# PruFed: Federated Fine-Tuning of Pruned LLMs

**Shrenik Bhansali, Alwin Jin, Tyler Lizzo, Larry Heck**
AI Virtual Assistant (AVA) Lab, Georgia Institute of Technology
{sbhansali8,alwinjin,lizzo,larryheck}@gatech.edu

## Abstract

Large language models (LLMs) have become indispensable in natural language processing (NLP) but face challenges regarding the high costs of fine-tuning and the invasive nature of data collection. Federated learning (FL) offers a solution by preserving user privacy while refining models. Current FL research focuses on homogeneous client models, neglecting the potential for heterogeneous federated systems (Zhang et al., 2023). We explore optimal LLM pruning strategies to reduce fine-tuning complexity and introduce a novel heterogeneous model aggregation approach for FL. Results show optimal pruning is task-dependent, and for heterogeneous FL, our approach outperforms traditional homogeneous setups for smaller models, reducing the gap between edge devices and full-sized models.

## 1 Introduction

Large Language Models (LLMs) represent a significant advance in Natural Language Processing (NLP) with their remarkable ability to generalize across queries and tasks. These models are typically fine-tuned using large, diverse datasets derived from high-quality instruction data (Gupta et al., 2022). However, developing such models requires the collection of vast amounts of data for fine-tuning and personalization, a process that results in significant time, labor, and accessibility constraints. Furthermore, the large size of robust models creates high computation costs for training, fine-tuning, and inference.

Federated learning (FL) is a promising solution to address these challenges(McMahan et al., 2017). It is a collaborative learning approach that allows client models to learn from users while preserving their privacy. LLMs are traditionally fine-tuned in a centralized manner, where data is aggregated from raw user interactions and shared globally to fine-tune a single global model. In contrast, FL utilizes distributed fine-tuning, with client models trained on user interactions and the global model created by aggregating client model weights. This preserves privacy by avoiding the need to share raw user data globally.

However, existing research on federating LLMs predominantly focuses on homogeneous federated systems, where all clients have identical LLMs (Zhang et al., 2023). This assumption is unrealistic, as different clients will have different devices of varying computational ability. We thus aim to address the unique challenges posed by heterogeneous federated systems, where LLMs differ among clients.

The major contributions of this work include:

- An analysis to determine an ideal pruning strategy for LLMs that maximizes knowledge distillation while reducing model size significantly. We determine that pruning strategies are task-dependent for each model.
- The development of a heterogeneous aggregation scheme that allows for knowledge transfer between LLMs of different sizes, allowing for more realistic FL.

## 2 Background

### 2.1 Model Pruning

In order to create compressed models that are less resource-intensive, we leverage model pruning to decrease parameter count while retaining performance. To avoid parameter-mismatching during aggregation, we utilize layer-pruning as opposed to individual parameter-pruning. We thus remove entire layers instead of specific parameters.

Prior work has demonstrated that specific layers can be removed from pretrained language models without significant performance degradation during the fine-tuning process (Sajjad et al., 2023). By pruning specific layers, we effectively distill smaller language models (LMs) from the original LLM. Since the remaining layers within this distilled LM are still present in the original LLM, we can perform aggregation across specific, overlapping layers between the models instead of averaging all parameters throughout the model. This layer-wise aggregation method enables a model-agnostic aggregation process, where we can aggregate models with similar architectures but different depths (number of layers) by aligning them at their shared layers.

This layer-pruning approach creates LMs with shallow depths, making them viable for more resource-constrained devices due to their smaller

size. As research on LLM efficiency and distillation continues to grow, scaling these shallow models to smaller and smaller devices becomes more feasible. This work will act as a link between the smaller models on the edge and the full-size global model they were distilled from.

### 2.2 Heterogeneity in FL

Model architecture heterogeneity presents significant challenges in model-agnostic FL. Differences in model architectures impede the use of standard aggregation techniques like FedAvg due to varying parameter sizes.

Previous work surrounding model-agnostic FL points towards using a proxy unlabeled public dataset to unify trained weights between different models (Huang et al., 2022). This allows constructing a cross-correlation matrix to learn a generalizable representation under domain shift.

However, due to the generality of LLMs, finding and using a large and diverse enough dataset to unify models distilled for specific downstream tasks is impractical. Instead, we must devise an aggregation method that operates without a unifying dataset, instead only operating on the model weights with no external information.

## 3 Method

Our work investigates two principle ideas: The best pruning strategy for LLMs and viable heterogeneous model aggregation schemes for the pruned LLMs. Using LLaMA 7B as our base model, we first perform a comprehensive analysis of various layer pruning strategies and test them across numerous bench marks. We then create an aggregation algorithm for an FL system leveraging the resulting pruned models.

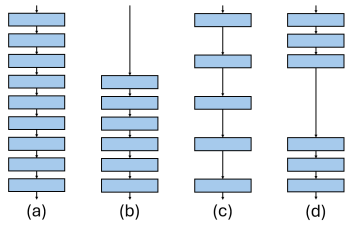


Figure 1: Various pruning strategies explored in this work: (a) full network, (b) top-layer dropping, (c) alternate dropping, (d) symmetric dropping.

We begin by evaluating multiple pruning strategies such as top-layer dropping, even-alternate dropping, odd-alternate dropping, parameter-based contribution dropping, magnitude-based dropping, and Symmetric dropping, as displayed in Figure 1. For each strategy, we establish baselines on various tasks, test the pruning strategy both before and after fine-tuning, and evaluate on the tasks.

To evaluate the transferabilty of model weights between pruned models, we insert the pruned models into a FL setup. Due to their differing depths, traditional aggregation techniques like FedAvg cannot be used. Therefore, we design a novel aggregation scheme that aggregates the model weights across each layer, and call it in place of FedAvg.

To measure performance, we compare each global model with layer amount $\theta^i$ to the global model outputted by a homogenous FL setup where all models in the system have layer amount $\theta^i$ fine-tuned over the same data.

**Algorithm 1** Model-Agnostic Federated Fine-Tuning of LLMs

Initialize $\theta^a$, $\theta^b$, $\theta^c$
**for** $n$ clients **do**:
  Initialize $U_i = (\theta^i, \Delta w)$
**end for**
**while** $k \leq K$ **do**:
  $U_k$ ▷ sample portion of users
  $U_k^a, U_k^b, U_k^c$ ▷ Group $U_k$ by model depth
  **for** device type $i \in \{a, b, c\}$ **do**
    **for** client $c \in U_k^i$ with adapter $\Delta w$ **do**
      $c$ = InstructionTuning($\Delta w$)
    **end for**
  **end for**
  $U_k$ = HeteAgg($U_k$)
**end while**

## 4 Results

Our results show that depending on the downstream task, the best pruning strategy changes accordingly. However, on average, top-layer dropping has minimal performance degradation and is on average the best strategy when there is no specific downstream task.

In a federated learning context, our pruned models in a heterogeneous FL system outperform their homogeneous FL counterparts at the cost of non-pruned, full size model's performance. The full-sized 32 layer model performs slightly worse than its homogeneous counterpart, but still outperforms all pruned models (homogeneous or heterogeneous).

When compared to a traditional homogenous FL strategy, our full sized non-pruned model performs slightly worse than the FedAvg model. However, when comparing the pruned models, our heterogenous FL setup enhances the performance of smaller models.

## References


Samyak Gupta, Yangsibo Huang, Zexuan Zhong, Tianyu Gao, Kai Li, and Danqi Chen. 2022. Recovering private text in federated learning of language models. *Advances in Neural Information Processing Systems*, 35:8130–8143.

Wenke Huang, Mang Ye, and Bo Du. 2022. Learn from others and be yourself in heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10143–10153.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2023. On the effect of dropping layers of pre-trained transformer models. *Computer Speech & Language*, 77:101429.

Jianyi Zhang, Saeed Vahidian, Martin Kuo, Chunyuan Li, Ruiyi Zhang, Guoyin Wang, and Yiran Chen. 2023. Towards building the federated gpt: Federated instruction tuning. *arXiv preprint arXiv:2305.05644*.