

基于动态优化客户端训练的联邦检索增强生成框架

张承业, 赵雁斌, 蓝 飘, 刘经友, 滚真灵

(广州广电运通信息科技有限公司, 广东 广州 510555)

摘要: 联邦学习与检索增强生成的协同架构为隐私保护下的分布式知识扩展提供了有效途径,但仍面临性能与效率优化的挑战。提出一种基于动态优化客户端训练的联邦检索增强生成框架DFRAG,通过多维度协同优化策略提升系统整体效能。首先,设计基于动态训练损失的客户端选择算法,显著降低系统通信开销;其次,通过动态调整检索数量,增强对复杂问题的处理能力;最后,通过基于损失的动态参数聚合算法,显著提升任务准确率。实验在Natural Questions和TriviaQA问答数据集上进行,结果表明,DFRAG提升了问答任务的准确率并显著优化了训练效率,3个优化模块也均展现出差异化贡献。验证了联邦学习与检索增强生成框架协同动态优化的可行性,为隐私敏感场景下的知识密集型任务提供了一种高效、可扩展的解决方案。

关键词: 联邦学习;检索增强生成;动态客户端选择;动态检索数量;动态参数聚合

DOI:10.11907/rjdk.251426

扫描二维码阅读全文:



中图分类号:TP391.1

文献标识码:A

文章编号:1672-7800(2025)012-0161-07

Federated Retrieval-Augmented Generation Framework Based on Dynamic Optimization of Client Training

ZHANG Chengye, ZHAO Yanbin, LAN Piao, LIU Jingyou, GUN Zhenling

(GRG Banking Equipment Co., Ltd, Guangzhou 510555, China)

Abstract: The synergistic architecture combining federated learning with retrieval-augmented generation provides an effective approach for privacy-preserving distributed knowledge expansion, while still facing challenges in performance and efficiency optimization. This study proposes DFRAG, a dynamically optimized client training-based federated retrieval-augmented generation framework that enhances overall system performance through multidimensional collaborative optimization strategies. First, a dynamic training loss-based client selection algorithm is designed to significantly reduce communication overhead. Second, dynamic retrieval quantity substantially enhances complex problem-solving capability. Finally, a loss-based dynamic parameter aggregation algorithm improves task accuracy. Experimental results on Natural Questions and TriviaQA datasets demonstrate that DFRAG improves question-answering accuracy while significantly optimizing training efficiency, with all three optimization modules showing differentiated contributions. This research validates the feasibility of collaborative dynamic optimization between federated learning and retrieval-augmented generation frameworks, providing an efficient and scalable solution for knowledge-intensive tasks in privacy-sensitive scenarios.

Key Words: federated learning; retrieval-augmented generation; dynamic client selection; dynamic retrieval quantity; dynamic parameter aggregation

0 引言

随着大语言模型(Large Language Model, LLM)在各领

域的深入应用,其性能表现与数据训练质量呈现显著相关性^[1]。然而在现实场景中,有价值的信息往往分散存储于不同的机构或个体,且普遍存在敏感信息,这使得传统集中式训练面临“数据孤岛”与隐私合规问题。联邦学习

收稿日期:2025-07-31

作者简介: 张承业(1985-),男,博士,广州广电运通信息科技有限公司高级工程师,研究方向为数据底座技术、大数据平台技术、政企数字化转型;赵雁斌(1998-),男,硕士,广州广电运通信息科技有限公司软件工程师,研究方向为自然语言处理、联邦学习。
本文通讯作者:赵雁斌。

(Federated Learning, FL)作为一种分布式协作训练范式,通过参数共享而非原始数据交换的方式,为解决隐私保护问题提供了可行方案^[2]。与此同时,检索增强生成(Retrieval-augmented Generation, RAG)技术通过引入外部知识库,显著缓解了LLM存在的幻觉问题并提升了时效性响应能力,尤其在知识密集型任务(如开放域问答、专业报告生成等)中表现突出^[3]。

近年来,联邦学习与检索增强生成的协同框架(FL+RAG)展现出独特优势。该框架通过分布式知识检索机制,在保护数据隐私的前提下同时增强模型生成能力。既弥补了传统联邦学习的知识缺失局限性,又规避了集中式RAG的隐私泄露风险,为跨机构知识共享提供了安全高效的解决方案。学术界已从多个维度推进该领域研究,并取得一系列研究成果。Mao等^[4]提出FedE4RAG框架,在联邦学习中结合知识蒸馏与同态加密实现隐私保护,同时提升检索效果。Jung等^[5]验证了FL+RAG在医疗领域提升模型泛化能力的有效性。Muhammed等^[6]提出CoRAG框架,利用客户端数据的非敏感内容构建协作知识库并搭配上公共知识库,探索了多客户端共建知识库的协作范式。Wu等^[7]提出FICAL,在联邦训练中直接利用知识概要代替参数传输,大幅降低通信开销。

这些研究共同表明,FL+RAG的融合架构有隐私安全与知识增强方面的双重价值,但现有工作普遍忽视了一个关键问题:客户端选择和动态权重策略对联邦检索框架性能与效率的影响尚未得到充分研究^[8]。传统随机或轮询式客户端选择可能导致低质量数据参与训练,而静态权重分配的聚合策略也难以满足联邦学习的公平性原则。尽管已有学者提出基于客户端质量选择的研究,例如Goetz等^[9]提出基于损失的效用值,优先选择效用值高的客户端,Cho等^[10]为选择损失值高的客户端提出理论基础,并提出POWER-OF-CHOICE策略以提升训练效率,Deng等^[11]提出使用强化学习选择客户端,自动学习既高效又有质量意识的客户端选择策略。但这些方法均未考虑到RAG场景特有的知识检索质量与数据分布特性,导致现有框架在实践过程中仍面临检索效率低下和知识利用不充分等问题,亟需针对性的优化方案。

本文基于Izacard等^[12]中的检索-生成训练范式,创新性地提出一种基于动态优化客户端训练的联邦检索增强生成框架(Dynamic Federated Retrieval-augmented Generation, DFRAG)。该框架通过3个维度的协同优化:动态客户端选择、动态检索数量、动态参数聚合,显著提升了系统性能。实验结果表明,在标准问答数据集Natural Questions上,DFRAG不仅提高了任务准确率,还优化了训练效率^[13]。进一步的有效性分析实验证明,DFRAG的3个优化模块在Natural Questions和TriviaQA问答数据集上均展现出差异化贡献^[14]。同时,通过参数敏感性分析验证了框架设计的鲁棒性优势。该工作为联邦检索增强生成框架

优化提供了新的技术路径。

1 相关工作

1.1 联邦学习

联邦学习由McMahan等^[2]首次提出,其核心思想是通过分布式协作训练解决数据隐私与合规性问题。与传统集中式训练不同,联邦学习允许数据保留在本地设备(如手机、IoT设备)上,并通过本地训练后,仅通过与服务端交换模型参数(而非原始数据),在服务端进行参数聚合,实现全局模型优化。该设计符合隐私保护需求,同时降低了数据传输成本。

本文是基于横向联邦学习(Horizontal FL)进行的研究,其也可称为按样本划分的联邦学习^[15]。横向联邦学习的常见结构是客户端-服务端结构,也称为主从结构。在这种系统中,有着多个分散的客户端 $\{c_1, c_2, \dots, c_n\}$,且每个客户端包含的数据为 $\{d_1, d_2, \dots, d_n\}$ 。横向联邦学习迭代流程如图1所示。

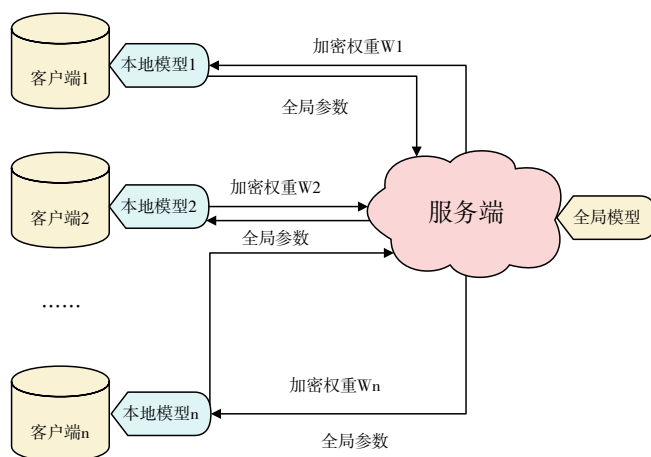


Fig. 1 Horizontal federated learning iterative process

图1 横向联邦学习迭代流程

训练过程通常包含3个步骤:①在本地训练阶段,各客户端设备从服务器下载当前全局模型,利用本地私有数据进行梯度计算和模型更新,并对梯度权重实施同态加密等隐私保护措施;②进入参数聚合阶段,服务端收集所有客户端上传的加密梯度或模型权重,采用FedAvg等加权平均算法生成新一代全局模型^[2];③在模型分发阶段,服务端将聚合后的模型参数安全分发给参与客户端,完成一轮联邦学习迭代。

以FedAvg算法为例,假设有 n 个客户端,第 n 个客户端表示为 c_n ,对应的数据表示为 d_n ,模型权重为 w_n ,那么在第 $t+1$ 轮时,对应客户端的模型权重计算公式如式(1)所示。

$$w_n^{t+1} \leftarrow w_n^t - \eta \nabla l(w_n^t; b) \quad (1)$$

其中, η 为学习率, ∇ 为梯度计算, $l(w_n^t; b)$ 为客户端损失函数。客户端更新后,一起将权重 w_n 返回给服务器进行

聚合。定义所有被选中的客户端总训练数据为 D , 则权重聚合公式如式(2)所示。

$$w^{t+1} \leftarrow \sum_{n=1}^N \frac{d_n}{D} w_n^t \quad (2)$$

1.2 检索增强生成

随着大语言模型 LLM 的日益流行,其在处理时效性强、专业性高以及上下文复杂的任务时仍存在显著局限性,主要表现为知识更新滞后、领域适应性不足,以及存在

事实性错误等问题。检索增强生成(RAG)作为一种创新的混合架构,通过整合参数化记忆(预训练语言模型的隐式知识表示)与非参数化记忆(外部知识库的显式文档检索),有效弥补了传统生成模型在知识密集型任务中的不足^[3]。其核心优势是通过实时检索相关文档片段以增强生成过程,从而提升生成内容的事实性、可解释性和知识更新能力。基础 RAG 流程如图 2 所示^[16]。

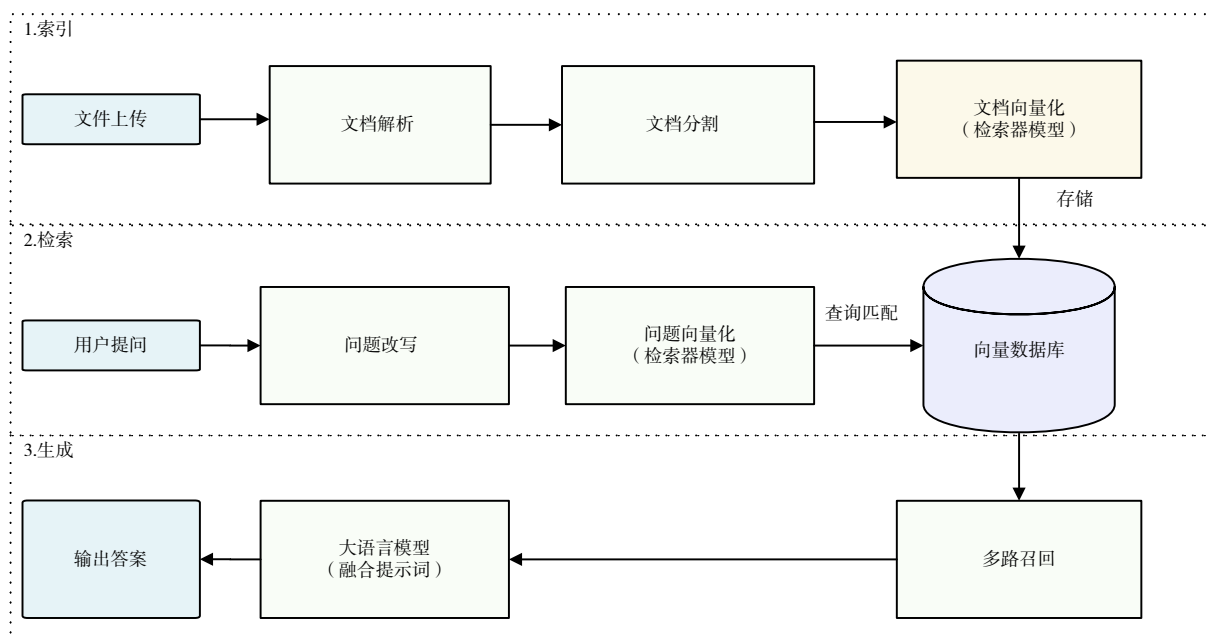


Fig. 2 Basic RAG process

图2 基础 RAG 流程

RAG 通常包含 3 个步骤:①在索引阶段,原始数据经过清洗、分块和向量化处理,通过检索器模型编码为稠密向量,并构建高效的向量索引结构;②进入检索阶段,用户查询被同一检索器模型编码为向量,通过近似最近邻搜索计算与索引库中文本块的相似度得分,按得分排序并返回 top- k 最相关文档;③在生成阶段,系统将用户问题与检索到的文档拼接为增强提示,输入大语言模型生成最终答案。

假设给定输入问题为 x ,检索器从向量库中检索出来的文档为 z ,查询到每个文档的概率记为 $p(z|x)$ 。大模型使用问题和文档生成的答案为 y ,搜索文档数量为 k , T 为生成的答案长度,则 RAG 模型的生成概率可形式化地记为 $p_{\text{RAG}}(y|x)$,计算公式如式(3)所示。

$$p_{\text{RAG}}(y|x) \approx \sum_{z \in \text{top-}k} p(z|x) \cdot \prod_{i=1}^T p_{\theta}(y_i|x, z, y_{1:i-1}) \quad (3)$$

2 DFRAG 框架

在联邦学习结合与检索增强生成协同优化的场景下,客户端选择机制和参数聚合策略是决定模型收敛效率与最终性能的关键因素,动态检索数量能有效增强对复杂问

题的处理能力。因此,本文在联邦学习结合 RAG 的应用场景下,从 3 个阶段对训练进行优化,分别为动态客户端选择、动态检索数量、动态参数聚合。通过在训练阶段联合优化检索器与生成器,实现客户端对知识的高效利用。DFRAG 流程如图 3 所示。

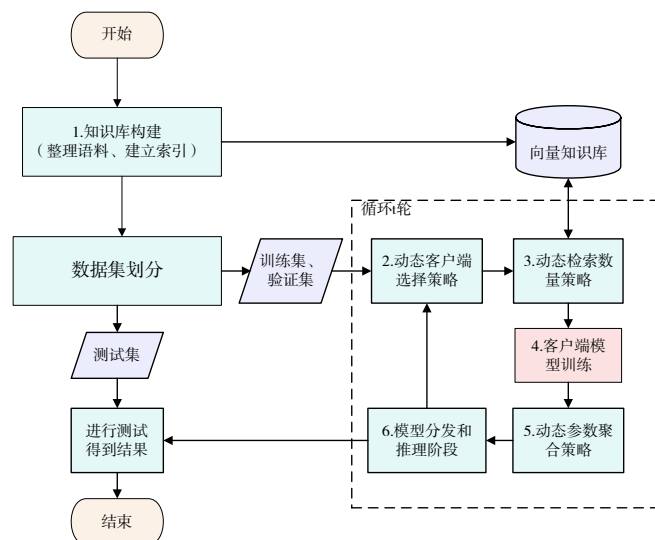


Fig. 3 DFRAG process

图3 DFRAG 流程

2.1 知识库与数据集划分

采用大规模通用语料库(如 Wikipedia)构建索引,基于 Flat 索引结构实现高效检索,并支持动态更新机制以适应领域知识的持续演进^[17]。同时,将问答数据集划分为训练集、验证集、测试集并分配给各客户端。选用预训练语言模型作为基础生成器,在联邦学习框架下进行分布式微调。在本地训练阶段,每个客户端 c_n 基于本地标注数据 $d_n = \{(x_n, y_n)\}$ 执行端到端的检索-生成联合优化。

2.2 动态客户端选择策略

基于 Goetz 等^[9]和 Cho 等^[10]的研究,采用标准化训练损失作为客户端筛选标准,能有效减少系统开销和提高模型性能。本文策略通过建立损失驱动的选择机制,优先选择当前训练效果欠佳的客户端参与联邦训练。为保障公平性,对连续未被选中的客户端实施渐进式补偿策略,通过动态调整其历史损失值逐步提升其后续训练轮次的参与概率。

具体实现如下:首轮训练选择全部客户端,从第二轮开始对上一轮的联合损失 L_{total} 使用 Softmax 归一化,使得损失缩放至 $[0, 1]$ 区间且总和为 1。第 n 个客户端归一化损失的计算公式如式(4)所示。

$$L_n^{norm} = \frac{e^{L_n}}{\sum_{i=1}^N e^{L_i}} \quad (4)$$

$$k^{(m,t+1)} = \begin{cases} \beta \cdot k, & \text{if } L_{retr}^{(m,t)} > \mu_{retr}^{(m)} + \sigma_{retr}^{(m)} \text{ or } L_{gen}^{(m,t)} > \mu_{gen}^{(m)} + \sigma_{gen}^{(m)} \\ \max\left(1, \left\lfloor \frac{k}{\beta} \right\rfloor\right), & \text{if } L_{retr}^{(m,t)} < \mu_{retr}^{(m)} - \sigma_{retr}^{(m)} \text{ or } L_{gen}^{(m,t)} < \mu_{gen}^{(m)} - \sigma_{gen}^{(m)} \\ k, & \text{otherwise} \end{cases} \quad (5)$$

2.4 客户端模型训练

在生成阶段,检索模型计算生成概率 $p_{RAG}(y|x)$,定义答案的真实分布为 $p^*(y)$,生成的答案长度为 T ,则使用交叉熵损失(Cross-Entropy)计算生成器损失 L_{gen} 的公式如式(6)所示。

$$L_{gen} = -\sum_{i=1}^T p^*(y_i) \log p_{RAG}(y_i|x) \quad (6)$$

同时,通过困惑度蒸馏(Perplexity Distillation, PDist)优化检索器,检索器的目标是让检索出来的文档概率 $p_{retr}(z_k|x)$ 尽可能接近文档重要性分布 p_k ^[12]。将问题 x 和每个检索文档 z_k 合并成多个字符串后,若生成答案 y 的概率为 $p_{LM}(y|z_k, x)$,则归一化后的文档重要性分布为 p_k 。 p_k 与检索器分布 $p_{retr}(z_k|x)$ 计算 KL 散度(Kullback-Leibler divergence),最后计算出的检索器损失为 L_{retr} 。 p_k 和 L_{retr} 的计算公式如式(7)、式(8)所示。

$$p_k = \frac{\exp(p_{LM}(y|z_k, x))}{\sum_{i=1}^k \exp(p_{LM}(y|z_i, x))} \quad (7)$$

$$L_{retr} = \sum_{i=1}^k p_k \log \frac{p_k}{p_{retr}(z_i|x)} \quad (8)$$

最后,将检索器 L_{retr} 与生成器 L_{gen} 的损失函数求和得

按照归一化后的损失 L_n^{norm} 进行排序后,选择损失值较高的 M 个客户端,对未选中的客户端,其历史损失值乘以鼓励系数 α 后参与后续轮次的归一化排序。

2.3 动态检索数量策略

Lin 等^[18]的研究表明,动态调整 top- k 检索数量是优化 RAG 系统效率与效果的核心,能有效平衡召回广度与计算开销、问题复杂度与知识覆盖率的矛盾。实验证明,在 Natural Questions 数据集上,将 k 值限定在 $[5, 10]$ 区间可覆盖多数单跳问题,同时降低计算开销。因此,本文针对 RAG 模块的特性,设计基于双指标反馈的动态调节机制。检索文档个数 k ,同时根据检索器损失 L_{retr} 和生成损失 L_{gen} 进行动态调整,建立以历史表现为基准的弹性策略。当检索性能显著偏离预期时,自动扩展或收缩检索文档数量,在稳定状态下则保持原有检索范围。

具体实现如下:从第三轮开始,对选中的 M 个客户端,从其本地知识库中检索 top- k 相关段落 $\{z_1, \dots, z_k\}$ 构成上下文集合 Z ,将查询和检索到的段落按格式拼接(如 question: $\{x\}$ context: $\{Z\}$)。同时,每轮初始 k 设定为固定值,设第 t 轮($t > 2$)训练时客户端 m 的检索器损失和生成损失分别为 $L_{retr}^{(m,t)}$ 和 $L_{gen}^{(m,t)}$,其历史均值和标准差分别为 $\mu_{retr}^{(m)}$ 、 $\sigma_{retr}^{(m)}$ 和 $\mu_{gen}^{(m)}$ 、 $\sigma_{gen}^{(m)}$ 。定义范围变化比例为 β ,则第 $t+1$ 轮的检索个数 $k^{(m,t+1)}$ 的调整策略公式如式(5)所示。

k , otherwise

到总损失 L_{total} 。

2.5 动态参数聚合策略

在 FedAvg 的基础上引入损失加权策略,综合考虑各客户端的本地数据规模与训练损失程度,通过动态加权实现精细化聚合。该策略赋予损失较大的模型更高权重,避免向特定客户端过拟合。在经过上一轮训练后,对新的联合损失 L_{total} 使用 Softmax 归一化得到 L^{norm} ,则 M 个客户端的全局模型更新公式如式(9)所示。

$$w^{t+1} \leftarrow \sum_{m=1}^M \frac{d_m}{D} w_m^t L_m^{norm} \quad (9)$$

2.6 模型分发与推理阶段

服务端将更新后的模型分发给各客户端进行推理验证。在经过 t 轮训练后,客户端加载全局模型 w_{t+1} ,在本地测试集上评估性能,并引入全局知识库生成答案并计算准确率。

3 实验与结果分析

3.1 模型与数据集

实验采用 NQ-finetuned Atlas-base^[12](约 220 M 参数)模型作为基础生成模型,该模型基于 T5 架构(Text-to-

Text Transfer Transformer)^[19]进行预训练,使用的语料集为 Wikipedia2021 年(3 750 万个段落)和 Common Crawl 网站数据的一个子集^[20]。选择较小的模型主要是基于联邦学习场景下实用性的考量,即在保证模型性能的同时有效控制通信开销和计算资源消耗。

实验中使用的数据集为 Natural Questions^[13]问答数据集和 TriviaQA 问答数据集^[14],将两个数据集划分为训练集、验证集、测试集,数据集细节如表 1 所示。

Table 1 Datasets details

表 1 数据集细节

数据集	训练集	验证集	测试集
Natural Questions	64	8 757	3 610
TriviaQA	64	8 837	11 313

在少样本训练的实验设定下,每个客户端的训练集仅包含 64 个训练 QA 对,这种训练设置旨在突出 RAG 的知识增强效果。验证集和测试集都采用均匀分配,以保证评估公平性。

在 RAG 部分,检索器使用通过无监督数据进行训练的 Contriever,分布式知识库使用 Wikipedia 2018 的精简语料库(约 40% 原始规模)^[21]。该设计通过以下机制保障隐私安全:知识库仅包含公开可获取的通用知识;检索过程不涉及原始训练数据;检索模型优化聚焦于查询编码方式而非知识内容。在训练、验证和测试阶段,都进行知识库检索辅助答案生成。

3.2 实验环境与细节

在配备双 GPU 节点的服务器上开展实验,操作系统为 Ubuntu(64 位),CPU 型号为 Intel Xeon Processor (Skylake) 2.19 Ghz,内存 32 G,显卡为 Tesla T4,显存为 16*2 G,使用分布式训练进行。软件环境上,使用 Python 3.8 作为编程语言,Pytorch 版本为 1.1.0。实验细节如表 2 所示。

Table 2 Experimental details

表 2 实验细节

超参数名称	参数值	超参数名称	参数值
客户端数量 N /个	10	生成模型学习率	$2e-5$
候选客户端数量 M /个	6	检索模型学习率	$4e-5$
输入文本最大长度/Tokens	512	每个 GPU 的 batch 数	2
初始检索数量 top- k /个	6	预热部署 warmup	4
随机失活 dropout	0.1	训练步数 step	20
训练精度	Fp32	权重鼓励系数 α	1.1
权重衰减	0.01	范围变化比例 β	1.5

3.3 评价指标与基线模型

在每轮联邦训练结束后,对所有客户端的验证集进行聚合评估,并保存全局最优模型。由于开放域问答(Open-Domain QA)任务的答案通常较短且结构化程度低,实验中采用平均精确匹配得分 EM(Exact Match)和 F1 Score 作为评价指标,同时记录训练时间以作对比。其中,EM 指标是生成答案与标准答案完全一致的比例。F1 Score 是精确率(Precision)和召回率(Recall)的调和平均数。精确率是预测正确的 Token 占总预测 Token 的比例,召回率是预测正

确的 Token 占真实答案 Token 的比例。

实验中采用的基线模型如下:①LocalRAG:基础 RAG 架构,完全未采用联邦学习机制;②FRAG:在 LocalRAG 的基础上引入标准联邦学习机制,表现为 FL+RAG 少样本训练,但未实施任何动态优化策略;③CoRAG-IRR^[6]:CoRAG 框架利用客户端非敏感内容构建协作知识库,并搭配公共知识库,CoRAG-IRR 是其中一种变体,协作知识库仅包含完全无关段落,公共知识库包含了 80% 的 Wikipedia 2018 数据集,同样进行 FL+RAG 的少样本训练;④CoRAG-SPLIT^[6]:是 CoRAG 的另一种变体,协作知识库不共享,客户端仅检索本领域段落和包含 10%Wikipedia 2018 数据集的公共段落,同样进行 FL+RAG 的少样本训练。

3.4 对比结果分析

DFRAG 与 4 个基线模型在 Natural Questions 问答数据集上的指标对比如表 3 所示。

Table 3 Baseline comparison

表 3 基线对比

模型	EM	F1	时间/h
LocalRAG	26.76	33.48	10.0
FRAG	29.68	36.49	12.0
CoRAG-IRR	30.44	-	-
CoRAG-SPLIT	40.06	-	-
DFRAG	41.46	49.62	8.0

可以看出,DFRAG 在 Natural Questions 数据集上的 EM 和 F1 指标均显著优于基线模型。相较于 4 个基线模型,DFRAG 在 EM 指标上达到 41.46,较表现次优的 CoRAG-SPLIT 稍有提升,较传统 FRAG 方法提升 39%。在 F1 指标上达到 49.62,较 FRAG 提升 36%。究其具体原因在于:①动态客户端选择机制优先筛选高损失节点参与训练,有效提升模型收敛效率,显著降低训练耗时;②动态检索数量优化,根据损失自适应调整检索数量 top- k ,显著增强了对复杂问题的处理能力,提升长尾问题覆盖;③基于训练损失的动态参数聚合策略平衡了各客户端贡献,加速了知识融合过程。

3.5 优化有效性分析

为了验证 DFRAG 各优化模块的有效性,本文设计了系统的消融实验,通过控制变量方法分析不同优化配置对模型性能的影响。实验中重点考察 3 个关键优化策略的效果差异:①动态客户端选择机制(60% 参与率)与全客户端参与对比;②动态检索文档数量调整与固定检索数量对比;③动态参数聚合策略与静态参数聚合对比。采用 Natural Questions 和 TriviaQA 两个标准数据集开展评估,同样对比 EM、F1 和时间指标。同时,为了研究各模块的耦合干扰或协同增益效应,在 Natural Questions 数据集上额外设置了单模块实验组,分别独立测试动态选择、动态检索和动态参数的单独作用效果。完整的实验结果如表 4 和表 5 所示。

消融实验结果表明,DFRAG 框架的优化模块有效性

Table 4 Effectiveness analysis on Natural Questions dataset**表4 Natural Questions数据集上的有效性分析**

框架	EM	F1	时间/h
DFRAG	41.46	49.62	8.0
w/o 动态选择	39.43	47.88	12.0
w/o 动态权重	36.92	44.86	8.0
w/o 动态检索	40.24	48.02	8.5
with 动态选择	37.12	46.27	8.5
with 动态权重	39.10	47.54	12.0
with 动态检索	32.68	41.82	11.0
FRAG	29.68	36.49	12.0

Table 5 Effectiveness analysis on TriviaQA dataset**表5 TriviaQA数据集上的有效性分析**

框架	EM	F1	时间/h
DFRAG	59.06	69.32	10.0
w/o 动态选择	56.75	65.86	14.0
w/o 动态权重	53.85	63.43	10.0
w/o 动态检索	56.03	64.89	10.5
FRAG	51.63	60.05	14.0

得到系统性验证:动态客户端选择、动态检索数量和动态参数聚合在效果和效率上均展现出差异化贡献,在两个数据集上的结果略有不同。具体而言,动态参数聚合模块对模型性能影响最为显著,在 Natural Questions 和 TriviaQA 数据集上禁用该模块分别导致 F1 值下降约 9% 和 8%,远超其他模块的影响幅度,充分证实了其对模型准确率的关键提升作用。动态客户端选择模块主要提升训练效率,在 Natural Questions 数据集上禁用后训练时间从 8 h 延长至 12 h (增幅 50%),在 TriviaQA 数据集上从 10 h 延长至 14 h (增幅 40%),且伴随轻微性能损失,这表明该模块能有效过滤低质量客户端的冗余计算。动态检索数量模块的作用呈现出明显的任务敏感性,在结构简单的 Natural Questions 数据集上作用较弱(禁用后 F1 仅下降 3.2%),但在较为复杂的 TriviaQA 数据集上效果显著(禁用后 F1 下降 6.4%),这一差异印证了该模块对复杂问题处理的重要性。

同时,表4的实验数据进一步揭示了各优化模块之间的协同效应。在仅启用单个优化模块的实验中,动态参数聚合模块展现出最强的独立性能提升能力,在 Natural Questions 数据集上 F1 达到 47.54,为 3 个优化中模块中最高。而动态选择客户端模块则表现出最优的时间效率,动态检索数量能略微提升效果和减少时间。值得注意的是,双模块组合效果显著优于单模块,未观察到明显的耦合干扰。最后,完整 DFRAG 在两个数据集上均为最优性能和最低耗时,F1 值相较于基线 FRAG 分别提升 36% 和 15%,且耗时平均缩减约 30%,证明三模块协同优化具有显著优势。

3.6 超参数敏感性分析

为了补充验证上述实验中超参数组合的全局最优性,本文通过系统性超参数敏感性实验,分析不同超参数组合的训练效果。通过系统性实验分析了 3 个关键超参数(候选客户端数量 M 、权重鼓励系数 α 、知识库搜索范围比例

β) 的影响,在 Natural Questions 数据集上进行实验。采用控制变量法:固定 $\alpha=1.1$ 和 $\beta=1.5$ 时测试 $M \in \{4, 6, 8\}$,固定 $M=6$ 和 $\beta=1.5$ 时测试 $\alpha \in \{1.05, 1.1, 1.2\}$,固定 $M=6$ 和 $\alpha=1.1$ 时测试 $\beta \in \{1.2, 1.5, 2.0\}$ 。其中,DFRAG 的参数组合是 $M=6$ 、 $\alpha=1.1$ 、 $\beta=1.5$,FRAG 不使用这些超参数,实验结果如表 6 所示。

Table 6 Influence of hyperparameter combination**表6 超参数组合的影响**

组合类别	M	α	β	EM	F1	时间/h
DFRAG	6	1.1	1.5	41.46	49.62	8.0
调整 M	8	1.1	1.5	40.68	48.23	10.0
	4	1.1	1.5	39.12	47.13	6.5
	6	1.05	1.5	40.58	48.68	8.0
调整 α	6	1.2	1.5	38.86	46.74	8.0
	6	1.1	1.2	40.32	48.00	8.0
调整 β	6	1.1	2.0	39.20	46.89	7.5
	6	1.1	1.5	40.32	48.00	8.0
FRAG	—	—	—	29.68	36.49	12.0

实验结果表明,DFRAG 框架的超参数组合($M=6$ 、 $\alpha=1.1$ 、 $\beta=1.5$)在效果和效率上达到了最优平衡。具体体现如下:首先,客户端数量 M 存在明显阈值效应,当 $M=4$ 时因覆盖不足导致 F1 下降,而 $M=8$ 虽能带来些许提升,但也增加了低质量客户端冗余,且计算时间成本增加了 25% (从 8 h 增至 10 h);其次,权重系数 α 对模型稳定性至关重要,当 $\alpha=1.2$ 时因过度激励导致客户端不断轮换,准确率下降过多,证实了适度激励策略的必要性;最后,知识检索个数范围 β 在 $[1.2-2.0]$ 区间变动时准确率波动较小,但是 $\beta=2.0$ 时由于搜索范围增大,可能会出现知识过少或者无关知识过多的极端情况,导致准确率略差。值得注意的是,大部分参数组合在效果、效率均优于表 3 中的基线模型,验证了 DFRAG 参数化设计的优势和鲁棒性。

4 结语

本文提出的 DFRAG 框架创新性地将动态优化机制引入联邦检索增强生成系统,通过动态客户端选择、动态检索数量和动态参数聚合的三维协同优化,有效解决了传统方法在知识利用效率和模型性能上的瓶颈问题。实验表明,该框架在保护数据隐私的前提下显著提升了知识增强效果和训练效率。

未来研究将进一步探索联邦学习联合检索增强生成的更深层次方案,探究该框架扩展至非独立同分布场景下对数据异质性和训练稳定性的影响,以及其他复杂任务中的泛化性能和可拓展性,推动隐私计算与知识工程的深度融合。

参考文献:

- [1] MINAE S, MIKLOV T, NIKZAD N, et al. Large language models: a survey[DB/OL]. <https://arxiv.org/abs/2402.06196v3>.
- [2] MCMAHAN B, MOORE E, RAMAGE D, et al. Communication-efficient

- learning of deep networks from decentralized data [C]//Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017: 1273–1282.
- [3] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks [J]. *Advances in Neural Information Processing Systems*, 2020, 33: 9459–9474.
- [4] MAO Q, ZHANG Q, HAO H, et al. Privacy-preserving federated embedding learning for localized retrieval-augmented generation [DB/OL]. <http://arxiv.org/abs/2504.19101>.
- [5] JUNG J, JEONG H, HUH E N. Federated learning and RAG integration: a scalable approach for medical large language models [C]//2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC), 2025: 0968–0973.
- [6] MUHAMED A, DIAB M, SMITH V. CoRAG: collaborative retrieval-augmented generation [DB/OL]. <http://arxiv.org/abs/2504.01883>.
- [7] WU P, LI K, NAN J, et al. Federated in-context LLM agent learning [DB/OL]. <http://arxiv.org/abs/2412.08054>.
- [8] ZHANG R L, DU J H, YIN H. Client selection algorithm in cross-device federated learning [J]. *Journal of Software*, 2024, 35(12): 5725–5740.
张瑞麟, 杜晋华, 尹浩. 跨设备联邦学习中的客户端选择算法 [J]. *软件学报*, 2024, 35(12): 5725–5740.
- [9] GOETZ J, MALIK K, BUI D, et al. Active federated learning [DB/OL]. <http://arxiv.org/abs/1909.12641>.
- [10] CHO Y J, GUPTA S, JOSHI G, et al. Bandit-based communication-efficient client selection strategies for federated learning [C]//2020 54th Asilomar Conference on Signals, Systems, and Computers, 2020: 1066–1069.
- [11] DENG Y, LYU F, REN J, et al. AUCTION: automated and quality-aware client selection framework for efficient federated learning [J]. *IEEE Transactions on Parallel and Distributed Systems*, 2022, 33(8): 1996–2009.
- [12] IZACARD G, LEWIS P, LOMELI M, et al. Atlas: few-shot learning with retrieval augmented language models [J]. *Journal of Machine Learning Research*, 2023, 24(1): 43–76.
- [13] KWIATKOWSKI T, PALOMAKI J, REDFIELD O, et al. Natural questions: a benchmark for question answering research [J]. *Transactions of the Association for Computational Linguistics*, 2019, 7: 452–466.
- [14] JOSHI M, CHOI E, WELD D, et al. TriviaQA: a large scale distantly supervised challenge dataset for reading comprehension [C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1601–1611.
- [15] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging [C]//Florida: The 20th International Conference on Artificial Intelligence and Statistics, 2017.
- [16] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: a survey [DB/OL]. <http://arxiv.org/abs/2312.10997>.
- [17] JOHNSON J, DOUZE M, JÉGOU H. Billion-scale similarity search with GPUs [J]. *IEEE Transactions on Big Data*, 2021, 7(3): 535–547.
- [18] LIN X, HUANG Z, ZHANG Z, et al. Explore what LLM does not know in complex question answering [C]//Proceedings of the AAAI Conference on Artificial Intelligence, 2025: 24585–24594.
- [19] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. *Journal of Machine Learning Research*, 2020, 21: 1–67.
- [20] WENZEL G, LACHAUX M A, CONNEAU A, et al. CCNet: extracting high quality monolingual datasets from web crawl data [C]//Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020: 4003–4012.
- [21] IZACARD G, CARON M, HOSSEINI L, et al. Towards unsupervised dense information retrieval with contrastive learning [DB/OL]. <https://arxiv.org/abs/2112.09118>.

(责任编辑: 孙 娟)