HFedMoE：基于混合专家的资源感知异构联邦学习

Zihan Fang、Zheng Lin、Senkang Hu、Yanan Ma、Yihang Tao、Yiqin Deng，*IEEE* 会员，Xianhao Chen，*IEEE* 会员 和 Yuguang Fang，*IEEE* 会士

arXiv:2601.00583v1 [cs.LG] 2 Jan 2026

摘要—**联邦学习（FL）能够在不损害数据隐私的前提下微调大语言模型（LLM），但 LLM 规模庞大，使得资源受限客户端（如移动设备）难以进行设备端训练。因此，混合专家（MoE）模型应运而生，成为一种计算高效的解决方案：模型训练期间仅激活稀疏专家子集，从而在不牺牲性能的前提下降低计算负担。尽管将 MoE 集成到联邦学习微调中具有巨大潜力，但仍面临三项关键挑战：i）由于缺乏可靠的指标来衡量每个专家对本地微调性能的影响，为客户端选择合适的专家仍然困难；ii）客户端之间异构的计算资源严重阻碍了基于 MoE 的大语言模型微调，因为不同输入样本会动态激活不同专家，可能使资源受限设备不堪重负；iii）客户端专属专家子集和路由偏好会削弱全局聚合效果，因为未对齐的专家更新和不一致的门控网络会引入破坏性干扰。为应对这些挑战，我们提出 HFed MoE，这是一种异构的基于 MoE 的联邦学习微调框架，为每个客户端定制专家子集，以实现计算高效的大语言模型微调。具体而言，HFedMoE 根据专家对微调性能的贡献确定其重要性，随后从信息瓶颈视角自适应地选择专家子集，使其符合各客户端的计算预算。此外，我们还设计了一种稀疏性感知模型聚合策略，以重要性加权的方式聚合主动微调的专家和门控参数。大量实验表明，在训练准确率和收敛速度方面，HFedMoE 均优于最先进基准。**

索引词—**联邦学习、混合专家、大规模语言模型、微调。**

一、引言

近年来，GPT [1]–[3]，LLaMA [4]，[5]，和 DeepSeek [6] 等大语言模型（LLM）因其他们在处理高复杂度、大规模数据集方面表现出卓越能力 [7]–[10]。大语言模型通常遵循两阶段训练流程。首先，在海量文本语料库（例如维基百科）上预训练大语言模型，以学习通用的语言和语义表征。其次，使用任务专用数据，使预训练后的大语言模型适配下游任务。然而，微调大语言模型所需的海量数据引发了严重的隐私问题，成为实施大语言模型微调的一大障碍。例如，客户端通常不愿共享其隐私敏感数据（如个人健康记录或财务信息） [11],来训练共享的大语言模型。

为解决上述问题，联邦学习（FL） [12]–[18] 已成为一种可行的替代方案，使客户端能够在不暴露原始数据的情况下协同训练。大语言模型微调的标准联邦学习流程包含三个阶段：i）每个客户端使用本地私有数据集独立微调其大语言模型；ii）所有客户端将本地更新后的大语言模型上传至中央服务器进行模型聚合（例如，通过加权平均）；iii）在下一训练轮次开始前，将聚合后的大语言模型分发回各客户端。尽管联邦学习具有保护隐私的特性，但在资源受限设备上通过联邦学习微调大语言模型的计算开销仍然令人望而却步 [19]–[23]。商用级设备端 GPU（例如 NVIDIA Jetson Orin）缺乏足够的处理能力，无法微调 LLaMA-2 7B 等大语言模型；这类模型需要近 60 GB 的 GPU 显存[1]，且处理每 4K 个词元需要 187.9 TFLOPs [27],，远超移动客户端设备的能力。为降低大语言模型微调开销，人们提出了 LoRA [28] 和适配器 [29] 等参数高效微调技术。虽然参数高效微调大幅减少了可训练参数的数量，但仍需对整个大语言模型进行完整的前向传播和反向传播。由于大语言模型规模庞大，客户端设备仍无法充分满足通过联邦学习进行参数高效微调的计算需求。即使采用 LoRA，微调 LLaMA-2 7B 也需要超过 20 GB 的 GPU 显存，这很容易使大多数移动客户端设备不堪重负 [30]。

混合专家（MoE） [27], [31], [32] 为通过联邦学习在资源受限的移动客户端设备上微调大语言模型提供了一种结构高效的解决方案。不同于参数高效微调方法

本文所述研究工作是在智慧城市JC STEM实验室开展的，该实验室由香港赛马会慈善信托基金根据合同 2023-0108 资助。本研究还部分获得香港特别行政区政府全球STEM教授与研究人才枢纽的支持，以及香港创新及科技委员会 InnoHK CIMDA项目的支持。Yiqin Deng 的工作部分获得国家自然科学基金委员会第 62301300 号资助。Xianhao Chen 的工作部分获得港大-SCF金融科技学院研发基金资助。

Z. Fang、S. Hu、Y. Ma、Y. Tao、Y. Deng 和 Y. Fang 任职于香港城市大学香港 JC SmartCity STEM 实验室和计算机科学系，地址为中国香港特别行政区九龙（电子邮箱：zihanfang3-c@my.cityu.edu.hk；senkang.forest@my.cityu.edu.hk；yananma8-c@my.cityu.edu.hk；yihang.tommy@my.cityu.edu.hk；yiqideng@cityu.edu.hk；my.fang@cityu.edu.hk）。

Z. Lin 和 X. Chen 就职于中国香港薄扶林的香港大学电机电子工程系（电子邮件：linzheng@eee.hku.hk；xchen@eee.hku.hk）。

[1]本文使用 GPU内存占用作为量化计算资源消耗的代理指标 [24]–[26]。由于中间状态内存存储和大规模矩阵乘法等内存密集型操作主导着大语言模型微调，因此 GPU内存占用能够反映大语言模型微调的计算需求。这使其成为衡量资源受限的客户端设备是否能够进行微调的有效指标。

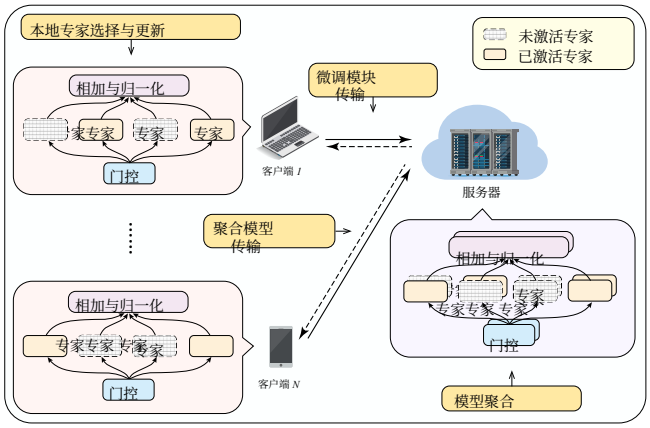


图1：通过联邦学习在多个客户端之间微调基于 MoE 的大语言模型的工作流程。

这些方法需要进行完整模型的前向传播和反向传播，而混合专家针对每个输入仅激活少量专家（例如，top-1 或 top-2），其余专家则保持非激活状态。这种稀疏激活在保留模型完整表征能力的同时，大幅降低了计算成本 [27], [32], [33], ，尤其是梯度计算和权重更新的成本；这两者在大语言模型微调中占据了主要训练成本 [34]。除计算优势外，混合专家也天然适用于联邦学习，因为依赖输入的专家路由能够实现客户端特定专门化，从而促进个性化模型适配。图1展示了混合专家如何与联邦学习结合进行大语言模型微调。每个客户端设备通过门控网络选择一组个性化专家并进行微调。随后，所有客户端设备将门控网络和专家上传至服务器，以进行模型聚合并更新大语言模型。这种设计能够达到与大型稠密模型（例如 LLaMA-2 7B）相当的性能，同时计算量不到其 40% [27]。

尽管将混合专家引入联邦学习前景广阔，但仍面临若干关键挑战。首先，由于客户端本地数据集存在异质性，为每个客户端选择合适的专家并非易事。不同专家在各客户端上的表现差异显著 [31], [35], 也就是说，某些专家可能对特定客户端数据集具有良好的泛化能力，但对其他客户端几乎没有帮助。因此，共享门控网络难以捕捉这些客户端特定的差异，导致专家利用不充分。这种不匹配不仅浪费计算资源，还会损害模型的整体性能。其次，客户端之间的异构计算资源严重限制了通过联邦学习部署基于混合专家的微调。不同输入词元所激活的专家各异，导致一个训练批次内同时激活的专家数量增加 [32], [36], 简单样本只需激活少数专家，而复杂样本则需要激活许多专家。这很容易超出资源受限的客户端设备的计算能力，导致微调失败，并大幅降低整体训练效率。第三，计算资源和数据分布各不相同，会导致客户端特定的专家选择和路由偏好，使基于混合专家的联邦学习中的模型聚合尤具挑战性。每个客户端只会微调与其本地数据相匹配的专家子集，使其他专家训练不足，并促使客户端的门控网络偏向不同的专家子集。这种不匹配会在聚合过程中造成严重干扰，从而降低全局模型的泛化能力。我们将在第II节通过实证测量研究探讨这些挑战。

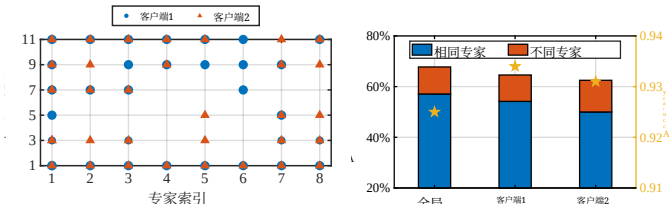
为应对上述挑战，我们提出了一种用于基于 MoE 的大语言模型微调的异构联邦学习框架，称为 HFedMoE，使异构边缘计算系统中的每个客户端都能选择性地微调专家子集。首先，为每个客户端设备选择合适的专家，我们提出一种专家重要性识别方案，用于量化各专家对不同客户端微调性能的贡献。其次，为适应客户端设备异构的计算预算，我们从信息瓶颈视角提出一种资源感知专家选择方法，动态选择关键专家子集，使其与各设备的计算预算相匹配，从而在不影响模型性能的前提下提升训练效率。最后，为缓解部分专家更新和不同路由偏好导致的聚合差异，我们设计了一种稀疏性感知模型聚合策略，仅更新客户端上实际参与训练的专家，同时按重要性加权贡献聚合门控参数，从而增强全局混合专家模型在结构异构情况下的泛化能力。本文的主要贡献总结如下。

- 我们提出 HFedMoE，这是一种基于混合专家的联邦学习微调框架，可为每个客户端定制用于模型训练的专家子集，从而在异构边缘计算能力下实现高效的大语言模型微调。
- 我们设计了一种专家重要性识别方案，用于量化专家对本地训练性能的贡献，从而优先选择关键专家进行微调。
- 我们设计了资源感知的专家选择机制，可在微调过程中动态选择关键专家子集，以适应每台设备的计算预算。
- 我们开发了一种稀疏性感知模型聚合策略，以显式处理部分专家更新和路由不一致，从而减轻客户端间结构异构性造成的性能下降。
- 我们通过大量实验对 HFedMoE 的微调性能进行了实证验证，结果表明，HFedMoE 在模型准确率和收敛速度方面均优于最先进的框架。

本文组织结构如下。第II节通过揭示在联邦学习网络中引入混合专家所面临的挑战，阐述 HFedMoE 的设计动机。第三节详细介绍框架设计，随后在第五节进行性能评估。第六节讨论相关工作和技术局限性。最后，第七节给出结论。

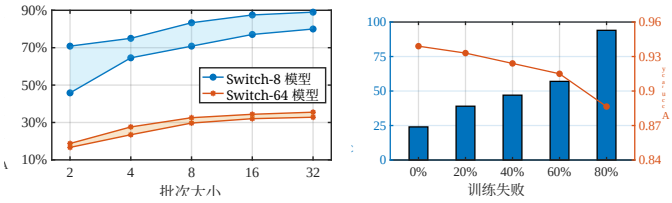
II. 挑战与动机

本节将开展广泛的初步测量研究，阐述在联邦学习中微调基于混合专家的大语言模型所面临的关键挑战，并以此为 HFedMoE 的设计提供依据。



(a) 客户端特定的专家选择　(b) 聚合后的比较

图 2：在批次大小为 4 的情况下，AGNews 数据集上各客户端在全局或本地专家路由下的专家激活情况。



(a) 激活专家比例　(b) 收敛性能

图 3：不同批次大小下的激活专家比例，以及不同训练故障率下的收敛性能（训练轮次和测试准确率）。

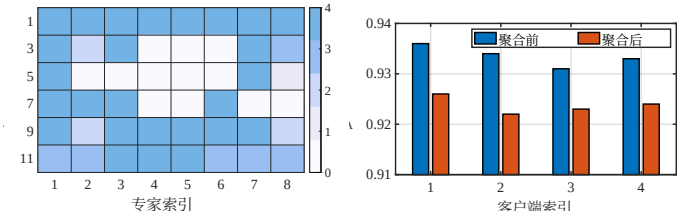
A. 次优专家利用

对于基于 MoE 的联邦学习微调，专家路由通常由共享全局门控网络 [37], [38]控制。尽管这种方式能够实现集中式协调，却无法捕捉客户端特定数据分布的异质性。由于每个专家都专注于不同的语义特征 [37],，其贡献（即所提供的任务相关信息）在不同客户端之间差异显著 [31], [35]。然而，全局门控网络是根据所有客户端的平均路由性能进行优化的，因而忽略了客户端特定偏好。这种不匹配会导致对某些客户端至关重要的专家利用不足，而频繁激活的专家可能对本地任务贡献甚微，最终造成专家利用次优，并降低训练性能。

为研究次优专家利用的影响，我们在 AGNew 数据集 [39] 上使用每层含有 8 个专家的 Switch Transformer [32]开展了一项动机性研究。我们比较了全局路由（即所有客户端共享单个门控网络）与客户端特定路由（即每个客户端独立训练其门控网络）。图 2a 显示，不同客户端的本地专家选择存在显著差异，表明专家对本地数据集具有敏感性。此外，图 2b 显示，门控网络聚合会改变本地专家选择，且无法准确地为特定客户端优先选择关键专家，从而降低本地微调准确率（图中的黄色星标）。这些观察结果凸显了识别专家重要性并针对每个客户端有选择地激活专家的必要性，以提升效率和个性化水平。

B. 异构计算资源

尽管混合专家模型通过稀疏专家激活实现高效的大语言模型微调，但客户端设备之间计算能力的异质性源于硬件配置和部署环境的差异 [17], [40]–[42],带来了重大挑战。尽管



（a）4 个客户端中的选定专家　（b）聚合性能

图4：AGNews 数据集上专家的激活频率及各客户端聚合性能对比。

每个词元仅激活排名第一的专家，但训练批次中不同样本的词元之间，专家选择往往各不相同，因此同时激活的专家数量也会变化 [32], [36]。这削弱了混合专家模型预期的稀疏性设计，并给资源受限客户端（例如配备商用级图形处理器的客户端）带来了沉重的计算负担。因此，资源需求与可用资源之间的不匹配，使得在联邦学习中进行设备端微调变得不可行。例如，微调 DeepSeekMoE-16B 模型，每处理 4K 个词元就需要 74.4T 次浮点运算 [27], ，这仍超出了大多数客户端图形处理器的处理能力，进而降低了整体训练效率。

为更好地了解客户端计算能力异构对联邦学习训练性能的影响，我们使用 Switch Transformer，在 AGNews 数据集上开展了探索性实验，其中每层包含8个或64个专家。通过分析不同批次大小下序列长度固定为128时的专家激活情况，图3a表明，即使采用 top-1 路由且批次大小仅为2，被激活的64个专家中仍有超过15%，远高于每层预期选择一个专家的比例。此外，图3b显示，资源有限的客户端往往无法完成本地大语言模型微调，而客户端之间计算能力的严重差异会显著降低协同训练整体的收敛性与稳定性。这些结果凸显了低效的逐 token 专家选择对资源受限的混合专家微调所造成的负面影响，也促使我们设计资源感知专家选择机制。

C. 模型聚合差异

在传统联邦学习中，模型聚合方法 [12], [22], [43] 假设模型结构统一，且所有参数都有意义的更新。然而，在基于混合专家的联邦学习中，每个客户端仅通过其门控网络微调针对本地数据分布定制的稀疏专家子集，导致模型结构存在差异，许多专家因此未被更新或从未被使用。此外，由于客户端的门控网络倾向于选择不同的专家子集，各客户端会形成不同的路由偏好，进而导致各客户端在专家选择方面的门控参数更新不一致。直接聚合此类异构模型结构会引发破坏性干扰，从而降低全局模型的泛化能力。

为实证展示模型差异如何影响聚合性能，我们使用包含 8 个专家的 Switch Transformer，在 AG-News 数据集上开展实验，其中

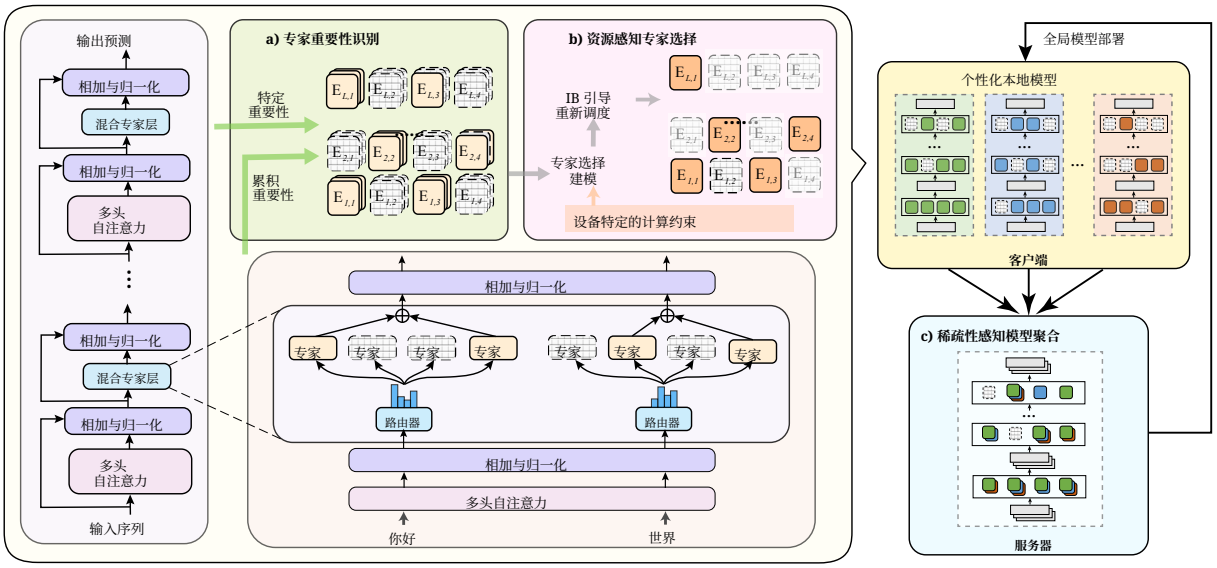


图5：HFedMoE框架概述。每个客户端在一个训练批次内量化各专家的重要性，并在设备特定计算约束下选择关键专家子集进行本地微调。只有已激活专家及其门控网络会被上传至服务器，服务器采用稀疏性感知聚合策略，仅聚合活跃专家，并显式对齐各客户端的路由偏好。

每个客户端独立选择与其本地数据相匹配的专家子集并进行微调。图4a显示，在客户端本地微调过程中，近 25% 的专家从未被激活，这表明参数更新存在专家差异。相比之下，超过 20% 的专家会被不同客户端选中，凸显了各客户端之间路由偏好的差异。此外，图4b显示，标准 FedAvg 聚合前后，客户端特定模型的性能存在显著差距，表明专家差异和门控网络更新不一致都会导致性能下降。这些观察结果凸显了采用稀疏性感知聚合策略的必要性：此类策略能够明确处理由客户端特定路由偏好和专家选择导致的模型差异，从而充分发挥基于 MoE 的大语言模型在联邦环境中的潜力。

III. 框架设计

本节介绍 HFedMoE，这是一种面向异构边缘计算系统的基于 MoE 的联邦学习微调框架，旨在动态选择关键专家子集。我们将首先概述系统，然后详细介绍HFedMoE框架。

A. 概述

为应对上述挑战，我们提出 HFedMoE，这是一种面向计算异构环境下基于 MoE 的大语言模型微调的资源高效联邦学习框架。我们的设计包含三个核心组件：专家重要性识别、资源感知专家选择和选择性模型聚合。首先，为了给每个客户端选择合适的专家，我们提出一种专家重要性识别方案，用于量化每个专家对本地微调性能的贡献（第III-C节）。其次，为了在微调期间适应客户端之间异构的计算资源，资源感知专家选择模块从信息瓶颈视角对专家选择进行建模（第III-D1节），并在客户端特定计算预算内重新安排最关键专家的激活（第III-D2节）。最后，我们设计了一种稀疏性感知模型聚合策略，在不同模型结构之间仅选择性上传活跃专家（第III-E1节），并根据重要性加权贡献自适应地聚合门控网络的参数（第III-E2节），从而增强全局泛化能力，同时保留关键本地更新。

如图5所示，HFedMoE 的工作流程分为三个步骤：i）在本地微调期间，每个客户端根据其本地数据确定专家的重要性。ii）利用估算出的专家重要性，客户端在本地计算预算范围内选择性地激活一部分影响较大的专家。iii）完成本地更新后，仅上传主动训练的专家和门控网络。随后，服务器执行稀疏性感知模型聚合，并将聚合模型分发给所有客户端，以进行下一训练轮次。

B. 系统模型

我们考虑一种联邦设置，其中 C 个客户端在异构计算资源约束下协同微调一个共享的基于 MoE 的大语言模型。每个客户端 $c \in\{1,\ldots,C\}$ 持有一个私有数据集 $\mathcal{D}_c$，该数据集从不同的数据分布中采样得到。全局模型由一个共享门控网络 $\mathcal{R}$ 和一个专家集 $\mathcal{E} = \{\mathbf{E}_{1,\,1},\ \ldots\ldots,\ \mathbf{E}_{L,\,S}\}$ 构成，其中每个 L 混合专家层包含 S 个并行专家。对于每个输入词元，门控网络都会计算路由分数，并在每层激活一个 top-k 专家子集（$k \ll S$）。这种稀疏激活使专家能够针对各客户端的本地数据集进行专门化，从而以较低的计算成本实现个性化模型适配。

在本地训练轮次 t 中，每个客户端 c 仅使用其本地数据更新选定的专家子集 $\mathcal{E}_c \subseteq \mathcal{E}$ 和门控网络 $\mathcal{R}_c$，并受其计算预算 B_c的约束。设

θ_c^t 表示由客户端 c 更新的本地模型参数，θ^t 表示全局模型。每个客户端最小化本地目标函数 $\mathcal{L}_c(\theta_c^t) = \mathbb{E}_{(x,y)\sim\mathcal{D}_c}\left[\ell(f(x;\theta_c^t),y)\right],$，其中 $\ell(\cdot)$ 是任务特定损失函数，$f(\cdot;\theta_c^t)$ 表示具有参数 θ_c^t 的混合专家模型的输出。本地训练完成后，客户端将异构模型参数 $\{\mathcal{E}_c,\mathcal{R}_c\}$ 上传至服务器，服务器随后聚合收到的稀疏更新，以更新全局模型 θ^{t+1} 全局。我们的目标是使每个客户端能够在计算约束下，根据其本地数据选择关键专家，同时协作参与构建一个在准确率与效率之间取得平衡的全局共享模型。

C. 专家重要性识别

回顾第 II-A 节，混合专家模型中的门控网络通常针对所有客户端的平均路由性能进行优化，因此往往无法准确优先选择对各个客户端最有益的专家。这会导致次优专家利用，并降低微调性能，凸显了估计专家重要性的必要性，以便更有效地进行选择。因此，我们提出了一种专家重要性识别方案，用于量化每个专家对本地微调性能的贡献，使客户端能够选择与其本地数据最相关的专家子集。

在标准混合专家模型中，每一层的门控网络会根据输入词元 x，为每个专家 e 计算路由分数 $G_e(x)$。经过 Softmax 归一化后，这些分数反映了各专家对当前输入的相对贡献。因此，选择分数最高的前 k 个专家进行微调，

$$\mathcal{E}_c = \text{Top-}k\left(\{Softmax(\mathbf{G}(x))\}_{e\in\mathcal{E}}\right), \tag{1}$$

最后，该混合专家层的输出 y 由选定专家的加权和计算得到：

$$y = \sum_{e\in\mathcal{E}_c} G_e(x)\cdot\mathbf{E}_e(x), \tag{2}$$

其中，$\mathbf{E}_e(x)$ 表示专家 e 的网络对输入词元 x 的输出。该路由分数 $G_e(x)$ 可用作专家重要性的估计值。

虽然这种逐词元路由机制支持动态选择专家，但它针对彼此孤立的词元进行操作，不会考虑专家在不同样本间持续激活的情况。然而，某些专家会在本地数据中频繁激活，并对特定客户端做出更大贡献。 [31], [35], 未能捕捉这些跨样本激活，会限制模型学习客户端特定专家偏好的能力。为此，我们通过分析一个批次中所有词元上的专家激活来量化专家重要性。由于小批量处理是大语言模型训练中的标准做法，这种方法几乎不会增加开销，同时支持更稳定、个性化的专家选择，从而实现资源高效的微调。为此，我们按如下方式识别在不同输入中持续发挥作用或对特定样本尤为重要的专家。

- **累积重要性：** 累积贡献较高的专家通常能够捕捉可泛化的特征 [33], ，从而为多样化样本提供有效的表征。优先选择持续提供信息的专家，可以最大限度地减少冗余激活，进而促进共享计算，同时不会牺牲模型性能。因此，我们将专家 e 的累积重要性定义为其在批次大小为 B的批次 x_i中所有样本上的平均激活值，表示为

$$s_b^{cumul}(e) = \frac{1}{B}\sum_{i=1}^{B} G_e(x_i). \tag{3}$$

该指标反映了专家 e 的激活频率和响应强度，可作为衡量其在整个批次中捕捉标签相关特征能力的替代指标。

- **特定重要性：** 仅依赖跨样本的累积贡献，无法捕捉那些激活频率较低、却能对特定样本（通常是难以分类的样本）做出关键响应的专家。例如，某个专家可能仅对少数几个具有挑战性的样本做出显著贡献，这表明它具有很强的类别特定表征能力。忽略此类专家，可能会降低模型在少数类或困难类别上的性能。为此，我们引入特定重要性，用于评估专家对单个样本的贡献，确保保留其对具有挑战性的输入所做出的关键响应。具体而言，我们使用每个专家在一个小批次所有样本中的最高路由分数，以捕捉其与任一样本的最大相关性及峰值影响，公式如下：

$$s_b^{specific}(e) = \max_{1\le i\le B} G_e(x_i). \tag{4}$$

该指标刻画了专家对单个样本的峰值影响，确保低频但影响大的专家在微调过程中得以保留。

最后，批次 b 中专家 e 的总体重要性 $s_b(e)$ 可表示为其累积重要性与特定重要性的加权组合：

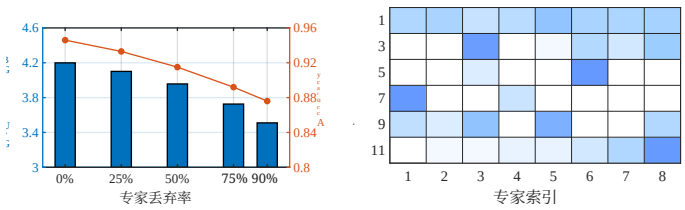
$$s_b(e) = \lambda\cdot s_b^{cumul}(e) + (1-\lambda)\cdot s_b^{specific}(e), \tag{5}$$

其中，λ 是一个超参数，用于控制激活多样性与特定性之间的权衡。这使我们能够在专家选择中平衡泛化与个性化，从而提高效率和有效性。

D. 资源感知专家选择

如第 II-B 节所述，在资源受限客户端上微调混合专家模型，会受到跨样本过度专家激活的限制。现有研究 [27], [31], [32] 主要根据路由分数为每个词元选择排名前 k 的专家，却忽略了一个批次内所有词元的累积计算负担，这往往会超出客户端能力。为解决这一瓶颈，我们设计了一种资源感知专家选择策略，根据每个客户端的计算预算，自适应地限制每个批次中激活的专家数量。

*1)*专家选择建模：要在资源受限客户端上实现高效微调，就必须在每个客户端的计算预算内选择性地激活专家。一种直接的方法是随机丢弃门控网络选出的若干专家。然而，图 6a 显示，虽然减少专家数量可以降低计算成本，但也会导致模型性能显著下降。



(a) 不同的专家数量　　(b) 专家激活

图 6：不同专家丢弃率对 AGNews 数据集上微调 8 专家 Switch Transformer 的影响。

专家可以降低计算成本，但也会导致模型性能显著下降。幸运的是，图 6b 表明，专家选择高度不均衡：少数专家带来了大部分性能提升，而其他专家的贡献则微乎其微。这些观察促使我们通过贡献引导选择策略缩小候选专家集，从而更好地适应资源限制，同时不损害模型性能。

受信息瓶颈（IB）原理 [44],的启发，专家紧凑子集可以被视为一种压缩表征：在有限计算预算下保留最大化的任务相关信息，其中每个专家的贡献由其任务相关信息量化，而输入冗余的最小化则促进稀疏专家激活。因此，我们引入一种基于 IB 的专家选择建模解释，从而提升可解释性以及对异构客户端约束的适应性。

具体而言，我们将一个批次上的专家激活模式重新解释为随机潜在表示 $\mathbf{z} = \{z_e, \forall e \in \mathcal{E}\}$，其中 $z_e \in \{0,1\}$ 表示对于给定输入 x，专家 e 是否被激活。从信息瓶颈的角度来看，专家激活模式$\mathbf{z}$充当输入 x 的压缩中间表征，用于预测目标标签 y。基于经典信息瓶颈公式 [44], ，我们的目标是通过最大化以下目标来选择专家的稀疏子集：

$$\max_{p(\mathbf{z}|x)} \; I(\mathbf{z};y) - \beta \cdot I(\mathbf{z};x) \tag{6}$$

其中， $I(\mathbf{z};y)$衡量专家激活模式$\mathbf{z}$中保留的任务相关信息，反映其对模型性能的贡献。相比之下， $I(\mathbf{z};x)$表示输入信息的冗余，可作为计算成本的代理指标。超参数 β 在任务相关信息量与资源压缩之间取得平衡。

由于混合专家模型中的 $I(\mathbf{z};y)$ 和 $I(\mathbf{z};x)$ 难以求解，我们采用变分近似 [45]:

$$I(\mathbf{z};x) \le \mathbb{E}_x \left[KL(p(\mathbf{z}|x)||p(\mathbf{z})) \right] \tag{7}$$

$$I(\mathbf{z};y) \ge \mathbb{E}_{(x,y)} \left[\mathbb{E}_{\mathbf{z}\sim p(\mathbf{z}|x)} (\log q(y|\mathbf{z})) \right] + H(y) \tag{8}$$

其中， $p(\mathbf{z}|x)$ 表示给定输入 x 时专家激活的条件分布， $p(\mathbf{z}) \sim \mathcal{N}(0, I)$ 表示先验分布，而 $q(y|\mathbf{z})$ 是用于近似标签似然的变分解码器。由于 $H(y)$ 与专家行为无关，因此在重要性比较中（即在所构建的优化中）可以将其省略。根据最优信息压缩原则，我们将原始问题中难以处理的联合优化信息瓶颈目标函数（方程（6））近似为：在一个批次 b 内分别评估每个专家的信息瓶颈贡献。对于每个专家e，其信息瓶颈贡献定义为

$$\begin{aligned} I_b(e) = I(z_e;y) - \beta I(z_e;x) \ge \mathbb{E}_{z_e\sim p(z_e|x)} \left[\log q(y|z_e)\right] \\ + H(y) - \beta \cdot KL(p(z_e|x)||p(z_e)) \end{aligned} \tag{9}$$

其中， $I_b(e)$ 表示专家 e 的信息量。当 $I_b(e)$ 较高时，专家 e被认为更为关键，因为这表明它与任务的相关性更高，且表征更加紧凑。

在混合专家框架中，专家激活 z_e 并未被显式建模，而是由门控网络隐式决定。由于路由分数 $G_e(x)$ 反映了输入 x 被分配给专家 e 的情况，因此激活分布 $p(z_e|x)$ 可以自然地近似为 $G_e(x)$。因此，边缘分布 $p(z_e)$ 可以估计为大小为 B 的批次中路由概率的平均值：

$$p(z_e|x) = G_e(x), \; p(z_e) = \frac{1}{B}\sum_{j=1}^{B} G_e(x_j), \tag{10}$$

因此， $I(z_e;x)$ 和 $I(z_e;y)$ 的近似值可直接根据路由行为表示为

$$KL(p(z_e|x)||p(z_e)) = \frac{1}{B}\sum_{i=1}^{B} G_e(x_i)\log\frac{G_e(x_i)}{p(z_e)}, \tag{11}$$

$$\mathbb{E}_{z_e\sim p(z_e|x)} \left[\log q(y|z_e)\right] = p(z_e|x)\cdot \log q(y|z_e). \tag{12}$$

然而，以专家激活 $q(y|z_e)$ 为条件的标签似然并未显式提供。为此，我们通过分析专家 e 的路由行为来估计 $I(z_e;y)$。关键在于，如果一组具有相同标签 y_i 的输入 x 中， $G_e(x)$ 始终较高，那么专家e 很可能捕获了与 y_i 相关的典型判别性模式。基于这一观察，路由行为本身间接反映了互信息：

$$I(z_e;y) \approx f(\{G_e(x)\}_{i=1}^{B}) \tag{13}$$

为实现高效且可解释的估计，我们转而推导一种可处理近似，将$I(z_e;y)$分解为两个可解释的组成部分：累积重要性$s_b^{cumul}(e)$和特定重要性$s_b^{specific}(e)$，其定义见公式（5）。为此，专家 e在整个批次 b上的信息瓶颈贡献 $I_b(e)$ 可表示为

$$I_b(e) = s_b(e) - \beta \cdot \frac{1}{B}\sum_{i=1}^{B} G_e(x_i)\log\frac{G_e(x_i)}{p(z_e)} \tag{14}$$

$I_b(e)$越高，表明专家 e越具有信息性且越紧凑，因此在计算资源约束下，它是进行选择性微调的理想候选。

2)信息瓶颈引导的专家重新调度： 给定客户端c的模型中所有专家 $e \in \{\mathbf{E}_1、\ldots\ldots、\mathbf{E}_S\}$ 的专家贡献 $I_b(e)$，我们的目标是为每个客户端c选择一个子集 $\mathcal{E}_{\text{active}}(c)$，其中最多包含 $C_{\text{budget}}(c)$个专家，以便在整个批次中激活，并优先激活信息量最大的专家。令 $\mathcal{E}_{\text{union}}(c) = \bigcup_{i=1}^{B} \mathcal{E}_i^{\text{route}}(c)$表示批次 b中所有样本的选定专家的并集，其中门控网络为每个样本选择排名前k 的专家 $\mathcal{E}_i^{\text{route}}$。在客户端 c的计算预算 $C_{\text{budget}}(c)$下（即每个批次可激活的专家数量上限），我们求解以下选择问题，其依据是方程（6）中的优化目标，

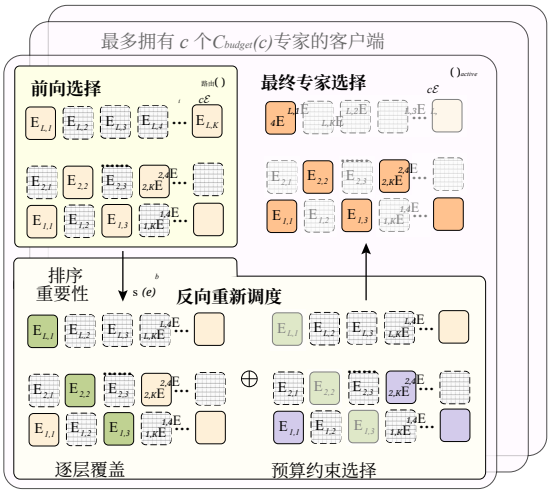


图 7：计算约束下的资源感知专家选择，其中专家激活会在反向传播阶段基于 IB 引导的重要性排序进行重新调度。

每个批次可激活的专家数量），我们根据方程 (6) 中的优化目标，求解以下选择问题：

$$\mathcal{E}_{\text{active}}(c) = \arg \max_{\substack{\mathcal{E}_{\text{active}}(c) \subseteq \mathcal{E}_{\text{union}}(c) \\ |\mathcal{E}_{\text{active}}(c)| \le C_{\text{budget}}(c)}} \sum_{e \in \mathcal{E}_{\text{active}}(c)} I_b(e) \tag{15}$$

为实现这一点，我们会对前向传播阶段选定的专家进行重新调度，使其仅在反向传播过程中激活关键专家，如图 7 所示。在前向传播阶段，对于批次中的每个样本 x_i，门控网络会根据路由分数为其选择排名前列的 k 个专家 $\mathcal{E}_i^{\text{route}}(c)$，从而保留原始的样本级路由决策。在反向传播阶段，我们会在整个批次范围内重新调度专家激活，以满足计算约束，具体采用以下资源感知流程：

- **基于重要性的排序：** 我们根据重要性分数 $s_b(e)$ 对所有 $e \in \mathcal{E}_{\text{并集}}(c)$ 按降序进行排序。
- **逐层覆盖：** 为确保所有混合专家层保持正常运行，我们首先选择每层中最重要的专家。
- **预算约束选择：** 在剩余预算内，我们继续选择得分最高的专家，而不考虑其所在的层，直到激活专家总数达到预算上限 $C_{\text{预算}}(c)$。

在反向传播过程中，只有 $\mathcal{E}_{\text{active}}$ 中选定的专家参与梯度计算和参数更新。所有其他非活跃专家的梯度都会被屏蔽，从而在不降低模型性能的情况下减少计算成本。专家选择过程会在每个训练轮次中重新执行，以确保动态适应本地数据。模型使用标准任务损失函数（例如交叉熵）进行优化，在专家选择阶段不引入任何额外的正则化。

E. 稀疏性感知模型聚合

如第 II-C 节所述，客户端特定约束会导致不同的路由偏好和个性化专家选择，从而引入破坏性干扰，并严重削弱全局模型的性能。这些差异会因异构计算资源而进一步加剧，因为资源感知的专家选择只会在每个客户端上激活重要的专家。为解决这一问题，如图 8 所示，我们提出一种稀疏性感知的模型聚合策略，仅更新在客户端上经过主动训练的专家，同时按重要性加权聚合门控参数，以确保全局路由器能够准确捕捉客户端特定路由行为。我们在算法 1 中总结了所提出的基于混合专家的大语言模型联邦学习微调方法。

1) 选择性专家聚合： 在基于MoE的联邦学习中，每个客户端根据其本地门控网络，仅更新专家的一个子集，导致许多专家仅得到部分训练，甚至完全保留随机初始化状态。然而，标准聚合方法（如 FedAvg） [12] 假设客户端之间具有一致的模型结构，并且所有参数都获得了有意义的更新。这种不匹配会在聚合无关或过时的专家权重时造成破坏性干扰，从而降低全局模型的泛化能力。

为了解决部分专家更新之间的差异，我们提出了一种选择性专家聚合策略。我们不再对所有专家一概进行直接聚合，而只更新在客户端上经过积极训练的专家。记 $u_c(e)$ 为专家 e 在客户端 c 的本地样本 D_c 上的归一化使用率，其估算方式如下：

$$u_c(e) = \frac{1}{|D_c|} \sum_{x_i \in D_c} G_e(x_i). \tag{16}$$

为过滤掉不可靠或非活跃专家的更新，我们引入使用率阈值 τ，使每个客户端仅上传使用率 $u_c(e) \ge \tau$ 的专家。随后，服务器保持非活跃专家不变，仅聚合这些活跃专家，具体如下：

$$E_e^{t+1} = \sum_{c=1}^{C} \frac{|D_c| \mathbf{1}(u_j(e) \ge \tau)}{\sum_{j=1}^{C} |D_j| \mathbf{1}(u_j(e) \ge \tau)} E_e^t(c) \cdot \mathbf{1}(u_c(e) \ge \tau), \tag{17}$$

其中，$E_e^t(c)$ 表示专家 e 在客户端 c 上经过第 t 轮本地微调后的参数。指示函数 $\mathbf{1}(\cdot)$ 确保客户端的专家 e 仅在经过积极训练时才参与聚合；否则，如果该专家未被使用或训练不足，则将其排除。该策略可避免与专家交互不足的客户端所产生的更新，从而提升参数一致性并保留专家的专门化能力。

2) 重要性加权门控聚合： 除了部分专家更新之外，客户端还会形成不同的路由偏好，因为它们的门控网络往往会优先选择不同的专家子集。这种差异导致各客户端的门控参数值不一致，反映出不同的专家激活模式。若直接聚合门控参数而不考虑客户端特定路由偏好，就会削弱本地学习到的专家选择模式，导致路由决策模糊不清。因此，聚合后的门控网络难以根据本地数据调整专家选择，从而损害训练稳定性并阻碍模型收敛。

为缓解这种不一致性，我们提出一种重要性加权门控聚合策略，明确纳入专家偏好和路由一致性。其核心思想是

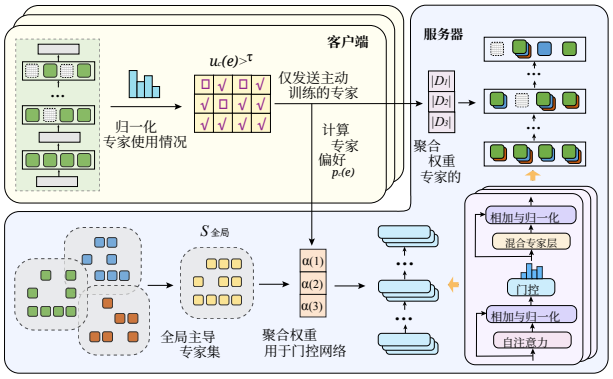


图 8：稀疏性感知模型聚合。选择性地聚合活跃专家，并按重要性加权门控参数，以对齐各客户端的路由行为。

根据各客户端的专家偏好及其与全局分布的一致性，调整其对全局门控网络的贡献。贡献较大的客户端将获得更大的聚合权重，从而确保全局门控网络不会因路由行为的差异而发生偏移。

为量化各客户端之间路由行为的一致性，我们关注能够体现每个客户端路由偏好的主导专家。对于每个客户端 c，其主导专家子集由本地数据下主动训练的专家确定，这些专家对路由决策的贡献最大。根据公式（16）中专家 e 的归一化使用率 $u_c(e)$，客户端 c 的主导专家子集可表示为 $\mathcal{S}_c = \{e|u_c(e) \geq \tau\}$，而全局主导专家集为 $\mathcal{S}_{\text{global}} = \cup\mathcal{S}_c$。随后，将客户端 c 的路由一致性 $r(c)$ 量化为其主导专家与全局主导专家集的重叠比例：

$$r(c) = \frac{|\mathcal{S}_c \cap \mathcal{S}_{\text{global}}|}{|\mathcal{S}_{\text{global}}|}, \tag{18}$$

其中，值越高，表示路由行为越一致、越可靠，因而更适合用于聚合。

虽然主导专家重叠能够反映客户端是否关注相似的专家子集，但它无法考虑这些专家的相对重要性，而后者在不同本地数据分布之间可能存在显著差异。为了进一步考虑各个专家的贡献，我们将专家重要性纳入聚合权重。由于公式（5）中的专家重要性分数 $s_b(e)$ 可估计专家在本地数据分布下的重要性，因此将其用作专家偏好，确保对本地性能影响更大的专家在全局门控网络的聚合中发挥更关键的作用。通过计算主导专家子集中的专家重要性，专家 e 对客户端 c 的偏好定义为

$$p_c(e) = u_c(e) \cdot s_b^c(e), \forall e \in S_c. \tag{19}$$

最后，客户端 c 的聚合权重定义为综合考虑其路由一致性 $r(c)$ 和专家偏好 $p_c(e)$：

$$\alpha(c) = \frac{\sum_{e\in\mathcal{S}_c} p_c(e)}{|\mathcal{S}_{\text{global}}|}, \tag{20}$$

这会突出那些既常被选中又在本地具有重要性的专家。随后，通过重要性加权平均更新全局门控网络参数：

$$G_{\text{gating}}^{t+1} = \sum_{c=1}^{C} \alpha(c) \cdot G_c^t, \tag{21}$$

其中，G_c^t 表示第 t个训练轮次中客户端c的门控网络参数。该门控聚合策略确保那些专家被频繁激活且持续使用的客户端在聚合中具有更大的影响力，从而提升基于混合专家的联邦学习训练的稳定性和收敛性。

算法 1： 联邦 MoE 微调 with 资源感知专家选择

输入： 总训练轮数 T，本地训练轮数 E，计算预算 C_{budget}
为所有专家初始化全局模型 θ^0 e;
对每一轮 $t = 1$ **至** T **执行**
  对每个客户端 c **执行**
    下载全局模型 θ_{global}^t 并初始化本地模型 $\theta_c \leftarrow \theta_{\text{global}}^t$;
    对于每个本地训练轮次 $e = 1$ **至** E **执行**
      对于每个小批次 $\mathcal{B}$ **执行**
        计算每个样本和专家的路由分数 $G_e(x_i)$ 样本 x_i 和专家 e;
        使用方程 (14) 计算专家重要性;
        通过以下方式选择专家 $\mathcal{E}_{\text{active}}$ 下的 $C_{\text{budget}}(c)$ 方程 (15);
        仅对选定的专家进行反向传播;
        使用随机梯度下降更新 θ_c ;
      end
    end
    根据方程 (16) 计算每个专家的 $u_c(e)$，并筛除非活跃专家;
    计算专家偏好 $\sum_{e\in\mathcal{S}_c} p_c(e)$ 的总和在主导专家子集之上;
    上传活跃专家、其他模型更新以及专家偏好之和至服务器;
  end
  服务器聚合专家和门控网络与方程 (17) 和方程 (21) 聚合专家与门控网络;
结束 输出： 最终全局模型 θ_{global}^T

IV. 仿真设置

本节详细介绍仿真设置，以评估我们提出的 HFedMoE 在异构基于混合专家的大语言模型微调中的表现。实验采用每层包含 64 个专家的 Switch Transformer 和 DeepSeek-MoE-16 模型。我们使用经过精心选择的超参数，将 HFedMoE 的性能与多种基线方法进行比较，以确保比较公平。

*1)*模型与数据集：在我们的实验中，我们采用了两种具有代表性的基于混合专家的大语言模型：Switch Transformer [32] 和 DeepSeek-MoE-16B [27]。Switch Transformer 对每层64个专家采用 Top-1门控机制，在仅激活11B参数的同时，可高效扩展至超过395B模型参数。DeepSeek-MoE-16B采用表达能力更强的骨干网络，并使用64个专家中的 Top-2 门控，其性能可与 LLaMA2-7B 模型相媲美。为评估模型性能，我们报告已收敛全局模型在测试集上的



图9：我们的 HFedMoE 原型和测试平台。

准确率，评估所用数据集广泛且应用普遍，涵盖从基础语义分类到高级知识推理的任务：（1）AGNews [39]：对新闻标题和描述进行四分类主题分类，类别涵盖世界、体育、商业和科学/技术领域。（2）PIQA [46]：用于物理常识推理的二选一数据集。给定一个简短场景，模型需要从两个解决方案中选出更合理的一项。（3）HellaSwag [47]：用于常识推理的四选一基准测试。每个样例都要求从给定上下文中选出最合理的续写。数据经过对抗性筛选，以避免模型仅进行表面模式匹配。（4）MMLU [48]：涵盖多个学术和专业领域的57项任务基准测试。每项任务均包含四选一问题，旨在测试跨领域推理能力。

2) 基线方法： 为研究我们的 HFed-MoE 框架的优势，我们将其与以下基线方法进行比较：

- **FedAvg [12]** 通过加权平均聚合模型权重，是联邦学习优化中的标准基线方法。
- **PFL-MoE [35]**将混合专家与个性化联邦学习相结合，允许每个客户端根据本地数据训练共享专家的一个子集，从而基于全局共享知识实现客户端特定的专家选择。
- **FedMoE [49]** 通过在初步微调期间识别次优子模型来降低每个客户端的计算负担，将专家激活限制在更小的子集内。
- **SEER-MoE [33]**通过采用门控正则化剪枝专家总数，提高混合专家模型中稀疏专家激活的效率，并通过更充分地利用专家资源来优化计算。

3) 超参数： 在实验中，我们使用 Switch-base-64 和 DeepSeek-MoE-16B 作为预训练骨干模型，实现了用于基于 MoE 的大语言模型指令微调的 HFedMoE 原型，并采用 8 秩 QLoRA 加载 DeepSeek-MoE-16B 模型。如图9所示，我们的分布式系统由一台配备 NVIDIA GeForce RTX 4090 GPU 的中央服务器和 $C = 4$ 个同步运行的客户端组成，每个客户端均部署在 NVIDIA Jetson AGX Xavier 套件上。每个客户端的学习率设为 0.0001，本地微调的批次大小设为 $B = 8$。我们将超参数设为 $\lambda = 0.9$、$\beta = 0.1$ 和 $\tau = 0.05$。为模拟计算异构性，我们限制

通过从 12GB 到 32GB 的范围内均匀采样，确定每个客户端可用于微调的最大计算资源。由于对 Switch-base-64 和 DeepSeek-MoE-16B 进行微调

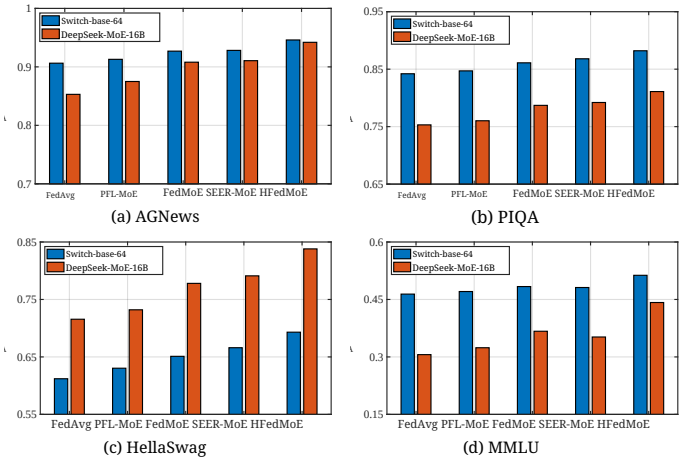


图 10：计算异构性条件下，采用 Switch-base-64 和 DeepSeek-MoE-16B 模型时，四个数据集在各项基准测试上的测试准确率。

至少需要 22GB，因此约有 50% 的客户端必须选择专家才能参与微调。

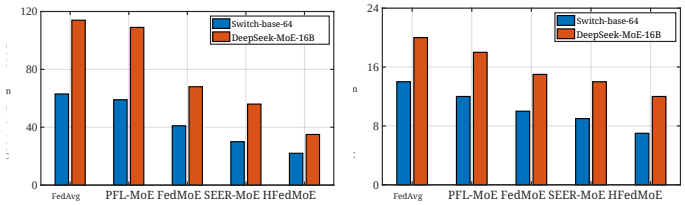
V. 性能评估

本节将 HFedMoE 与各种基准方法进行比较，评估其整体性能和计算效率。我们还将研究 HFedMoE框架中不同超参数设置的影响。此外，我们还分析了 HFedMoE 中各个精心设计组件的作用，以阐明它们在所提出框架中的独立贡献。

A. 整体性能

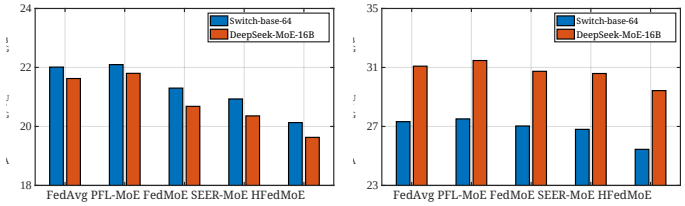
1) 测试准确率： 图10展示了在计算异构性条件下，HFedMoE及其他基准方法在四个数据集上的测试准确率。在各类基准测试中，HFedMoE均表现出优异性能，在 AGNews、PIQA、HellaSwag 和 MMLU 上的测试准确率分别超过 94%、81%、72% 和 45%。与 FedMoE 和 PFL-MoE 等联邦混合专家微调方法相比，HFedMoE 凭借所提出的稀疏性感知模型聚合策略，大幅提升了模型性能。该策略能够协调客户端特定的路由偏好，并聚合采用选择性专家激活的异构 MoE 子模型。尽管 FedMoE 和 SEER-MoE 也减少了激活专家的数量，但由于缺乏准确的专家重要性识别能力，且无法在专家选择过程中处理并发专家激活，其性能显著下降，导致模型更新效果不佳、泛化能力减弱。相比之下，HFedMoE 通过识别不同样本中的专家重要性，并动态选择关键专家以适应各客户端的计算预算，实现资源感知专家选择，从而在不牺牲性能的前提下高效开展本地微调。

2) 收敛性能：图11比较了 HFedMoE 与其他四种基准方法在 AGNews（分类）和 MMLU（推理）任务上的收敛时间，这些任务使用了 Switch-base-64 和 DeepSeek-MoE-16B 模型，并处于



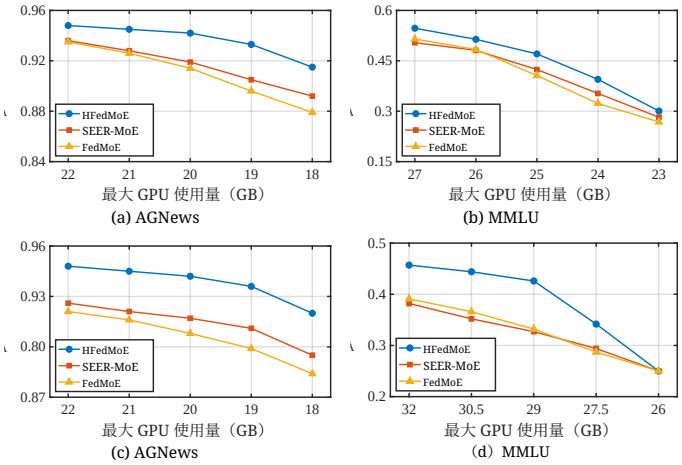
(a) AGNews (b) MMLU

图11：计算异构性下 AGNews 和 MMLU 数据集的收敛时间。



(a) AGNews (b) MMLU

图12：计算异构性下 AGNews 和 MMLU 数据集的平均 GPU 使用量。



(a) AGNews (b) MMLU (c) AGNews (d) MMLU

图13：Switch-base-64（图13(a)-(b)）和 DeepSeek-MoE-16B（图13(c)-(d)）模型在计算限制下的性能下降。通过限制具有异构计算资源的每个客户端的最大GPU使用量来施加计算限制。

异构计算约束下。在两个数据集和两种模型上，HFedMoE 均始终实现最快收敛；在 Switch-base-64 和 DeepSeek-MoE-16B 模型上，速度分别提升了 1.4 倍以上和 1.6 倍以上。这一改进源于 HFedMoE 的资源感知专家选择，该机制既能缓解计算过载导致的本地微调失败，又能在微调过程中保留信息量最大的专家激活。尽管 SEER-MoE 也会有选择地激活关键专家以减轻计算负担，但由于缺乏聚合设计，其专家选择会受到严重干扰，显著降低收敛速度。另一方面，PFL-MoE 和 FedMoE 支持客户端特定的专家选择并聚合共享专家，但仍忽略了客户端之间的路由分歧，导致各客户端用于专家选择的门控参数更新不一致，从而造成收敛不稳定。相比之下，HFedMoE 中的稀疏性感知聚合会明确考虑客户端特定路由偏好和专家选择之间的差异，因此在异构环境下能够实现更快收敛。

3) 计算效率： 图12比较了在异构计算约束下，使用 AGNews 和 MMLU 数据集对 Switch-base-64 和 DeepSeek-MoE-16B 模型进行微调时，HFedMoE 与其他四种基准方法的平均 GPU 使用率。如图所示，HFedMoE 展现出更高的计算效率：除稀疏专家激活本身带来的效率优势外，与不受计算约束的 FedAvg 和 PFL-MoE 基线方法相比，它还将平均 GPU 使用率降低了 10%。这一优势源于 HFedMoE 的专家重要性识别机制，该机制量化专家对本地性能的贡献，确保有限的计算资源分配给最关键的专家，尤其是在计算预算有限的情况下。通过剪除低影响专家带来的冗余计算，HFedMoE 的资源感知专家选择策略能够在本地计算约束下动态调整活跃专家子集，使拥有异构资源的客户端能够有效参与，同时保留具有代表性的专家多样性。相比之下，尽管 FedMoE 和 SEER-MoE 在选出紧凑的专家子集后显著减轻了客户端的计算负担，但获取稳定的专家子集仍面临计算方面的挑战，导致其平均 GPU 使用率明显更高。

4) 计算限制的影响：图13报告了在 AGNews 和 MMLU 任务上，计算限制对 HFedMoE 收敛后测试准确率的影响。虽然所有基线方法在计算能力降低时都不可避免地出现精度下降，但 HFedMoE 展现出显著的鲁棒性：即使一半的客户端面临严重的资源限制，其准确率仍能保持在完整性能的95%以上。这主要归因于 HFedMoE 的资源感知专家选择策略，该策略能够有效地将有限资源分配给信息量最大的本地更新；同时，其针对专家和门控网络的稀疏性感知聚合能够显式处理部分专家更新和不同的路由偏好。尽管 FedMoE 和 SEER-MoE 也采用了专家重要性估计来减少内存占用，但其粗粒度的重要性估计无法考虑不同样本中的输入特定贡献，而初步训练期间跨样本的并发专家激活又大幅增加了客户端负担，导致在计算预算紧张时准确率严重下降。

B. 最优超参数

1) 专家重要性的不同加权组合：表 I 展示了 HFedMoE 中不同加权系数 λ 的影响，该系数用于控制专家重要性识别中累积重要性与特定重要性之间的平衡。在使用 Switch-base-64 和 DeepSeek-MoE-16B 模型对 AGNews 和 MMLU 数据集进行的实验中，收敛时间和测试准确率明显取决于 λ 的取值。当 λ 较

λ	AGNews				MMLU			
	Switch-base-64		DeepSeek-MoE-16B		Switch-base-64		DeepSeek-MoE-16B	
	收敛	准确率	收敛	准确率	收敛	准确率	收敛	准确率
1.0	21	0.9434	38	0.9391	8	0.4871	11	0.3963
0.9	23	0.9453	36	0.9409	**7**	**0.5132**	**13**	**0.4421**
0.8	**22**	**0.9460**	**35**	**0.9419**	11	0.5001	16	0.4290
0.7	24	0.9445	36	0.9401	12	0.4861	20	0.4421
0.6	26	0.9416	39	0.9412	16	0.5097	18	0.4355

Tabl表 I：采用不同专家重要性加权组合的 HFedMoE 性能 e.

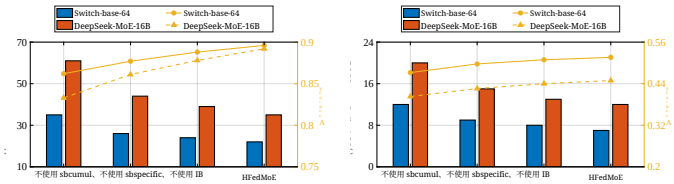
τ	AGNews				MMLU			
	Switch-base-64		DeepSeek-MoE-16B		Switch-base-64		DeepSeek-MoE-16B	
	收敛	准确率	收敛	准确率	收敛	准确率	收敛	准确率
0.0	20	0.9417	37	0.9401	9	0.5001	22	0.4421
0.05	**22**	**0.9460**	**35**	**0.9419**	**7**	**0.5132**	**13**	**0.4421**
0.1	24	0.9437	40	0.9385	8	0.4871	17	0.4291
0.15	24	0.9408	42	0.9397	13	0.4936	20	0.4421
0.2	27	0.9395	49	0.9364	19	0.4796	8	0.3137

表 II：HFedMoE 在主动训练专家使用率阈值不同时的性能 ts.

小时，专家选择会更加依赖特定重要性，侧重于适应特定且通常难以分类的样本。这会在微调过程中保留关键响应，从而加快初始收敛，但也会增加过拟合至次优路由模式的风险，导致准确率降低。相反，较大的 λ 会提高累积重要性的权重，促进模型在不同样本上的有效表征并提升泛化能力，但由于专家多样性降低，收敛速度会变慢。此外，我们发现最佳 λ 值因数据集而异：较小的值（$\lambda = 0.8$）在 AGNews 上表现更好，而较大的设置（$\lambda = 0.9$）则在 MMLU 上实现了收敛与准确率之间的最佳权衡。这表明，异质性较高的数据集可通过提高累积重要性来稳定专家选择。

2) 主动训练专家的使用率阈值变化：表 II 展示了在 AG-News 和 MMLU 数据集上聚合 Switch-base-64 和 DeepSeek-MoE-16B 模型时，不同使用率阈值τ 对收敛时间和测试准确率的影响。结果凸显了将训练不足的专家排除在全局聚合之外的优势。若不进行选择性专家聚合（$\tau = 0$），部分训练或完全未参与训练的专家仍会被不加区分地聚合——

从而在优化不一致的专家之间引入干扰，降低全局模型的泛化能力。这种聚合阻碍了客户端之间专家专门化的对齐，导致全局模型收敛速度变慢、泛化能力下降。相比之下，将 τ 设置得过高会将聚合限制在少量专家上，限制全局知识共享并削弱学习到的表示的多样性，从而导致准确率不理想。适中的阈值（$\tau = 0.05$）通过过滤掉未主动训练的专家，同时保留多样且主动训练的专家参与聚合，实现了最佳平衡，进而在 AGNews 和 MMLU 上带来更稳定的收敛和最高准确率。



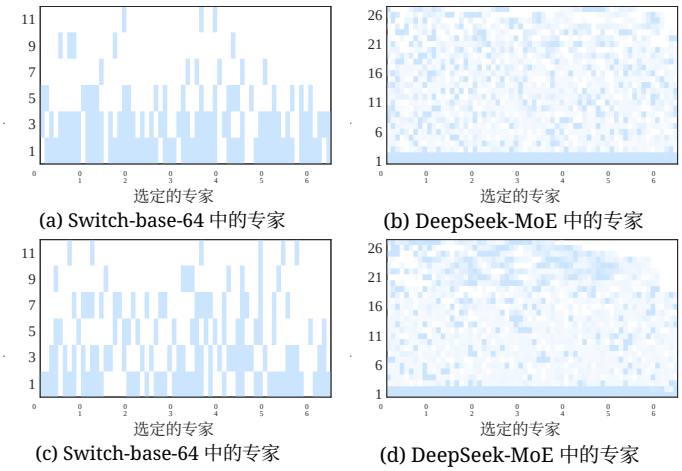
(a) AGNews (b) MMLU

图14：AGNews 和 MMLU 数据集上的专家重要性识别。

C. 消融实验

*1)*专家重要性识别： 图14展示了专家重要性估计对微调 Switch-base-64 和 DeepSeek-MoE-16B 模型的收敛时间和测试准确率的影响。结果表明，专家的累积重要性和特定重要性均显著促进了性能提升，而缺少 IB优化目标会导致专家选择的区分性降低，并使准确率下降。具体而言，累积重要性能够捕捉不同输入样本之间专家使用情况的重叠，从而实现显著的计算共享并减少冗余计算，使模型在计算资源受限的情况下将收敛速度提升至 x1.35。进一步纳入特定重要性后，还能捕捉每个专家对本地数据的独特贡献，从而增强适应性，更好地对齐客户端特定的数据分布。这使速度进一步提升 x1.2，测试准确率提高约 2%，展现了所提出专家重要性指标的优越性能。

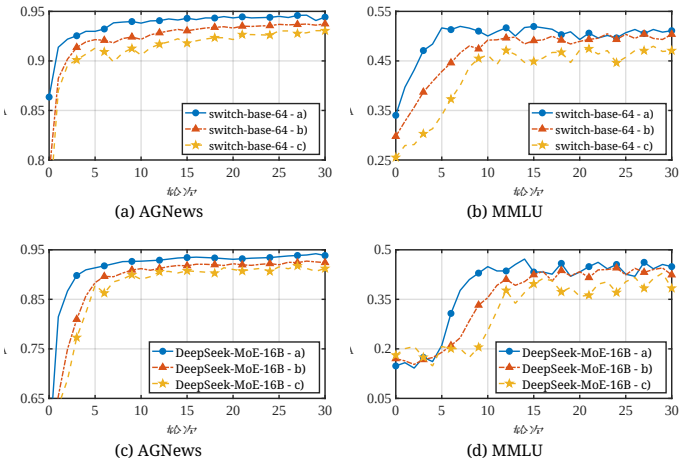
2) 资源感知专家选择：图15展示了在计算资源受限的情况下，微调期间激活的专家数量及其分布。在 Switch-base-64 和 DeepSeek-MoE-16B 模型上进行微调时，HFedMoE 分别激活了约 25% 和 78% 的专家，同时保持了其原始测试准确率的 95% 以上。这些结果突显了 HFedMoE 能够以极小的性能下降实现高效的资源利用率，彰显了其在资源



(a) Switch-base-64 中的专家 (b) DeepSeek-MoE 中的专家

(c) Switch-base-64 中的专家 (d) DeepSeek-MoE 中的专家

图1 5：使用 Switch-base-64 和 DeepSeek-MoE-16B 模型，在 AGNews（图1 5(a)-(b)）和 MMLU（图1 5(c)-(d)）数据集上进行的资源感知专家选择。



(a) AGNews (b) MMLU

(c) AGNews (d) MMLU

图16：使用 Switch-base-64（图16(a)–(b)）和 DeepSeek-MoE-16B（图16(c)–(d)）模型在 AGNews 和 MMLU 数据集上进行的稀疏性感知模型聚合。

受限设备上部署的潜力。这种效率得益于所提出的资源感知专家选择方法：该方法优先选择影响较大的专家，同时减少对性能影响较小的专家所带来的冗余。此外，研究发现，专家选择并非在各层中均匀进行：较深层往往保留较少的专家，而功能上更为关键的层则保持较高的激活程度。这种自适应分配最大限度地提高了计算效率，并与先验观察一致 [50] ，即大型模型的较深层通常存在更多冗余。

3) 稀疏性感知模型聚合： 图16展示了在异构计算资源条件下，Switch-base-64 和 DeepSeek-MoE-16B 模型的模型聚合微调性能。比较了三种聚合条件：a）选择性聚合专家，同时按照重要性加权贡献聚合门控参数；b）选择性聚合专家，但不考虑客户端之间存在差异的路由偏好；以及 c）标准 FedAvg 聚合，即在所有客户端之间统一聚合所有专家和门控参数。与假设模型结构一致且所有参数都具有有意义更新的标准聚合方法相比，HFedMoE 在 Switch-base-64 和 DeepSeek-MoE-16B 上均实现了更高的准确率和更快的收敛。这一提升源于所提出的稀疏性感知模型聚合方法：该方法会过滤不可靠的专家更新，仅聚合活跃专家，从而有效缓解客户端之间部分专家更新所造成的干扰。此外，重要性加权门控聚合根据各客户端的专家偏好和路由一致性，调整其对全局门控网络的贡献，从而缓解路由行为不一致的问题，进一步提升测试准确率并加快收敛。

VI. 相关工作

联邦学习中的基于 MoE 的大语言模型：通过选择性激活专家子集，混合专家为降低计算开销的同时保留模型的表征能力。借助这种效率，混合专家（MoE）已被广泛采用为大语言模型（LLM）开发中的基础框架， [27], [32], [51], 以较低的计算成本实现卓越性能。进一步的研究将混合专家（MoE）扩展至个性化联邦学习（FL）， [35], [38], [52], 通过客户端特定的专家激活，使每个客户端能够仅微调与其本地数据最相关的专家。尽管这些方法提升了本地适应能力，但它们通常依赖共享门控网络进行专家选择，忽视了每个专家对关键本地特征的贡献，从而导致性能欠佳。此外，现有大多数研究未能考虑客户端设备有限的计算资源，严重限制了基于 MoE 的大语言模型在异构联邦环境中的部署。

基于 MoE 的大语言模型微调以降低计算开销：为进一步降低基于 MoE 的大语言模型的计算开销，近期研究开始探索专门的专家选择策略。FedMoE [49], 等现有资源受限的联邦学习方法通过预训练来识别客户端专属专家子集，从而减轻客户端在后续微调过程中的内存压力。然而，由于不同样本并发激活专家会带来较高的计算负担，用于获取稳定专家子集的预训练阶段仍然计算密集，因此限制了其缓解计算资源约束的能力。此外，未能处理客户端间的路由分歧并准确估计专家重要性，也妨碍了获得可靠且最优的结构，而这种结构对于有效缓解资源限制至关重要。另一类基于剪枝的方法 [31], [33], [53], [54] 通过移除平均激活频率较低的专家，降低微调期间的资源消耗。然而，这类方法同样未能在不同输入下的专家选择中考虑并发激活，也无法适应客户端设备之间异构的计算能力，因此难以在联邦场景中实现性能与效率之间的最佳权衡，尤其是在微调期间受到客户端专属资源约束时。

联邦MoE微调中的模型聚合：尽管将混合专家模型整合到联邦学习中的研究日益增多，但如何在保持全局性能的同时有效聚合结构异构的本地模型，仍面临重大挑战。现有联邦学习方法 [12], [22], [43]通常假设所有客户端共享相同的模型结构，但不同的专家子集和路由偏好会使简单的参数聚合无法奏效。近期的一些研究 [35], [38], [53]尝试缓解专家之间的结构异构性，但仍忽视了聚合过程中的不同专家激活模式，因此无法处理部分专家更新和不一致的门控网络，导致基于 MoE 的联邦学习性能下降。尽管某些方法 [55]–[57]通过保留私有门控网络并仅聚合专家参数来维持本地适应性，但客户端间由此产生的部分专家更新会导致优化错位，进而削弱专家专门化并损害模型泛化能力。

VII. 结论

本文提出了一种名为 HFedMoE 的异构联邦学习框架，旨在提升基于 MoE 的大语言模型微调在具有不同计算能力的异构客户端上的有效性。首先，专家重要性识别方案量化了每个专家对不同客户端微调性能的贡献。其次，基于每个专家的重要性，资源感知专家选择方法确保在每个客户端的计算预算下动态选择关键专家。最后，选择性专家聚合策略进一步增强了模型泛化能力，减轻了专家结构异构性的影响。大量实验结果表明，HFedMoE 优于最先进的基准方法，在准确率下降极小的情况下显著加快了收敛。

参考文献

[1] J. Achiam、S. Adler、S. Agarwal、L. Ahmad、I. Akkaya、F. L. Aleman、D. Almeida、J. Altenschmidt、S. Altman、S. Anadkat 等，“GPT-4技术报告”，*arXiv* 预印本 *arXiv:2303.08774*，2023年3月。[2] Z. Lin、G. Qu、Q. Chen、X. Chen、Z. Chen 和 K. Huang，“将大语言模型推向 6G边缘：愿景、挑战与机遇”，*IEEE*通信杂志，2023年。[3] T. Duan、Z. Zhang、S. Guo、D. Huang、Y. Zhao、Z. Lin、Z. Fang、D. Luan、H. Cui 和 Y. Cui，“Leed：一个面向多智能体强化学习的高效且可扩展的大语言模型赋能专家示范框架”，*arXiv* 预印本 *arXiv:2509.14680*，2025年。[4] H. Touvron、T. Lavril、G. Izacard、X. Martinet、M.-A. Lachaux、T. Lacroix、B. Rozi`ere、N. Goyal、E. Hambro、F. Azhar 等，“LLaMA：开放且高效的基础语言模型”，*arXiv* 预印本 *arXiv:2302.13971*，2023年2月。[5] Z. Lin、X. Hu、Y. Zhang、Z. Chen、Z. Fang、X. Chen、A. Li、P. Vepakomma 和 Y. Gao，“SplitLoRA：一种面向大语言模型的拆分式参数高效微调框架”，*arXiv* 预印本 *arXiv:2407.00952*，2024年。[6] A. Liu、B. Feng、B. Xue、B. Wang、B. Wu、C. Lu、C. Zhao、C. Deng、C. Zhang、C. Ruan 等，“DeepSeek-V3技术报告”，*arXiv* 预印本 *arXiv:2412.19437*，2025年2月。[7] Z. Fang、Z. Lin、Z. Chen、X. Chen、Y. Gao 和 Y. Fang，“面向大语言模型参数高效微调的自动化联邦流水线”，*IEEE*移动计算汇刊，2026年。[8] Z. Lin、Y. Zhang、Z. Chen、Z. Fang、X. Chen、P. Vepakomma、W. Ni、J. Luo 和 Y. Gao，“HSplitLoRA：一种面向大语言模型的异构拆分式参数高效微调框架”，*arXiv* 预印本 *arXiv:2505.02795*，2025年。

[9] T. Duan, Z. Zhang, Z. Lin, S. Guo, X. Guan, G. Wu, Z. Fang, H. Meng, X. Du, J.-Z. Zhou 等，《面向移动系统中多智能体强化学习的大语言模型驱动稳定性感知专家示范》，arXiv 预印本 arXiv:2511.19368，2025年。[10] Z. Fang, Z. Lin, S. Hu, Y. Tao, Y. Deng, X. Chen 和 Y. Fang，《面向户外健康监测的动态不确定性感知多模态融合》，arXiv 预印本 arXiv:2508.09085，2025年。[11] A. Karargyris, R. Umeton, M. J. Sheller, A. Aristizabal, J. George, A. Wuest, S. Pati, H. Kassem, M. Zenk, U. Baid 等，《借助 MedPerf 对医疗人工智能进行联邦基准测试》，Nature Machine Intelligence，第5卷，第7期，第799–810页，2023年7月。[12] B. McMahan, E. Moore, D. Ramage, S. Hampson 和 B. A. y. Arcas，《基于去中心化数据的深度网络通信高效学习》，载于 AISTATS 论文集，第54卷。PMLR，2017年4月，第1273–1282页。[13] B. Luo, P. Han, P. Sun, X. Ouyang, J. Huang 和 N. Ding，《异构6G网络中联邦学习的优化设计》，IEEE Netw.，第37卷，第2期，第38–43页，2023年4月。[14] M. Hu, J. Zhang, X. Wang, S. Liu 和 Z. Lin，《通过模型分割加速边缘网络中的联邦学习》，IEEE Trans. Green Commun. Netw.，2024年。[15] Y. Zhang, Z. Lin, Z. Chen, Z. Fang, W. Zhu, X. Chen, J. Zhao 和 Y. Gao，《Satfed：一种资源高效的低轨卫星辅助异构联邦学习框架》，Engineering，2024年。[16] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečný, S. Mazzocchi, B. McMahan 等，《迈向大规模联邦学习：系统设计》，SysML 论文集，第1卷，第374–388页，2019年3月。[17] Z. Lin, Z. Chen, Z. Fang, X. Chen, X. Wang 和 Y. Gao，《Fedsn：一种面向异构低轨卫星网络的联邦学习框架》，IEEE移动计算汇刊，2024年。[18] Y. Zhang, H. Chen, Z. Lin, Z. Chen 和 J. Zhao，《LCFed：一种面向异构数据的高效聚类联邦学习框架》，arXiv 预印本 arXiv:2501.01850，2025年。[19] Z. Wang, Z. Shen, Y. He, G. Sun, H. Wang, L. Lyu 和 A. Li，《Flora：采用异构低秩适配的大语言模型联邦微调》，NeurIPS 论文集，第37卷，第22 513–22 533页，2024年12月。[20] Z. Lin, G. Qu, W. Wei, X. Chen 和 K. K. Leung，《Adaptsfl：资源受限边缘网络中的自适应拆分联邦学习》，IEEE Trans. Netw.，2024年。[21] Z. Jiang, Y. Xu, H. Xu, Z. Wang, J. Liu, Q. Chen 和 C. Qiao，《通过自适应模型剪枝实现计算与通信高效的联邦学习》，IEEE移动计算汇刊，第23卷，第3期，第2003–2021页，2023年3月。[22] Z. Lin, G. Zhu, Y. Deng, X. Chen, Y. Gao, K. Huang 和 Y. Fang，《资源受限无线边缘网络中的高效并行拆分学习》，IEEE移动计算汇刊，第23卷，第10期，第9224–9239页，2024年。[23] Z. Lin, Z. Chen, X. Chen, W. Ni 和 Y. Gao，《HASFL：边缘计算系统中感知异质性的拆分联邦学习》，arXiv 预印本 arXiv:2506.08426，2025年。[24] N. Dhar, B. Deng, D. Lo, X. Wu, L. Zhao 和 K. Suo，《在边缘设备上部署大语言模型的实证分析与资源占用研究》，载于 ACMSE 论文集，2024年4月，第69–76页。[25] D. Xu, W. Yin, H. Zhang, X. Jin, Y. Zhang, S. Wei, M. Xu 和 X. Liu，《EdgeLLM：通过推测解码实现快速设备端大语言模型推理》，IEEE移动计算汇刊，第24卷，第4期，第3256–3273页，2025年。[26] G. Qu, Q. Chen, W. Wei, Z. Lin, X. Chen 和 K. Huang，《大语言模型的移动边缘智能：当代综述》，IEEE Commun. Surv. Tutor.，2025年。[27] D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu 等，《DeepSeekMoE：迈向混合专家语言模型中的极致专家专门化》，载于 ACL 论文集（第1卷：长篇论文），2024年，第1280–1297页。[28] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang 和 W. Chen，《LoRA：大语言模型的低秩适配》，载于 ICLR 论文集，2022年1月。[29] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan 和 S. Gelly，《面向自然语言处理的参数高效迁移学习》，载于 ICML 论文集，2019年6月，第2790–2799页。[30] Z. Lin, W. Wei, Z. Chen, C.-T. Lam, X. Chen, Y. Gao 和 J. Luo，《分层拆分联邦学习：收敛分析与系统优化》，IEEE移动计算汇刊，2025年。[31] X. Lu, Q. Liu, Y. Xu, A. Zhou, S. Huang, B. Zhang, J. Yan 和 H. Li，《专家并非一视同仁：混合专家大语言模型的高效专家剪枝与跳过》

载于 ACL 论文集（第1卷：长篇论文），2024年，第6159–6172页。[32] W. Fedus, B. Zoph 和 N. Shazeer，《Switch Transformers：通过简单高效的稀疏性扩展至万亿参数模型》，机器学习研究期刊，第23卷，第120期，第1–39页，2022年4月。[33] A. Muzio, A. Sun 和 C. He，《SEER-MoE：通过正则化提升混合专家的稀疏专家效率》，arXiv 预印本 arXiv:2404.05089，2024年4月。[34] Z. Du, S. Li, Y. Wu, X. Jiang, J. Sun, Q. Zheng, Y. Wu, A. Li, H. Li 和 Y. Chen，《SIDA-MOE：面向高效且可扩展的大型混合专家模型的稀疏性驱动数据感知服务》，MLSys 论文集，第6卷，第224–238页，2024年4月。[35] B. Guo, Y. Mei, D. Xiao 和 W. Wu，《PFL-MoE：基于混合专家的个性化联邦学习》，载于 APWeb-WAIM 论文集。Springer，2021年8月，第480–486页。[36] M. Lewis, S. Bhosale, T. Dettmers, N. Goyal 和 L. Zettlemoyer，《基础层：简化大型稀疏模型的训练》，载于 ICML 论文集。PMLR，2021年7月，第6265–6274页。[37] M. Reisser, C. Louizos, E. Gavves 和 M. Welling，《联邦混合专家》，arXiv 预印本 arXiv:2107.06724，2021年7月。[38] C. Dun, M. H. Garcia, G. Zheng, A. H. Awadallah, R. Sim, A. Kyrillidis 和 D. Dimitriadis，《Fedjets：借助联邦混合专家实现高效即时个性化》，arXiv 预印本 arXiv:2306.08586，2023年6月。[39] X. Zhang, J. Zhao 和 Y. LeCun，《用于文本分类的字符级卷积网络》，NeurIPS 论文集，第28卷，2015年12月。[40] C.-J. Jhang, C.-X. Xue, J.-M. Hung, F.-C. Chang 和 M.-F. Chang，《面向人工智能边缘设备的基于静态随机存取存储器的存内计算：挑战与趋势》，IEEE Trans. Circuits Syst. I, Reg. Pap.，第68卷，第5期，第1773–1786页，2021年5月。[41] X. Zhang 和 S. Debroy，《移动边缘计算中的资源管理：全面综述》，ACM Comput. Surv.，第55卷，第13s期，第1–37页，2023年7月。[42] G. Zhu, Y. Deng, X. Chen, H. Zhang, Y. Fang 和 T. F. Wong，《Esfl：资源受限异构无线设备上的高效分割联邦学习》，物联网期刊，第11卷，第16期，第27 153–27 166页，2024年。[43] J. Zhang, S. Vahidian, M. Kuo, C. Li, R. Zhang, T. Yu, G. Wang 和 Y. Chen，《迈向构建 FederatedGPT：联邦指令微调》，载于 ICASSP 论文集。IEEE，2024年4月，第6915–6919页。[44] N. Tishby, F. C. Pereira 和 W. Bialek，《信息瓶颈方法》，arXiv 预印本 physics/0004057，2000年4月。[45] A. A. Alemi, I. Fischer, J. V. Dillon 和 K. Murphy，《深度变分信息瓶颈》，arXiv 预印本 arXiv:1612.00410，2019年10月。[46] Y. Bisk, R. Zellers, J. Gao, Y. Choi 等，《PIQA：自然语言中的物理常识推理》，载于 AAAI 论文集，第34卷，第05期，2020年2月，第7432–7439页。[47] R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi 和 Y. Choi，《HellaSwag：机器真的能续写你的句子吗？》，载于 ACL 论文集，2019年，第4791–4800页。[48] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song 和 J. Steinhardt，《大规模多任务语言理解的测量》，arXiv 预印本 arXiv:2009.03300，2021年1月。[49] H. Mei, D. Cai, A. Zhou, S. Wang 和 M. Xu，《Fedmoe：通过异构混合专家实现个性化联邦学习》，arXiv 预印本 arXiv:2408.11304，2024年。[50] X. Men, M. Xu, Q. Zhang, B. Wang, H. Lin, Y. Lu, X. Han 和 W. Chen，《Shortgpt：大语言模型中的层比你想象的冗余得多》，arXiv 预印本 arXiv:2403.03853，2024年10月。[51] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bam-ford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand 等，《Mixtral of Experts》，arXiv 预印本 arXiv:2401.04088，2024年1月。[52] Y. Feng, Y.-a. Geng, Y. Zhu, Z. Han, X. Yu, K. Xue, H. Luo, M. Sun, G. Zhang 和 M. Song，《PM-MOE：面向个性化联邦学习的私有模型参数混合专家》，载于 WWW 论文集，2025年4月，第134–146页。[53] L. Jin, Y. Zhang, Y. Li, S. Wang, H. H. Yang, J. Wu 和 M. Zhang，《MoE^2：优化边缘大语言模型的协同推理》，arXiv 预印本 arXiv:2501.09410，2025年1月。[54] S. He, D. Dong, L. Ding 和 A. Li，《迈向高效混合专家：压缩技术的全面研究》，arXiv 预印本 arXiv:2406.02500，2025年3月。[55] Z. Zhan, W. Zhao, Y. Li, W. Liu, X. Zhang, C. W. Tan, C. Wu, D. Guo 和 X. Chen，《FedMoE-DA：通过领域感知的细粒度聚合实现联邦混合专家》，载于2024年第20届国际移动、感知与网络会议（MSN）论文集。IEEE，2024年，第122–129页。[56] G. Hu、Y. Teng、P. Wu 和 N. Wang，“Fft-moe：在异构边缘环境下通过大规模稀疏混合专家实现基础模型的高效联邦微调”，*arXiv* 预印本 *arXiv:2508.18663*，2025年。[57] L. Yi、H. Yu、C. Ren、H. Zhang、G. Wang、X. Liu 和 X. Li，“pfedmoe：面向模型异构个性化联邦学习的混合专家数据级个性化”，*arXiv* 预印本 *arXiv:2402.01350*，2024年。