NetGPT：超越个性化生成服务的内生智能网络架构

NetGPT: An AI-Native Network Architecture for Provisioning Beyond Personalized Generative Services

陈宇轩/CHEN Yuxuan，李荣鹏/LI Rongpeng，
张宏纲/ZHANG Honggang

（浙江大学，中国 杭州 310007）
(Zhejiang University, Hangzhou 310007, China)

DOI：10.12142/ZTETJ.202305011
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20231016.1554.016.html
网络出版日期：2023-10-17
收稿日期：2023-08-02

摘要：提出了基于边缘和云端部署相匹配大型语言模型（LLM）的内生智能网络架构NetGPT方案。边缘LLM可以有效地利用基于位置的信息进行个性化的补充，从而与云端LLM进行有效交互。通过在边缘和云端部署开源LLM，验证了NetGPT的可行性。认为面向NetGPT的内生智能网络架构的工作重点是通信和计算资源的深度集成以及AI逻辑工作流的灵活设计。认为NetGPT是一种可提供个性化的生成式服务的、有前途的内生智能网络架构。

关键词：LLM；内生智能网络架构；云边协同；个性化生成服务

Abstract: The NetGPT framework, which is founded upon the alignment of large language models (LLMs) tailored for both edge and cloud deployments, is introduced. Edge-oriented LLMs harness location-based data to effectively personalize content augmentation, facilitating seamless interactions with their cloud-based counterparts. The viability of the NetGPT paradigm is empirically substantiated through the deployment of open-source LLMs at both the edge and cloud strata. It is believed that within the realm of endogenous intelligent network architectures designed to support NetGPT, the central emphasis rests on the profound integration of communication and computational resources, coupled with the adaptability in the design of AI logic workflows.

Keywords: LLM; AI-native network architecture; edge-cloud collaboration; personalized generative services

引用格式：陈宇轩，李荣鹏，张宏纲. NetGPT：超越个性化生成服务内生智能网络架构 [J]. 中兴通讯技术，2023，29(5)：68-75. DOI：10.12142/ZTETJ.202305011

Citation: CHEN Y X, LI R P, ZHANG H G. NetGPT: an AI-native network architecture for provisioning beyond personalized generative services [J]. ZTE technology journal, 2023, 29(5): 68-75. DOI: 10.12142/ZTETJ.202305011

随着深度学习从AlphaGo到ChatGPT应用的转变，人工智能（AI）的作用将在6G网络中不断体现。一方面，边缘计算能力的提升将会使网络资源得到有效安排，服务质量（QoS）得到改善，以AI为核心的高效服务供给研究也受到普遍重视。另一方面，一个AI智能模型的应用往往局限于某些场景或任务。例如，大型语言模型（LLMs）[1]在各种自然语言处理（NLP）和计算机视觉任务中表现出色，而在实际场景中，要使预训练LLM遵循人类意图生成个性化输出，则需要对LLM进行微调[2]。仅在集中式云端部署LLM以寻求模型的个性化微调将为云端带来多个完整模型参数副本，因此是一种低效的做法。

为了改善LLM的个性化问题，找到合适的云边协同方法至关重要[3]。与仅在云端部署LLM的方法相比，用云边协同的方式来部署大模型有多重优点。这种方法能赋予边缘服务器较大的自由度，从而可以部署多种微调的LLM，适应环境差异以实现服务个性化、定制化。同时，这种云边协同可以将数据丰富的生成式设备连接到更多邻近服务器上，从而减少向更多远程云服务器上传数据的延迟，并节省通信开销。把生成式LLM融入边缘网络有望促进通信和计算（C&C）资源的高效使用。

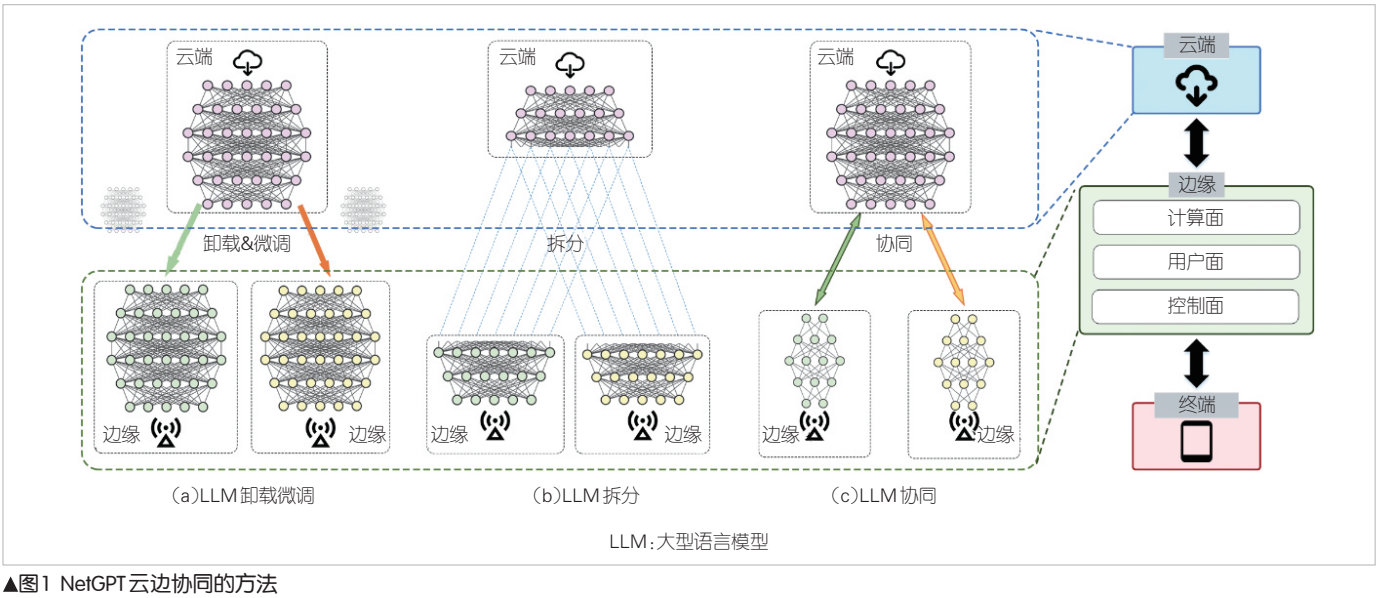
基金项目：国家自然科学基金项目（62071425）；浙江省“领雁”计划项目（2022C01093）；浙江省杰出青年基金项目（LR23F010005）

如图1所示，有几种方法可以实现LLM的云边协同部署，如本地微调、模型拆分等。具体而言，本地边缘服务器可通过卸载云端训练的LLM来定制LLM，实现个性化、定制化服务，满足用户喜好及场景需求。在此场景下，联邦学习或并行训练可作为一种辅助手段来实现调优[2, 4]。但对完整LLM的重复微调意味着庞大的计算，并有可能会给模型开发人员带来知识产权上的困扰，因此该方法在实际应用中也存在着诸多问题。同时，对边缘处的LLM整体进行强制拟合可能会使边缘服务器受限，计算资源紧张，从而导致边缘计算所需开销过于庞大。另外，卸载LLM也会产生明显的通信开销。另一种方案是将LLM拆分后部署到云端和边缘服务器，通过在边缘部署一些大规模深度神经网络（DNN）层，将剩余的层留给云端，从而有效平衡边缘服务器和云端服务器间的计算资源。在模型划分中，如何有效地将DNNs从边缘到云端进行划分是最具挑战性的问题之一，这需要在最小化端到端时延的同时，为边缘服务器保留足够小的模型尺寸[4]。考虑到典型的LLM中有数十亿个参数，这样的模型划分可能会非常复杂，LLM中广泛采用的参差链接也可能会限制合适划分点的选择。另外，LLM可能会泄露训练数据中的隐私细节[5]，因此直接以局部微调与模型分割的方法来实现云边协同，也会存在一定的挑战。

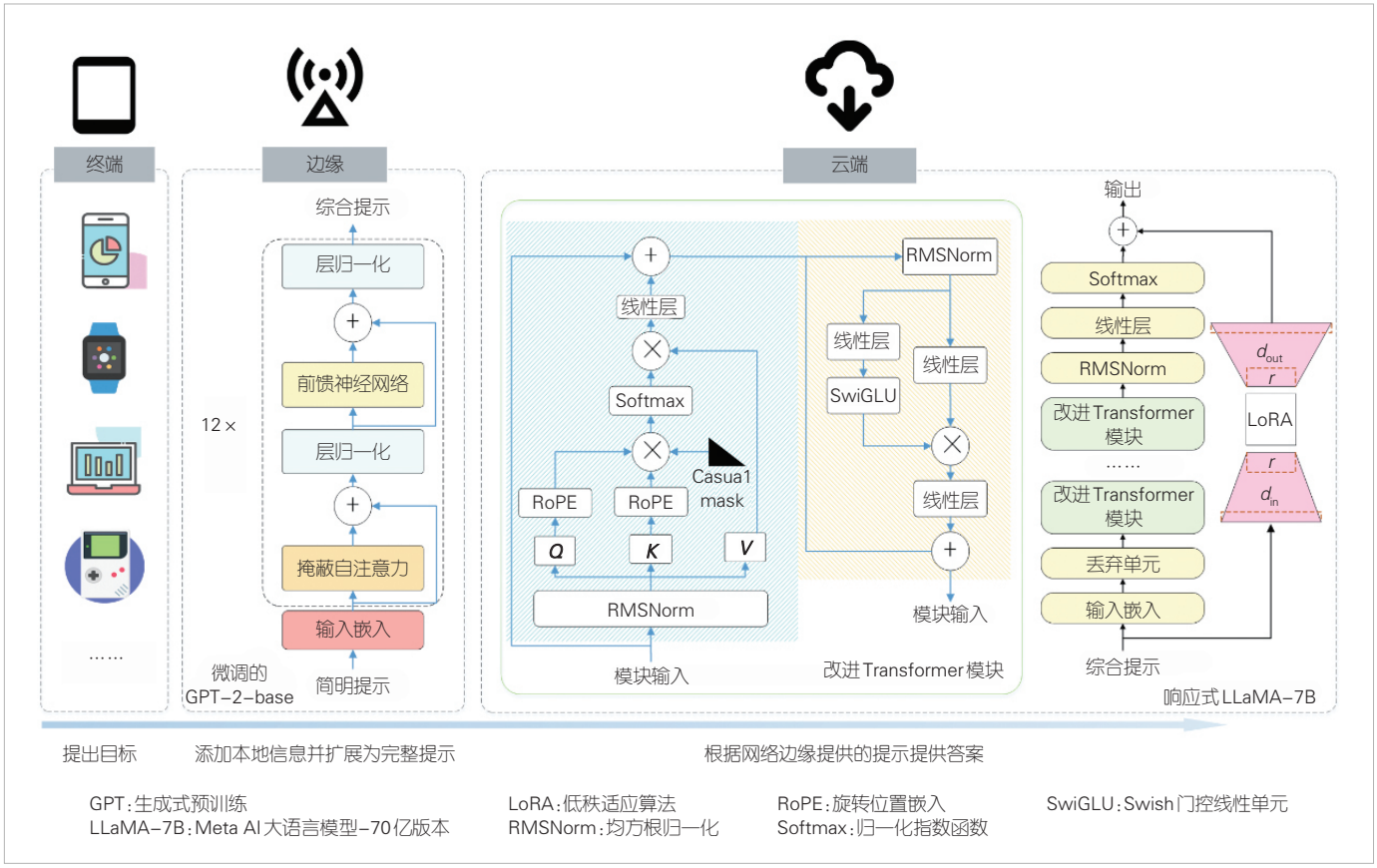
本文中，我们提出内生智能网络架构NetGPT，基于云边不均衡的资源分布，实现了边缘与云端之间不同尺寸功能性LLM的协同。与具有解耦C&C资源的AI外生网络明显不同，NetGPT能够使用融合C&C对边缘部署更小的LLM，对云端部署更大的LLM，以进行有目的的云边协同计算，从而提供个性化的内容生成服务。此外，NetGPT还集成并发展了有逻辑的AI工作流，以识别具有相同性能的通信链路。例如，在NetGPT中，假设边缘LLM提供满意的内容，那么性能驱动的通信链路会在边缘终止以加速响应。否则，在即时学习[6]理念的影响下，边缘LLM能够推断上下文并主动附加（或填充）部分个性化信息，从而在云端得到更加全面的效果。同时，边缘LLM有助于为智能网络管理和调度（例如用户意图推断、流行度预测等）提供统一的解决方案。因此，NetGPT符合C&C深度融合的发展趋势，代表了一种由LLM驱动的内生智能网络架构。

1 NetGPT的实现

我们提出了一个云边协同框架，具体如图2所示。通过在云端和边缘（如基站）使用不同的预训练LLM，可以完成个性化的生成服务。受限于开源LLM的可用性，我们选择并部署了LLaMA-7B模型（Meta AI大语言模型-70亿版本）[7]和基于GPT-2的模型。这两个模型分别由大约67亿个和1亿个参数组成，并部署在云端和边缘。值得注意的是，NetGPT可以根据需要使用其他LLM。基于此，我们对云边LLM协同NetGPT的实现细节进行了深入研究。首先，我们对作为LLM基础的Transformer进行了总体概述，并列出了LLaMA-7B模型和GPT2基础模型两个LLM的详细DNN结构。随后，我们探讨了在计算受限的设备上微调LLM的有效途径，并且展示了其对基于位置的个性化生成服务的效果。



▲图1 NetGPT云边协同的方法



▲图2 面向NetGPT的云边协同计算框架

1.1 Transformer概述

Transformer已被广泛用作LLM中多层解码器的基础模型。Transformer是通过使用多层自注意力和前馈神经网络（FNN）来构建DNN结构。自注意力依赖于由query、key和value矩阵（即$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$）定义的参数化注意力头，通过推导不同的权重并将其分配给序列中的不同位置来计算输入序列内的内部相关性。在FNN中，每个位置的表示使用的是非线性变换。此外，Transformer采用层归一化等技术来缓解梯度消失的问题。

1.2 边缘和云端LLM的DNN结构

1.2.1 GPT-2-base模型的DNN结构

GPT-2-base模型是GPT-2系列中最小参数的版本，包含了原Transformer结构的12层堆叠（即8头自注意力子层和FNN子层）。该模型利用正余弦位置的固定绝对位置编码方式来预变换输入序列。此外，GPT-2使用修正线性单元（ReLU）来激活函数。该模型具有相对优异的性能和较低的计算要求，因此适合部署在网络边缘。

1.2.2 LLaMA模型的DNN结构

LLaMA经过大量无标签数据的训练，非常适合下游任务的微调，同时也有多种参数版本[7]。与GPT-3相比，LLaMA结合了多项特定增强功能，从而在保持相似性能的同时显著减少了参数数量[7]。为了提高训练稳定性，LLaMA采用了对各子层的输入进行归一化而非对输出进行归一化的方式。此外，LLaMA使用一种简化的替代方案，即采用均方根层归一化（RMSNorm）函数[8]，利用均方根而非标准差进行归一化处理。此外，RMSNorm引入了可学习的缩放因子进行自适应特征缩放，从而增强具有不同值域的各种特征的归一化效果。其次，LLaMA用Swish-Gated线性单元（SwiGLU）[9]取代了ReLU激活函数。该激活函数将Swish函数（即$f_{\text{Swish}}(x)=x\cdot\sigma(\beta x)$，其中$\sigma(x)=\frac{1}{1+e^{-x}}$，$\beta$为可训练参数）和门控线性单元（GLU）（即$f_{\text{GLU}}(x)=x\cdot\sigma(Wx+b)$，$W$和$b$为可训练参数）相结合，从而可以根据输入以更有选择性的方式激活神经元，在输入发生改变时更加平滑，以有效捕获复杂的非线性关系。最后，LLaMA引入了旋转位置嵌入（RoPE）[10]，利用预设的旋转矩阵对位置信息进行编

码，自然地将显式相对位置依赖性纳入自注意力公式中。相较于给序列中每个位置分配独特编码表示的句对位置编码方式，RoPE中所采用的相对位置编码方式能更有效地对上下文信息中的远程依赖关系进行建模，易于直观理解，并且在实验中表现出优秀的性能。

1.3 微调技术

1.3.1 基于低秩自适应的轻量级微调

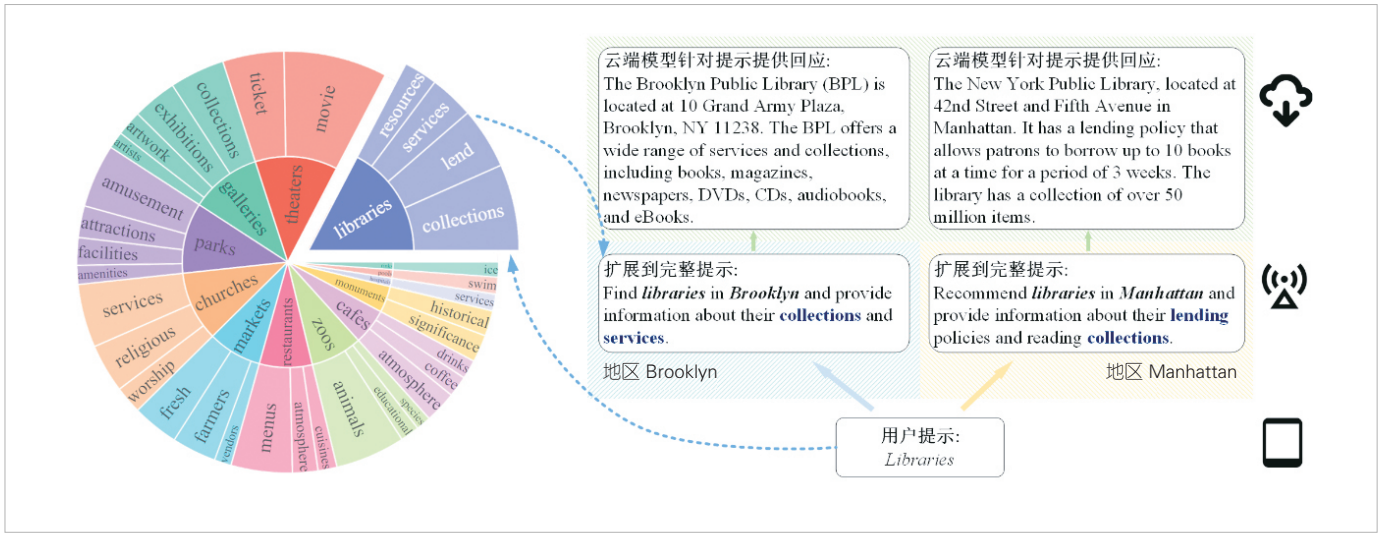
LLaMA缺乏生成响应式文本的能力[7]，因此仍需进行微调，但直接微调会耗费大量的计算资源。例如，直接微调LLaMA-7B模型需要112 GB视频随机存取存储器（VRAM），这超过NVIDIA A100 Tensor Core GPU的容量。因此，我们采用低秩适应（LoRA）技术[11]在消费级硬件上进行参数高效的微调。为了微调完整的参数矩阵 $\boldsymbol{W} \in \mathbb{R}^{d_{in} \times d_{out}}$，LoRA添加了一条旁路路径，通过使用两个具有内在秩 r 的下游参数矩阵 $\boldsymbol{A} \in \mathbb{R}^{d_{in} \times r}$ 和 $\boldsymbol{B} \in \mathbb{R}^{r \times d_{out}}$ 来模拟矩阵更新 $\Delta \boldsymbol{W}$。也就是说，在 $r \ll min(\boldsymbol{d}_{in}, \boldsymbol{d}_{out})$ 的条件下，LoRA成功地将大型参数矩阵 ΔW 转换为 $\Delta \boldsymbol{W} \approx AB$ 的低秩矩阵。实验证明，微调后的LLaMA-7B模型只消耗了28 GB VRAM而没有明显增加训练时间。另外，微调后需要的存储空间由12.55 GB明显缩小至50 MB左右[1]。基于LoRA，我们使用Stanford Alpaca数据集[12]对LLaMA-7B模型进行微调，成功得到响应式LLaMA-7B模型。

1.3.2 在LLM引导下收集数据

为了实现个性化边缘LLM，GPT-2-base模型需要通过附加基于位置的额外信息对“简明提示”进行扩展。本质上，基于5G接入与移动管理功能（AMF）所储存的附属BS位置能够很容易得到定位信息。同时，为了补充更全面的信息，本文中我们用self-instruct方法[13]，即使用OpenAI的TextDavinci-003模型来生成有效的文本样本。在每个地区，我们使用一组手动编写的位置相关提示与OpenAI的Text-Davinci-003模型进行交互，并将生成响应文本作为“综合提示”。与此相对应，可以收集一系列“简明提示”和“综合提示”之间的映射关系。考虑到边缘LLM的规模和任务复杂性，我们收集了一个包含大约4 000个样本的数据集，将GPT-2的模型微调为即时完成模型。值得注意的是，针对高通用性的场景，可以使用更大规模的LLM来增强边缘LLM，也可以采用LoRA等微调技术。

1.4 性能展示

图3显示的是NetGPT的性能。其中，边缘LLM能够根据左侧的图表补充“简明提示”，从而产生每一个对应的“综合提示”。不同基站也加入基于位置的个性化信息来满足不同需求。随后，边缘LLM对用户提出的“简明提示”进行处理，并向云端反馈补充提示信息，之后云端LLM再给出完整的回应。从图3右侧可看出，NetGPT能产生基于位置的不同回应，这反映出通过对云边进行高效的协同处理可以实现个性化的服务生成。



▲图3 NetGPT基于位置的个性化生成服务的性能展示

1 此类统计数据是在 $r = 8$ 且与学习率相关的标量因子等于16的配置下获得的。

2 面向NetGPT的内生智能网络架构

我们认为，NetGPT为蜂窝网络向内生智能网络架构的转变提供了机会，给用户带来个性化、网络化以及包容性智能，并赋予用户更多权限，以随时随地访问生成服务。尽管如此，这一转变是需要付出代价的。这需要对网络架构做出实质上的改变，不仅仅是在边缘位置安装服务器机架，以及在本地中断流量以进行边缘处理。相对于传统面向连接通信系统，典型服务在两个具体终端间建立了连接，通信源与目的地是最终用户清晰界定的。NetGPT需通过更加隐性的方式来建立产生性能驱动的连接。此外，NetGPT涉及更频繁的数据收集和处理模块，因此计算资源应保持一致。如图4所示，NetGPT需要在无线接入网络（RANs）中设计深度融合的C&C。在这些创新功能之外，NetGPT还需要设计逻辑AI工作流来创建（超越）生成服务的编排。

2.1 RAN中的融合C&C

为了有效地组织同时覆盖地面通信和非地面通信的异构资源，用于NetGPT的RAN首先必须提供用于无缝连接控制的控制面（CP），以支持用户平面（UP）中的提示和生成信息传输。这些组件可按5G及5G的先进技术来开发。此外，引入一个独立的计算平面（CmP）来协调计算资源，并执行AI相关功能以促进生成服务的部署和更新也是非常必要的。

2.2 数据处理和隐私保护

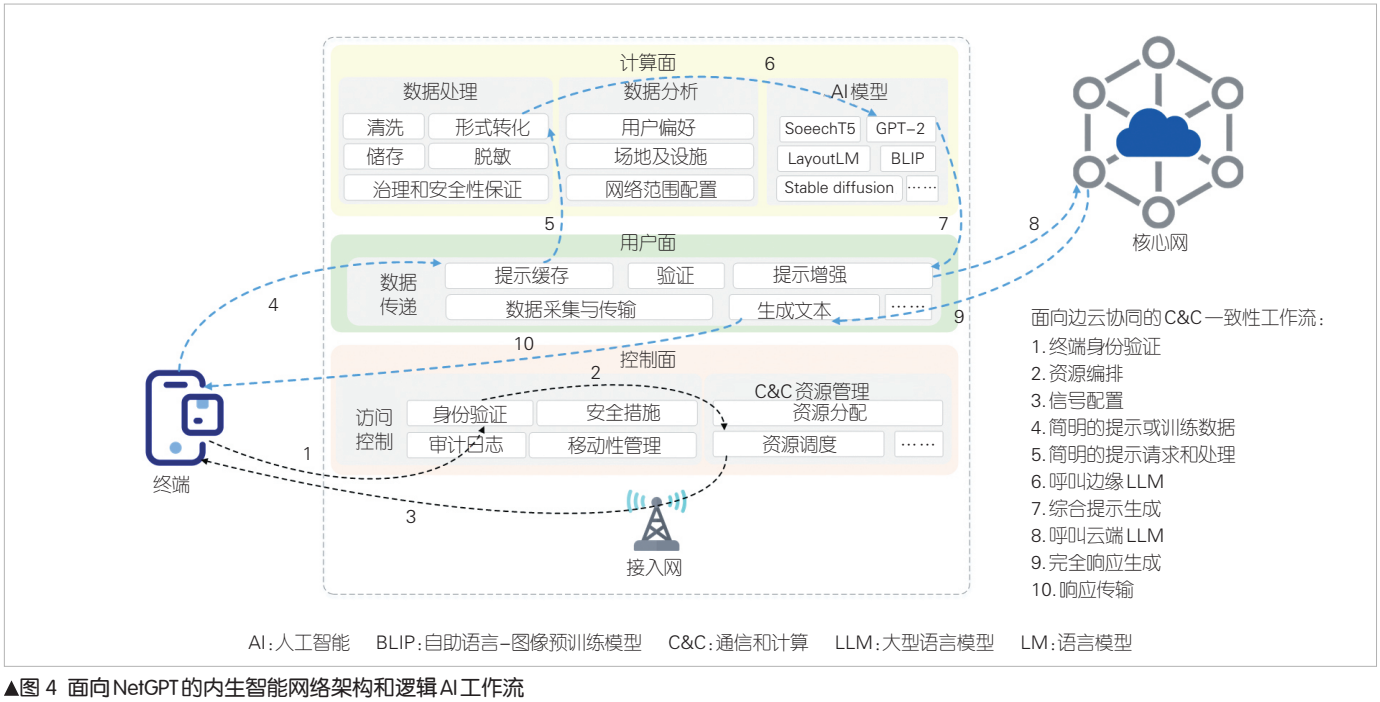
如1.2节所述，数据处理（如数据收集和微调）技术的广泛使用为LLM的产生提供了依据。在采集并储存数据的同时，引入数据脱敏模块也非常重要，这样可以规避隐私风险并对嵌入数据进行隐私防护。同时，数据策略执行模块将默认执行，并根据监管和非监管规则（如地理限制）处理数据，以确保数据处理的完整性和合法性。另外，基于监管与数据使用政策，可设计若干数据处理模型库，并设置适当的访问权限控制，使得有需要的实体能够调用这些数据服务。

2.3 个性化分析

为了创建高度自定义NetGPT，应增强对面向位置相关信息的分析，以支持个性化生成AI服务的定义和操作。例如，一方面，可以专门收集该地区的场地和设施信息来训练边缘LLM；另一方面，用户服务节点（USN）也可以包含最终用户级别的定制服务，以满足多样化的客户需求，同时也能进一步支持用户档案的建立和所联系终端的表征。

2.4 C&C资源管理

NetGPT资源编排是未来蜂窝网络供应服务的组成部分，它与其他网络服务资源编排具有一定的相似性，包括建立连接和计算资源分配。但由于其涉及的资源范围跨越云到终端等分布式节点，这也带来了一些挑战。因此，需要在无线控制信令或无线数据协议上携带新的协议栈来传输AI生成消



▲图4 面向NetGPT的内生智能网络架构和逻辑AI工作流

息并实现模型的更新和分发。

2.5 逻辑AI工作流

为了有效地提供AI服务，需要首先制定部分逻辑AI工作流以解析并安排NetGPT服务。值得注意的是，逻辑AI工作流能够协调分散在边缘与云端的一系列网络功能，连续地提供“简明提示”“综合提示”和“生成响应”，规范数据处理与分析，从而在CmP中进行个性化LLM训练。另外，还需要在服务部署过程中将逻辑AI工作流映射为物理资源，来满足有关服务QoS需求。由于工作流程涉及许多网络功能，其处理过程可以建立在串行或有向无环图上。另一方面，逻辑AI工作流并不局限于产生服务。通过更为可定制的途径，逻辑AI工作流极大地促进了QoS的改进。

3 基于LLM的网络管理与协调统一解决方案

除了提供个性化的生成服务外，面向NetGPT的内生智能网络架构还可以在边缘LLM上为智能网络管理和编排提供统一的解决方案。

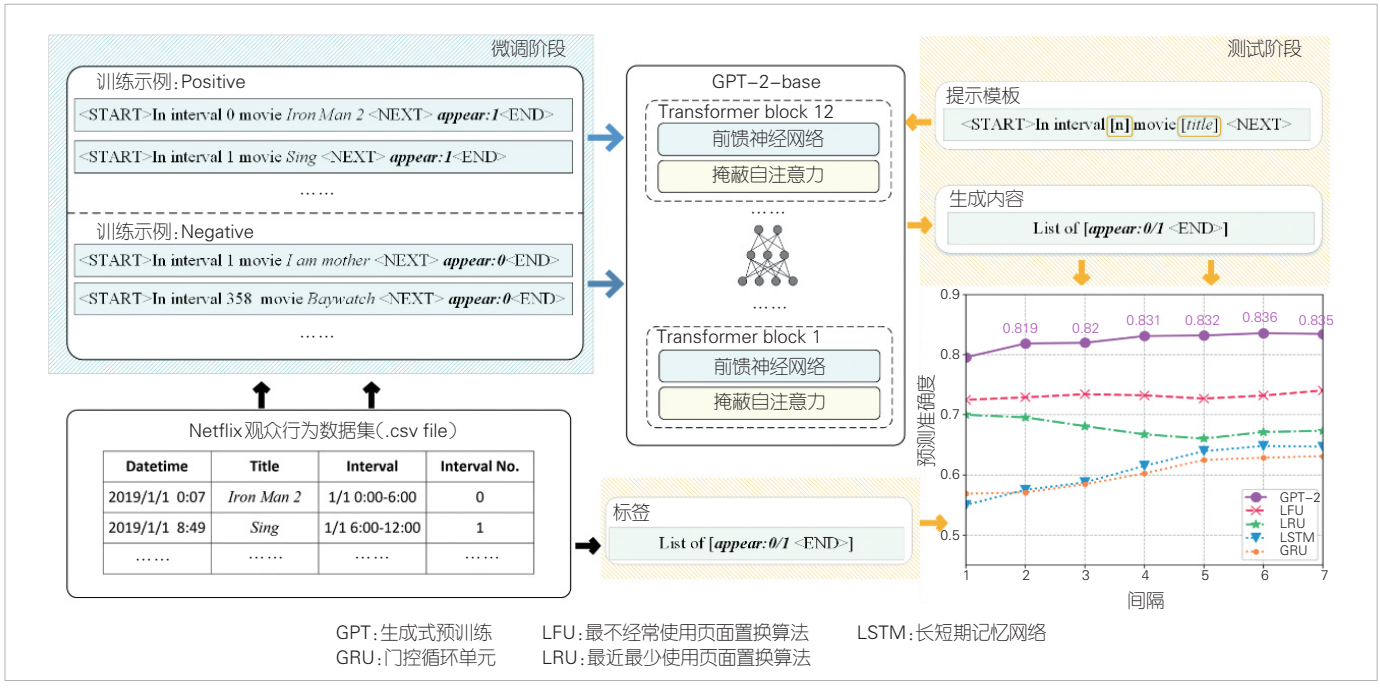
3.1 流行度预测

通过让C&C资源适应预测的需求，流行度预测可以显著提高网络效率[14]。鉴于DNN结构基本原理，GPT-2能够从RAN的附属终端历史访问记录中解释用户的偏好。另外，通过结合特定地点的数据，边缘LLM可以更好地捕捉每个地区独有的个性化特征。

为了证明边缘LLM（即GPT-2-base模型）的预测能力，我们以Netflix观众行为数据集为例进行了实验。结果表明，该方法能够有效地预测观众行为。为了解决数据稀疏性问题，我们先根据6 h循环对时间范围进行区间划分和编号标注；然后，根据每一部电影在特定区间内的存在情况，选取频率最高的20部电影进行标注。得益于CmP中的数据格式功能，相关的历史信息被转换为符合特定模板的一些自然语言。例如，“In interval 1，movie ‘Iron Man 2’ appear：1”表示电影《Iron Man 2》出现在区间1，与图5左下方所给的某一具体日期时间相对应。同时我们添加了特殊的标记来创建提示模板，从而帮助语言模型进行信息理解和生成响应。边缘LLM可在直接微调后根据提示模板的格式生成标签，即电影是否出现在间隔下。此外，为了增强模型的通用性，我们专门利用数据集中上半年的数据进行实验，并将95%的内容作为训练集，剩余的5%作为测试集。图5最后给出了边缘LLM的预测精度。可以观察到，针对这一任务，GPT-2显示了一个可以接受的精度水平，并明显优于其他经典算法。

3.2 意图推断

意图驱动网络旨在解决基于模板的服务在垂直业务中应



GPT：生成式预训练　LFU：最不经常使用页面置换算法　LSTM：长短期记忆网络
GRU：门控循环单元　LRU：最近最少使用页面置换算法

▲图5 流行度预测的边缘大型语言模型

用时难度增加的问题。意图驱动网络需要在取代人工配置网络和对网络问题做出反应之前，充分理解客户的实时需求[15]。如何准确地理解客户的意图，并将其转化为可行的网络配置，是最根本的问题之一。边缘LLM能够满足这种意图识别过程[15]，并精确理解口头陈述。通过采用类似于上文提到的微调方法，边缘LLM可以理解和提取任意自然语言输入中的关键词。值得注意的是，这样的微调可以很容易地完成。在实验中，我们仅使用了4 000个输入关键字对即可实现优秀的效果。图6展示了GPT-2-base识别客户意图的相应能力。从图6可以看出，当一个用户想要建立从Access 1到Cloud 2的10 Gbit/s连接并提供流量保护时，GPT-2-base模型能够很方便地提取出所需的关键词，且不受语句差异的影响。因此，GPT-2-base模型避免了繁琐的模板设计和客户学习过程。相对于传统NLP工具，LLM显示了更加强大的意图驱动网络实现功能。

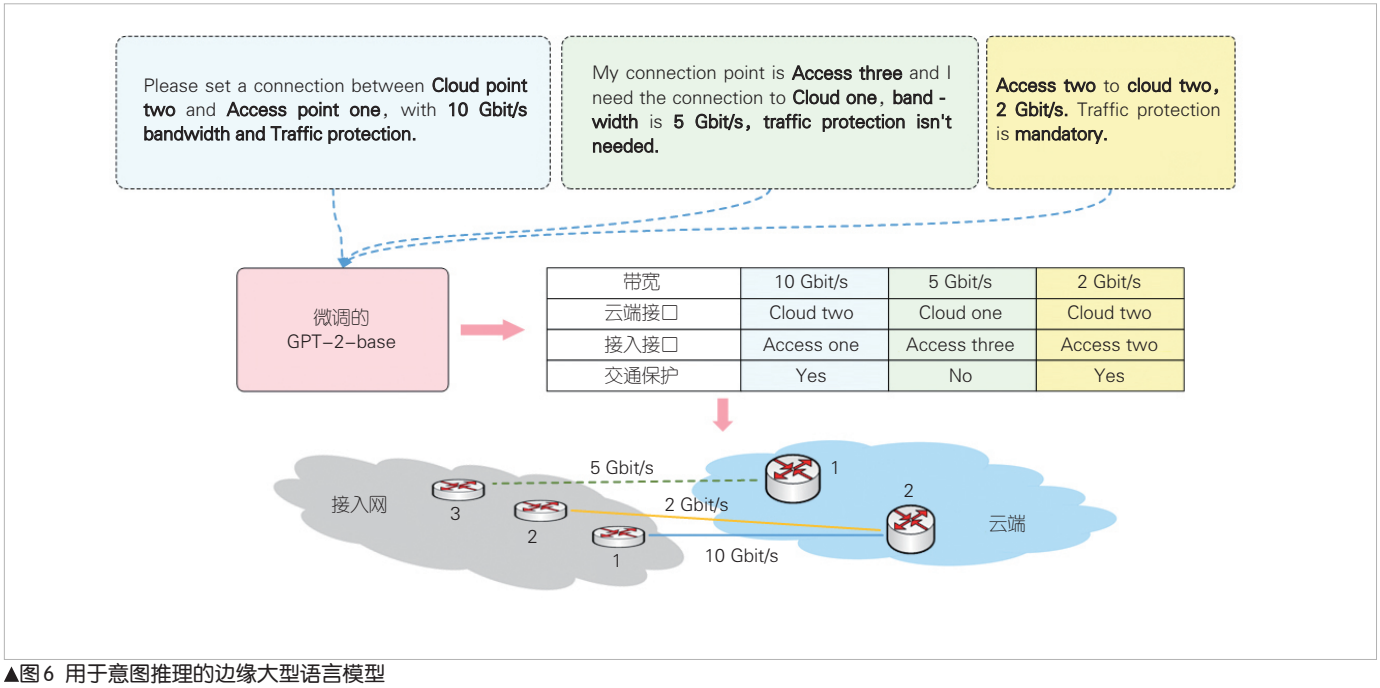
4 结束语

本文中，基于LLM，我们提出了一种内生智能网络架构NetGPT，从而提供超越个性化生成内容的网络服务。通过有效的云边协同，我们在边缘和云端部署一些具有代表性的开源LLM（如GPT-2-base模型和LLaMA模型）。在采用基于LoRA的参数高效微调技术后，我们再对开源LLM的一致性进行评估。实验证明，NetGPT用于基于位置的个性化服务是可行的。除此之外，我们还强调了NetGPT所需的一些实质性架构更改（例如，深度C&C集成和逻辑AI工作流）。在此基础上，我们提出一个可能的边缘LLM许可网络管理与编排统一人工智能解决方案。

尽管NetGPT是一种很有前景的架构，但在本文中我们并未讨论其面临的研究挑战。要成功地部署NetGPT，必须解决如下几个问题：

- 如何在设备终端上实现推理和微调，以符合受限成本中计算能力的实质性约束？
- 如何模仿新必应[2]实现LLM在线学习来满足边缘无线环境动态性要求？
- 由于数值推理的灵敏度有限，可能存在幻觉效应，如何进一步提高LLM的严谨性，从最新的LLM中可以得到哪些启示？同时如何将LLM评估指标与逻辑AI工作流相结合导出适当的工作流？
- 除了更高层的网络优化，如何开发基于LLM的物理和无线电链路层，体现LLM的优势？如何使用LLM实现低延迟、超可靠通信？如何对无线通信系统进行优化，以实现LLM在将来的网络上的有效部署与操作？

我们期待着未来有更多的研究工作能在这个领域展开，从而为LLM和网络架构的结合以及内生智能网络的构建带



▲图6 用于意图推理的边缘大型语言模型

2 新必应是指GPT授权的搜索引擎，可通过https://www.Bing.com/New访问。

来更多的创新和突破。

参考文献

[1] OpenAI. GPT-4 technical report [EB/OL]. [2023-08-10]. http://arxiv.org/abs/2303.08774

[2] ZHANG J Y, VAHIDIAN S, KUO M, et al. Towards building the federated GPT: federated instruction tuning [EB/OL]. (2023-05-09) [2023-08-10]. http://arxiv.org/abs/2305.05644

[3] XU M, DU H, DUST N, et al. Unleashing the power of edge-cloud generative AI in mobile networks: a survey of AIGC services [EB/OL].(2023-03-28)[2023-08-10]. http://arxiv.org/abs/2303. 16129

[4] WU W, LI M S, QU K G, 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. DOI: 10.1109/JSAC.2023.3242704

[5] KANDPAL N, ERIC W, COLIN R, et al. Deduplicating training data mitigates privacy risks in language models [EB/OL]. (2022-02-14) [2023-08-10]. https://arxiv.org/abs/2202.06539

[6] LIU P F, YUAN W Z, FU J L, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing [J]. ACM computing surveys, 55(9): 1-35. DOI: 10.1145/3560815

[7] TOUVRON H, THIBAUT L, GAYTUER L, et al. LLaMA: open and efficient foundation language models [EB/OL]. [2023-08-10]. http://arxiv. org/abs/2302.13971

[8] ZHANG B, SENNRICH R. Root mean square layer normalization [EB/OL]. (2019-10-16)[2023-08-12]. https://arxiv.org/abs/1910.07467

[9] SHAZEER N. GLU variants improve transformer [EB/OL]. (2020-02-12) [2023-08-10]. https://arxiv.org/abs/2002.05202

[10] SU J, LU Y, PAN S, et al. RoFormer: Enhanced transformer with rotary position embedding [EB/OL]. (2022-08-09)[2023-08-12]. https://arxiv.org/abs/2104.09864

[11] SU J, LU Y, PAN S, et al. LoRA: low-rank adaptation of large language models [EB/OL]. (2021-04-20) [2023-08-12]. https://arxiv. org/abs/2104.09864

[12] TAORI R, GULRAJANI I. Stanford alpaca: an instruction-following llama model [EB/OL]. [2023-08-10]. https://github. com/tatsu-lab/stanford alpaca

[13] WANG Y, KORDI Y. Self-instruct: aligning language model with self generated instructions [EB/OL]. (2022-12-20) [2023-08-12]. https://arxiv.org/abs/2212.10560

[14] ZHU J H, LI R P, DING G, et al. AoI-based temporal attention graph neural network for popularity prediction and content caching [EB/OL]. (2022-08-18)[2023-08-10]. https://arxiv.org/abs/2208.08606v1

[15] PANG L, YANG C, CHEN D, et al. A survey on intent-driven networks [J]. IEEE acess, 2020, 8: 22862 - 22873. DOI: 10.1109/ACCESS.2020.2969208

作 者 简 介

陈宇轩，浙江大学在读博士研究生；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，并获得吴文俊人工智能优秀青年奖、江苏省科学技术一等奖等。

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》最佳论文奖等；发表论文265余篇，拥有IEEE 802.15UWB国际标准提案16项、国际专利3项。