞，潘广进，余涛，等. AI使能的高能效无线通信技术[J]. 移动通信，2023，47(6)：77-82.
[11] 蒋秋萍. 基于开源无线通信的异构通算方法研究与实现[D]. 北京：北京邮电大学，2023.
[12] 周勇，吴瑕，狄宏林. 基于人工智能和大数据的网络故障诊断可视化平台研究[J]. 佳木斯大学学报（自然科学版），2024，42(6)：26-29，71.
[13] 承楠，陈芳炯，陈文，等. 6G全场景按需服务：愿景、技术与展望[J]. 中国科学：信息科学，2024，54(5)：1025-1054.
[14] 朱宏，邓程，王瑜，等. 基于人工智能的运营商故障分析能力提升研究[J]. 邮电设计技术，2024(6)：72-77.
[15] 王浩宇. 面向运营商业务的智能营销客服系统设计与实现[D]. 哈尔滨：哈尔滨工业大学，2023.
[16] 李露，李福昌，高谦. 5G-A/6G无线网智能化技术研究[J]. 信息通信技术，2024，18(1)：38-43.
[17] 李福昌，李露，高谦. 无线网络智算融合需求及技术研究[J]. 移动通信，2024，48(8)：2-7.

作者简介：
李露，工程师，硕士，主要从事5G/6G移动通信网络、人工智能等方面的研究工作；李福昌，教授级高级工程师，博士，主要从事无线通信网络研究等工作。