通信网络与大模型的融合与协同

Integration and Collaboration of Communication Networks and Large Models

任天骐/REN Tianqi[1]，李荣鹏/LI Rongpeng[1]，
张宏纲/ZHANG Honggang[2]
（1. 浙江大学，中国 杭州 310007；
2. 之江实验室，中国 杭州 310012）
(1. Zhejiang University, Hangzhou 310007, China;
2. Zhijiang Lab, Hangzhou 310012, China)

DOI：10.12142/ZTETJ.202402005
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.tn.20240407.1930.004.html
网络出版日期：2024-04-10
收稿日期：2024-02-20

摘要：随着以大模型（LM）为代表的生成式人工智能（AI）的兴起，将LM应用于通信网络的研究引起了学术界和工业界的广泛关注。回顾了目前LM的主流神经网络架构及其能力涌现机理，然后从AI与通信的双向协同、网络大模型部署两方面，深入探讨了通信网络LM研究的主要进展。还分析了网络LM NetGPT将要面临的挑战以及未来的发展方向。考虑到基于AI/机器学习（ML）的通信模型相较于传统模型获得的出色性能，认为将通信网络与LM进行融合并使二者协同工作，能进一步地提升系统的性能。要实现通信网络与LM的融合与协同，本质上是要构建好网络LM，云边协同就提供了一种很好的网络LM部署方案。

关键词：LM；生成式AI；网络智能；NetGPT；模型协同

Abstract: Along with the springing up of generative artificial intelligence (AI), notably epitomized by large models (LM), the incorporation of these LMs within communication networks has attracted extensive attention in both academia and industry. An overview of the dominant deep neural network (DNN) architecture of LMs and its emerging capabilities is introduced. The significant advancements achieved by applying LMs for communication networks from two aspects are discussed, namely, the mutual collaboration between AI and communications, and the deployment of network generative pre-trained transformer (NetGPT). Additionally, the imminent challenges and further work are also discussed. Considering the outstanding performance of AI/machine learning (ML)-based communication models compared to traditional models, it is believed that integrating communication networks with large models and enabling them to work together can further enhance system performance. To realize the integration and collaboration of communication networks and large models, it is essentially necessary to build NetGPT properly. Edge-cloud collaboration provides a good deployment solution for NetGPT.

Keywords: LM; generative AI; network intelligence; NetGPT; model collaboration

引用格式：任天骐, 李荣鹏, 张宏纲. 通信网络与大模型的融合与协同 [J]. 中兴通讯技术, 2024, 30(2): 29-36. DOI: 10.12142/ZTETJ.202402005
Citation: REN T Q, LI R P, ZHANG H G. Integration and collaboration of communication networks and large models [J]. ZTE technology journal, 2024, 30(2): 29-36. DOI: 10.12142/ZTETJ.202402005

随着移动通信网络复杂度的显著增加和通信业务生态的多样化，通信网络面临着越来越多复杂场景的挑战。因此，通信网络既要满足高速、高质量的通信需求，又要向用户提供颇具差异性的业务体验，还要考虑稳定性和安全性，这对通信网络的设计、运营和维护提出了更高的要求。在这样的背景下，人工智能（AI）技术的出现为解决这些问题带来了新希望。现代AI建立在机器学习的基础上，在众多机器学习模型中，Transformer架构[1]因其独特的自注意力机制而脱颖而出，它能够处理长距离依赖关系，这在自然语言处理（NLP）等领域尤为重要。

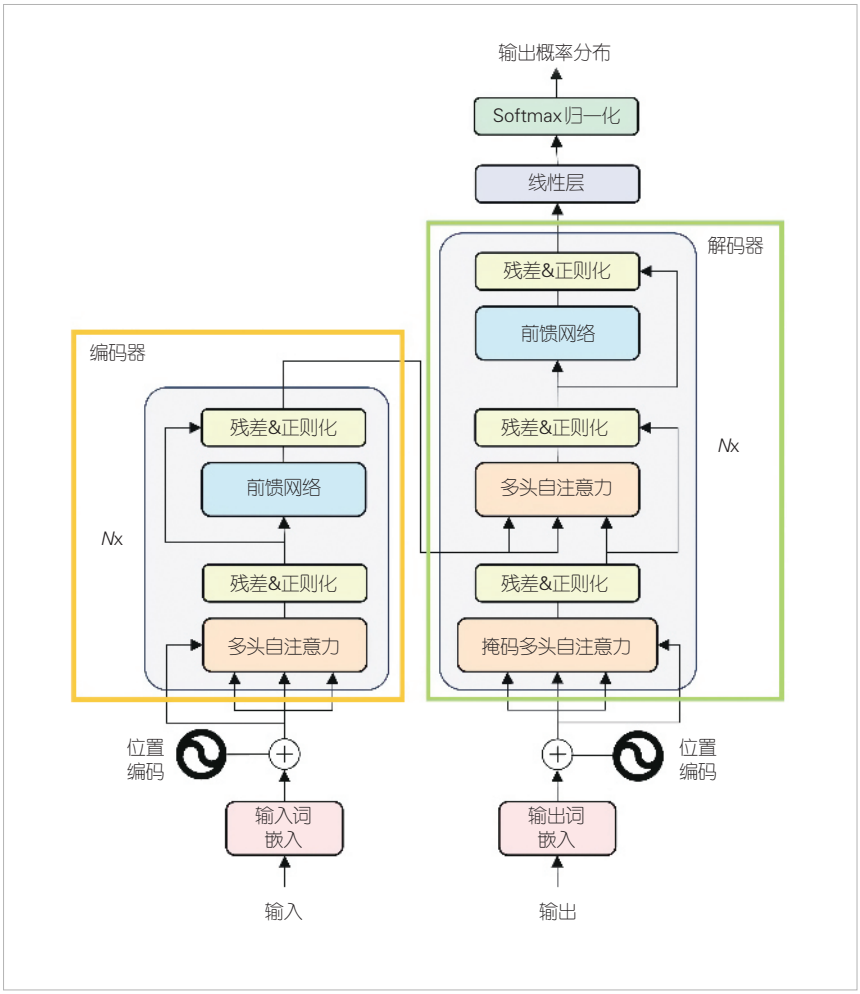
Transformer模型为大型预训练模型的构建提供了基础架构。2018年，谷歌提出的BERT[2]是基于Transformer架构的第一个突破，其革新的双向训练策略极大地提升了模型对文本深层次语义理解能力，在多项语言任务中取得优秀的性能。紧随其后的GPT-2[3]采用了Transformer的解码器结构，通过学习大量的文本数据，能够生成多样且逻辑合理的文本。随着计算资源的提升和优化，以及大模型能力标度率和具体涌现能力的发现，模型的规模正在迅速扩张[4-6]。从BERT模型1.1亿个参数到GPT-3[7]和GPT-4[8]的数百亿乃至

基金项目：国家自然科学基金项目（62071425）；浙江省“领雁”计划项目（2022C01093）；浙江省杰出青年基金项目（LR23F010005）

数万亿个参数。此外，通过预训练和微调，大模型中还融合多模态技术[9-10]，使得大模型在自然语言处理、计算机视觉[11-12]、自动驾驶[13-14]等多个领域展现出强大的潜力。

同时，数据、算力与模型构成了实现AI的三大基石，而6G成为“通、感、算、智、存”集于一体的超级基础设施平台，为融合AI提供了充足的条件，因此基于内生智能的新型网络架构应运而生。内生智能网络不仅要引入AI来构建网络，还需要充分利用网络节点的通信、计算和感知能力，并将通过分布式学习、群智式协同以及云边端一体化算法部署，原生支持各类AI应用，为各行业用户提供实时AI服务和实时计算类新业务[15-16]。

本文中，我们将首先探讨大语言模型（LLM）的基础原理，包括Transformer结构、标度率和涌现能力，以及LLM的预训练与微调过程。进一步地，我们将分析AI，特别是LLM在通信网络中的应用及其带来的双向增益。同时，也将探讨大模型发展面临的问题与挑战，以及如何更好地利用AI技术来优化并实现通信网络的转型。



▲图1 Transformer模型架构[1]

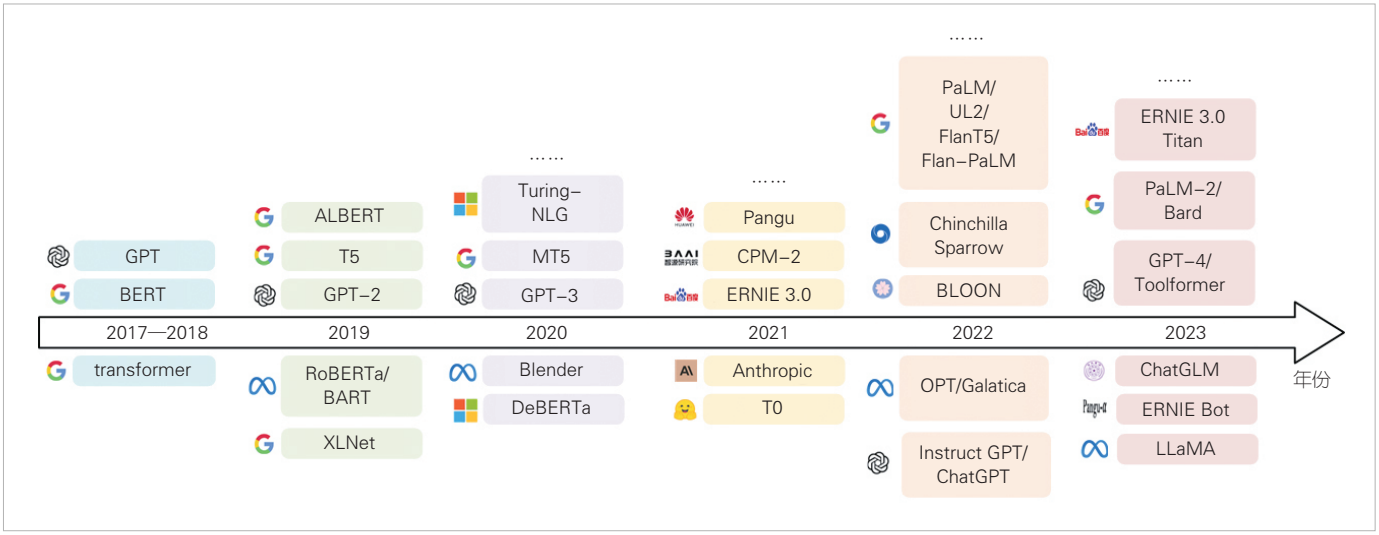
1 大模型的理论与技术

1.1 大模型架构

现有LLM的进步主要得益于Transformer的发展[1]。Transformer模型完全摒弃了传统语言模型广泛使用的循环神经网络（RNN）和长短期记忆网络（LSTM）模型，全面采用自注意力机制来处理序列。

如图1所示，Transformer模型包含编码器和解码器，两者均由N个（原文中N=6）相同的层堆叠而成。编码器负责理解输入文本并构造语义表示，而解码器则使用编码器的输出来生成目标序列。编码器中的每个层由多头自注意力层和全连接前馈网（FFN）两个子层构成，而解码器相比编码器多出一个掩码多头自注意力层。注意力机制的引入使得Transformer模型在处理序列的每个元素时，能够考虑到整个序列的上下文信息，从而在NLP任务中表现出并行化训练和性能优异的特点[18]。例如，Transformers架构通过自注意力机制解决了长距离依赖问题，使模型能够直接关注到序列中任意两个位置之间的关系。同时，Transformer架构允许比RNN更多的并行化，这使得图形处理器（GPU）上的大量数据上有效地预训练非常大的语言模型成为可能。

Transformer模型的提出极大地推动了LLM的发展。LLM的发展历程如图2所示。基于Transformer架构，LLM演化为3种主要架构：仅编码器（encoder-only）、仅解码器（decoder-only）、编码器-解码器（encoder-decoder）。目前，最主流的是仅解码器架构，代表性的LLM有GPT系列[3, 7-8]、LLaMA[18-19]、PaLM[20]等。仅编码器架构模型的代表是BERT系列，包括BERT[2]、RoBERTa[21]和ALBERT[22]等；编码器-解码器架构的代表模型有谷歌的T5模型[23]、Meta AI的BART模型[24]和华为的Pangu大模型等。3种架构各有优劣：仅解码器架构更多关注于从已有的信息扩展出新的内容，适合文本生成和扩展类型的任务，但需要大量的训练数据来提高生成文本的质量和多样性；仅编码器架构能更好地理解输入文本的语义和上下文信息，适合理解和分析类型的任务，缺点是无法直接生成文本输出；编码器-解码器架构能更好



▲图2 大型语言模型发展时间线

地处理输入序列和输出序列之间的关系，适合需要理解输入内容并生成相关响应的任务，如机器翻译、生成式问答等，但模型复杂度较高，训练时间和计算资源消耗较大。

1.2 标度率和涌现能力

大模型的标度率是OpenAI在2020年提出的概念[5]，是AI模型训练过程中的一个重要的经验性发现。在传统的小模型中，其性能往往会随着训练次数的增加而趋于稳定，甚至出现过拟合而导致性能下降。大模型的标度率则揭示了一个不同的现象：随着模型规模、数据集大小以及训练计算量的扩增，模型性能够获得持续提升。具体而言，当不受其他两个因素制约时，模型性能与每个单独的因素呈幂律关系。进一步的研究揭示[6]，当前的LLM实际上训练不足，而为了实现最佳性能，模型规模和训练数据集大小应以大致相同的速度扩增。此外，除了数据集大小，数据质量也被认为是影响模型性能的关键因素。

标度率提出后，可以预见：随着模型参数量的增加，模型在大部分任务中表现出的性能较为稳定。而随着模型规模的持续扩大，研究者发现[5]，对于特定的任务和模型来说，在模型规模小于某个阈值之前，模型基本不具备任务解决能力；但当模型规模大到一定程度时，模型性能显著提高。这被称为大模型的涌现能力。

1.3 大语言模型的预训练、微调与对齐

在大语言模型的预训练阶段，自监督学习发挥着核心的作用。该方法使模型能够在无需人工标注的数据集上学习并理解语言的丰富特征。自监督学习通过构建任务，如掩码语言模型（MLM）或自回归预测，使模型能够从大规模未标注文本中抽取和学习复杂的语言结构和语义信息。这种自监督机制的广泛应用源于其赋予模型从大量的文本数据中学习到通用语言表示的能力，这为模型后续进行特定任务的微调奠定了坚实基础。

预训练完成后，LLM可以获得处理各种任务的通用能力。为了将LLM适配到特定领域的任务，需要对LLM进行微调。LLM的微调，通常采用监督学习的技术路线。由于使用的训练数据通常包含标签或特定任务的指导信息，监督学习能使已经预训练过的模型针对具体的应用进行优化，提高了特定任务上的表现。近期，指令微调作为一种先进的微调策略，允许模型通过理解并执行明确的任务指令来调整其行为，进一步增强了模型对不同任务的适应能力和灵活性。

预训练和微调的策略反映了一种互补性：前者通过自监督学习为模型提供广泛的语言理解能力，而后者则确保模型在前者的基础上针对特定任务实现优化。这种互补性策略极大地提升了模型在多种自然语言处理任务中的泛化能力。

LLM预训练使用了语言建模的目标，但却没有考虑到人类的价值观或偏好，可能产生有害的、误导性的或有偏见的表达，因此需要一些对齐技术来使LLM的行为符合人类期望。为此，InstructGPT[25]利用基于人类反馈的强化学习（RLHF）技术[26]，通过学习奖励模型使LLM适配人类反馈，并将人类纳入训练的循环中来得到对齐良好的LLM[27]。

2 通信网络大模型的研究与发展

6G对网络架构提出了“万物智联，数字孪生”的总体愿景，强调智慧内生是6G网络应当具备的一大特征[15]。这

一特征意味着6G网络将内嵌AI能力，实现架构级智慧内生。6G网络对内能够利用智能来优化网络性能，增强用户体验，自动化网络运营，即使用AI来构建网络；对外能够抽取和封装网络智能，为各行各业用户提供网络和AI结合的通信和计算服务，即网络赋能AI[16]。AI构建网络和网络赋能AI两个概念共同构成了通信网络与大模型融合协同的框架。

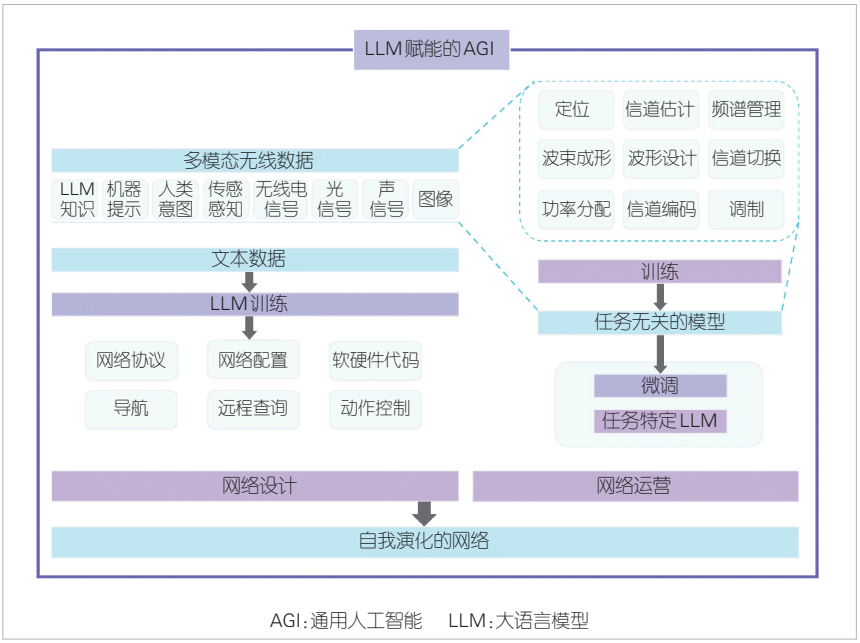
2.1 AI与通信网络的双向协同：构建与赋能

目前，通信网络的AI应用主要涉及机器学习的各个领域，包括监督学习、非监督学习和强化学习等，而生成式AI与通信网络的深度融合还处于起步阶段。这些技术构成了通信网络中机器学习的基础，致力于学习网络参数以优化网络性能。

近年来，LLM作为生成式AI的典型代表，在通信网络中的作用开始受到业界关注。LLM通过在大规模语料库上进行预训练，进而在多种下游任务中微调，从而展现出电信语言的理解能力。L. BARIAH等[28]通过微调LLM来识别第三代合作伙伴项目（3GPP）技术文档中的规范类别，证实了LLM在电信领域的应用价值。此外，LLM还被用于辅助网络运营（NetOps）和增强网络管理，如LLM可以作为NetOps中的常识性知识和推理能力的良好工具[29]。尽管在直接操作网络拓扑方面，LLM依然存在可靠性、可解释性等问题，但S. K. MANI等[30]提出的新框架通过生成自定义代码来解决这些问题，推进了LLM的网络管理实践。

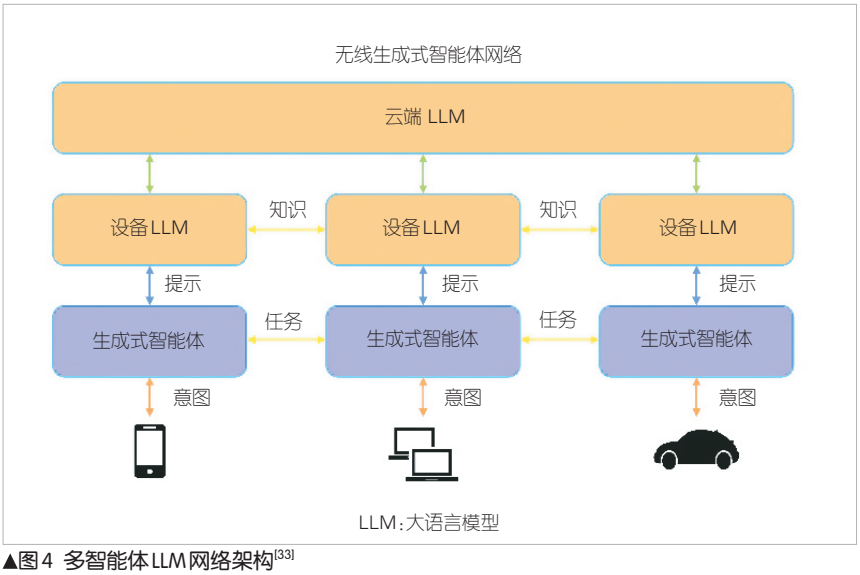
此外，大型生成式AI（GenAI）模型的发展为通信网络带来了新机遇。这些模型通过集成多模态数据，展示出在预训练基础模型、改善无线传感和传输方面的强大能力。L. BARIAH等[31]的研究深入探讨了GenAI模型与电信数据融合的策略，显示出大型GenAI模型在推动网络向自我演化方面的关键作用。图3中给出了通用人工智能（AGI）赋能无线网络的架构。总的来说，尽管LLM和GenAI模型在通信网络中的应用仍面临挑战，但它们在推动电信行业自动化和智能化发展方面的潜力是巨大的[32]。

在网络赋能AI方面，生成式AI正在推动无线设备实现集体智能，这对6G网络中的知识转移计算结构至关重要。该结构的目标在于利用云中的大型生成式AI模型，促进其向分布式集体智能过渡[31]。LLM巨大的计算和存储需求使其难以直接部署在边缘计算环境中。但通过在多个边缘设备上的部署，可以实现多智能体间的协同规划和任务决策。ZOU H.等[33]提出的多智能体LLM网络架构充分展示了这一点，如图4所示，其中无线生成式智能体不仅作为感知环境的传感器，也参与执行决策，这体现了生成式AI、边缘网络和多智能体系统之间创新性的协同效应。



AGI：通用人工智能　LLM：大语言模型

▲图3 AGI赋能的无线网络[31]



LLM：大语言模型

▲图4 多智能体LLM网络架构[33]

2.2 构建网络大模型的实践

在《网络大模型的十大问题》[34]中，网络大模型（NetGPT）被定义为无线网络中部署的大模型，其架构如图5所示。要实现通信网络与大模型的融合与协同，本质上是要构建好网络大模型。

在构建网络大模型的实践方面，WANG Y. C.[35]等调研了如何利用边缘云计算范式构建大规模GenAI系统。边缘云计算是指计算和存储资源靠近数据源或终端设备，将计算功能从传统的云数据中心推向网络边缘。边缘云计算利用了云服务器中强大的计算资源以及边缘服务器中高效的数据管理和通信。相比于云计算和多址边缘计算，边缘云计算在满足计算要求和低延迟要求上展现出优势，同时具有良好的可扩展性和数据安全性。

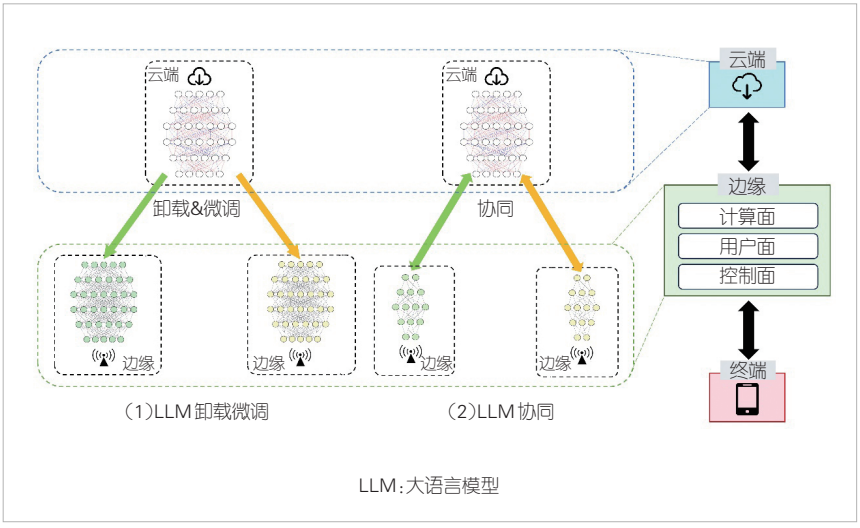
然而，将计算功能推向网络边缘意味着，在边缘端模型需要从云端进行计算卸载，且边缘和云端将缺乏一定的关联性。为了缓解这个问题，CHEN Y.等[36]提出了一种云边协同的部署方案，通过在云端与边缘端部署不同规模的模型协同作业来实现目的。在此架构中，边缘端部署的LLM是轻量级的，专门优化以适应边缘计算的资源限制，并能够利用位置相关信息增强个性化服务，以满足区域特定的需求。相对而言，云端的设备由于其更强大的计算能力和更大的存储空间，部署了完整版本的LLM，负责处理更复杂的全局任务。图6中给出了LLM卸载微调和LLM协同的两种部署方案。

在云边协同的架构中，边缘节点上的LLM负责收集并预处理来自本地区域的请求，包括将简单请求扩展为含有丰富区域特征的完整请求，并执行请求的去重整合。这些处理过的请求随后被发送到云端的LLM，后者利用其强大的计算能力生成高质量、个性化的回答。此过程不仅展现了通信网络在赋能AI方面的作用，还能通过边缘与云端LLM的高效协作，提升AI生成内容的质量和个性化程度；而AI对通信网络的增益则体现在通过边缘节点的LLM实现请求的有效预处理和减少冗余传输，这能够降低通信成本并优化网络时延。

在云端和边缘端部署LLM都需要基于预训练的LLM向通信任务进行迁移。CHEN Y.等[36]在工作中选择并部署了LLaMA-7B模型和GPT-2-base模型，并对部署的LLM进行微调来适应任务需求。在云端部署的LLaMA-7B模型，无法直接生成响应式文本，因此选择基于低秩适应（LoRA）的技术[37]，使用Standford Al - paca数据集[38]进行参数高效的微调。在边缘端部署的GPT-2-base模型，需要附加基于位置的信息来扩展提示，以实现个性化，因此选择self-instruct方法[39]，使用手动编写的位置相关提示与OpenAI的TextDavinci-003模型进行交互，来生成有效的文本样本作为“综合提示”。

云边协同部署网络大模型的工作流程主要集中于协调边缘和云端的一系列网络功能，并优化数据处理和分析的过程。当用户请求生成特定内容时，该架构通过先进的逻辑AI工作流来解析和编排服务，根据用户需求和网络状况的动态变化，选择是在边缘端进行初步处理还是在云端执行深

▲图5 网络大模型NetGPT的3层架构[34]



▲图6 网络大模型部署方案[36]

度处理。在服务部署阶段，逻辑AI工作流将根据服务质量需求映射到相应的物理资源。在融合通信和计算（C&C）资源管理层面，我们不仅需要考虑控制面的无缝连接和用户平面中的信息传输可靠性，还需要在计算平面中有效地协调异构计算资源。此外，该架构还引入新的协议栈以传输AI生成的消息，并实时更新和分发模型，同时考虑引入新的标识符来为实时AI工作流优先分配网络资源。

总体而言，网络大模型实现了C&C资源的深度融合，并通过个性化的云边大模型耦合更新机制促进了云边协同以提高服务质量。此外，通过在边缘处理私有数据，在云端处理大数据的分割机制，达成了计算效率和数据安全的最优平衡。

3 问题与挑战

目前，将AI算法融入通信系统已经有众多应用场景，如AI赋能物理层、AI赋能高层、AI赋能应用层等。此前，研究人员在这些应用场景做出了许多有益的尝试[40]，包括但不限于基于AI的信道估计及反馈[41–42]、基于AI的多输入多输出（MIMO）检测[43–44]、基于AI的资源和功率分配[45]和基于AI的传输层拥塞控制技术[46]等。这些研究都证明，与传统通信模型相比，基于AI/机器学习（ML）的通信模型可以获得更出色的性能。目前的研究大多采用传统的AI算法或神经网络结构，但根据标度率的发现以及大模型在众多领域展示出的卓越性能，我们有理由相信，将大模型应用于这些任务中，将会获得更大的增益。然而网络大模型领域的研究依然面临着一系列基础性问题的挑战。这些挑战主要涉及大模型本身的设计类问题和网络设计如何支撑大模型应用类问题[34]，主要如下：

1）模型协同：在不同规模和部署位置的模型之间实现有效的数据和参数协同是一个主要挑战。此外，不同任务类型对模型推理协同的需求也有所不同。针对跨域任务，L0全网通用大模型需要协同多个L1网络专业大模型进行处理，并提供通用知识；而针对单域任务，L1网络专业大模型需要和多个L2网络小模型进行协同处理，并提供专业知识。总的来说，实现网络内不同规模模型的协同进化，以及明确各自的职责，是解决这一挑战的关键策略。

2）网络架构设计：引入NetGPT优化网络服务需要考虑如何利用NetGPT的自然语言理解能力为应用程序生成专有的网络服务，并处理模型更新导致的计算负担。此外，考虑到当前网络的基于字符串的接口协议可能被基于模型间的协作接口取代，为了保证网络性能的实时性、稳定性和可靠性，需要把NetGPT深度融入6G网络架构，推动网元的智能化。

3）分布式学习与部署：在分布式网络中，考虑到节点计算资源和存储能力的差异，模型需要分布式拆分和自适应调整。在模型学习算法层面，现有的模型并行和数据并行方式存在局限性，因此还需要我们深入探索分布式训练方法。此外，分布式节点间的通信瓶颈是制约模型性能的关键因素，这就需要从算法和网络设计两方面同时入手，进行模型压缩，如剪枝和量化等，在网络内设计高效的节点间通信机制。此外，数据隐私与数据异质性、以及如何降低通信开销，也是需要关注的问题。

4）全生命周期管控和编排：在生命周期管控方面，不仅要选择适当的拆分策略，还要设计高效的更新和维护策略以应对计算开销和时间成本的显著增加。同时，考虑到NetGPT的知识产权保护，还需要建立平衡的协同管理机制。在编排方面，需要对计算任务和网络资源进行合理的识别、编排和反馈，以提高系统性能和资源利用率，实现面向动态需求的NetGPT闭环控制。

4 结束语

大模型作为当前最热门的研究热点，毫无疑问将成为AI与通信融合的关键组成部分，在提高网络中AI的通用性和多任务处理能力等方面发挥重要作用。本文中，我们首先从大模型的架构、标度率和涌现能力以及LLM的训练微调与对齐3个方面回顾了大模型的理论与技术，之后探讨了LLM和生成式AI在通信网络中的应用及其带来的双向增益。接下来，强调了AI与通信网络的双向协同，包括AI构建网络和网络赋能AI的概念，以及构建网络大模型NetGPT的实践。网络大模型作为一种内生智能的新型网络架构展现出巨大潜力，但要成功地部署网络大模型仍然存在一定的挑战。我们期待在该领域能有更多的前瞻性研究工作，为通信网络与大模型的融合与协同带来创新和突破。

致谢

感谢浙江大学陈宇轩和鲁芝琳在本文撰写过程中给予的帮助和支持。

参考文献

[1] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 6000–6010. DOI: 10.5555/3295222.3295349

[2] DEFLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep

bidirectional transformers for language understanding [EB/OL]. [2024−03−04]. https://aclanthology.org/N19−1423/
[3] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners [EB/OL]. [2024−03−04]. https://cdn. openai. com/better−language−models/language_models_are_unsupervised_multitask_learners.pdf
[4] KAPLAN J, McCANDLISH S, HENIGHAN. Scaling laws for neural language models [EB/OL]. (2020−01−23) [2024−03−04]. https://arxiv.org/abs/2001.08361
[5] WEI J, TAY Y, BOMMASANI R, et al. Emergent abilities of large language models [EB/OL]. (2022−01−15) [2024−03−04]. https://arxiv.org/abs/2206.07682
[6] HOFFMANN J, BORGEAUD S, MENSCH A, et al. Training compute−optimal large language models [EB/OL]. (2022−03−29) [2024−03−04]. https://arxiv.org/abs/2203.15556
[7] BROWN T B, MANN B, RYDER N, et al. Language models are few−shot learners [C]//Proceedings of the 34th International Conference on Neural Information Processing Systems. ACM, 2020: 1877−1901. DOI: 10.5555/3495724.3495883
[8] ACHIAM J, ADLER S, AGARWAL S, et al. GPT−4 technical report [EB/OL]. (2023−03−15) [2024−03−04]. https://arxiv. org/abs/2303.08774
[9] DONG R, HAN C, PENG Y, et al. DreamLLM: synergistic multimodal comprehension and creation [EB/OL]. (2023−09−20) [2024−03−04]. https://arxiv.org/abs/2309.11499
[10] LIN Z Q, YU S, KUANG Z Y, et al. Multimodality helps unimodality: cross−modal few−shot learning with multimodal models [C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2023: 19325−19337. DOI: 10.1109/cvpr52729.2023.01852
[11] LI J, LI D, XIONG C, et al. Blip: bootstrapping language−image pre−training for unified vision−language understanding and generation [C]//Proceedings of the 39th International Conference on Machine Learning. JMLR, 2022: 12888−12900. DOI: 10.48550/arXiv.2201.12086
[12] LI J, LI D, SAVARESE S, et al. Blip−2: bootstrapping language−image pre−training with frozen image encoders and large language models [C]//Proceedings of the 40th International Conference on Machine Learning. JMLR, 2023: 19730−19742. DOI: 10.5555/3618408.3619222
[13] WANG H W, XIE J, HU C Y, et al. Drivemlm: aligning multi−modal large language models with behavioral planning states for autonomous driving [EB/OL]. (2023−12−14)[2024−03−04]. https://arxiv.org/abs/2312.09245
[14] CUI C, MA Y, CAO X, et al. A survey on multimodal large language models for autonomous driving [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2024: 958−979. DOI: 10.48550/arXiv.2311.12320
[15] IMT−2030(6G)推进组. 6G 总体愿景和潜在关键技术 [EB/OL]. (2022−02−18)[2024−03−02]. https://www.eet−china.com/news/202106090412.html
[16] IMT−2030(6G)推进组. 6G总体网络架构愿景和关键技术展望 [EB/OL]. (2021−09−16) [2024−03−02]. https://cloud. tencent. com/developer/news/857663
[17] ALMMAR J. The illustrated transformer [EB/OL]. (2018−06−27) [2024−03−04]. https://jalammar.github.io/illustrated−transformer/
[18] TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: open and efficient foundation language models [EB/OL]. (2023−02−27) [2024−03−04]. https://arxiv.org/abs/2302.13971
[19] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine−tuned chat models [EB/OL]. (2023−06−18) [2024−03−04]. https://arxiv.org/abs/2307.09288
[20] CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: Scaling language modeling with pathways [J]. Journal of machine learning research, 2023, 24(240): 1−113
[21] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach [EB/OL]. (2019−7−26) [2024−03−02]. https://arXiv preprint arXiv:1907.11692, 2019
[22] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite bert for self−supervised learning of language representations [EB/OL]. (2020−02−09) [2024−03−02]. https://arXiv preprint arXiv: 1909.11942, 2019
[23] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text−to−text transformer [J]. The journal of machine learning research, 2020, 21(140): 1−67. DOI: 10.48550/arXiv.1910.10683
[24] LEWIS M, LIU Y, GOYAL N, et al. Bart: denoising sequence−to−sequence pre−training for natural language generation, translation, and comprehension [EB/OL]. (2019−10−29) [2024−03−02]. https://arXiv preprint arXiv:1910.13461, 2019.
[25] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [C]//Advances in neural information processing systems 35 (NeurIPS 2022). Curran Associates, 2022: 27730−27744. DOI: 10.48550/arXiv.2203.02155
[26] CHRISTIANO P F, LEIKE J, BROWN T B, et al. Deep reinforcement learning from human preferences [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. ACM, 2017: 4302 – 4310. DOI: 10.5555/3294996.3295184
[27] WANG Y, ZHONG W, LI L, et al. Aligning large language models with human: a survey [EB/OL]. (2023−07−24) [2024−03−04]. https://arxiv.org/abs/2307.12966
[28] BARIAH L, ZOU H, ZHAO Q, et al. Understanding telecom language through large language models [EB/OL]. [2024−03−04]. https://arxiv.org/pdf/2306.07933v1.pdf
[29] MIAO Y K, BAI Y, CHEN L, et al. An empirical study of NetOps capability of pre−trained large language models [EB/OL]. (2023−09−11)[2024−03−05]. https://arxiv.org/abs/2309.05557
[30] MANI S K, ZHOU Y J, HSIEH K, et al. Enhancing network management using code generated by large language models [C]//Proceedings of the 22nd ACM Workshop on Hot Topics in Networks. ACM, 2023: 196−204. DOI: 10.1145/3626111.3628183
[31] BARIAH L, ZHAO Q Y, ZOU H, et al. Large generative AI models for telecom: the next big thing? [J]. IEEE communications magazine, 2024: 1−7. DOI: 10.1109/mcom.001.2300364
[32] MAATOUK A, PIOVESAN N, AYED F, et al. Large language models for telecom: Forthcoming impact on the industry [EB/OL]. (2023−08−11) [2024−03−05]. https://arxiv. org/abs/2308.06013
[33] ZOU H, ZHAO Q, BARIAH L, et al. Wireless multi−agent generative AI: from connected intelligence to collective intelligence [EB/OL]. (2023−07−06) [2024−03−05]. https://arxiv.org/abs/2307.02757
[34] TONG W, PENG C, YANG T, et al. Ten issues of NetGPT [EB/OL]. [2024−03−05]. https://arxiv.org/pdf/2311.13106.pdf
[35] WANG Y C, XUE J T, WEI C W, et al. An overview on generative AI at scale with edge−cloud computing [J]. IEEE open journal of the communications society, 2023, (4): 2952−2971. DOI: 10.1109/ojcoms.2023.3320646
[36] CHEN Y, LI R, ZHAO Z, et al. NetGPT: a native−AI network architecture beyond provisioning personalized generative services [EB/OL]. [2024−03−05]. https://ieeexplore. ieee. org/document/10466747
[37] SU J, LU Y, PAN S, et al. LoRA: low−rank adaptation of large language models [EB/OL]. (2021−04−20) [2024−03−02]. https://

arxiv. org/abs/2104.09864
[38] TAORI R, GULRAJANI I. Stanford alpaca: an instruction-following llama model [EB/OL]. [2024-03-02]. https://github.com/tatsu-lab/stanford alpaca
[39] WANG Y, KORDI Y. Self-instruct: aligning language model with self generated instructions [EB/OL]. (2022-12-20) [2024-03-02]. https://arxiv.org/abs/2212.10560
[40] SUN Y, PENG M, ZHOU Y, et al. Application of machine learning in wireless networks: key techniques and open issues [J]. IEEE communications surveys & tutorials, 2019, 21(4): 3072-3108. DOI: 10.1109/COMST.2019.2924243
[41] GAO J, ZHONG C, LI G Y, et al. Deep learning-based channel estimation for massive MIMO with hybrid transceivers [J]. IEEE transactions on wireless communications, 2021, 21(7): 5162-5174. DOI: 10.1109/TWC.2021.3137354
[42] MA X, GAO Z, GAO F, et al. Model-driven deep learning based channel estimation and feedback for millimeter-wave massive hybrid MIMO systems [J]. IEEE journal on selected areas in communications, 2021, 39(8): 2388-2406. DOI: 10.1109/JSAC.2021.3087269
[43] YUN S, MOON S, JEON Y S, et al. Intelligent MIMO detection with momentum-induced unfolded layers [J]. IEEE wireless communications letters, 2024, 13(3): 879-883. DOI: 10.1109/LWC.2023.3348933
[44] HE H, WEN C K, JIN S, et al. Model-driven deep learning for MIMO detection [J]. IEEE transactions on signal processing, 2020, 68: 1702-1715. DOI: 10.1109/TSP.2020.2976585
[45] KARAKS E K, GEMICI Ö F, HOKELEK İ, et al. Work-in-progress: AI based resource and power allocation for NOMA systems [C]//2023 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom). IEEE, 2023: 402-407. DOI: 10.1109/BlackSeaCom58138.2023.10299756
[46] PILLAI B, CHHABRA G. TCP-CNNLSTM: congestion control scheme for MANET using AI Technologies [C]//2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS). IEEE, 2023: 63-69. DOI: 10.1109/ICAISS58487.2023.10250756

作 者 简 介

任天骐，浙江大学在读本科生；研究方向为大型语言模型在通信场景中的应用及语义通信。

李荣鹏，浙江大学信息与电子工程学院副教授、博士生导师；主要研究方向为智能通信网络、网络智能、网络切片等；曾入选首批博士后创新人才支持计划，获得浙江省杰出青年基金项目资助，并获吴文俊人工智能优秀青年奖、江苏省科学技术奖一等奖等。

张宏纲，浙江大学兼任教授、博士生导师；长期从事无线通信与网络、人工智能、认知通信、绿色通信、复杂网络等领域的研究；曾获2021年IEEE通信学会杰出论文奖、IEEE Internet of Things Journal（IoT-J）最佳论文奖等；发表论文265篇，拥有IEEE 802.15 UWB国际标准提案16项、国际专利3项。