

Compressing Large Language Models by Streamlining the Unimportant Layer

Xiaodong Chen^{1†}, Yuxuan Hu^{2†}, Jing Zhang^{2*}

¹ Xi'an Jiaotong University, China

² Renmin University of China, China

chenxiaodong6677@163.com

{huyuxuan1999, zhang-jing}@ruc.edu.cn

Abstract

Large language models (LLM) have been extensively applied in various natural language tasks and domains, but their applicability is constrained by the large number of parameters of the models. Consequently, there is an increasing emphasis on compact models that exhibit high performance. In this study, we observe that different layers in LLM have varying degrees of perturbation on the hidden states, which allows us to identify less important layers. Based on this phenomenon, we propose LLM-Streamline, which consists of two parts: layer pruning, where we remove a set of consecutive layers with the lowest importance in the model according to the target sparsity; and layer replacement, where we train a lightweight model to substitute the pruned layers, thereby mitigating the performance degradation caused by pruning. In our experiments, we utilize structures such as a multi-layer perceptron (MLP) and a transformer layer as lightweight models and ultimately demonstrate that a single MLP can effectively fit the pruned layers. Comprehensive experiments show that our proposed method, LLM-Streamline, outperforms previous state-of-the-art (SOTA) model pruning methods.

1 Introduction

The recent development of large language models (LLM) based on the transformer architecture [1] has been impressive, leading to their widespread application across various tasks and domains. However, the prevailing trend in LLM research is towards larger scales, with an increasing number of parameters of the models raising hardware requirements, which significantly constrains their applicability and hinders their deployment in real-world scenarios.

To alleviate the hardware demands for deploying LLM, there is a growing emphasis on compact models that exhibit high performance, typically derived from model compression techniques [2, 3]. These compact models reduce the size of the original model, substantially decreasing memory consumption and computation time, and greatly reducing the cost of LLM deployment. Existing model compression techniques can be broadly categorized into knowledge distillation [4–8], model quantization [9–11], and model pruning [12–17]. Model quantization compresses the model by modifying the parameter storage format, however, it does not reduce the number of parameters. Model distillation, as opposed to quantization, compresses model parameters but comes with a substantial computational and storage cost. Model pruning can more effectively compress model parameters, but existing pruning techniques usually result in irregular model structures, which makes them inconvenient for practical use.

[†]These authors contributed equally.

^{*}Jing Zhang is the corresponding author.

Meanwhile, existing pruning methods focus more on pruning dense matrices [18], attention heads [19, 20], filters [21, 22], and hidden dimensions [17, 23] and less on layer pruning, which may be due to layer pruning is not as effective as other pruning techniques in experiments with medium-sized language models. However, we find that this result does not hold for large language models. Through our experiments, we find that some successive layers in the LLM have only small perturbations to the hidden states, which suggests that these layers are significantly less important than others, and provides the opportunity for effective layer pruning.

Therefore, based on these empirical observations, we propose LLM-Streamline, a simple and effective layer-wise pruning algorithm for LLM. Our algorithm comprises two parts: layer pruning, where we remove several consecutive layers with the lowest importance in the original model according to the target sparsity; and layer replacement, where we train a lightweight model to substitute the pruned layers to recover the performance degradation caused by pruning.

In our experiments, we use various structures as lightweight models to replace the pruned layers, including multi-layer perceptron (MLP), transformer layer, sharing-parameter MLP, and sharing-parameter transformer layer. We find that although transformer layer or sharing-parameter methods introduce more parameters or computational costs than MLP, they do not yield superior performance. A single MLP can effectively fit the pruned layers. Additionally, we attempt to directly remove these pruned layers without using any lightweight model for replacement, which results in a significant drop in model performance, thus demonstrating the necessity of using a lightweight model to fit the pruned layers. The main contributions of this paper are as follows:

- We analyze the perturbation of different layers in large language models (LLM) to the hidden states and notice that some consecutive layers in LLM have minimal perturbation, which inspires us to remove less important layers in LLM.
- We propose using the cosine similarity between the layer input and output hidden states as a criterion for measuring layer importance and prune layers with lower importance based on this.
- We demonstrate that an MLP trained with only tens of thousands of data points can recover the model’s performance. Comprehensive experiments show that for models with 7B parameters, With a pruning rate of 25%, we can maintain 92% of their performance on classification tasks and 68% of their performance on generation tasks.

2 Related Work

Large language models are being used in different scenarios and their high computational cost makes model compression techniques more necessary than before. Existing model compression techniques include quantization [9–11], distillation [4–8], pruning [12–17], low-rank factorization [24–26], and other techniques. This work focuses on compressing large language models by the pruning method.

Model Pruning. Model pruning is an effective way of compressing large language models, which can be categorized into unstructured pruning [12, 14–16] and structured pruning [13, 17, 18, 27]. Unstructured pruning compresses the model by converting each model’s dense matrices into sparse matrices. While unstructured pruning can maintain performance at large compression rates, it requires specialized sparse operators, which limits its use. Structured pruning, on the other hand, ensures that the matrix is dense when compressing the model and is, therefore, more hardware-friendly. Recent structured pruning approaches for large language models focus more on compressing filters [21, 22], attention heads [19, 20], and hidden dimensions of the model [17, 23]. In contrast to these approaches, we focus on removing unimportant layers from the model.

Knowledge Distillation. Knowledge distillation is another way of compressing large language models by transferring knowledge from a large teacher model to a small student model. Distillation of language models can be divided into two categories of approaches based on the transparency of the teacher model. For non-transparent teacher models, existing work has explored the use of different features for knowledge transfer, including model outputs, hidden states, and attention matrices. Whereas for transparent teacher models, e.g., LLMs that can only be accessed through APIs [6–8, 28],

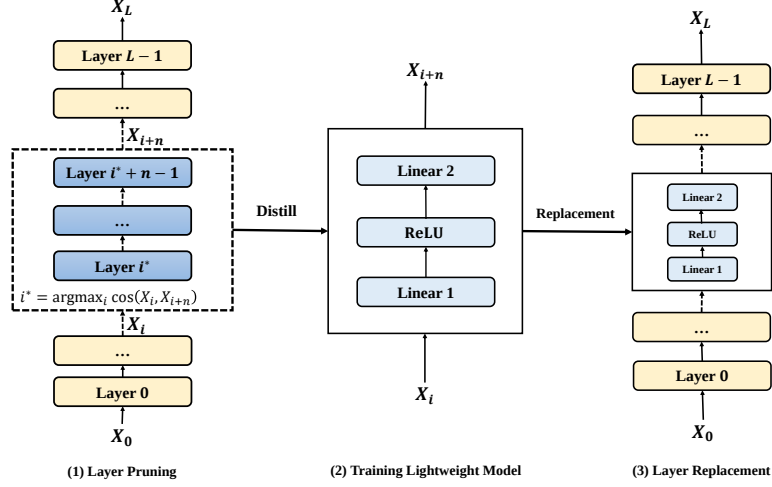


Figure 1: Illustration of the workflow of LLM-Streamline

existing work focuses on using teacher models to automatically synthesize high-quality training data to train student models.

3 LLM-Streamline

In this section, we will provide a brief overview of LLM and highlight the redundancy within its layers. Leveraging the observed layer redundancy, we will introduce the LLM-Streamline framework for LLM layer pruning, along with its basic workflow. Specifically, as depicted in Figure 1, the workflow comprises two main steps: layer pruning and layer replacement. Initially, we prune redundant layers from the LLM. Subsequently, we train a lightweight network to replace the pruned layers, thereby restoring the model’s performance.

3.1 Preliminary

Current LLM implementations predominantly rely on transformer architectures, where a typical LLM comprises a series of transformer decoder layers. Each decoder layer consists of a Multi-Head Attention (MHA) module and a Feed Forward Network (FFN) module. During both training and inference, the i -th decoder layer takes hidden states $X_i \in \mathbb{R}^{d \times n}$ as input and generates $X_{i+1} \in \mathbb{R}^{d \times n}$ as output, where n represents the sequence length and d signifies the dimensionality of the hidden states. This process can be formally expressed as follows:

$$X_i^M = \text{MHA}(\text{LN}(X_i)) + X_i, \quad (1)$$

$$X_{i+1} = \text{FFN}(\text{LN}(X_i^M)) + X_i^M, \quad (2)$$

where LN denotes Layer Normalization, and $\text{MHA}(\cdot)$ and $\text{FFN}(\cdot)$ represent the MHA module and FFN module of the i -th decoder layer L_i . The entire process is abbreviated as:

$$X_{i+1} = L_i(X_i) + X_i. \quad (3)$$

3.2 Redundancy of Layers in LLM

In Formula 3, each layer in the LLM contributes an increment $L_i(X_i)$ to the input X_i . Consequently, we propose that the importance of each layer in the LLM can be measured by assessing its impact on

the inputs. To achieve this, we employ the cosine similarity between inputs X_i and outputs X_{i+1} as a metric. Essentially, a higher cosine similarity between the inputs and outputs of a layer suggests lower importance, and conversely. This interpretation arises from the observation that a higher cosine similarity between the inputs and outputs of a layer implies that the layer may be approximated as a scaling factor.

To calculate the cosine similarity metric, we randomly sample data from the wikitext-2 dataset and feed it into the LLM. We then extract the hidden states produced by each layer of the LLM and compute the cosine similarity between the input and output hidden states of each layer. Our experiments are conducted on the OPT families (OPT-1.3B, OPT-2.7B, OPT-6.7B) [29] and Llama2-7B [30]. The results are illustrated in Figure 2. From the experimental findings, it is evident that several consecutive layers in all models exhibit a high cosine similarity.

Given that these less significant layers only introduce minimal perturbations in the direction of the hidden states, we contend that employing a transformer layer to capture such straightforward relationships lacks justification. Instead, it is imperative to explore simpler, lightweight models as alternatives to these transformer layers. Furthermore, it is important to mention that substituting each of these layers with an individual lightweight model would not substantially decrease the number of parameters and might even aggravate cumulative errors. As illustrated in Figure 2, layers identified as less important (those exhibiting a high cosine similarity between input and output hidden states) tend to cluster together. Therefore, we propose the fitting of a group of consecutive layers with a single lightweight model as a more viable approach.

3.3 Layer Pruning

We need to pinpoint a consecutive sequence of layers that can be most effectively pruned. The number of layers to prune, denoted as n , is determined by the target sparsity. Following the metric outlined in Section 3.2, we randomly select samples from the training set and feed them into the model. For the i -th layer, we calculate the cosine similarity between the input hidden states of that layer and the output hidden states of the $(i + n)$ -th layer. We identify the sequence of consecutive layers with the highest cosine similarity as the most easily adaptable. The computation of cosine similarity can be expressed by the following formula:

$$\cos(X_i, X_{i+n}) = \frac{1}{N} \sum_{j=1}^N \frac{X_{i,j}^T X_{i+n,j}}{\|X_{i,j}\|_2 \|X_{i+n,j}\|_2}, \quad (4)$$

where $X_i \in \mathbb{R}^{d \times N}$ denotes the hidden states of all tokens that input to i -th layer, $X_{i,j}$ is the j -th column of X_i , d signifies the dimensionality of the hidden states, and N is the total number of tokens. We identify the value of i that maximize the cosine similarity as i^* , which identifies the layer to be pruned,

$$i^* = \arg \max_i \cos(X_i, X_{i+n}). \quad (5)$$

3.4 Layer Replacement

After identifying the initial layer for pruning, denoted as i^* , we collect samples from the training set and feed them into the model to obtain the input hidden states of the i^* -th layer and the input hidden states of the $(i^* + n)$ -th layer, which acts as training data for the lightweight model. In our experiments, we employ MLPs, transformer layers, sharing-parameter MLPs, and sharing-parameter transformer layers to capture the input-output mappings of these consecutive layers. Following the training of the lightweight model, we employ it to replace the pruned layers within the original model.

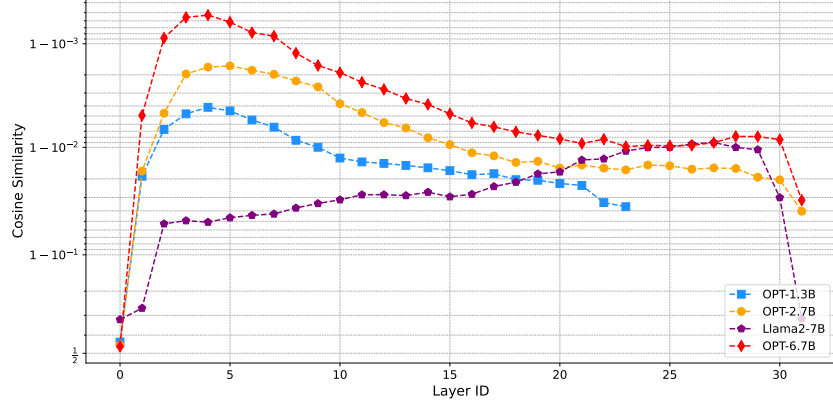


Figure 2: The cosine similarity between the input and output hidden states of the layers in OPT-1.3B, OPT-2.7B, OPT-6.7B, and Llama2-7B.

4 Experiments

4.1 Setup

To validate the effectiveness of our proposed model, we conduct experiments on popular open-source language models, including OPT families (OPT-1.3B, OPT-2.7B, OPT-6.7B) [29] and Llama2-7B [30]. In our approach, for Llama2-7B and OPT-6.7B, we utilize the SlimPajama-6B dataset and follow the method of Sheared-LLama [17] to determine the proportion of different domains. Based on this proportion, we randomly extract 20,000 samples for MLP training of Llama2-7B and 10,000 samples for MLP training of OPT-6.7B. We use a smaller dataset for the MLP training of OPT-6.7B because the layers in OPT-6.7B have higher cosine similarity. For OPT-1.3B and OPT-2.7B, we extract 3,000 samples from wikitext-2 for MLP training.

4.2 Benchmark

To comprehensively assess the changes in capabilities of LLM before and after pruning, we follow the evaluation method of LaCo [31]. On OPT-6.7B and Llama2-7B, assessments are conducted in five domains: reasoning, language, knowledge, examination, and understanding. We select several benchmarks from each category. For reasoning: **CMNLI** [32], **HellaSwag**(HeSw) [33], **PIQA** [34]. For language: **CHID** [35], **WSC** [36]. For knowledge: **CommonSenseQA**(CoQA) [37], **BoolQ** [38]. For examination: **MMLU** [39], **CMMLU** [40]. For understanding: **Race-High/Middle** [41], **XSum** [42], **C3** [43]. In addition, to more fully demonstrate the changes in the capabilities of LLM after pruning, we include two generation-type tasks: **GSM8K** [44] and **StrategyQA** [45], to showcase the LLM’s performance on generation tasks after pruning. On OPT-1.3B and OPT-2.7B, we follow the evaluation method of Llama [30], categorizing tasks based on common sense reasoning datasets: **PIQA**, **HellaSwag**, **WinoGrande** [46], **ARC-easy** [47], **ARC-challenge** [47], and **OpenbookQA** [48]. We use the official scripts from OpenCompass [49] for evaluation, employing a zero-shot or few-shot approach. Two evaluation modes are utilized: perplexity (PPL) and generation(GEN). For XSum, GSM8K, and StrategyQA, we use the GEN mode. The remaining benchmarks are evaluated using the PPL mode.

4.3 Baseline

To evaluate the effectiveness of our method, we compare several structured pruning methods for large language models following LaCo [31] and ShortGPT [50].

LLM-Pruner [27] utilizes gradients to estimate the importance of model weights and prunes the less significant coupled structures within the model based on this estimation.

LLM	Method	Ratio	Benchmarks												Ave.	Per.
			CMNLI	HeSw	PIQA	CHID	WSC	CoQA	BoolQ	Race-H	Race-M	C3	MMLU	CMMLU		
Llama2-7B	Dense	0.00%	33.0	71.3	78.1	41.6	37.5	66.7	70.8	35.5	33.1	43.8	46.8	31.8	49.2	100.0
	LLMPruner	27.0%	34.3	56.5	71.2	25.3	36.5	42.5	55.2	22.6	22.4	25.6	23.3	25.3	36.7	74.6
	SliceGPT	26.4%	31.7	50.3	66.2	20.8	36.5	41.4	38.3	21.1	21.7	39.8	28.9	25.4	35.2	71.5
	LaCo	27.1%	34.4	55.7	69.8	36.1	40.4	45.7	64.1	22.6	23.6	39.7	26.5	25.2	40.3	81.9
	ShortGPT	27.1%	33.0	53.0	66.4	24.7	52.5	48.0	74.7	32.3	35.2	39.6	44.0	32.2	44.6	90.7
	ShortGPT*	27.1%	33.1	58.2	66.9	22.3	36.5	48.8	65.2	36.9	35.5	41.3	45.4	29.2	43.3	88.0
	Ours	26.2%	33.0	59.2	70.1	24.2	36.5	57.3	65.0	37.5	37.1	43.2	47.0	31.9	45.2	91.9
OPT-6.7B	Dense	0.00%	32.9	63.3	75.4	21.6	41.4	54.8	64.6	25.4	25.1	38.7	24.7	25.5	41.1	100.0
	ShortGPT*	27.0%	32.6	25.7	50.8	14.5	38.5	18.8	43.6	22.7	22.0	26.6	24.0	25.5	28.8	70.1
	Ours		32.6	52.4	73.1	19.4	39.4	38.0	63.0	22.0	23.3	35.9	24.6	25.0	37.4	91.0

Table 1: Comparison of pruning methods on PPL benchmarks. "ShortGPT*" is our reproduction.

LLM	Method	Ratio	Benchmarks			Ave.	Per.
			Xsum	GSM8K	StrategyQA		
Llama2-7B	Dense	0.00%	19.4	16.5	60.2	32.0	100.0
	LLMPruner	27.0%	11.5	-	-	-	-
	SliceGPT	26.4%	4.9	-	-	-	-
	LaCo	27.1%	15.6	-	-	-	-
	ShortGPT*	27.1%	0.7	1.7	35.1	12.5	39.1
	Ours	26.2%	18.0	2.3	44.9	21.7	67.8
OPT-6.7B	Dense	0.00%	13.4	2.2	54.3	23.3	100.0
	ShortGPT*	27.0%	3.7	0.2	35.3	13.1	56.2
	Ours	26.0%	13.1	1.0	47.3	20.5	88.0

Table 2: Comparison of pruning methods on GEN benchmarks. "ShortGPT*" is our reproduction.

SliceGPT [18] inserts dimensionality reduction matrices into the model and employs PCA to initialize and trim the unimportant components in these matrices. It then merges the reduced-dimensional matrices with the model’s original weight matrices, thereby reducing the model’s parameter count.

LaCo [31] is a layer pruning method that prunes layers by merging similar layers.

ShortGPT [50] is another layer pruning method that evaluates the importance of layers using the Block Influence metric and directly removes the less significant layers for pruning.

LLM-Pruner and SliceGPT are both post-training methods, whereas LaCo and ShortGPT do not involve post-training. For fairness, we do not conduct post-training on any method for OPT-6.7B and Llama2-7B. However, for OPT-1.3B and OPT-2.7B, we apply post-training to all methods to enhance the performance of the pruned models.

4.4 Main Results

To validate the effectiveness of our proposed method, we conduct comparative experiments on OPT-6.7B and Llama2-7B. Following previous work, we carry out experiments with approximately 25% of the parameters pruned. The experimental results are presented in Table 1 and Table 2. The results indicate that the model pruned using our proposed pruning method, LLM-Streamline, outperforms the baseline method in overall performance, managing to retain most of the capabilities of the large language model across a variety of tasks. Specifically, in comparison to the previous SOTA method, ShortGPT, LLM-Streamline surpasses ShortGPT in both classification and generation tasks and outperforms ShortGPT by 30 points in generation tasks.

To present a more comprehensive view of the results of our proposed LLM-Streamline under different pruning ratios, we also conduct experiments on OPT-1.3B and OPT-2.7B with pruning rates of 15-20%, 20-25%, and 25-30%. As shown in Table 3, our pruning method outperforms the baselines across all pruning rates on both models. When pruning at a rate of 15-20%, OPT-1.3B maintains 92% of the original model’s performance, and OPT-2.7B maintains 96% of the original model’s performance. This indicates that our pruning method remains highly effective even on smaller-scale models. In the experiments of OPT-1.3B and OPT-2.7B, LLM-Streamline is post-trained using the wikitext-2 dataset. To perform post-training more efficiently, we freeze the parameters of the original model and only train the inserted MLP. To enhance the effectiveness of SliceGPT, we follow the method in the SliceGPT paper and employ the Alpaca dataset [51] for post-training SliceGPT.

LLM	Method	Ratio	Benchmarks						Ave.	Per.
			PIQA	WinoGrande	HellaSwag	ARC-easy	ARC-challenge	OpenBookQA		
OPT-1.3B	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	SliceGPT	18.1%	67.6	53.6	35.7	51.1	23.1	20.2	41.8	86.7
	Ours	18.1%	68.8	58.4	39.1	54.3	23.3	23.3	44.5	92.3
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	SliceGPT	25.8%	65.5	52.8	34.2	48.8	24.4	17.0	40.5	84.0
	Ours	25.8%	66.4	56.0	36.8	51.6	22.2	21.0	42.3	87.8
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	SliceGPT	33.6%	62.4	52.6	32.2	45.4	23.1	16.6	38.7	80.3
	Ours	33.6%	62.9	52.1	33.9	48.3	20.8	20.6	39.8	82.6
OPT-2.7B	Dense	0.00%	73.8	61.0	45.9	60.9	26.8	25.0	48.9	100.0
	SliceGPT	16.8%	69.6	56.3	40.4	56.2	27.5	20.2	45.0	92.0
	Ours	16.8%	70.7	60.4	42.9	57.8	25.3	24.4	46.9	95.9
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	SliceGPT	25.7%	69.1	55.0	37.9	53.9	26.7	18.2	43.5	84.0
	Ours	25.7%	67.0	59.5	40.3	54.6	24.7	22.2	44.7	91.4
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	SliceGPT	34.6%	64.8	54.1	35.6	50.0	26.5	18.0	41.5	84.9
	Ours	34.6%	65.3	55.3	36.3	51.4	24.5	21.0	41.9	85.7

Table 3: Comparing pruning methods on common sense reasoning benchmarks for OPT-1.3B and OPT-2.7B

LLM	Training Data Quantities	Benchmarks											Ave.	Per.
		CMNLI	HeSw	PIQA	CHID	WSC	CoQA	BoolQ	Race-H	Race-M	C3	MMLU	CMMLU	
Llama2-7B	5000	32.4	57.7	69.0	17.9	36.5	55.4	67.5	38.4	39.5	40.4	46.5	29.0	44.2 89.8
	10000	32.8	57.7	69.9	22.8	36.5	57.3	64.7	39.5	37.7	39.0	47.1	31.2	44.7 90.9
	15000	33.0	58.5	70.1	21.2	36.5	57.3	65.8	37.7	39.1	41.0	47.3	31.4	44.9 91.3
	20000	33.0	59.2	70.1	24.2	36.5	57.3	65.0	37.5	37.1	43.2	47.0	31.9	45.2 91.9
	25000	33.0	58.6	70.4	17.7	36.5	57.0	63.2	39.5	39.6	38.5	45.8	29.0	44.1 89.6
OPT-6.7B	5000	32.0	50.3	72.3	19.7	37.5	35.5	61.5	21.8	23.3	35.7	24.1	24.9	36.6 89.1
	10000	32.6	52.4	73.1	19.4	39.4	38.0	63.0	22.0	23.3	35.9	24.6	25.0	37.4 91.0
	15000	31.9	45.0	70.4	18.6	37.5	32.6	59.1	21.8	22.1	36.2	25.1	24.7	35.4 86.1

Table 4: Performance of Llama2-7B and OPT-6.7B with different training data quantities.

5 Analysis

5.1 Effect of the Amount of Training Data

We conduct a comprehensive analysis of the performance of the MLP model during training on Llama2-7B and OPT-6.7B architectures. The results in Table 4 cover all 12 PPL benchmarks under various training data quantities. According to the results in Table 4, we can observe a clear trend: the accuracy initially shows improvements, followed by a slight decline, reaching a peak of around 20,000 data points for Llama2-7B (10,000 data points for OPT-6.7B). This indicates that the MLP has converged at 20,000 data points (10,000 data points for OPT-6.7B).

5.2 Effect of Different Lightweight Models

To validate the necessity and simplicity of employing an MLP as a lightweight model, we conduct experiments on OPT-1.3B and OPT-2.7B, wherein we replace the utilized MLP with a transformer layer. In addition, we investigate the impact of the parameter-sharing technique on the performance of the lightweight model, i.e., we let the lightweight model compute n times during the forward computation (n is the number of layers removed). Finally, we attempt to remove the pruned layers without substituting them with any lightweight model. The MLPs and transformer layers utilized in our experiments are trained on 3,000 samples from the wikitext-2 dataset. Consistent with Section 4.4, we conduct post-training on the pruned models, concentrating solely on adjusting the parameters of the lightweight models. For the approach of directly removing pruned layers, we employ LoRa [52] to ensure that the number of parameters involved in the training is equivalent to that of an MLP. The experimental results are presented in Table 5.

LLM	Lightweight Model	Ratio	Benchmarks						Ave.	Per.
			PIQA	WinoGrande	HellaSwag	ARC-easy	ARC-challenge	OpenBookQA		
OPT-1.3B	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	MLP w/o params sharing	18.1%	68.8	58.4	39.1	54.3	23.3	23.2	44.5	92.3
	MLP w/ params sharing	18.1%	68.6	56.8	39.1	54.2	21.9	23.0	43.9	91.1
	Transformer Layer w/o params sharing	15.5%	58.1	54.4	30.1	35.1	22.4	17.2	36.2	75.1
	Transformer Layer w/ params sharing	15.5%	67.1	54.9	37.9	52.1	22.6	20.0	42.4	88.0
	None	19.4%	57.2	51.7	29.1	32.5	22.7	13.2	34.4	71.4
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	MLP w/o params sharing	25.8%	66.4	56.0	36.8	51.6	22.2	21.0	42.3	87.8
	MLP w/ params sharing	25.8%	65.8	54.1	36.4	51.8	21.9	21.6	41.9	86.9
	Transformer Layer w/o params sharing	23.2%	53.1	50.3	26.7	29.7	20.1	14.8	32.5	67.4
	Transformer Layer w/ params sharing	23.2%	62.2	53.5	32.8	47.8	20.2	18.4	39.2	81.3
	None	27.1%	52.2	51.1	25.7	26.6	20.5	14.0	31.7	65.1
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	MLP w/o params sharing	33.6%	62.9	52.1	33.9	48.3	20.8	20.6	39.8	82.6
	MLP w/ params sharing	33.6%	62.3	53.4	33.0	48.1	20.9	20.4	39.7	82.4
	Transformer Layer w/o params sharing	31.0%	53.4	53.0	26.1	26.9	19.7	14.4	32.3	67.0
	Transformer Layer w/ params sharing	31.0%	59.4	51.5	29.1	43.0	17.8	15.4	36.0	74.7
	None	34.8%	50.5	51.5	25.8	26.2	20.3	14.6	31.5	65.4

Table 5: The influence of different lightweight models on pruning performance.

LLM	Lightweight Model	Ratio	Benchmarks						Ave.	Per.
			PIQA	WinoGrande	HellaSwag	ARC-easy	ARC-challenge	OpenBookQA		
OPT-1.3B	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	MLP	18.1%	68.8	58.4	39.1	54.3	23.3	23.2	44.5	92.3
	Multiple MLP	12.9%	66.3	55.9	37.5	52.3	23.0	20.4	42.6	88.4
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	MLP	25.8%	66.4	56.0	36.8	51.6	22.2	21.0	42.3	87.8
	Multiple MLP	18.1%	64.1	53.0	34.5	48.5	20.6	19.0	40.0	83.0
	Dense	0.00%	72.4	59.3	53.7	51.0	29.5	23.4	48.2	100.0
	MLP	33.6%	62.9	52.1	33.9	48.3	20.8	20.6	39.8	82.6
	Multiple MLP	23.2%	60.4	51.0	30.6	46.4	19.5	15.6	37.3	77.4

Table 6: The influence of one MLP or multiple MLP on pruning performance.

We can see that the parameter-sharing MLP introduces more computation but does not ultimately outperform the MLP. Replacing the MLP with a transformer layer results in a significant decrease in model performance. Unlike MLP, the utilization of parameter sharing enhances the transformer layer’s efficiency but also does not exceed the performance of MLP. In addition, not using any lightweight models also leads to a significant decrease in model performance. Therefore, we conclude that using a simple MLP instead of a pruning layer is sufficient.

5.3 Effect of the Number of Lightweight Models

Due to the tendency for less important layers to cluster together in LLM, we propose employing a lightweight model to substitute the pruned layers and restore model performance. To demonstrate that a lightweight model is the most parameter-friendly and overall performance-optimal, we conduct the following experiment on OPT-1.3B and OPT-2.7B: for each layer that requires pruning, we train an MLP to replace that layer. The results of the experiment are presented in Table 6. From the experimental results, it is evident that although using multiple MLPs results in a significantly larger number of parameters, its performance does not surpass that achieved by using just one MLP.

6 Conclusion

This work introduces a novel model pruning method, LLM-Streamline, which measures the importance of layers based on the cosine similarity between the hidden states of their inputs and outputs and prunes those layers that are deemed less important. The method then trains a lightweight model to replace the pruned layers. Our studies indicate that certain layers in LLM have high cosine similarity between their input and output hidden states, suggesting that the mapping relationships represented by these layers are simple and can be approximated by lightweight models. Experiments have shown that using only an MLP to replace multiple consecutive transformer layers can maintain up to 92% of the LLM’s classification performance and 68% of the LLM’s generation performance while reducing

the model’s parameter count to about 25%, outperforming previous state-of-the-art (SOTA) pruning methods. We have validated LLM-Streamline across models with 1B, 3B, and 7B parameter scales, achieving promising performance in each case. Finally, we explore the impact of different lightweight models and varying training data quantities on the performance of LLM-Streamline.

7 Limitation

Although our model can outperform the existing baseline, the model obtained from pruning still has a large performance gap with the model that satisfies the scaling law. Meanwhile, for the generation task, although our method has a large performance improvement compared to the baseline method, it still has a large performance loss compared to the original model. Therefore, we still need to further investigate how to compress the model and guarantee its performance while using limited data.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [2] Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models, 2023.
- [3] Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. Model compression and efficient inference for large language models: A survey, 2024.
- [4] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [5] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819, 2021.
- [6] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*, 2022.
- [7] Yukun Huang, Yanda Chen, Zhou Yu, and Kathleen McKeown. In-context learning distillation: Transferring few-shot learning ability of pre-trained language models. *arXiv preprint arXiv:2212.10670*, 2022.
- [8] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.
- [9] Zhenhua Liu, Yunhe Wang, Kai Han, Wei Zhang, Siwei Ma, and Wen Gao. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems*, 34:28092–28103, 2021.
- [10] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-Power Computer Vision*, pages 291–326. Chapman and Hall/CRC, 2022.
- [11] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.
- [12] Christos Louizos, Max Welling, and Diederik P Kingma. Learning sparse neural networks through l_0 regularization. *arXiv preprint arXiv:1712.01312*, 2017.
- [13] Tianyi Chen, Tianyu Ding, Badal Yadav, Ilya Zharkov, and Luming Liang. Lorashear: Efficient large language model structured pruning and knowledge recovery. *arXiv preprint arXiv:2310.18356*, 2023.
- [14] Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR, 2023.
- [15] Rocktim Jyoti Das, Liqun Ma, and Zhiqiang Shen. Beyond size: How gradients shape pruning decisions in large language models. *arXiv preprint arXiv:2311.04902*, 2023.

- [16] Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*, 2023.
- [17] Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. Sheared llama: Accelerating language model pre-training via structured pruning. *arXiv preprint arXiv:2310.06694*, 2023.
- [18] Saleh Ashkboos, Maximilian L Croci, Marcelo Gennari do Nascimento, Torsten Hoeffler, and James Hensman. Slicept: Compress large language models by deleting rows and columns. *arXiv preprint arXiv:2401.15024*, 2024.
- [19] Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.
- [20] Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned, 2019.
- [21] JS McCarley, Rishav Chakravarti, and Avirup Sil. Structured pruning of a bert-based question answering model. *arXiv preprint arXiv:1910.06360*, 2019.
- [22] Sai Prasanna, Anna Rogers, and Anna Rumshisky. When bert plays the lottery, all tickets are winning. *arXiv preprint arXiv:2005.00561*, 2020.
- [23] Tycho FA van der Ouderaa, Markus Nagel, Mart Van Baalen, Yuki M Asano, and Tijmen Blankevoort. The llm surgeon. *arXiv preprint arXiv:2312.17244*, 2023.
- [24] Ye Liu and Michael K Ng. Deep neural network compression by tucker decomposition with nonlinear response. *Knowledge-Based Systems*, 241:108171, 2022.
- [25] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur. Semi-orthogonal low-rank matrix factorization for deep neural networks. In *Interspeech*, pages 3743–3747, 2018.
- [26] Mingxue Xu, Yao Lei Xu, and Danilo P Mandic. Tensorgpt: Efficient compression of the embedding layer in llms based on the tensor-train decomposition. *arXiv preprint arXiv:2307.00526*, 2023.
- [27] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.
- [28] Yuxin Jiang, Chunkit Chan, Mingyang Chen, and Wei Wang. Lion: Adversarial distillation of closed-source large language model. *arXiv preprint arXiv:2305.12870*, 2023.
- [29] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.
- [30] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
- [31] Yifei Yang, Zouying Cao, and Hai Zhao. Laco: Large language model pruning via layer collapse. *arXiv preprint arXiv:2402.11187*, 2024.
- [32] Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*, 2020.
- [33] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- [34] Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.
- [35] Chujie Zheng, Minlie Huang, and Aixin Sun. Chid: A large-scale chinese idiom dataset for cloze test. *arXiv preprint arXiv:1906.01265*, 2019.
- [36] Hector Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Thirteenth international conference on the principles of knowledge representation and reasoning*, 2012.

- [37] Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.
- [38] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*, 2019.
- [39] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.
- [40] Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese. *arXiv preprint arXiv:2306.09212*, 2023.
- [41] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*, 2017.
- [42] Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.
- [43] Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. Investigating prior knowledge for challenging chinese machine reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:141–155, 2020.
- [44] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [45] Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021.
- [46] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.
- [47] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- [48] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*, 2018.
- [49] OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. *GitHub repository*, 2023.
- [50] Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.
- [51] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models*. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, 3(6):7, 2023.
- [52] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

A The proportion of different domains randomly selected from the SlimPajama-6B dataset

	CC	GitHub	Book	StackExchange	Wiki	ArXiv	C4
SlimPajama-6B	54.1%	4.2%	3.7%	2.8%	3.1%	3.4%	28.7%
Ours	36.1%	0.8%	9.1%	1.0%	3.1%	0.7%	49.2%

Table 7: The proportion of different domains randomly selected from the SlimPajama-6B dataset.

B MLP Training Implementation Details

As mentioned in Section 4.3, for OPT-1.3B and OPT-2.7B, we train the MLP using 3,000 randomly selected samples from the wikitext-2 dataset. The batch size is set to 512, the learning rate to $4e-4$, and the number of epochs to 5, which could be completed in just 15 minutes on a single RTX 3090. For OPT-6.7B and Llama2-7B, we train the MLP using 10,000 or 20,000 randomly select samples from the SlimPajama-6B dataset. The batch size is set to 2048, the learning rate to $1e-3$, and the number of epochs to 5, requiring approximately 2.5 hours to complete the training on a single A100.