面向 AI 大模型的网络使能技术

Network Enabling Technologies for Artificial Intelligence Large Models

目录

摘要

随着大模型和智能体（Artificial intelligence agent, AI agent）技术的发展，未来越来越多的工作将被基于大模型的智能体所取代。一方面，由于大模型对数据和算力的需求巨大，资源受限的终端将难以满足模型训练和推理的需求。另一方面，未来第六代移动通信（Six generation, 6G）网络存在大量低时延需求的价值场景，例如无人驾驶、虚拟和增强现实等，云端大模型难以满足这些场景用户的需求。因此，向无线网络寻求算力和数据的支撑将成为大模型时代的必然。本文介绍了大模型时代下网络使能人工智能（Artificial intelligence, AI）技术的需求和驱动力，详细阐述了未来 6G 网络能为大模型提供的 AI 服务，包括数据感知、分布式训练、指令优化、端边云协同推理和模型优化等，通过案例分析说明了相关技术的实践应用，并总结了未来可能的研究方向和所需要面对的挑战。

一、AI 大模型发展概述

（一）发展历程

随着深度学习技术的应用范围不断拓展和人工智能的快速发展，在大数据、高算力和强算法等关键技术的共同推动下，以 ChatGPT 为代表的 AI 大模型大量涌现，提供了高度智能化的人机交互体验和极富创造力的内容生成能力，改变了人们的工作和生活方式，实现了 AI 技术从“量变”到“质变”的跨越。

AI 大模型是指拥有超大规模参数、超强计算资源的机器学习模型，能够处理海量数据，并完成各种复杂任务。AI 大模型的发展可以追溯到 20 世纪 50 年代。此后，从卷积神经网络（CNN）到循环神经网络（RNN），再到 Transformer 架构，模型的性能不断提升。总的来说，AI 大模型的发展历程主要可以分为四个阶段，如图 1 所示。

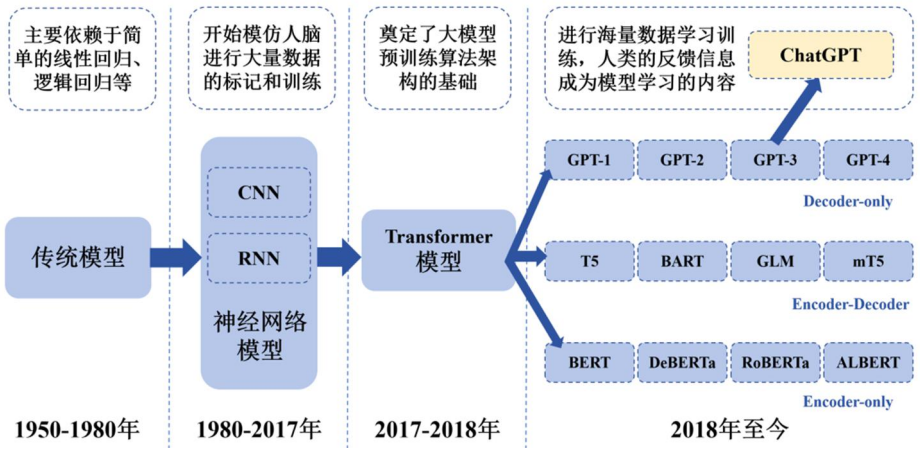

图 1. AI 大模型的发展历程

►传统模型（1950-1980）：在 AI 发展的早期，传统模型主要依赖于简单的线性回归、逻辑回归等方法。这些模型能够处理分类和回归等基本任务，但在处理复杂数据和任务时表现有限。

►神经网络模型（1980-2017）：1980 年，卷积神经网络的雏形 CNN 诞生。2000 年代初期，有学者开始研究神经网络模型，开始模仿人脑进行大量数据的标记和训练，并尝试解决简单的问题，如手写数字识别等。

►Transformer 模型（2017-2018）：2017 年，Google 颠覆性地提出了基于自注意力机制的 Transformer 架构，奠定了大模型预训练算法架构的基础。2018 年，OpenAI 和 Google 分别发布了 GPT-1 与 BERT 大模型，使得 NLP 领域的大模型性能得到了质的飞跃。

►现代 AI 大模型（2018 至今）：2022 年，聊天机器人 ChatGPT 横空出世，迅速引爆互联网。此后发布的多模态预训练大模型 GPT-4，再次引发了生成式 AI 的热潮。目前各类大模型正持续涌现，性能也在不断提升。

（二）发展趋势

1. 多模态能力提升，应用场景范围扩大

单模态模型通常只能处理一种类型的数据，例如文本、图像或声音，缺乏对复杂环境的全面理解。而具有多模态能力的 AI 模型能够同时处理多种类型的数据，例如将视觉和语言信息相结合，以实现更深层次的理解和交互，并在更广泛的场景中得到应用。

2. 模型轻量化部署，资源需求成本降低

在 AI 技术快速发展的当下，智能手机等移动设备在人机交互、语音交流等功能方面的需求不断提升，将大模型轻量化部署到终端设备也正成为一个重要的研究方向和发展趋势。利用端侧 AI 可以更好地为用户提供个性化的服务和支持，帮助用户进行自我管理，实现更加智能和高效的设备互联。

3. 外部工具相结合，交互方式更加智能

传统的小模型通常专注于特定的任务，缺乏与外部环境交互的能力。结合外部工具调用、记忆和规划功能的 AI 大模型，可以被视为智能代理（Agent），它们能够执行更加复杂的任务，如自主决策、规划和学习。这种模型的交互方式更加智能，能够根据用户的需求和偏好进行自我调整，提供更加个性化的服务。

这些发展趋势不仅预示着 AI 技术的不断进步，也反映了用户对于更加智能、个性化服务的需求。随着研究的深入和技术的成熟，我们可以期待 AI 大模型在未来将在更多领域发挥关键作用，改善人们的生活和工作效率。

二、网络使能大模型的需求和驱动力

（一） 未来 6G 网络的通算智融合趋势

人工智能已成为新一轮产业升级的核心驱动力，各行业的自动化、数字化、智能化需要泛在智能，许多高价值的 AI 场景，例如 AI 手机、自动驾驶、智能制造、移动机器人等，具有移动性、实时性、边端协同、隐私性等要求，需要网络这一 AI 服务基础设施进行支持。而随着大模型技术在上述场景中的深入应用，终端对于网络侧算力和数据资源支撑的需求将进一步扩大。

ITU 将 6G 场景扩展到包括通信与 AI 融合在内的智慧泛在，需要将 AI 打造成 6G 通信网络的新能力和新服务，实现 AI 即服务(AI as a service，AIaaS)。这要求 6G 网络能够随时随地提供 AI 服务、支持低时延的推理和训练服务、支持移动式 AI、保障 AI 服务质量、提供安全隐私保护。

（二）网络使能大模型价值场景

1. AI 手机

在当今科技快速发展的时代，手机大模型已成为各大厂商竞相研发的热点。各大手机厂商纷纷推出了自家的大模型，为用户带来更加智能化的体验。如表 1 所示，手机大模型的功能主要包括文字类和图像类。在文字类功能方面，用户可以享受到智能问答、文本创作、文本总结、通话摘要等便捷服务，这些功能的响应时延通常在 1 秒之内，让用户感受到即时的互动体验。而图像类功能包括文生图、图像消除、图片问答等，其中，文生图响应时间较长，一般在 5 秒以上。在模型部署方面，目前主要有端侧部署和云端部署两种方式。端侧部署的大模型参数量通常不超过 10B，这种部署方式可以更好地保护用户隐私，同时降低对网络环境的依赖。而云端部署的大模型参数量可达 100B 以上，这种部署方式可以充分利用云端强大的计算资源，提供更加复杂和强大的功能，但需要较为稳定的网络环境支持。

表 1. 各厂商大模型手机调研信息

品牌	大模型功能	大模型性能	参数量	部署位置
vivo[1]	-智能问答 -文本创作 -文本总结 -逻辑推理	-智能问答首词响应 1s -文本总结首词响应 ms 级	1B/7B	端侧
			70B/130B/175B	云端
OPPO[2]	-智能问答 -通话摘要 -文本总结 -图像消除 -文生图	-智能问答首字响应 0.2s -512*512 生图时长 6s	7B	端侧
			70B/180B	云端
荣耀[3]	- 智慧成片 - 一语查图	暂无公布数据	7B	端侧

小米[4]	- 智能问答 - 文生图 - 图片问答 - 图像消除/扩图	暂无公布数据	暂无公布数据	暂无公布数据
苹果[5]	-智能问答 -文本摘要 -重要消息置顶 -文生图 -图像消除 -跨应用操作	暂无公布数据	暂无公布数据	端侧
				云端

基于上述分析，手机大模型主要分为终端推理和云端推理两类。因此，6G 网络使能手机大模型也可以相应地分为使能终端推理和使能云端推理两类。

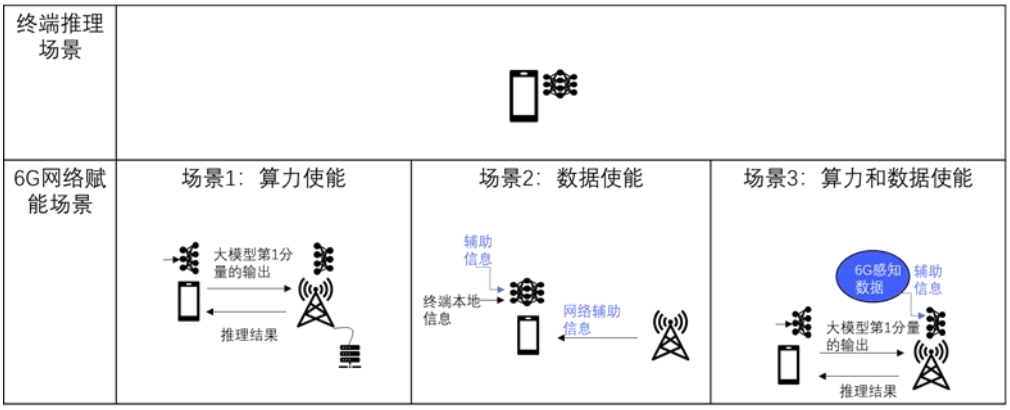

图 2. 6G 网络赋能大模型终端推理场景

如图 2 所示，6G 网络使能终端推理可以包括算力使能、数据使能以及算力和数据使能 3 种场景。考虑到目前手机大模型中文生图的时延较长的痛点，价值场景 1 是 6G 网络通过算力卸载的方式，将终端算力全部或部分卸载到 6G 网络内，通过对通信资源和算力资源的协同调度，可以降低响应时延，并降低终端推理功耗。而价值场景 2 则是 6G 网络通过例如感知获得价值数据，并将该价值数据作为终端推理的辅助信息，以提升推理精度。至于价值场景 3，则是网络同时提供算力和数据服务，从而可以降低终端推理的响应时延和功耗，并提升推理

准确度。

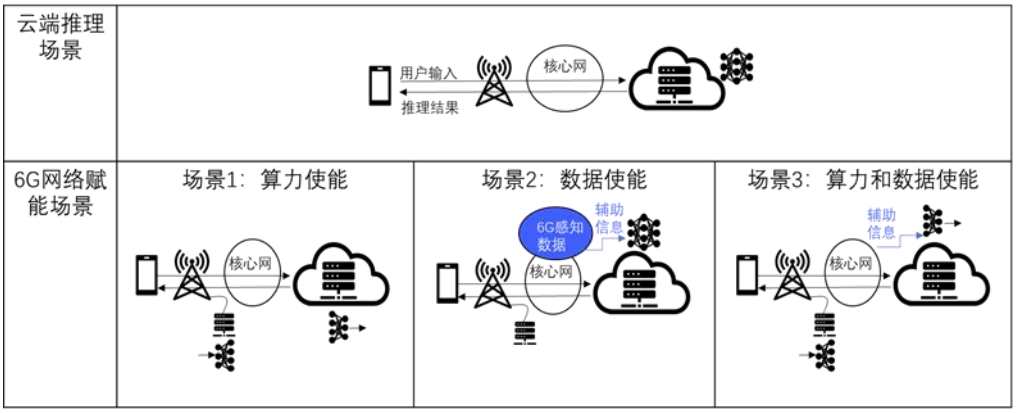

图 3. 6G 网络赋能大模型云端推理场景

如图 3 所示，6G 网络使能云端推理也可以包括算力使能、数据使能以及算力和数据使能 3 种场景。在价值场景 1 中，6G 网络通过算力卸载的方式，将云端算力全部或部分卸载到 6G 网络内，通过对通信资源和算力资源的协同调度，并通过更短的传输路径，可以显著降低响应时延，提升用户体验。而价值场景 2 则是 6G 网络通过例如感知获得价值数据，并将该价值数据作为云端推理的辅助信息，以提升推理精度。至于价值场景 3，则是网络同时提供算力和数据服务，可以同时降低云端推理时延，并提升云端推理精度，为用户带来更加高效和智能的服务体验。

2. 自动驾驶

自动驾驶车辆通过传感器（如摄像头、雷达、LIDAR）采集到大量感知周围环境数据，需实时处理和分析、进行路径规划和驾驶决策。将连接的车云系统扩展到分布式网络节点/基站环境中，使数据和应用程序可以更靠近车辆，提供快速的道路侧相关功能。终端设备采集传感器数据，进行初步处理和特征提取。在车辆附近的分布式边缘节点进行实时数据处理，如环境感知和初步路径规划，利用 6G 网络的低延迟特性，快速传播危险警告和延迟敏感信息，确保实时响应。

在中央网络节点/云端进行大规模模型训练和全局优化，利用大数据提升模型的准确性和鲁棒性。根据车辆位置和网络状况，可动态调整分布式网络各节点计算资源，确保高效运行。

3. 智能医疗

可实时监测患者的健康的医疗设备和穿戴设备收集大量患者体征数据，通过医疗大模型训练和推理，进行疾病预测和诊断。穿戴设备和医疗传感器采集生理数据，进行初步处理和传输。通过分布在医疗机构的边缘节点进行实时数据分析和初步诊断，减轻中央网络节点负担。中央网络节点进行复杂的医疗数据分析和模型训练，支持远程诊断和治疗方案的优化，通过高可靠性和低延迟的通信网络赋能医疗数据的实时传输和处理。除了实时传输能力和边缘节点部署能力，6G 网络还提供了高可靠和加密的数据隐私保护机制，保障患者的数据隐私和安全。

4. 工业 4.0

工业 4.0 要求智能工厂通过物联网设备进行设备监控、生产管理和质量控制，需要高精度、低延迟的数据传输和处理。工业传感器和设备采集生产数据，进行初步处理和传输。工厂内部的分布式网络节点部署计算，提供本地化的生产监控和实时优化能力，进行设备监控和故障预测。在中央网络节点进行大规模数据分析和模型训练，提升生产效率和产品质量。大带宽和低延迟的 6G 网络确保了生产数据在传感器、边缘网络节点及中央网络节点之间的实时传输和处理，高可靠性网络连接保障了生产过程的连续性和稳定性。

5. 工业元宇宙

工业元宇宙打造与现实工业映射和交互的全数字化虚拟世界，构建工业全生命周期的虚实共生、相互操作及高效闭环的工业体系新范式，推动传统行业数字

化智能化转型，是新质生产力的数字底座。在工业元宇宙中，虚拟世界与物理世界的深度融合是实现其全部潜力的关键。虚拟世界不仅要能够感知和接收来自物理世界的数据，还需要能够理解这些数据背后的意图，并据此做出合理的决策和控制。这一过程中，大模型显著提升了工业元宇宙的智能化和自主化水平。

虚拟世界对物理世界的理解是工业元宇宙虚实交互的核心任务之一。工业元宇宙需要处理海量的数据，包括物联网设备传感器的数据、生产线监控信息、供应链的实时动态等。通过传统的规则模板解析、机器学习算法和深度学习算法，虚拟系统可以分析物理设备的数据并做出响应。然而，这些方法通常需要大量的规则和参数配置，灵活性较差。大模型的引入，尤其是基于大模型的生成式人工智能，使得意图识别和理解更加灵活和高效。大模型通过自然语言处理和深度学习技术，能够高效地解析、分析和处理这些数据。通过对海量文本、图像和其他数据的训练，能够在没有明确规则的情况下识别出复杂场景中的意图，将其转化为可执行的操作指令或预测性分析结果。例如，在一个智能工厂中，生产设备通过传感器反馈数据，虚拟系统不仅能够监测设备运行状态，还可以理解操作员的工作意图，从而调整生产部署等。

大模型在虚拟世界的构建过程中起到了加速器的作用。工业元宇宙的构建不仅仅依赖于物理世界的数据输入，还需要大量的虚拟内容生成，诸如虚拟场景、产品设计、生产流程模拟等。文本、音频、视频等不同类型的数据可以被自动生成，这极大地提升了虚拟世界的丰富性和细节表现。在产品设计过程中，大模型使能的设计软件可以生成大量的设计方案和模型，大幅缩短了产品设计周期，同时提高了创新性。不仅加速了产品的迭代，还能推动工业设计从传统的线性流程向智能化、迭代式的流程转变。虚拟现实（VR）和增强现实（AR）技术是工业元宇宙的重要组成部分，而构建真实感强、细节丰富的虚拟场景往往需要大量的人工干预和资源投入。通过大模型和 AIGC 技术，虚拟场景的生成可以更加自动

化、智能化，极大提升了开发效率。在一个虚拟工厂中，AIGC 可以基于物理工厂的布局自动生成相应的三维模型，并根据实时数据动态调整场景的布局和功能。这种虚实交互和自动化生成能力，提升了虚拟世界的沉浸感，使得企业能够更灵活地进行生产规划和调整。

决策和控制是工业元宇宙的核心之一，大模型的自主学习和决策能力提升了工业元宇宙的智能化水平。在工业生产过程中，生产环境和工艺流程通常非常复杂，需要根据实时数据动态调整。大模型可以基于大规模的数据训练，学习到各种复杂场景下的最优策略，并通过持续学习不断优化，使得工业元宇宙中的虚拟系统与物理世界紧密互动，优化资源分配，最终实现更高效工业部署和生产。

单一的大模型往往难以全面覆盖所有工业元宇宙场景需求，需要 AI 大模型与小模型融合，形成更全面的智能工业元宇宙系统。视觉引擎、语音引擎和机器人控制引擎等不同领域的 AI 小模型可以与大模型协同工作，补充其在特定任务中的不足，形成一个多功能的、全覆盖的 AI 使能的工业元宇宙系统，适应更加复杂多变的工业环境。

三、网络使能大模型服务

表 2 给出了网络使能大模型服务和一般 AI 模型服务的对比，可以看出 6G 网络使能的大模型服务在实时性、动态调整和高可靠性方面的显著优势，能够更好地满足不同应用场景的需求，提高系统的整体性能和用户体验。

表 2. 网络使能大模型服务和一般 AI 模型服务的对比

对比项		**网络使能大模型服务**	**一般 AI 模型服务**
需求	**带宽**	处理的数据量更大，有更高的带宽需求：	数据传输量相对较小，带宽需求较低
	实时性	具有超低延迟的应用需求，在自动驾驶、实时视频处理等场景中，低延迟是关键	多数场景对实时性的要求较低，相对较高的延迟容忍性

能力	**大带宽**	提供更大传输带宽	现有资源带宽难以提升
	低延迟	利用 6G 网络的高带宽和低延迟，实现数据的实时传输和处理，支持实时分析和决策	主要依赖固定网络基础设施，数据传输和处理的实时性较低
	动态调度	利用智能调度系统，动态调整计算资源和任务分配	计算资源分配相对固定，难以动态调整和优化
	分布训练	广泛使用分布式数据并行和模型并行技术	通常在单个计算节点上完成训练和推理
	边缘计算	充分利用边缘计算能力，提高实时性和响应速度，减轻云端压力	边缘计算支持较少，主要依赖云端进行数据处理和模型训练

（一）数据感知服务

在未来的 6G 网络时代，一个引人注目的趋势是网络中将会部署大量的传感设备。这些传感设备以其高灵敏度、高精度和高覆盖率的特点，将实现对物理世界的全面感知和实时监测。无论是环境参数的测量、人体健康指标的监测，还是物体位置和运动状态的追踪，传感设备都能提供详尽而准确的数据。与此同时，随着人工智能技术的飞速发展，大型模型在各个领域的应用越来越广泛。然而，这些大模型的训练和推断过程都需要大量的数据作为支撑。数据是模型学习的基石，是提升模型性能的关键。

在 6G 网络中，传感设备所收集的数据正好能够满足这一需求。在模型推断阶段，传感数据可以作为输入信息，增强大模型的推理精度。由于传感设备能够实时获取和传输数据，模型可以基于最新的数据进行推断，从而更准确地反映实际情况。这不仅提高了模型的实用性，还增强了其应对复杂和多变环境的能力。而在模型训练阶段，传感数据同样具有不可替代的价值。通过将传感数据作为训练数据的一部分，可以丰富模型的训练样本，提高模型的泛化能力。同时，利用传感数据进行数据增强，还可以进一步提升模型的训练效果，使其在各种应用场景中都能表现出色。因此，未来 6G 网络中的传感设备将成为大型模型训练和推

断的重要数据源，为人工智能技术的发展提供强有力的支持。

（二）分布式训练服务

1. 分布式机器学习理论

随着“大数据”概念的兴起，数据量爆炸式增长，数据和算法双驱动的模型逐渐受到工业界和企业界的重视。大数据的特征可以概括为大数据量、多类型、低价值密度和数据在线。其中，大数据量指的是数据集的规模非常庞大，通常达到TB甚至PB级别，这种规模的数据量远超传统数据处理工具和单机系统的处理能力，同时如此规模的数据不仅自身数据量庞大，并且存在大量非结构化数据（如图片、视频），需要复杂的数据处理和整合方法。数据在线是指数据实时更新和变化，大规模数据自身数据量庞大的同时自身数据增长的速度也非常快，存在大量衍生数据，需要实时的监测、分析和处理。大数据量和数据在线是使得传统机器学习不能适应当前环境的主要因素。

传统机器学习即在单机内进行数据处理和计算，注重在单机内处理数据的速度，由于内存和单机算力的限制，大数据条件下庞大的数据存储和计算是无法在单机中做到的，因此，将计算模型分布式地部署到多台、多类型的机器上进行同时计算是一种必要的发展趋势。

基于以上原因，提出了分布式机器学习的概念。分布式机器学习研究将具有大规模数据量和计算量的任务分布式地部署到多台机器上，其核心思想是“分而治之”，即将数据集或是计算任务分解成多个小数据集或计算任务，分配到不同的计算节点上处理，有效提高了大规模数据计算的速度并节省了开销。

分布式机器学习在概念的提出时就展现了独特的优势，包括针对大数据量问题的处理海量数据和针对数据在线问题的实时数据处理。在其发展和成熟的过程中，在其他的一些方面也展现了优势，首先，分布式架构支持动态扩展计算资源，

可以根据具体的计算需求的变化灵活地调整计算节点的数量和计算任务的分配，切薄系统的高效运行。其次，分布式系统的架构就保证了整个系统具有较强的鲁棒性，能够在某个节点发生故障时自动进行任务再分配，避免了计算过程的前功尽弃，保证了计算过程的稳定性和连续性。最后，分布式系统联合了大量低成本的硬件资源和计算资源去解决复杂的梯度计算，显著降低了能源和资源成本。

分布式机器学习分为面向扩展性的分布式机器学习和面向隐私保护的分布式机器学习，这种分类主要是针对传统机器学习的不同限制因素进行改进。

1）面向扩展性的分布式机器学习

在近年的研究中，训练的数据规模和模型参数规模以指数形式增长，以卷积神经网络为代表的神经网络使用大量训练数据训练一个参数为千万量级甚至上亿的模型，所使用的计算资源和所消耗的时间成本不是单机所能够做到的。面向扩展性的分布式机器学习是这种情况的一个可行的解决方案，它专注于将机器学习的算法扩展到多个计算单元（如 GPU），尝试将无论是廉价但低效的计算资源或是高昂但高效的计算资源纳入自身体系中，通过并行和分布式计算来处理大规模的数据集和复杂的模型。

面向扩展性的分布式机器学习主要通过数据并行、模型并行或是混合并行的策略实现它的处理任务，不同策略的使用则是基于不同的任务，有着各自的优势和缺陷，下面将一一介绍。

在分布式机器学习技术乃至大数据技术中，数据并行都是最为常见的一种并行方式。在数据并行的策略中，数据集被分割成若干个子数据集，并加载在若干个训练设备中（如 GPU）。因此，数据并行主要需要实现数据集分割以及训练后的模型参数同步两个部分的设计，其中后者是设计时关注的关键问题。数据并行是一种实现简单，扩展性好，且对于绝大多数深度学习任务都适用的可以应用于分布式机器学习的策略。然而，数据并行的通信成本很大，对于中心节点（参数

服务器）的通信压力也较大。

模型并行是基于单个节点无法容纳完整模型的问题所提出的，它将模型的不同层或不同参数分配到不同的节点上，每个节点只计算模型的一部分。需要频繁的设备间通信来传递中间结果。模型并行的提出和使用都是基于模型参数量庞大，例如卷积神经网络，但是它的实现较为复杂，通信开销较大。流水线并行（管道并行），通常认为是模型并行的一种特殊形式，它将模型按层或模块顺序切分成多个阶段，每个阶段分配到不同的计算节点上，形成流水线。管道并行通过分阶段处理和数据流动，减少了单个节点的内存占用，但是对比于数据并行，实现上较为复杂。

2）面向隐私保护的分布式机器学习

面向隐私保护的分布式机器学习的主要目的则是保护用户隐私和数据安全，在面向隐私保护的 DML 中，数据的来源是多个参与方。有研究提到，在需要分布式机器学习技术来利用每个参与方的训练数据的时候，不同参与方的数据集可能具有不同的数据特征，所以实际中经常遇到的是训练数据的纵向划分，因此面向隐私保护的 DML 适用于具有纵向划分数据集的场景。

目前来看，其实不必纠结于面向隐私保护的 DML 适用于纵向划分的数据集还是横向划分的数据集，面向隐私保护的 DML 实际上提供了隐私保护的多条思路，在方法和实现上具有更大的覆盖范围，因此无论数据集的划分方式如何，都可以使用面向隐私保护的 DML 提供的隐私保护思路和方法。一方面，对于数据传输成本大的环境，如企业对用户（B2C），而数据敏感程度相对于政府部门和公司机密较低的情况，可以采用不直接共享数据的情况下进行机器学习模型的实现；另一方面，对于隐私要求严格，但对于传输环境安全要求不严格的情况，可以使用差分隐私等技术模糊化处理；如果对于隐私和数据安全都具有严格要求的情况下，如金融服务或是国家安全，可以采用同态加密和安全多方计算等技术。

从这个角度来说，联邦学习即是 DML 在隐私保护领域对于某一方面需求的继承与进一步发展。

2. 分布式机器学习平台和算法设计

分布式机器学习的主要研究方向包括分布式机器平台研究和分布式机器学习算法设计，在实际应用中，平台研究要结合算法的可行性，算法设计需要考虑在平台上的执行效果。

简单的例子是，一个由若干个计算节点和一个参数聚合服务器组成的分布式机器学习系统，每个计算节点都是一台机器，训练数据被分成若干个数据分片并发送给各个计算节点，计算节点在本地执行随机梯度下降算法，计算节点将梯度或者模型参数发送至参数服务器，参数服务器对收到的梯度或者模型参数进行聚合，从而得到全局梯度或者全局模型参数。

分布式机器学习的算法主要包括服务器如何将计算任务分配给每一个计算节点、计算节点在本地执行什么样的算法、参数服务器的全局参数如何聚合等。而分布式机器平台研究主要目的是搭建一个可以应用于多类型设备，搭载了分布式机器学习算法，并且具有优异性能的分布式深度学习框架。

1）分布式机器学习平台

分布式机器学习平台研究起步的时间实际上较早。2005 年，Apache 实现了 Hadoop 分布式系统的基础架构。在经过接近 20 年的发展后，出现了大量成熟的分布式机器学习平台。分布式机器学习平台包括基于数据流模型、基于参数服务器、以及基于混合模型的三类。

(1)基于数据流模型

数据流模型通常把计算抽象成数据流图，关于数据流图，它是一种由节点和边组成的有向无环图。在数据流模型种，每个节点代表一个计算操作，边则表示

数据流动的方向，定义了数据处理的顺序和中间过程的依赖关系。在数据流图种存在源节点和汇节点，它们分别代表数据输入和数据输出。需要注意的是，源节点和汇节点并不必须是唯一的，每个源节点都可以代表一个独立的数据输入源，它们可以是数据集、数据流或是数据库，同样，每个汇节点也可以对应一个独立的外部存储或者其他系统。

数据流图的节点-边结构用于设计分布式机器学习平台是极为合适的，首先，数据流图支持并行处理，加载分布式学习算法后，划分的数据子集可以作为多个独立的源节点，每个计算单元可以作为数据流图上的一个或数个节点，大大提高了处理效率。此外，数据流模型使得分布式机器学习不仅在物理上是一个“分而治之”的系统，在计算逻辑上也成为了一个由若干个小的、可管理的节点组成的处理系统，从而由数据流模型实现的分布式机器学习平台具有较好的模块化属性和更高的可扩展性。

Spark 是一个具有代表性的基于数据流模型的分布式处理系统，虽然它主要被设计用来进行大规模的数据处理。它的一个关键特性是内存计算，将数据尽可能地存储在内存中进行计算，避免频繁的磁盘 I/O 操作，从而显著地提高了数据处理的熟读，尤其适用于机器学习这样的迭代计算任务。

2010 年，Google 的研究人员最早提出关于参数服务器的概念，Google 在 2012 年发布了一个为大规模分布式训练深度神经网络设计的框架，即 DistBelief，它也是 Tensorflow 的前身。

（2）基于参数服务器

基于参数服务器的分布式机器学习平台的主要组成部分包括参数服务器和工作节点，参数服务器负责存储和管理模型的参数，管理工作节点的生命周期以及分配计算任务，工作节点则负责数据处理和梯度计算。参数服务器模型将参数划分给各个工作节点，提高了大规模参数模型训练的性能。

数据流模型和参数服务器模型的出现和发展不是继承关系，而是两种不同的设计思想，这两种设计思想各有利弊，因此衍生出了混合模型，集成了数据流模型的实时处理能力、高并行性以及参数服务器灵活的参数更新策略、适合大规模机器学习的优势。简略地说，混合模型仍然采用节点代表计算操作，边代表节点之间的依赖关系，但是使用参数服务器的思想进行训练，将训练任务抽象成数据流图后，使用参数服务器进行任务调度，使用工作节点进行计算任务。

（3）基于混合模型

混合模型的代表分布式机器学习处理系统主要有 TensorFlow 和 PyTorch，它们都将网络模型的符号表达式抽象成计算图。

Google Brain 的团队在 DisBelief 的基础上研发了 TensorFlow。它将数据流和参数服务器搭配使用，从而取得了更快的速度、更高的移植性和灵活性。早期的 TensorFlow 使用的是静态计算图，这种方式在优化和部署时会具有一定的优势，后续 TensorFlow 引入了 Eager Execution，从而使得默认情况下计算图时动态的，便于机器学习的调试和开发。从 2015 年发布至今，TensorFlow 进行了多次更新，支持在各类环境下执行分布式机器学习程序，包括移动设备、Windows 和 CPU、GPU。

Pytorch 则更加适合于小规模的项目，它使用了动态计算图。它的主要特点是：使用了类似于 Numpy 的 N 维 tensor，从而在 GPU 加速上取得了杰出的成就。其次，它使用的自动微分方法可快速构建和训练神经网络，在前向和后向传播都取得了较好的效果。最后，动态计算图的使用可以加速模型收敛，便于将计算分布式地部署在 GPU 和其他机器上。

2)分布式机器学习算法设计

为了实现分布式机器学习的具体应用，通常需要在机器学习平台上加载分布式机器学习算法。然而机器学习算法已经是一个十分成熟的门类，按照学习方式

可以分为监督学习、半监督学习、非监督学习和强化学习，因此，将机器学习算法移植到分布式机器学习上不仅关注机器学习算法的分布式实现，而且关注分布式的机器学习算法针对通信延迟、一致性和容错性问题的优化。

无论是监督学习、半监督学习、非监督学习还是强化学习，都可以移植到分布式机器学习平台上，如分布式梯度下降、分布式聚类、分布式强化学习。不同的机器学习算法进行分布式实现时需要考虑的优化方面不同，主要可以分为数据分割和预处理（如果存在数据集的话）、计算负载平衡、扩展性和可维护性考虑。而几乎所有分布式机器学习都需要考虑通信成本和一致性问题。一致性问题的解决策略关注模型的参数更新方式，包括同步更新和异步更新，同步更新可以保证算法的收敛率，但由于计算资源的层次不齐，由于短板效应的存在导致资源利用率较低；异步更新的资源利用率很高，但无法保证收敛效果。

大模型能够捕捉复杂模式，提供全面的知识，在各类任务中展现出卓越的性能。然而，如何在海量数据和复杂任务中实现自我优化和演进成为进一步提高大模型泛化能力的关键问题。这一过程不仅依赖于初始的预训练数据，还需要在实际应用中不断学习和适应来自用户的新数据与任务。分布式学习由于可以充分利用分布式算力完成大规模机器学习模型的协作训练，成为高效学习和适应新数据与新任务的有效手段。在分布式学习架构下，移动设备可以部署为网络中参与学习任务的计算节点甚至是智能体在本地执行训练和推理任务，并通过交互中间训练结果来完成大模型的全局微调训练，这样不仅可以增强算力，加速数据处理，还可以驱动传统人工智能服务、AI 生成服务等网络智能化任务。然而，分布式架构下全参数训练过程中需要频繁进行参数更新、参数同步和梯度交换，这对网络传输、计算、存储等能力提出极高的要求。一个有效的解决思路是借助高效微调技术（比如高效参数微调，提示微调）来实现大模型的分布式微调训练。比如利用参数高效微调技术，用户可以训练额外小规模的微调模型（如 LoRA）捕捉

其本地任务或数据相关的知识，在保持接近全参数训练性能的同时，大幅降低了计算和存储的需求，而共享和聚合小规模的微调模型对网络来说也相对资源友好。同时，在分布式网络系统中，还需要考虑到每个用户的计算资源和存储能力都有所不同，例如，一些用户可能使用高性能的服务器或专业的计算设备，而其他用户可能只能访问基本的智能手机或小型边缘设备。不同用户在本地进行大模型微调时可承受的微调模型的参数量也会有所不同，直接导致了在不同用户间微调模型的异构性。因此，还需进一步探索端侧资源异构性场景下高效的分布式训练服务的解决方案，例如微调模型的知识蒸馏、异构微调模型对齐等。此外，在分布式网络中提供分布式训练服务的时候还需要关注用户的一致性问题，针对不同网络场景设计高效的共识机制、分布式一致性算法，在训练过程中需要确保用户模型的同步和一致性，从而保证大模型可以正确有效的收敛。（电子科技大学：刘贻静，汪云翔）

3. 分布式训练框架

分布式训练的关键技术包括云边协同计算和分布式训练框架的实现：

- 分布式协同计算：分布式网络节点在靠近数据源的地方进行数据处理和分析，可以减少传输延迟和带宽占用，提高实时性和响应速度，适用于需要快速决策的场景。中央网络节点则借助强大的计算和存储资源，进行复杂的数据处理和模型训练，支持大规模数据分析和全局优化，提升整体系统性能。分布式网络节点和中央网络节点协同通过在分布式网络节点进行初步的数据处理和训练任务，减轻中央网络节点压力；利用 6G 网络的低延迟特点，实现分布式节点与中央网络节点的高效协同，提高整体效率。
- 分布式训练框架：分布式训练框架包括并行计算和智能调度，其中并行计算主要方法为数据并行和模型并行，共享网络中的计算节点实现。

数据并行技术将训练数据集分割成多个子集，每个子集在不同的计算节点上独立训练，这些节点共享同一个模型副本，但独立计算梯度。节点间同步更新模型参数，利用 6G 网络的高带宽进行快速数据传输和梯度同步。数据并行通过以下过程实现：1）数据分割：将训练数据集分成若干子集，分配到不同的计算节点；2）独立训练：每个节点使用自己的数据子集进行前向和后向传播，计算梯度；3）梯度汇聚：所有节点的梯度通过网络进行汇聚（例如使用参数服务器或全局同步），然后更新模型参数。数据并行能够处理非常大的数据集，并且适用于大多数深度学习框架，但通信开销较大，尤其是在节点数量很多时，会导致同步瓶颈。

模型并行技术将模型本身分割成多个部分，不同的计算节点负责不同部分的计算，节点间传递中间结果。适用于超大规模模型，利用 6G 网络的低延迟进行高效通信。这对于超大规模的模型特别有效。具体实现上，模型分割将模型的各个层或模块分配到不同的计算节点；通过并行计算，每个节点只负责自己部分的前向和后向计算，节点之间通过网络传递中间结果和梯度信息。通过模型并行可以处理单个节点无法容纳的大模型，但需要仔细规划模型的分割策略和节点间的通信，以减少延迟和通信开销。

智能调度技术则动态调整计算资源和任务分配，根据网络状况和计算需求，优化分布式训练过程。利用 AI 算法动态调整计算资源和任务分配，从而提高资源利用率和训练效率。如图 4 所示，分布式训练服务的部署和内涵应从如下步骤进行考虑：

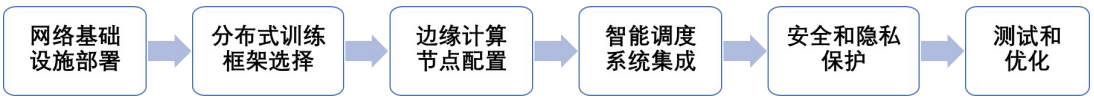

图 4. 分布式训练服务部署步骤

- 部署 6G 网络基础设施，确保高带宽、低延迟和高可靠性的网络环境。
- 选择适合的分布式训练框架，支持数据并行和模型并行。

- 部署边缘计算节点，配置高性能计算和存储资源，确保边缘节点具备足够的处理能力。
- 集成智能调度系统，动态调整计算资源和任务分配，优化训练过程。
- 实施数据加密、访问控制等措施，确保分布式训练过程中的数据安全和隐私保护。
- 进行全面的测试和优化，确保分布式训练系统的性能和稳定性，满足实际应用需求。

4. 联邦学习

根据 scaling law，大模型的性能是和模型参数、数据大小、计算量成正比的。6G 网络使能大模型分布式训练的优势包括宝贵的数据源、海量的闲置算力。其中，联邦学习是一种实现网络使能大模型分布式训练的隐私保护范例。通常支持联邦学习的无线网络由多个本地客户端和一个边缘服务器组成。联邦学习进行分布式训练包括如下五个迭代步骤：

1）全局模型初始化：在中央服务器上初始化一个全局模型w_0，并将其分发给 K个参与训练的客户端。

2）本地模型训练：每个客户端根据本地数据进行模型训练，计算更新后的模型参数。对每个客户端 K，训练的目标是最小化其本地损失函数：$|F_k(w) = \frac{1}{n_k}\sum_{i=1}^{n_k} f(w; x_i, y_i)|$。其中，$(x_i, y_i)$是客户端 k上的数据样本，$f(\bullet)$是损失函数，w是模型参数，n_k是客户端 k的本地数据量。

3）本地模型上传：客户端在完成本地训练后，将模型参数上传到中央服务器。此时，每个客户端 k提交的模型可以表示为w_k。

4）全局模型聚合：中央服务器接收每个客户端上传的模型参数后，进行加权聚合来更新全局模型。最常用的聚合方法是 FedAvg，其公式为：$w' =$

$\sum_{k=1}^{K} \frac{n_k}{n} w_k$。其中，$n$是所有客户端的数据总量，$K$是参与训练的客户端数，$w'$是更新后的全局模型参数。

5）模型迭代：服务器将新的全局模型w'分发给所有客户端，重复上述过程，直到模型收敛。

结合上述的训练过程以及不同的客户端设备，联邦学习的具体训练场景分跨孤岛（cross-silo）和跨设备（cross-device）两种：

1）跨孤岛：每个参与的孤岛（silo）通常是一个具有相对较强计算能力的本地计算实体，比如数据中心、公司内部的服务器、企业专有的计算资源等。通过跨多个这种具有独立计算能力的单位或组织进行协作式分布式学习，这些单位或组织之间不会直接交换数据，而是通过本地模型的更新与全局模型的聚合来完成训练。

2）跨设备：参与的设备通常是功能有限的智能设备，例如手机、IoT（物联网）设备等，计算能力不如 silo 中的服务器强大，这些设备之间通过无线网络进行协作，但网络连接可能较为不稳定。

因此，将大模型训练集成到支持联邦学习的无线网络之前，我们必须考虑不同联邦学习场景所施加的限制与大模型计算/存储/通信密集型要求之间的冲突[6]，所面临的主要挑战包括：

1）高功耗：由于训练所需的数据和模型参数数量庞大，大模型的训练过程对计算硬件和能耗都有很大的要求。当以合理可持续的方式部署高耗能的大模型时，能源效率变得至关重要。

2）有限且异质的算力资源：无线网络出了提供分布式训练服务之外，还需要承担基础的连接任务，硬件算力资源有限。为了将集成到无线网络中，通常需要专门的硬件来进行 AI 计算加速，例如 GPU、TPU 等。并且，客户端之间计算

资源的异质性，使得联邦训练大模型遭受更多空闲时间的影响，需要对算力资源建立统筹有效的编排和调度机制。

3）高存储和内存要求：为了满足大模型训练要求，必须大幅增加存储和内存来处理流式收集/生成的数据以及训练期间模型参数的更新。典型的网络架构可能不一定满足此类存储和内存要求。

4）高通信开销和时延：由于模型规模巨大，基于联邦学习的从头训练需要持续、大量的通信资源，这一过程将非常耗时且占用带宽。例如，通过 100Mbps 通道（5G 中用户体验的数据速率）传输一次 GPT2-XL（约 5.8 GB 的中型 LLM）大约需要 470 秒[7]，而从头训练可能需要对数千个 GPU 进行长达几个月的连续训练。虽然 5G 及以上网络有严格的延迟要求。目前尚不清楚将大模型集成到支持联邦学习的网络中如何满足如此严格的延迟要求。

5）数据问题：针对各客户端数据特征分布（非独立同分布，Non-IID）、标签分布以及样本量等方面的差异，微调大模型时必须设计有效的策略来应对这种数据异构性；不同客户端数据质量的差异（含噪声、标签错误或者缺失值等）会直接影响大模型的微调效果，一些必要的数据预处理如数据清洗、异常检测等步骤也需要进一步的考虑。

在以上挑战中，最关键的问题在于联邦学习设备算力有限条件下如何降低通信开销和优化内存。

降低通信开销方面的研究包括在传输层面以及在算法层面的优化两种：传输层面可以采用参数高效 PEFT 方法[8]减少需要传输参数量，如 FedPETuning[9]；降低通信轮次，如 DiLoCo[10]；压缩；量化等；算法层面可以采用更高效的联邦聚合算法，如 OpenFedLLM[11]，或者传输内容更少的其他优化算法，如 FwdLLM[12]、FedKSeed[13]用零阶有限差分估计梯度。

内存优化的维度包括采用参数高效 PEFT 方法减少可训练参数量；使用梯度

累积、激活值重计算等技术；优化器的选择，SGD、带动量的 SGD、Adam、AdamW 等；结合模型分割 Split learning 技术[14]，但代价是会加剧通信开销；只训练部分层/层冻结，如 AutoFreeze[15]、SmartFRZ[16]、FedOT/FedOST[17]等。此外，针对内存的优化还可以采用混合精度训练、ZeRO 零冗余优化器[18]等技术。

其次，对于算力异质性的研究，FATE-LLM[19]提出很多种架构的可能性，允许算力异质性允许终端使用不同的本地模型，知识蒸馏出一个通用于联邦聚合的模型。FedIT[20]提出每个设备可以采用不同的 Lora 配置，即层级最优秩(Layer-wise Optimal Rank Adaptation)思想。

5. 联邦大模型

大模型的参数规模极为庞大，且各大厂商也在持续刷新大模型参数量的上限。以 GPT 系列为例，从 GPT-1 到 GPT-4，模型的参数量从 1.1 亿增长至 1.8 万亿，由模型规模带来的性能提升出现边际递减效应[8]。目前，针对特定任务的模型微调（Fine-tuning，FT）已成为利用大模型的主要方法[21]，但直接微调对算力、内存都提出了更高的要求，AI 硬件（GPU）内存难以跟上模型扩大的需求。为了解决算力增速不足的问题，研究者考虑用多节点集群进行分布式训练，将训练扩展到多个 AI 硬件上（如 GPU），从而突破于单个硬件内存容量和带宽的限制，支撑更大规模模型的训练。目前较多分布式训练架构主要有两种模式：集合通信（collective communication，CC）模式和参数服务器（parameter server，PS）模式。

NLP、CV、多模态、科学计算等领域的模型往往具有模型结构复杂、参数稠密的特点，集合通信训练模式可以很好地支持此类模型的训练。集合通信模式对计算芯片的算力和芯片之间的网络互联要求较高，如高性能计算的 GPU、芯片

之间的高速网络互联 NVLink 和 InfiniBand 等。搜索、推荐等场景的模型往往数据量巨大，特征维度高且高度稀疏化。参数服务器训练模式可以很好地支持此类模型的训练。参数服务器训练模式可以同时做到对数据和模型的并行训练，对于存储超大规模模型参数的训练场景十分友好，常被用于训练拥有海量稀疏参数的搜索、推荐领域模型。

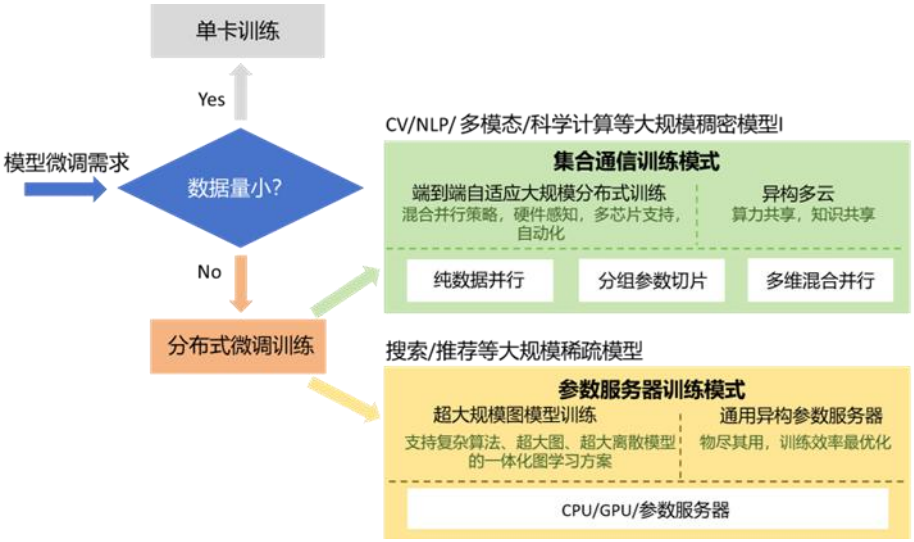

图 5. 分布式微调训练

然而，传统公开的可用数据集无法满足大模型微调的需求[22]，特别是大规模中文数据集十分缺乏，对中文大模型以及业界模型的中文支持都有很大的影响。收集多样化、高质量的指令数据仍面临挑战，尤其是在隐私敏感领域，往往禁止收集、融合使用数据到不同的地方进行 AI 处理，本地数据不足或微调和预训练数据集之间存在显著差异可能导致模型的泛化性能不佳。

为了解决隐私数据给用户安全和模型泛化性能带来的挑战，联邦学习（Federated learning， FL）[22]作为一种分布式框架被引入。其中，联邦分割学习（Federated split learning，FSL）框架[23]将模型分割成多个部分，在边缘用户设备上仅针对部分模型基于本地任务数据进行训练，训练完成后上传模

型参数 至服务器进行聚合，而无需共享原始数据。随后，服务器训练的全局模型参数 被发送回所有客户。在联邦分割学习框架下微调大模型，可以有效降低边缘设备和服务器之间传递的参数数量，从而减少各设备的计算和通信开销，提高了效率和隐私安全，且微调后的模型在客户端之间共享，促进协作同时保障隐私。通过利用多样化的数据集，充分释放大模型的潜力并提高大模型的泛化性能。

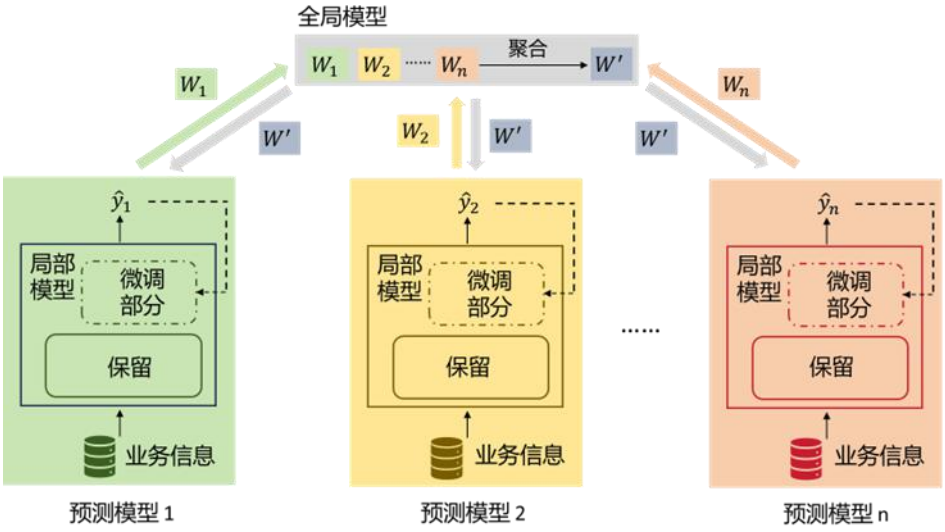

图 6. 联邦学习

此外，Jianyi Zhang 等人[24]将 FL 用于大模型的指令微调，保护隐私的同时，在一定程度上提升了模型性能。Jing Jiang 等人[25]提出了低参数联邦学习方法 LP-FL，旨在对大模型任务流进行微调，实现了在有限计算和通信资源下的高效学习。Tao Fan 等人[26]提出了用于工业级大模型的联邦学习框架 FATE-LLM，使用一方的私人数据对另一方的私有大模型进行有效微调，解决了微调大模型所需计算资源庞大和高质量数据分散的问题，并保护了模型知识产权和数据隐私。Chaochao Chen 等人[27]考虑到计算和通信两个方面，提出将参数高效微调方法整合到联邦学习框架中，例如适配器微调、前缀微调、提示微调和低秩适应。通过最小化参数梯度计算和减少聚合参数的数量，有效降低了计算和通信成本，这种方法在保留近似性能和显著减少计算和通信负担之间取得了平衡。

（三）指令优化服务

指令优化是提高大模型整体泛化性能的有效技术，特别是在零样本场景下经过指令优化的大模型比参数高效微调的大模型具有更强的泛化能力，而后者更多是针对特定下游用户任务而设计的。不同于参数高效微调方法中冻结大模型参数，指令优化过程一般涉及到优化整个大模型的参数，这意味着它将对网络多维资源提出更高的需求。尽管指令优化对于指令导向的语言任务（例如自然语言推理、问答、翻译等）普遍有效，但为了使其在推理任务上也能比传统大模型微调表现得更好，一方面考虑在指令优化的监督样本中包含结构化提示（比如思维链提示，思维图提示，思维树提示等）。结构化提示通过将复杂推理问题分解成多个中间推理步骤，可以从多维度多角度有效增强大模型的逻辑推理能力。

然而，复杂问题的精准推理过程会生成多种不同的提示推理路径，其中任意逻辑推理步骤都有可能出错，这会导致差异化的训练和推理结果，并且降低模型训练和推理性能。同时，基于结构化提示的推理通常要求模型的规模足够大以提供全面的信息，但需要消耗巨大的计算能力和推理成本，而网络稀缺的计算、传输等多维度资源将限制推理性能。因此，需要设计适用于网络的高效结构化提示优化方案（比如，思维链提示优化方案），使得网络即使在多维度资源受限时，仍然能生成并且判别合理的提示样本来引导较大模型的训练和推理过程，同时保证训练和推理的合理性与准确性。另一方面，可以考虑人类反馈的强化学习（RLHF），通过在指令优化步骤之后增加给定指令的人类反馈促使大模型的输出更贴近期望的行为模式。但对大量数据进行基于 RLHF 的大模型训练，实时获取和处理人类反馈样本增加了网络负担；同时，分布式网络中不同用户提供的反馈样本可能存在质量差异导致不同评估者间一致性较低。因此，也需要探索高效的基于 RLHF 的指令优化方案，例如将利用边缘计算在靠近用户侧处进行反馈样本

缓存和预处理，保证网络带宽和延迟需求，以确保反馈样本的及时传输和处理；其次，设计合理的反馈机制和评价标准，避免不同评估者可能存在的偏好偏差以及低质量反馈样本影响大模型的训练效果。

（四）端边云协同推理服务

随着人工智能技术的飞速发展，尤其是大规模语言模型（LLMs）在自然语言处理和多模态理解任务中的广泛应用，端边云协同推理服务成为提升 AI 服务性能和用户体验的重要手段。传统的云计算模式尽管拥有强大的计算能力，但随着大量终端设备的接入，数据传输带宽和延迟问题愈发突出。因此，结合终端设备、边缘节点与云端资源的协同推理架构逐渐成为解决上述问题的有效途径。

1. 端边云大模型部署策略

在这种协同架构中，边缘节点通常部署经过压缩和优化的小型化 LLMs，用于处理地理位置相关的数据以及个性化请求。例如，边缘 LLMs 可以根据用户所在位置生成更加精准的提示（prompt），这些提示随后被传输到云端，结合云端强大的计算能力，生成更加复杂且符合用户需求的响应。这种模式显著缩短了数据传输的延迟，减少了通信带宽的占用，并提升了整体的服务质量。

然而，端边云协同推理服务面临的一个关键挑战是如何在不同的计算节点之间进行模型的合理切分。对于大规模 LLMs 来说，随着模型层数的增加，计算复杂度显著提升，这对用户设备的计算能力提出了更高的要求。为了在保持推理性能的同时减轻 UE 的计算负担，合理的模型切分点显得尤为重要。通过应用模型强化学习等先进技术，可以在动态变化的信道环境中自适应地调整模型的切分点，确保在不稳定的无线网络条件下仍能实现最优的推理效果。

为了提升终端设备上的模型推理性能，研究者提出了多种协同学习策略。例如，CD-CCA 框架通过云端的大型 MLLMs 增强设备端压缩模型的泛化能力。该

框架包括设备到云端的上行链路、云端知识蒸馏和云到设备的下行链路优化。在上行链路中，利用不确定性引导的令牌采样策略（UTS）来过滤设备端产生的分布外令牌，减少上行传输成本并提高训练效率。在云端，基于适配器的知识蒸馏（AKD）方法将大型 MLLMs 的知识传递给设备端的小型模型，并通过动态权重更新压缩策略（DWC）对设备端模型的更新参数进行自适应选择和量化，提升传输效率和模型性能。

为了应对各种动态环境下的推理任务，诸如 FlexInfer 框架等还提出了多种优化算法，能够根据任务需求灵活调整推理输入和模型配置，从而实现精度与延迟之间的平衡[28]。例如，通过在设备端处理简单任务并将复杂任务传输至云端，系统能够优化资源使用，同时满足高效、低延迟的推理需求。这种灵活的配置机制使得边缘和云端能够针对不同任务提供优化的解决方案，确保了推理任务在多个节点之间的高效协作。

此外，LoRA（低秩适应）技术的引入，使得终端设备能够在计算资源有限的条件下进行模型微调。这种技术通过在模型参数矩阵中添加低秩路径，显著减少了微调所需的存储空间和计算资源，从而实现了在边缘设备上高效的推理和定制化服务[29]。结合这些技术，边缘节点可以快速响应用户请求，生成初步的推理结果，而复杂的分析则由云端完成，确保了系统的响应速度和计算资源的高效利用。

在数据隐私保护方面，端边云协同推理服务也展现了独特的优势。通过在边缘节点进行数据的初步处理和过滤，可以有效降低敏感数据在传输过程中被截获的风险。此外，边缘计算能够结合本地数据进行模型训练和推理，减少了将所有数据上传至云端的必要性，从而进一步增强了数据隐私保护的力度。

综上所述，端边云协同推理服务为大规模 AI 模型的高效部署提供了技术支撑，通过智能的任务分配和优化机制，实现了服务的高效性、定制化和安全性。

这一框架不仅可以应对复杂多变的网络环境，还能在资源受限的条件下提供高质量的 AI 服务。随着技术的不断进步，端边云协同推理服务将在更多场景中展现其优势，为人工智能技术在未来的广泛应用奠定坚实的基础。

2. 端边云大小模型协同推理策略

随着大模型参数规模的增加，从几百亿到几千亿到现在 GPT4 的万亿参数，其对计算和存储成本的消耗也越来越大。在网络内部署大模型时，需要考虑可能合适的部署方式。大模型捕捉复杂模式与知识的能力展现出更强的泛化能力，但其运行需耗费大量网络及计算成本；相较之下，小模型更专注于特定场景或者任务，相对而言参数较少、结构简单、计算量较少，适用于处理规模较小、简单的数据集，但其智能能力可能受到模型大小、规模等因素的限制。由于端边侧的自身的计算、存储和模型资源限制，可以在网络的端边部署和运行小模型或者轻量化的大模型（例如采用量化、模型剪枝、模型拆分等轻量化技术对大模型做小型化处理），以在尽可能保持模型性能的前提下降低网络、计算资源开销，同时在网络资源充足的云数据中心部署大模型。因此，考虑探索端边云大小模型的协同推理来充分发挥大小模型各自的优势。每一次协同推理服务过程，首先由端边侧小模型执行简单推理任务，进而根据初步推理结果或者数据的复杂度，决定将部分或者全部数据发送至边缘设备、云数据中心做进一步处理；云端大模型或者边缘侧大模型通过多模态数据对齐、多模态模型融合等技术来对接收到的数据进行深度协同推理和优化模型，并将更新结果发送回设备端。同时，可以考虑在端边缓存频繁处理任务和数据，减少冗余的计算和推理。另外，也可以考虑大小模型的迁移学习，利用云端大模型在不同任务上学到的知识，通过知识迁移、特征迁移、增量式迁移等技术，将大模型学到的先验知识、特征等迁移到边缘侧或者终端的轻量化模型，帮助模型更快、更准确地适应新任务或新环境。

（1）服务需求背景

大模型一路发展下来，无论是 OpenAI 的 ChatGPT，还是国内的文心一言，Kimi 和豆包等，通常都部署在云端。这些大模型体积庞大，适用于复杂任务的高级推理、数据分析和上下文理解。然而，云端模型不仅需要大量计算资源，还要求用户上传数据。出于成本和数据隐私安全的考虑，越来越多的模型和应用厂商开始选择将大模型部署在端侧。

实际上，Meta、谷歌、微软等大型科技公司在发布大模型时，通常会提供大中小三种模型套装，其中最小的模型参数基本都在 10B 以下，并且有些模型专门为手机等大众终端设备进行了适配。例如，微软在开发者大会上发布了专为手机端侧推理设计的 Phi-3 系列模型；Apple 也在 WWDC 开发者大会上发布了一个 3B 的端侧小模型。此外，终端厂商也在积极尝试将模型集成到 PC 和手机中，这就是目前热议的“AI PC”和“AI 手机”等概念。而在 6G 时代，机器人、无人车等新型终端对于智能的要求也会越来越高，内置的端侧模型能够帮助这些终端实时处理本地的任务推理需求。

然而，如前所述，受限于端侧的算力、内存和局部观测条件，端侧大模型在实际运行中依然存在很多挑战。6G 网络正可以通过以边助端，云边端协同来提升端侧模型生成、更新效率和推理的准确度，以实现智能普惠的愿景。

（2）端边云协同推理服务的操作类型

对于端侧大模型来说，简单的推理任务可以直接在本地闭环，一些复杂的、推理要求高的任务就需要通过网络边侧进行协同推理，通过网络的内生智能能力来进一步增强，甚至对一些实时性不敏感，例如长期规划类的任务还可以进一步提交给云端进行协同推理。

通过 6G 网络以边助端来增强端侧大模型推理的操作，可以总结为以下几类：

1) 增强端侧大模型推理置信度。端侧大模型在进行推理时，如果当前的推

理置信度不足，则通过网络上部署的更强大的模型来进行增强，提供更准确的推理结果。对于没有实时性要求的推理任务，这个增强大模型可以部署到云端；但是网络提供的增强大模型可以提供更实时性的服务，并且在隐私保护和能耗上都更加具有优势。

2) 提升端侧大模型推理效率。对于现在主流的生成式大模型来说，推理时需要逐个 token 进行计算输出，效率比较低。为此，网络可以提供推理效率提升的服务，以随机推理为例，在网络上部署端侧大模型的拷贝模型（可以通过蒸馏，剪枝，量化等方式生成，也可以是独立生成），利用网络更强大的计算和并行能力提前生成更多 token，最后提交给端侧大模型来一次性进行校验和最后输出，从而提升端侧大模型的推理效率。

3) 降低端侧大模型的部署门槛。考虑到有些终端的算力和内存相对比较弱，例如 3GPP 定义的 RedCap（Reduced Capability）所覆盖的终端，其在部署大模型上面临较大的门槛。为此，网络可以和终端进行协同部署，利用网络算力和终端进行协同拆分推理，并且模型分拆的比例规模还可以根据端侧的算力和通信状态进行调整。相比在云上进行模型拆分，网络可以更高效的掌握终端和信道状态变化，从而优化拆分模式，提供最优的协同推理效果。

4) 使能多个端侧智能体协同。部署了大模型的终端智能体，其个体行为之间存在需要协作的场景，例如工厂机器人之间共同完成一个搬运工作，无人车辆之间的避让和转向选择等。然而，终端智能体受限于其单体的观测空间，无法做到全局的最优规划，或者就需要更加复杂和长时间的交互。利用网络的全局视角，结合 6G 的通感一体能力，可以为多个端侧智能体间的协同推理提供更多的全局信息，通过全局规划给出最优的决策空间。

（3）网络需要具备的能力

网络为端侧大模型提供协同推理服务，是在不同厂家的异构设备之间进行的，

为此需要设计标准化的智能协作框架，支持协作服务的各类操作类型，对端边云协同推理进行统一管理和配置，包括协作实例的创建、协作类型的确定和配置，协作流程的监控和优化等。

3. 面向 AI Agent 的端边云协同方案

随着大模型在自然语言处理、计算机视觉和机器人控制等领域所涌现出的强大的通用能力，各行各业都已经开展了大模型的应用研究。由于大模型预训练需要庞大的算力和海量的数据资源，早期的大模型均部署在云端。随着存储和计算芯片的发展，将预训练好的大模型轻量化后在终端侧部署推理也逐渐成为业界大模型应用的趋势。相对于云端大模型，终端大模型有利于降低端到端时延，提升用户体验。然而，终端设备的计算和存储资源始终有限，轻量化后的大模型在推理性能上会受到一定程度的影响。为了解决这一问题，现有技术方案采用模型分割的方法来进行端云或者端边的协同推理，将端侧的计算量分流至网络边缘或者云服务器中，实现准确性和端到端时延的折中。

现有的基于模型分割的端边云协同大模型轻量化方法虽然能够在推理准确性和端到端时延性能上取得折中，但存在以下技术问题：

(1)模型分割算法复杂度高，分割点寻找困难。现有方案中的模型分割往往采用遍历深度神经网络中的所有隐藏层的方法，来获取最佳分割方案以保证在达到所要求的端边计算卸载量的同时、端到端时延最小，而大模型的隐藏层层数通常较深，所以这种穷搜的方法复杂度较高。

(2)基于分割后的模型在进行端边或端云协同推理时需要传输中间隐藏层的高维特征，由于推理性能对隐藏层特征的变化可能十分敏感，当无线信道环境较差时，将导致推理性能严重下降。

本文提出了一种基于端边云协同的大模型轻量化部署方法，该方法通过在终

端部署剪枝后的大模型实现轻量化部署，并通过在网络边缘部署与终端大模型进行协同推理的模块来保证推理性能。一方面，由于不需要进行模型分割，该方法避免了复杂度高的分割点寻找算法；另一方面，端边协同推理时通过网络传输的是端侧轻量化模型的推理结果，其维度相对于中间隐藏层高维特征来说往往较小，因此最终推理性能对其的变化具有一定鲁棒性，所以本方案具有一定抗干扰能力。

所提方法的部署框架如图 7 所示，包括在云端部署的预训练和性能监控模块、在终端部署的信息采集和推理模块以及在网络边缘部署的推理模块。

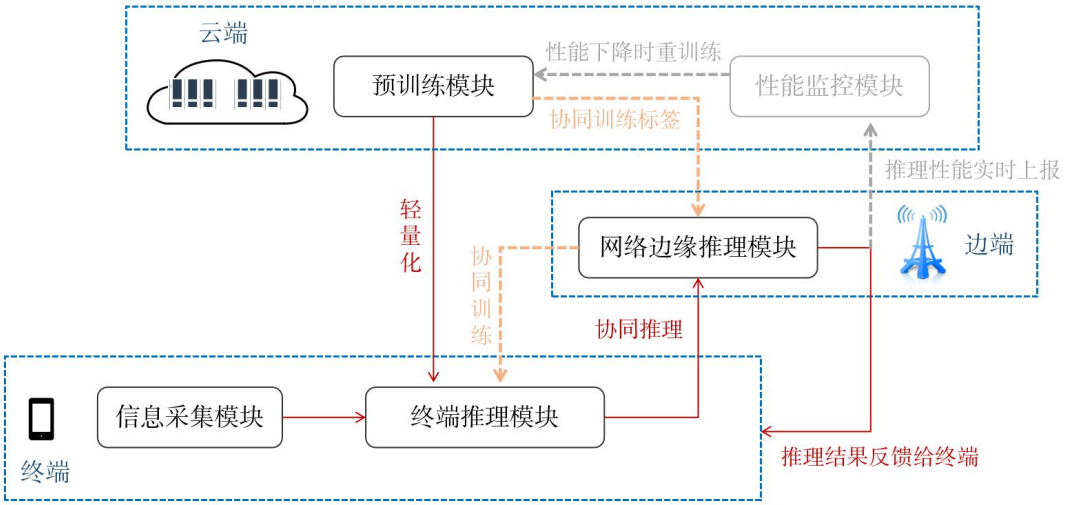

图 7. 端边云协同大模型轻量化部署框架

（五）模型优化服务

现有的针对无线通信 AI 模型的生命周期管理方案采用的是外挂叠加式的方法，即通过在终端设备中添加处理单元执行对模型的性能监控，这种方案存在以下技术问题：仅针对特定的 AI 任务，对于不同的任务需要额外添加监控处理单元，成本较高，无法满足未来 6G 网络多样的 AI 任务场景需求；模型的重训练没有考虑基于端边云协同的任务卸载，当网络算力资源不足时可能无法满足模型优化需求。

本文提出了一种面向 AI 大模型的优化服务，通过在基站端部署性能监控和算力检查模块，实时监控模型性能和网络剩余算力资源，在模型性能下降时灵活利用网络或者云端算力重新训练模型，通过引入 AI 编排模块来针对不同任务统一调度网络资源，能够满足 6G 网络 AI 任务场景多样化的需求。

所提方法的架构如图 8 所示，包括 AI 任务编排、性能监控、算力检查、信息采集、推断和训练控制模块。

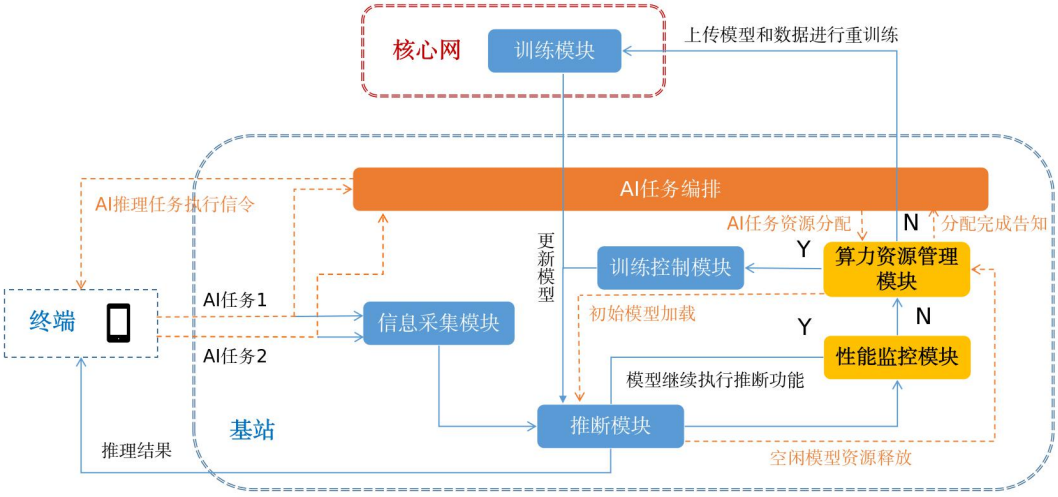

图 8 内生 AI 模型的生命周期管理架构

四、案例分析

生成式 AI 在语义通信系统中的应用

1. 生成式 AI 与语义通信概述

生成式 AI 是一种利用深度学习模型来生成新内容的人工智能技术。这种技术不仅仅局限于特定的任务，如图像识别或自然语言处理，而是致力于展示全面的认知能力，包括逻辑推理、问题解决、感知和语言理解等[30]。通过深入学习

历史数据的复杂模式，生成式 AI 模型能够自主地生成全新的内容，这些内容可以是文本、图像、音频或视频等多种形式。生成式 AI 不是简单地根据给定的规则或数据生成输出，而是模仿人类的创造力，生成具有创新性和实用性的新内容，从而形成了人工智能生成内容（AI-Generated Content，AIGC）的概念。AIGC 的优势在于其低延迟、创造力、高效率、可扩展性和个性化定制。这些特点使得 AIGC 在数字艺术、音乐创作、广告设计和产品创新等多个领域具有广泛的应用前景。通过生成式 AI 的深度学习模型和算法，AIGC 能够快速生成高质量的内容，同时保持创意和独特性，满足用户的不同需求。其核心技术包括生成对抗网络（Generative Adversarial Network，AIGN）、变分自编码器（Variational Autoencoder，VAE）、循环神经网络（Recurrent Neural Network，RRN）等。这些技术通过不同的机制学习数据的分布，并生成新的数据实例。生成式 AI 的工作过程是一个迭代的过程，需要不断地调整模型和评估生成结果，从而得到更好的生成效果。随着大数据和计算机处理能力的不断增强，生成式 AI 技术取得了重大进展，并在自然语言处理、计算机视觉和音频处理等多个领域产生了巨大的影响。

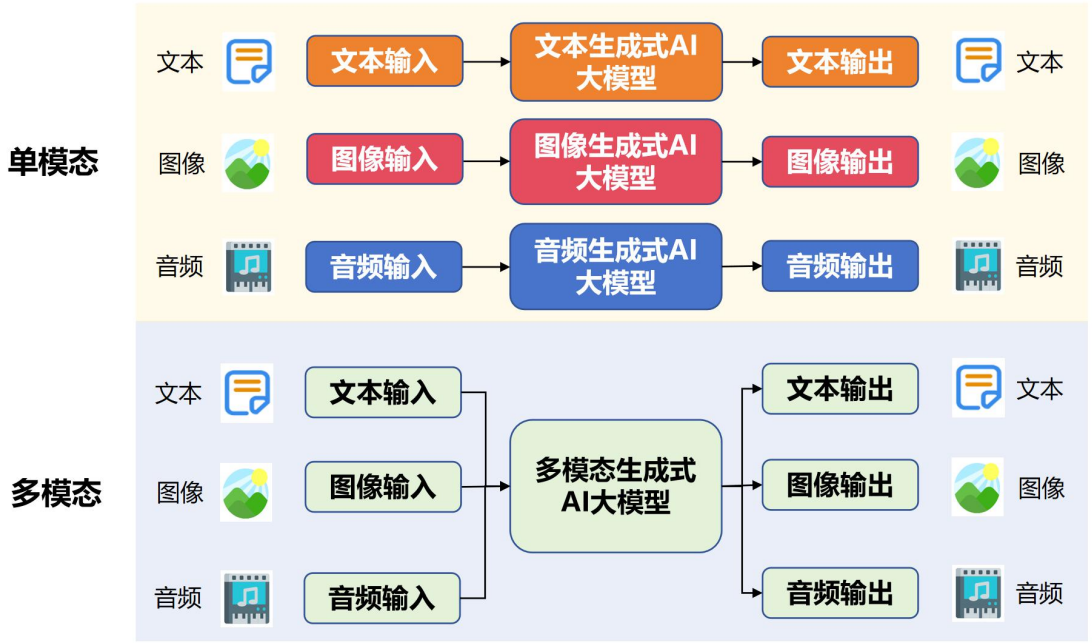

图 9. 两类生成式 AI 大模型:单模态和多模态

表 3. AIGC 应用和模型概览[33]

模型类型			AIGC 应用	AI 大模型	神经网络架构
单模态	文本转文本		ChatGPT-3, Bing AI	GPT-3, T5	Transformer, LSTM, CNN, RNN, GAN, VAE
	图像转图像		PaintMe.AI, Vizcom	StyleGAN, VQ-VAE	GAN, VAE, CNN, Transformer, RNN, 扩散模型
	音频转音频		Murf.AI, Resemble.AI	WaveGAN, SpecGAN	GAN, VAE, RNN, CNN, Transformer
多模态	文本转 X	文本转图像	DALL-E 2, NightCafe	DALL-E, CLIP	RNN, CNN, Transformer, GAN, VAE, 扩散模型
		文本转视频	Synthesia, Pictory	CogVideo, Phenaki	Transformer, VAE, CNN, RNN, GAN, VAE, FCN
		文本转音频	Murf AI, PlayHT	WaveNet, AudioLM	CNN, LSTM, RNN, FCN
	X 转文本	图像转文本	Transkribus	NIC, CLIP	Transformer, RNN, LSTM, CNN
		视频转文本	Google Cloud Video Intelligence API	VideoCLIP, VideoBERT	Transformer, RNN, LSTM, CNN
		音频转文本	Speak AI	DeepSpeech	Transformer, RNN, CNN

语义通信技术是一种前沿且创新的通信技术，它不仅强调对信息的精准处理，还注重在通信过程中传输语义层面的信息。它打破了模块分离优化的约束，采用端到端贯通式优化和信源信道联合设计的技术手段，获得通信系统的整体优化[31]。这种技术以任务为主体，遵循“先理解，后传输”的原则，从而大幅提升了通信系统的传输效率和可靠性。在语义通信过程中，信息的发送者和接收者会共享一个语义理解的环境，使得信息的传输不仅限于数据本身，更包括数据背后的含义[32]。为实现这一目标，语义通信的发送端和接收端均具备强大的语义处理能力，从而实现信息的精准提取、高效编码与智能解码。在信息的发端，语义通信系统首先对原始信号进行语义分析。这一步骤的关键在于特征提取，即识别并提取出对后续任务至关重要的语义特征。为了降低传输数据量同时确保信息的完整性和准确性，系统会进一步对这些特征进行压缩和优化。在信息的收端，语义通信系统则利用相似的机器学习技术，对接收到的数字信号进行解码和语义重构。这一过程包括智能分析接收到的信号，根据发送方和接收方共享的语义理解环境，恢复出原始信息中的语义特征。随后，系统会对这些特征进行重构，生成与发送方原始信息尽可能一致的语义表示，以确保信息的精准解码和重构。

2. 生成式 AI 大模型驱动的语义通信系统

生成式 AI 大模型驱动的语义通信系统是一种将生成式 AI 模型与语义通信相结合的创新网络架构，整合了生成式 AI 的广泛认知能力和语义通信的高效语义信息传输[33]。该网络架构主要包括物理基础设施、数据平面和网络控制平面三个部分：

- 物理基础设施：由多个无线终端设备、接入点、基站、边缘服务器和中央云服务器等组成。这些实体不仅执行传统通信功能，还配备了生成式 AI 模型和知识库，以支持 AIGC 服务。无线终端设备可以上传数据，并通过接入点

和基站下载知识和训练好的模型，实现知识的整合和知识库的更新。边缘节点能够利用自身和连接的无线终端设备的知识对生成式 AI 模型进行预训练和微调，并将训练好的模型部署到相应的无线终端设备上[34]。中央云服务器则拥有庞大的存储和计算资源，用于大规模生成式 AI 模型的预训练和提供全局的 AIGC 服务。

- 数据平面：AIGC 数据在该网络的数据平面上生成、传输和评估。生成式 AI 模型负责生成包括单模态和多模态在内的 AIGC 信息。这些信息通过语义通信方式进行传输，利用语义编码器和信道编码器提取并压缩关键语义信息，随后通过无线信道进行传输。接收端则通过信道解码器和语义解码器对数据进行恢复，以减少传输过程中的失真[31]。此外，数据平面还负责从任务完成度和相关性等多个维度评估 AIGC 信息的有效性。
- 网络控制平面：由语义层和生成层组成，分别负责处理语义信息和管理生成式 AI 模型。在训练生成式 AI 模型和语义通信模型的过程中，知识管理扮演着至关重要的角色。构建、共享和更新知识库是核心环节，这涉及到原始数据的收集、分类、编码，以及对知识库的持续监测和更新，以确保其动态更新和可靠性[35]。网络控制平面还承担着网络管理的职责，以适应语义通信的需求，并充分发挥 AIGC 的优势，如低延迟、创造力、效率、可扩展性和个性化，为用户提供高效、智能的通信服务。

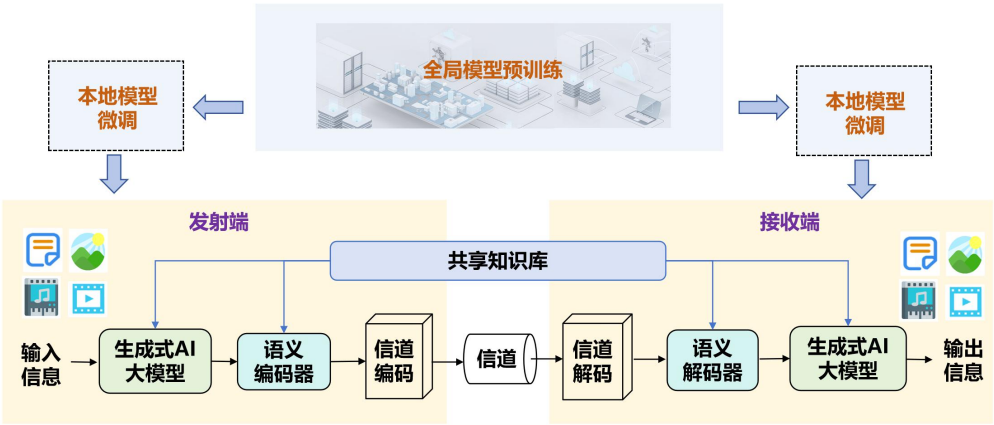

图 10. 生成式 AI 大模型驱动的语义通信系统

在生成式 AI 大模型驱动的语义通信系统中，资源分配策略需要综合考虑计算资源和通信资源，根据实时网络状态、应用需求以及数据语义值来实现网络资源的智能化与动态化管理。

1) 计算资源分配：计算资源涵盖了处理能力、存储以及内存等多个方面，构成了生成式 AI 大模型驱动的语义通信系统的核心支撑。由于网络流量的波动性和负载的动态性，静态的资源分配方式已无法满足当前复杂多变的网络环境。AI 技术的引入，使得系统能够根据实时网络需求进行动态计算资源分配，实现高效响应。AI 模型通过分析历史网络行为，利用机器学习算法预测未来的资源消耗模式，从而做出精准的资源分配决策[36]。这种基于预测的分配方式，不仅能够提高资源利用率，还能减少资源浪费。在多样化的无线网络环境中，不同的应用程序对计算资源的需求各不相同。例如，高清视频流处理需要强大的计算能力和充足的内存，而物联网传感器数据传输则可能只需较小的计算工作量，但对存储的要求较高。AI 模型能够实时监控网络状态和应用程序的资源需求，并依据这些实时和历史数据，智能地重新分配计算资源[35]。通过动态调整计算资源，生成式 AI 大模型驱动的语义通信系统能够有效应对网络流量变化，提升整体网络性能和用户满意度。

2) 通信资源分配：在语义通信系统中，通信资源分配的策略也呈现出智能化、上下文感知和用户中心化的特点[37]。传统的通信资源分配主要关注数据吞吐量和带宽效率等位相关的指标，而生成式 AI 大模型驱动的语义通信系统则更加关注数据的语义或含义，旨在实现更高效、更精准的信息传递。为了实现这一目标，AI 模型被用于理解和解释数据，并根据数据的语义值对其进行优先级排序。这种基于语义的优先级排序，能够确保重要的、相关的语义信息得到优先传输，从而提高信息传递的及时性和可靠性。在基于置信度的

资源分配方案中，系统会根据数据的置信度来决定其传输优先级，从而确保高置信度的数据得到优先传输[38]。而在多模态提示方面，系统则会利用视觉和文本等提示信息来恢复和增强数据的语义上下文，提高消息恢复的稳定性和准确性[39]。此外，由于语义通信中不同的知识匹配程度会导致移动用户观察到不同的语义性能，因此，基于知识匹配度的资源分配策略应运而生[40]。这种策略旨在通过优化收发器之间的知识匹配度，来提高语义通信系统的资源管理效率。同时，还有一些研究探讨了生成式 AI 大模型驱动的语义通信系统中 AIGC 服务的资源分配问题，旨在通过平衡资源使用和系统需求来优化服务效率。这些创新性的通信资源分配策略，为生成式 AI 大模型驱动的语义通信系统提供了更强大的通信支持，使得系统能够在各种复杂网络环境中实现高效、可靠的语义信息传递。

在文献[34]中，研究人员针对生成式 AI 大模型驱动的语义通信系统进行了仿真实验。在实验中，研究人员结合了 ViT 和 GPT-2 的图像-文本模型来实现图像到文本的转换,以及关键词提取和目标识别。此外,还采用了 Stable Diffusion 2.1 模型作为全局 GAI，用于从接收到的提示中生成 AI 图像。语义通信部分则采用了深度卷积网络和 Transformer 驱动的语义解码器进行语义分割和恢复。所有模型都在信噪比为 0 dB 的 AWGN 信道上进行训练，以传输 300 张不同内容的图像。使用 Adam 优化器对神经网络进行训练，初始学习率为 5×10^{-4}。

表 4：在下行链路传输 300 张图像所需的比特数以及 PSNR 性能[34]

不同的图像传输方案	**下行链路所需比特数**	**PSNR（峰值信噪比）**
集成生成式 AI 的传统通信	1.28×10^{5}	28.05
语义通信[41]	5.99×10^{4}	28.25

生成式 AI 大模型驱动的语义通信	3.03×10^4	28.64

表 4 中的仿真结果显示，生成式 AI 大模型驱动的语义通信系统在传输 300 张 1024x1024 像素的图像时，仅需要 3.03×10^4 比特，相较于文献[41]中的语义通信方案减少了 2.96×10^4 比特，与集成生成式 AI 的传统通信方案相比更是减少了 9.77×10^4 比特。在图像传输质量方面，生成式 AI 大模型驱动的语义通信系统的 PSNR 得分为 28.64，略高于其他两种方法，表明该框架在减少比特数的同时，还能保持较高的图像传输质量。也就是说，通过利用生成式 AI 大模型，语义通信系统可以显著减少图像传输所需的比特数，从而有效节省了带宽资源。这些结果表明生成式 AI 大模型驱动的语义通信系统在提供高质量语义通信服务的同时，能够实现准确的语义传输，为图像传输提供了一种新的高效方法。

五、未来展望

随着科技的飞速发展，未来的通信网络正逐步迈向一个全新的时代。在这个时代，AI Agent 将有望颠覆传统的通信模式，引领一场前所未有的变革。然而，这一变革并非一蹴而就，需要我们在架构设计和协议制定上做出深入的探索和研究。AI Agent 作为未来网络的核心元素，其潜力巨大，能够自主决策、优化网络性能，并为用户提供更加个性化的服务。但要实现这一愿景，我们必须设计出与之相匹配的架构和协议，以确保智能体能够在复杂的网络环境中高效、稳定地运行。这需要我们深入研究智能体的工作原理，理解其与传统网络架构的差异，从而构建出更加灵活、可扩展的网络体系。与此同时，未来的网络将提供泛在的 AI 服务，这意味着云、核心网、无线网和终端等各个层面都需要实现紧密的协同。这种协同不仅要求技术上的无缝对接，更需要我们在策略、管理和运维等方面做出全面的优化。只有这样，我们才能确保 AI 服务能够覆盖到网络的每一个角落，为用户提供无处不在的智能体验。

网络数字孪生作为解决 AI 概率性和网络高可靠性之间矛盾的主要途径，正逐渐受到业界的广泛关注。通过将网络实体与数字模型进行实时映射和交互，我们可以更加精准地预测和应对网络中的潜在风险，从而提高网络的稳定性和可靠性。同时，数字孪生技术还能够与大模型和 AI Agent 技术相互融合、相辅相成，共同推动未来网络的发展。

从端到端的视角来看，我们需要体系化地设计面向智能体的网络架构、安全可信机制及通信组网机制。这要求我们不仅要关注单个技术点的突破，更要从整体上把握网络的发展趋势，确保各个层面之间的协同和一致性。只有这样，我们才能构建出一个真正智能、高效、安全的未来网络。

智能体作为未来 6G 网络发展的新动能，其重要性不言而喻。我们期待与业界合作伙伴携手共进，共同推进 6G 智能体的系统创新、标准制定及产业试点应用。通过我们的共同努力，相信未来的通信网络将变得更加智能、更加高效，为人类社会的发展注入新的活力。

六、参考文献

[1] 2023 年 11 月 vivo 开发者大会，https://dev.vivo.com.cn/vdc/2023/index.html#/

[2] https://www.oppo.com/cn/events/find-x7-series-launch/

[3] https://www.honor.com/cn/phones/honor-magic6/

[4] https://xiaoai.mi.com/

[5] https://developer.apple.com/cn/videos/play/wwdc2024/101/

[6] Wang G, Liu J, Li C, et al. Cloud-Device Collaborative Learning for Multimodal Large Language Models[C]//Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 12646-12655.

[7] Lin Z, Qu G, Chen Q, et al. Pushing large language models to the 6G edge: Vision, challenges, and opportunities. arXiv preprint arXiv:2309.16739, 2023.

[8] Ding N, Qin Y, Yang G, et al. Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature Machine Intelligence, 2023, 5(3): 220-235.

[9] Zhang Z, Yang Y, Dai Y, et al. Fedpetuning: When federated learning meets the parameter-efficient tuning methods of pre-trained language models. ACL, 2023: 9963-9977.

[10] Douillard A, Feng Q, Rusu A A, et al. DiLoCo: Distributed Low-Communication Training of Language Models. arXiv preprint arXiv:2311.08105, 2023.

[11] Ye R, Wang W, Chai J, et al. OpenFedLLM: Training Large Language Models on Decentralized Private Data via Federated Learning. arXiv preprint arXiv:2402.06954, 2024.

[12] Xu M, Cai D, Wu Y, et al. Fwdllm: Efficient fedllm using forward gradient. arXiv:2308.13894, 2024.

[13] Z. Qin, D. Chen, B. Qian, B. Ding, Y. Li, and S. Deng, Federated fullparameter tuning of billion-sized language models with communication cost under 18 kilobytes, in ICML, 2024.

[14] Lin Z, Zhu G, Deng Y, et al. Efficient parallel split learning over resource-constrained wireless edge networks. IEEE Transactions

on Mobile Computing, 2024.

[15]Liu Y, Agarwal S, Venkataraman S. Autofreeze: Automatically freezing model blocks to accelerate fine-tuning (2021). arXiv preprint arXiv:2102.01386.

[16]Li S, Yuan G, Dai Y, et al. Smartfrz: An efficient training framework using attention-based layer freezing. arXiv preprint arXiv:2401.16720, 2024.

[17]Kuang W, Qian B, Li Z, et al. Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning. arXiv preprint arXiv:2309.00363, 2023.

[18]Rajbhandari S, Rasley J, Ruwase O, et al. Zero: Memory optimizations toward training trillion parameter models,SC20: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2020: 1-16.

[19]Fan T, Kang Y, Ma G, et al. Fate-llm: A industrial grade federated learning framework for large language models. arXiv preprint arXiv:2310.10049, 2023.

[20]Zhang J, Vahidian S, Kuo M, et al. Towards building the federatedGPT: Federated instruction tuning. IEEE ICASSP, 2024: 6915-6919.

[21]Jiang Feibo et al. “Personalized wireless federated learning for large language models.” ArXiv abs/2404.13238 (2024): n. pag.

[22]McMahan, H. B. et al. “Communication-efficient learning of deep networks from decentralized data.” International Conference on

Artificial Intelligence and Statistics (2016).

[23]Thapa Chandra et al. “SplitFed: When federated learning meets split learning.” ArXiv abs/2004.12088 (2020): n. pag.

[24]Zhang Jianyi et al. “Towards Building the Federated GPT: Federated Instruction Tuning.” ArXiv abs/2305.05644 (2023): n. pag.

[25]Jiang Jing et al. “Low-parameter federated learning with large language models.” ArXiv abs/2307.13896 (2023): n. pag.

[26]Fan Tao et al. "FATE-LLM: A industrial grade federated learning framework for large language models", ArXiv abs/2310.10049 (2023): n. pag.

[27]Chen Chaochao et al. “Federated large language model: A position paper.” ArXiv abs/2307.08925 (2023): n. pag.

[28]Yang Z, Ji W, Wang Z. Adaptive joint configuration optimization for collaborative inference in edge-cloud systems[J]. Science China Information Sciences, 2024, 67(4): 149103.

[29]Chen Y, Li R, Yu X, et al. Adaptive Layer Splitting for Wireless LLM Inference in Edge Computing: A Model-Based Reinforcement Learning Approach[J]. arXiv preprint arXiv:2406.02616, 2024.

[30]R. Gozalo-Brizuela and E. C. Garrido-Merchan, “A Survey of Generative AI Applications,” arXiv preprint arXiv:2306.02781, 2023.

[31]H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, “Deep Learning Enabled Semantic Communication Systems,” IEEE Transactions on Signal Processing, vol. 69, pp. 2663–2675, 2021.

[32]Q. Lan, D. Wen, Z. Zhang, Q. Zeng, X. Chen, P. Popovski, and K. Huang, “What is Semantic Communication? A View on Conveying Meaning in the Era of Machine Intelligence,” Journal of Communications and Information Networks, vol. 6, no. 4, pp. 336–371, 2021.

[33]C. Liang et al., "Generative AI-driven Semantic Communication Networks: Architecture, Technologies and Applications," in IEEE Transactions on Cognitive Communications and Networking, 2024, doi: 10.1109/TCCN.2024.3435524.

[34]L. Xia, Y. Sun, C. Liang, L. Zhang, M. A. Imran, and D. Niyato, “Generative AI for Semantic Communication: Architecture, Challenges, and Outlook,” arXiv preprint arXiv:2308.15483, 2023.

[35]H. Du, R. Zhang, Y. Liu, J. Wang, Y. Lin, Z. Li, D. Niyato, J. Kang, Z. Xiong, S. Cui et al., “Beyond Deep Reinforcement Learning: A Tutorial on Generative Diffusion Models in Network Optimization,” arXiv preprint arXiv:2308.05384, 2023.

[36]G. Liu, H. Du, D. Niyato, J. Kang, Z. Xiong, and B. H. Soong, “Vision based Semantic Communications for Metaverse Services: A Contest Theoretic Approach,” arXiv preprint arXiv:2308.07618, 2023.

[37]Y. Xu, G. Gui, H. Gacanin, and F. Adachi, “A Survey on Resource Allocation for 5G Heterogeneous Networks: Current Research, Future Trends, and Challenges,” IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 668–695, 2021.

[38]B. Du, H. Du, H. Liu, D. Niyato, P. Xin, J. Yu, M. Qi, and Y. Tang,“YOLO-based Semantic Communication with Generative AI-aided

Resource Allocation for Digital Twins Construction," arXiv preprint arXiv:2306.14138, 2023.

[39]H. Du, G. Liu, D. Niyato, J. Zhang, J. Kang, Z. Xiong, B. Ai, and D. I. Kim, "Generative AI-aided Joint Training-free Secure Semantic Communications via Multi-modal Prompts," arXiv preprint arXiv:2309.02616, 2023.

[40]L. Xia, Y. Sun, D. Niyato, X. Li, and M. A. Imran, "Joint User Association and Bandwidth Allocation in Semantic Communication Networks," IEEE Transactions on Vehicular Technology, 2023.

[41]D. Huang, X. Tao, F. Gao, and J. Lu, "Deep Learning-Based Image Semantic Coding for Semantic Communications," in 2021 IEEE Global Communications Conference (GLOBECOM). IEEE, 2021, pp. 1–6.

七、主要贡献单位和编写人员

贡献者	单位
吴佳骏、陈天骄、崔莹萍、邓娟	中国移动
王飞、黄欢欢、彭程晖	华为
孙万飞	中信科移动
周通	vivo
沈钢	上海诺基亚贝尔
倪万里、秦志金、侯玮琛	清华大学
于小雪、李荣鹏、张宏纲	浙江大学
刘贻静、汪云翔	电子科技大学
程乐、章辉	南开大学
杨一帆、杨铮、马瑛、曾捷	北京理工大学