Head-wise Shareable Attention for Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) suffer from huge number of parameters, which restricts their deployment on edge devices. Weight sharing is one promising solution that encourages weight reuse, effectively reducing memory usage with less performance drop. However, current weight sharing techniques primarily focus on small-scale models like BERT and employ coarse-grained sharing rules, e.g., layer-wise. This becomes limiting given the prevalence of LLMs and sharing an entire layer or block obviously diminishes the flexibility of weight sharing. In this paper, we present a perspective on ***head-wise shareable attention for large language models***. We further propose two memory-efficient methods that share parameters across attention heads, with a specific focus on LLMs. Both of them use the same dynamic strategy to select the shared weight matrices. The first method directly reuses the pre-trained weights without retraining, denoted as **DirectShare**. The second method first post-trains with constraint on weight matrix similarity and then shares, denoted as **PostShare**. Experimental results reveal our head-wise shared models still maintain satisfactory capabilities, demonstrating the feasibility of fine-grained weight sharing applied to LLMs.

1 Introduction

Large Language Models (LLMs) have achieved breakthrough performance in a variety of natural language processing tasks (Wei et al., 2022; Bubeck et al., 2023; Zhao et al., 2023). However, such remarkable capability typically comes at the cost of a substantial increase in the model size (Kaplan et al., 2020). Thus, LLMs with billions of parameters (Brown et al., 2020; Touvron et al., 2023) are more resource-hungry despite a wide margin of superiority over small-scale models (Devlin et al., 2018; Liu et al., 2019). This can also pose challenges for deployment on low-capability devices due to limited storage and GPU memory.

To address the high memory requirements of models, weight sharing (Takase and Kiyono, 2021; Liu et al., 2023) aims to reuse the same parameters to achieve memory- and storage-efficiency while preserving model performance. For small-scale models, e.g., BERT, it is known that several techniques (Lan et al., 2019; Liu et al., 2023) are proposed to explore across-layer parameter sharing. While, Zhang et al. (2022) demonstrate identical weights across different layers are the main cause of training instability and performance degradation. Moreover, the effective of similar techniques at the scale of LLMs remains uncertain.

Thus, we strive to solve this central question: ***Can we design fine-grained weight sharing strategy that can smoothly apply to large language models*?** For an effective memory-efficient weight sharing method tailored to LLMs, two key challenges must be tackled: a) the choice of shared modules whose weights are reused; b) the trade-off between reducing memory footprint and preserving diverse capabilities.

In the preliminary work, we empirically evaluate the feasibility of weight sharing across the attention heads in LLMs inspired by attention map (i.e., attention scores) reuse. Subsequently, we introduce our design of head-wise shareable attention strategy. It is a simple and intuitive technique for parameter sharing that can be implemented in a few minutes. Specifically, given the pre-trained weight matrices, we concatenate the weight matrix W^q and W^k for each head to measure the cosine similarity that determines which heads can be shared. Meanwhile, head-wise weight sharing promotes parameter diversity in the layers, and thus its performance degradation is acceptable when the number of shared parameters is below 30%. Even as weight sharing ratio increases rapidly, our proposed constrained post-training method can narrow the performance drop, which may necessitate additional time.

In summary, our key contributions include:

- We investigate the feasibility of head-wise weight sharing for large language models and propose two corresponding methods named **DirectShare** and **PostShare**.
- The proposed **DirectShare** is time-efficient and retain a large portion of the performance when sharing ratio is below 30%. Complementarily, **PostShare** yields satisfactory performance via post-training, especially under large ratios.
- Experiments show our proposal achieves comparable performance to the competitive memory-efficient methods. Additional analysis also indicates its efficiency in small-scale models.

2 Related Works

2.1 Memory-efficient Approaches for LLMs

With the growing size of language models, several memory-efficient techniques are proposed to solve. One line to reducing the memory footprint involves network compression, like quantization (Bai et al., 2020; Tao et al., 2022), pruning (Yang et al., 2022; Tao et al., 2023) and knowledge distillation (Wu et al., 2023; Tan et al., 2023). However, when applied to LLMs, many approaches have become infeasible (Frantar and Alistarh, 2023). To recover accuracy, they require extensive post-training of the model (Dettmers et al., 2023; Sun et al., 2023).

In addition to these conventional methods, researchers have also investigated more efficient variations of the self-attention mechanism for LLMs (Kitaev et al., 2020; Lv et al., 2023). Reformer (Kitaev et al., 2020) leverages sparsity in the attention layers to improve the efficiency on long sequences and with small memory use. Lightformer (Lv et al., 2023) deploys SVD weight transfer and parameter sharing, which can significantly reduce the parameters on the premise of ensuring model performance. In this paper, our focus is on weight sharing across attention heads.

2.2 Weight Sharing

Weight sharing is a widely used technique (Lan et al., 2019; Liu et al., 2023; Lv et al., 2023; Xu and McAuley, 2023) that aims to improve parameter efficiency and reduce inference memory footprint. Weight sharing enables model compression by eliminating redundant parameters and decouples computation and parameters by reusing the same parameters for multiple computations.

Task-oriented Weight Sharing. One of the prevalent tasks using weight sharing mechanisms is nerual machine translation (NMT). Tied Transformer (Xia et al., 2019) considers model-level sharing and shares the weights of the encoder and decoder of an NMT model. Dabre and Fujita (2019) proposes a method, which shares the weights across all Transformer layers and keeps performance in NMT. Besides, Chi et al. (2021) bring the idea of ALBERT (Lan et al., 2019) to the speech recognition task.

Layer-wise Weight Sharing. Universal Transformer (Dehghani et al., 2018) shares the weights across all layers with a dynamic halting mechanism and improves accuracy on several tasks. Subformer (Reid et al., 2021) utilizes sandwich-style parameter sharing, which only shares the central layers while leaving the first and last layers independent. Takase and Kiyono (2021) study strategies to explore the best way to prepare parameters of M layers and assign them into N layers (1≤M≤N).

3 Motivation and Empirical Analysis

In this section, we analyze the feasibility of head-wise weight sharing from the perspective of attention map reuse.

3.1 Attention Map Similarity: From Layer-wise to Head-wise

Prior researches (Xiao et al., 2019; Ying et al., 2021; Bhojanapalli et al., 2021) demonstrate the effectiveness of attention map reuse due to the high similarity of attention scores between different layers (especially for adjacency layers). Motivated by this, we delves into attention map similarity, specifically transitioning from layer-wise to head-wise analysis. To measure the evolution of the attention maps over layers and heads, we use the cosine similarity $\mathcal{S}_{cos}$. When $\mathcal{S}_{cos}$ equals one, it means that the attention maps are perfectly similar. Considering two specific self-attention layers, the cosine similarity is calculated as follows:

$$\mathcal{S}_{cos}(\mathbf{A}_p, \mathbf{A}_q) = \frac{\mathbf{A}_p^T \mathbf{A}_q}{\|\mathbf{A}_p\| \|\mathbf{A}_q\|} \quad (1)$$

where $\mathbf{A}_p, \mathbf{A}_q$ denote the attention map of layers p and q.

We visualize the layer-wise and head-wise attention map similarity across three task-specific datasets: WMT14 (En-Fr) (Bojar et al., 2014), CommonsenceQA (Talmor et al., 2019) and WSC

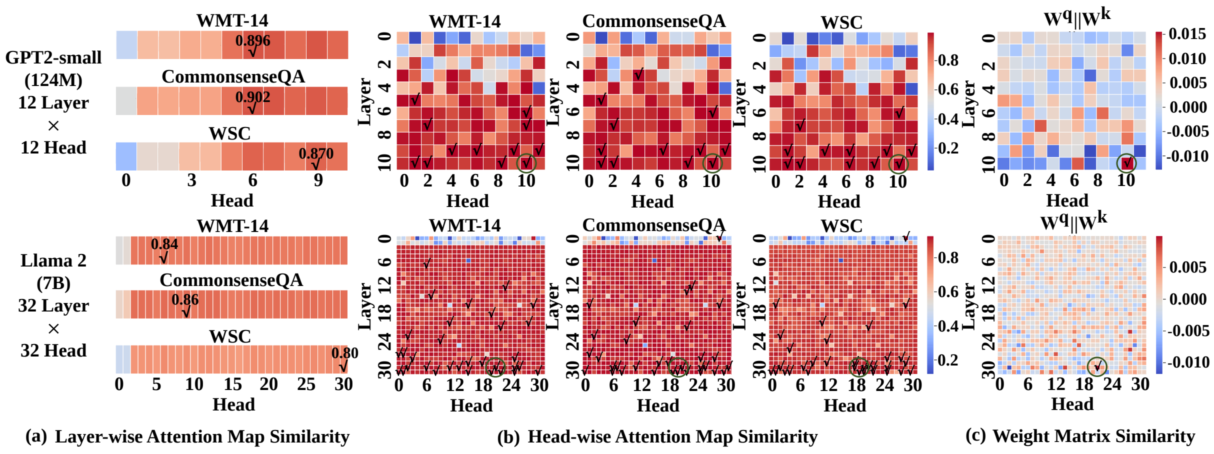


Figure 1: (a) Layer-wise Attention Map Similarity. Taking the last layer as an example, the most similar attention layer with it is marked with $\surd$. (b) Head-wise Attention Map Similarity. $\surd$ mark the top n heads whose attention maps that are most similar to the 6-th head in the last layer(n=the number of heads per layer). (c) Weight Matrix Similarity. ◯ mark the connection between attention map similarity and weight similarity.

(Levesque et al., 2012). As shown in Fig. 1(a) and (b), the degrees of similarity in attention scores computed in different layers and heads present a certain level of consistency across different tasks. In addition, we find that the cosine similarity values for pairs with high similarity are higher among different heads compared to different layers. Specifically, the most similar self-attention layers reach a cosine similarity value of approximately 0.90, while in the case of head-wise comparisons, several pairs have a remarkable similarity of nearly 0.99.

One observation is that as the number of parameters increases, modules with high similarity exhibit variations, particularly in the fine-grained (e.g., head-wise) comparisons within large-scale pre-trained language models. Existing approaches employ "learning to share" techniques to dynamically adjust the sharing strategy (Xiao et al., 2019) or use a uniform sharing strategy but train the modified model from scratch (Ying et al., 2021; Shim et al., 2023). However, such strategies pay little attention on reusing attention map among heads and incur high computational costs for LLMs.

3.2 From Attention Map Similarity to Weight Matrix Similarity

Attention weight matrix similarity provides a complementary perspective to attention map similarity, since the attention scores are calculated based on the weight matrices W^q, W^k. Weight sharing is traditionally based on the assumption that overparameterization is evident in large-scale Transformer models, i.e., the difference in weights decreases as model size increases (Li et al., 2020). In this paper, we explore a potential relationship between attention map similarity and weight similarity.

As mentioned in Section 3.1, head-wise attention map similarity is higher than the cross-layer similarity, while to the best of our knowledge, head-wise attention map reuse is yet to be explored. This might be attributed to the difficulty in finding an optimal dynamic head-wise sharing strategy across different tasks. One intuitive solution is to first measure the attention map similarity between every pair of heads in each dataset separately, and then choose the overlapping modules to share.

Combined with the analysis of weight matrix similarity, we have made a key discovery: given a pre-trained LLM, by concatenating the weight matrix W^q and W^k for each head to measure the cosine similarity, ***the most similar weight matrix corresponds to the overlapping modules with highly similar attention maps observed across different datasets***. As illustrated in Fig. 1(b) and (c), deep green circles mark the connection between attention map similarity and weight similarity.

This finding implies that attention heads with high weight matrix similarity also demonstrate analogous attention map similarity regardless of the datasets and model size. Furthermore, since different heads within the layer present sufficient diversity (Zhou et al., 2021; Vig, 2019), we suppose that weight sharing among these heads can result in higher model behavior consistency compared to layer-wise weight sharing. Thus, we further pro-

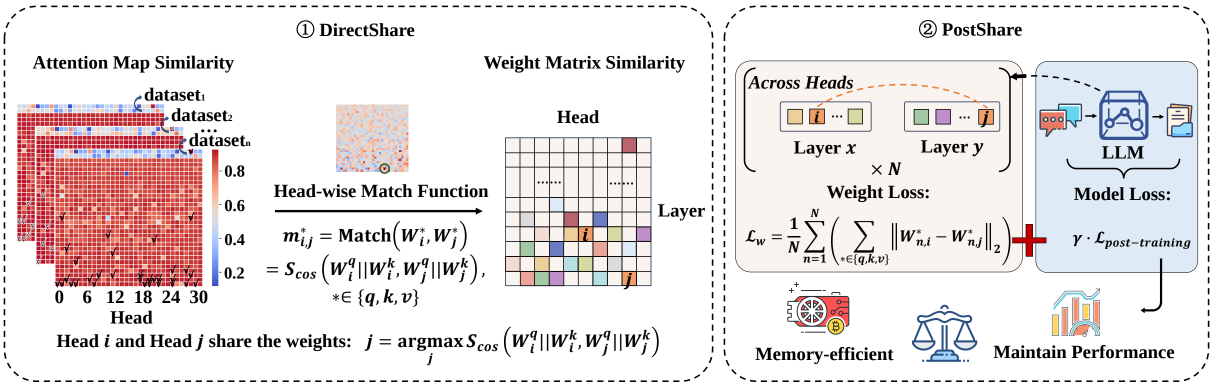

① DirectShare
Attention Map Similarity
dataset₁
dataset₂
datasetₙ
0.8
0.6
0.4
0.2
0 6 12 18 24 30
Head
Head-wise Match Function
$m^*_{i,j} = \text{Match}(W^*_i, W^*_j)$
$= S_{cos}(W^q_i||W^k_i, W^q_j||W^k_j)$,
$* \in \{q,k,v\}$
Weight Matrix Similarity
Head
Layer
Head i and Head j share the weights: $j = \text{argmax}_j\, S_{cos}(W^q_i||W^k_i, W^q_j||W^k_j)$
② PostShare
Across Heads
Layer x
Layer y
$\times N$
Weight Loss:
$\mathcal{L}_w = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{*\in\{q,k,v\}} \|W^*_{n,i} - W^*_{n,j}\|_2\right)$
LLM
Model Loss:
$\gamma \cdot \mathcal{L}_{post-training}$
Memory-efficient
Maintain Performance

Figure 2: ① **DirectShare**: Inspired by attention map reuse, directly share weight matrices across different heads based on cosine similarity; ② **PostShare**: To balance the memory usage and the performance, implement post-training with the constraint of weight matrix similarity and then share.

pose a simple yet effective method for head-wise weight sharing, especially validating its feasibility in large-scale models.

4 Head-wise Shareable Attention

Inspired by Section 3, we present a perspective on head-wise shareable attention for LLMs. Based on one straightforward yet effective weight sharing strategy, we propose two complementary methods, named **DirectShare** and **PostShare**. The overview of our proposal is presented in Figure 2.

4.1 Head-wise Weight Sharing Strategy

Multi-Head Attention (MHA) block is essentially a procedure that computes the relevance of each token in a sentence with respect to all other tokens. Let L be the number of input tokens and M be the number of attention heads in total. Given the input $X \in \mathbb{R}^{L\times D}$, we can obtain queries, keys, and values in the i-th ($1 \le i \le M$) head via three weight matrices, denoted by $W^q_i \in \mathbb{R}^{L\times d_q}$, $W^k_i \in \mathbb{R}^{L\times d_k}$ and $W^v_i \in \mathbb{R}^{L\times d_v}$, respectively. D is the embedding dimension, and $d_q, d_k (= d_q), d_v$ represent the dimensions of three weight matrices, respectively.

To investigate the strategy of weight sharing applied to all the above three weight matrices across heads for LLMs, we perform preliminary experiments in the choice of head-wise match functions **Match**$(\cdot,\cdot)$. For the match functions, inputs are the weight matrices of head $\boldsymbol{i}, \boldsymbol{j}$ and outputs are called matching scores $\boldsymbol{m}$. The higher the score, the more likely it is to share parameters across the heads.

$$\boldsymbol{m}^*_{\boldsymbol{i},\boldsymbol{j}} = \textbf{Match}(W^*_i, W^*_j), * \in \{q,k,v\} \tag{2}$$

Based on our intuitive analysis in Section 3.2, we choose the cosine similarity between **the concatenation matrix of W^q_i and W^k_i**:

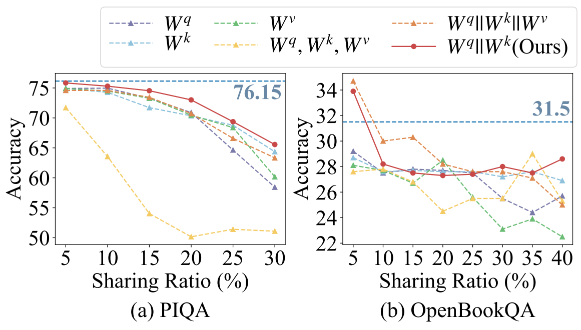


Figure 3: Experiments performed on PIQA and OpenBookQA using different head-wise match functions for Baichuan 2-7B model.

$$\boldsymbol{m}^q_{\boldsymbol{i},\boldsymbol{j}} = \boldsymbol{m}^k_{\boldsymbol{i},\boldsymbol{j}} = \boldsymbol{m}^v_{\boldsymbol{i},\boldsymbol{j}} = \mathcal{S}_{cos}(W^q_i||W^k_i, W^q_j||W^k_j) \tag{3}$$

Besides, we try another five match functions to compare: (1) **Only** W^q_i used to measure the cosine similarity, i.e., $\boldsymbol{m}^*_{\boldsymbol{i},\boldsymbol{j}} = \mathcal{S}_{cos}(W^q_i, W^q_j)$; (2) **Only** W^k_i used to measure the cosine similarity, i.e., $\boldsymbol{m}^*_{\boldsymbol{i},\boldsymbol{j}} = \mathcal{S}_{cos}(W^k_i, W^k_j)$; (3) **Only** W^v_i used to measure the cosine similarity, i.e., $\boldsymbol{m}^*_{\boldsymbol{i},\boldsymbol{j}} = \mathcal{S}_{cos}(W^v_i, W^v_j)$; (4) **Concatenate all the three matrices** and then calculate the cosine similarity, i.e., $\boldsymbol{m}^*_{\boldsymbol{i},\boldsymbol{j}} = \mathcal{S}_{cos}(W^q_i||W^k_i||W^v_i, W^q_j||W^k_j||W^v_j)$; (5) **Separately use** W^q_i, W^k_i, W^v_i to measure the cosine similarity and do weight sharing respectively, i.e., $\boldsymbol{m}^*_{\boldsymbol{i},\boldsymbol{j}} = \mathcal{S}_{cos}(W^*_i, W^*_j)$ and again $* \in \{q,k,v\}$.

Figure 3 shows the results of our exploratory study via **DirectShare**. As evidenced by the performance curve, using separately weight sharing causes a significant decline in performance compared with sharing the three weight matrices together. And it is enough to do head-wise weight sharing focusing only on the concatenation matrix of W_i^q and W_i^k, since it achieves a favorable trade-off between reducing memory footprint and maintaining performance.

4.2 DirectShare

In practice, we traverse all head pairs to compute matching scores on Equation 3 and for each head, select the one with the highest score to match. When M candidate head pairs prepared, we select the top-N pairs in descending order according to the desired sharing ratio. Finally, we can share the weight matrices together between each selected attention head pairs. A detailed algorithm for our **DirectShare** is presented in Algorithm 1.

Algorithm 1: DirectShare using Head-wise Weight Sharing Strategy

```
Input: Sharing Ratio $\alpha$, Weight Matrices of
       the MHA $\{W^*\}, * \in \{q,k,v\}$
Output: The LLM after weight sharing
1  for $layer_i \leftarrow 2$ to $l$ do
2      for $i \leftarrow 1$ to $head\_num$ do
3          $match\_index \leftarrow$ -1
4          $match\_score \leftarrow$ -1
5          for $layer_j \leftarrow 1$ to $layer_i - 1$ do
6              for $j \leftarrow 1$ to $head\_num$ do
7                  Compute $\boldsymbol{m}^*_{i,j}$ using Eq. 3;
8                  if $\boldsymbol{m}^*_{i,j} > match\_score$ then
9                      $match\_score \leftarrow \boldsymbol{m}^*_{i,j}$
10                     $match\_index \leftarrow (layer_j, j)$
11         Get one candidate head pair
           $< i, match\_index >$ for sharing;
12 Sort the matching scores in descending
   order;
13 Select the top-$N$ pairs according to $\alpha$;
14 Share the weight matrices together between
   each selected attention head pairs.
```

4.3 PostShare

Although **DirectShare** demonstrates effectiveness in our experiments, we have also encountered noticeable performance drop in minor reading comprehension datasets. To alleviate this problem, we propose **PostShare**, softly aligning model weights during the post-training process.

With the same sharing strategy (Section 4.1), **PostShare** first selects the set of weight matrices to share. Next, we incorporate a regularization term into the loss function to constrain our post-training process, encouraging selected weight matrices more similar:

$$\mathcal{L}_w = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{*\in\{q,k,v\}}\left\|W^*_{n,i} - W^*_{n,j}\right\|_2\right) \quad (4)$$

where N is the number of selected attention head pairs <i,j> for sharing. With this regularization weight loss, the proposed **PostShare** learn model weights W by minimizing the following combined loss function:

$$\min_W \mathcal{L}_{post-training} + \gamma \times \mathcal{L}_w \quad (5)$$

where $\mathcal{L}_{post-training}$ is the original post-training loss, γ controls the strength of $\mathcal{L}_w$. After the post-training process, the corresponding weight matrices can be shared as **DirectShare** does. Although post-training indeed increases the time cost of weight sharing, **PostShare** achieves stable and satisfactory performance across different tasks when reducing memory usage.

5 Experiments

5.1 Experimental Settings

Backbone Models. We evaluate **DirectShare** and **PostShare** on two open-source LLMs: Llama 2 (Touvron et al., 2023) and Baichuan 2 (Baichuan, 2023) with 7B and 13B parameters. In **PostShare**, we use English Wikipedia (Foundation) to post-train the backbone models for weight sharing.

Evaluation. To comprehensively evaluate the model capabilities, we experiment on five distinct tasks: reasoning, understanding, language, knowledge and examination. For consistent comparisons, we deploy open-source LLM evaluation platform OpenCompass (Contributors, 2023).

Baselines. Since existing weight sharing techniques do not support LLMs, we compare **DirectShare** against **Magnitude Pruning** (Zhu and Gupta, 2017) and **LLM-Pruner** (Ma et al., 2023), two influential works for model pruning. Certainly, they are not directly comparable. To ensure fairness in the experiments, both of them only prune the multi-head attention module and thus compare when the same number of parameters is reduced. See Appendix A for additional information.

5.2 Main Results

5.2.1 Evaluation on DirectShare

Logical and Common Sense Reasoning. In the domain of reasoning, we consider two Chinese natural language inference benchmarks and three English benchmarks: CMNLI (Xu et al., 2020), OCNLI (Hu et al., 2020), along with AX-b, AX-g and RTE from SuperGLUE (Wang et al., 2020).

In Table 1, we show the results on the above five tasks of memory-efficient Llama 2 models. The corresponding results for Baichuan 2 models can be found in Appendix B.1. When applying a 30% parameter sharing to Llama 2-7B, our **DirectShare** can still maintain an average performance of 99.51% across the five benchmarks, compared to the base model. With the same setting, the shared Llama 2-13B retains 99.21% performance. This suggests our finding of head-wise shareable attention for LLMs indeed can work without significant performance degradation in reasoning tasks.

The overall efficacy of our **DirectShare** rivals with the structured pruning results of LLM-Pruner, without any training. Moreover, our method is quite simple and fast, independent on the original training corpus, while structured pruning will nearly fail in the zero-shot generation tasks without dependencies (Ma et al., 2023).

Ratio	Method	CMNLI	OCNLI	AX-b	AX-g	RTE
0%	**Llama 2-7B**	32.98	33.12	53.53	55.34	49.82
10%	Magnitude	32.99	30.63	56.70	49.44	47.29
	LLM-Pruner	32.99	**33.75**	**57.61**	50.00	48.38
	DirectShare	**33.00**	32.50	54.17	**51.97**	**50.90**
30%	Magnitude	33.16	**35.00**	54.71	50.56	46.93
	LLM-Pruner	32.99	31.25	56.34	**52.53**	48.74
	DirectShare	**33.33**	32.50	**57.07**	51.69	**49.10**
0%	**Llama 2-13B**	32.99	35.00	58.81	50.56	47.29
10%	Magnitude	32.82	33.12	51.99	**50.56**	**48.38**
	LLM-Pruner	**32.99**	**36.25**	**58.70**	50.00	46.93
	DirectShare	**32.99**	**36.25**	57.61	50.00	47.29
30%	Magnitude	**33.78**	33.75	46.65	50.00	**51.99**
	LLM-Pruner	32.99	34.38	57.16	**54.21**	45.85
	DirectShare	32.99	**35.00**	**58.33**	50.00	46.57

Table 1: Evaluation Results on Reasoning of the Memory-efficient Llama 2-7B & Llama 2-13B.

Natural Language Understanding (NLU). In this field, we cover multiple tasks, including RACE (Lai et al., 2017) and OpenBookQA (Mihaylov et al., 2018) for reading comprehension, CSL (Li et al., 2022) for content summary and TNEWS (Xu et al., 2020) for content analysis.

Table 2 shows the detailed results of **DirectShare** applied in Llama 2 model family on these benchmarks. We provide a more comprehensive evaluation on Baichuan 2 models in Appendix B.2. Compared to reasoning tasks, our experimental results unveil a notable performance decrease of approximately 30% in large-scale reading comprehension datasets when applying **DirectShare** to Llama 2-7B model. Beyond this, we discover that on content summary and analysis tasks, **DirectShare** manages to retain 94.23% of the performance exhibited by the base model. The evaluation results of Llama 2-13B align with those of Llama 2-7B and we find the accuracy gap is larger as model size increases. This trend also exists in Magnitude Pruning and LLM-Pruner, even the performance drop is larger: LLM-Pruner drops ≈ 3 points more than ours on average while Magnitude Pruning is outperformed by ours by a large margin.

To mitigate this degradation, some post-training pruning methods like SparseGPT (Frantar and Alistarh, 2023) preserves accuracy via the weight update procedure. Similarly, LLM-Pruner uses the low-rank approximation (LoRA, Hu et al., 2021) to post-train the pruned model. Motivated by this, our **PostShare** proves to be beneficial, substantially improving 17.80% accuracy, albeit at a certain amount of time cost. For more details refer to Section 5.2.2. However, this does not diminish the significance of our **DirectShare**. The absence of post-training allows us to better understand the feasibility of head-wise weight sharing for LLMs.

Ratio	Method	RACE-middle	RACE-high	OBQA	CSL	TNEWS
0%	**Llama 2-7B**	33.15	35.51	31.80	55.62	20.22
10%	Magnitude	25.42	26.47	**28.20**	49.38	14.85
	LLM-Pruner	28.20	**30.73**	27.20	53.12	19.76
	DirectShare	**28.34**	28.96	**28.20**	**54.37**	**20.86**
30%	Magnitude	**21.80**	21.53	25.00	45.62	7.01
	LLM-Pruner	21.52	**22.21**	**26.80**	50.00	10.20
	DirectShare	21.45	21.53	26.00	**51.25**	**20.22**
0%	**Llama 2-13B**	60.24	58.03	42.40	58.75	22.13
10%	Magnitude	22.42	21.78	27.40	51.25	15.39
	LLM-Pruner	51.46	50.80	**47.00**	56.25	**20.95**
	DirectShare	**54.04**	**55.63**	39.40	**56.88**	17.94
30%	Magnitude	21.80	22.01	**28.80**	46.25	4.19
	LLM-Pruner	23.96	25.33	26.40	53.75	**16.76**
	DirectShare	**26.53**	**27.53**	27.40	**59.38**	16.12

Table 2: NLU Abilities of the Memory-efficient Models.

Knowledge-related Tasks. We perform evaluations regarding knowledge on various datasets: WinoGrande (Levesque et al., 2012) about language, BoolQ (Clark et al., 2019) testing knowl-

Ratio	Method	WinoGrande	BoolQ	C-Eval	MMLU	RACE-middle	RACE-high	OBQA	OBQA-fact
0%	**Llama 2-7B**	54.04	70.67	32.20	46.69	33.15	35.51	31.8	42.2
30%	DirectShare	50.18	54.43	26.24	26.53	21.45	21.53	26.00	27.60
	PostShare	52.98 ↑ 2.80	66.57 ↑ 12.14	26.38 ↑ 0.14	33.36 ↑ 6.83	29.81 ↑ 8.36	29.45 ↑ 7.92	27.60 ↑ 1.60	33.60 ↑ 6.00

Table 3: Performance of Memory-efficient Llama 2-7B via **PostShare**. See Appendix C for results on Llama 2-13B.

edge question answering, C-Eval (Huang et al., 2023) and MMLU (Hendrycks et al., 2021) standing for two comprehensive examination benchmarks. Table 4 summarizes the mean accuracies on those tasks after **DirectShare** applied to Llama 2 models. See Appendix B.3 for the results based on Baichuan 2 models.

As depicted in Table 4, **DirectShare** takes a clear advantage over other approaches in the field of examination. Our chosen C-Eval and MMLU span diverse disciplines to test both world knowledge and problem solving ability exclusively in a Chinese and English context, respectively. To make this more concrete, Figure 4 vividly contrasts the performance across different subjects based on Llama 2-7B on C-Eval and MMLU. But we have to admit directly do weight sharing across attention heads results in a obvious decline in knowledge-related abilities, which can be solved in **PostShare**.

Ratio	Method	WinoGrande	BoolQ	C-Eval	MMLU
0%	**Llama 2-7B**	54.04	70.67	32.20	46.69
10%	Magnitude	51.58	60.80	22.16	28.20
	LLM-Pruner	**52.98**	66.09	22.31	38.11
	DirectShare	52.63	**67.74**	**28.75**	**43.43**
30%	Magnitude	**50.88**	44.59	24.38	23.15
	LLM-Pruner	**50.88**	**54.77**	22.82	25.16
	DirectShare	50.18	54.43	**26.24**	**26.53**
0%	**Llama 2-13B**	55.44	71.50	40.17	55.81
10%	Magnitude	49.82	62.32	22.52	27.54
	LLM-Pruner	**55.44**	68.07	30.25	51.45
	DirectShare	54.39	**69.45**	**37.17**	**52.81**
30%	Magnitude	49.12	56.45	**23.99**	22.86
	LLM-Pruner	**51.58**	**63.21**	22.17	27.22
	DirectShare	50.18	59.36	22.30	**30.79**

Table 4: Results on Knowledge-related Tasks of the Memory-efficient Models.

5.2.2 Evaluation on PostShare

Based on the evaluation conducted on **DirectShare**, we experiment on **PostShare**, with a special focus on those benchmarks where **DirectShare** experiences a large accuracy degradation.

Table 3 reports how the performance improves with only 0.5 training epoch for Llama 2-7B model. Specifically, in the reading comprehension and

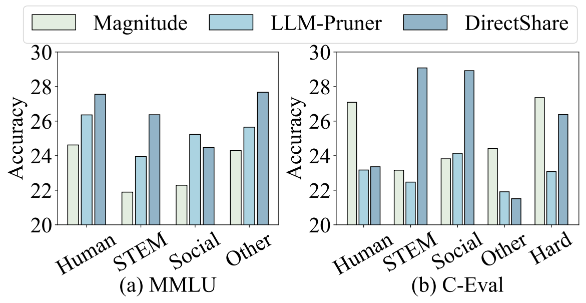


Figure 4: Performance across Different Subjects based on Llama 2-7B on C-Eval and MMLU.

knowledge-related tasks mentioned above, **PostShare** achieves 87.53% of the overall accuracy attained by the original model. Most of the gap between models after **DirectShare** and the original counterparts can be narrowed via **PostShare**, especially in BoolQ and RACE datasets.

Last, it is important to emphasize that here we perform post-training with limited training corpus and thus it runs the risk of overfitting when training only for one epoch. For example, PostShare achieves the higher accuracy in BoolQ at 0.3 epoch than at 0.5 epoch (68.29 vs. 66.57). In contrast, as the training epoch increases from 0.5 to 0.9, the accuracy in WinoGrande rises (52.98 vs. 54.39). It means that due to the domain-constrained corpus, overfitting to one specific dataset will potentially compromise the capabilities in other tasks. The in-depth analysis is provided in Appendix D.1.

5.3 Additional Analysis

Ablation on Impact of Different Head-wise Matching Functions. For weight sharing, the choice of shared heads is critical. In Figure 3, we plot the performance curve on PIQA (Bisk et al., 2019) and OpenBookQA using different head-wise match functions for Baichuan 2-7B model. And the corresponding detailed results are presented in Appendix D.2. Notably, using the cosine similarity between the concatenation matrix of W^q and W^k attains the most favorable outcomes. This may be because it guarantees the maximum similarities between attention maps from the model before and after weight sharing. Also, this choice is much more stable and robust in some tasks like reading

Method Ratio=30%	CMNLI	OCNLI	AX-b	AX-g	RTE	Wino-Grande	BoolQ	C-Eval	MMLU	RACE-middle	RACE-high	OBQA	OBQA-fact	CSL
DirectShare	33.33	32.50	57.07	51.69	49.10	50.18	54.43	26.24	26.53	21.45	21.53	26.00	27.60	51.25
DirectShare + 4bit GPTQ	34.61 ↑1.28	30.63 ↓1.87	57.79 ↑0.72	47.47 ↓4.22	49.82 ↑0.72	49.12 ↓1.06	51.95 ↓2.48	21.88 ↓4.34	25.38 ↓1.15	21.24 ↓0.21	21.33 ↓0.20	23.40 ↓2.60	26.60 ↓1.00	50.00 ↓1.25

Table 5: Weight Sharing and Quantization on Llama 2-7B.

comprehension(e.g., OpenBookQA).

Robustness on the Model Size. In previous experiments, we adopt our approach in LLMs settings. Since small-scale models are not highly overparameterized as large-scale models (Gao et al., 2023), we further verify the effectiveness of our method on smaller models like BERT-base, GPT2-small. For this analysis, we set the sharing ratio from 0% to 50% with a step of 10% for the fine-tuned GPT-small model on WMT-14 En-Fr dataset. As shown in Table 6, at a 50% sharing ratio, the GPT-small can still yield a BLEU score of 39.44 without any post-training. Such kind of variance in performance is acceptable that to some degree proves our method is also suitable for small-scale models.

Sharing Ratio	0%	10%	20%	30%	40%	50%
BLEU	**43.62**	42.49	41.95	41.34	39.96	39.44
Meteor	**42.33**	40.75	40.18	38.43	37.21	36.62

Table 6: Robustness on the Model Size via **PostShare** (Performed on GPT2-small using WMT-14 En-Fr).

Combine Weight Sharing with Quantization. In terms of saving memory consumption, post-quantization employs the strategy of reducing precision in the LLM parameters, while weight sharing aims to reduce the number of parameters. From these two different directions, we suppose integrating weight sharing and quantization may help towards even more memory reduction of LLMs. Hence, we choose GPTQ (Frantar et al., 2022) as a representative and test the effectiveness of applying two techniques in tandem. Specifically, we quantize Llama 2-7B model after 30% **DirectShare** to 4 bit precision. As is reported in Table 5, they can be effectively combined with no more than 5 points performance drop.

Combine PostShare with DirectShare. Another interesting research finding is the combination of our DirectShare and PostShare, where PostShare can play a role in fast performance recovery for DirectShare. Specifically, if we set the sharing ratio to 30% and post-train only 0.5 epoch, the combination based on Llama 2-7B performs on par with the PostShare, as Figure 5 shows. It can also

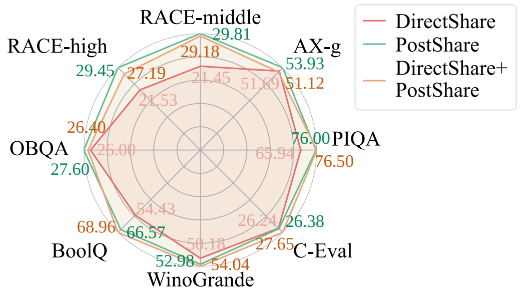


Figure 5: Results in Various Benchmarks via **DirectShare+PostShare** based on Llama 2-7B model.

be seen that DirectShare+PostShare outperforms in some specific datasets like BoolQ and WinoGrande, which we speculate is due to the mitigation of over-fitting problem in PostShare to some extent.

Visualization Study on the Shared Weights. To provide a more detailed explanation of our rationale behind head-wise weight sharing, we conduct a visualization study on the ratios of weight sharing across the MHA layers in two models of different scales (see Appendix D.3). Results indicate the shareable weights distribution across attention heads is similar regardless of the sharing ratio. We also observe a relative balanced sharing ratio across MHA layers than layer-wise weight sharing, which may seem counter-intuitive. However, we find such fine-grained operation on weights has already been used in model pruning (Sun et al., 2023; Ma et al., 2023), constantly superior to layer-wise pruning.

6 Conclusion

In this paper, we illustrate the feasibility of fine-grained weight sharing strategy applied in LLMs, namely, head-wise shareable attention. Consequently, we propose two methods for head-wise weight sharing called **DirectShare** and **PostShare**, which are complementary in terms of time and performance. Our DirectShare concentrates on a simple, no-training yet effective sharing strategy, performing competitively with one of the state-of-the-art model pruning methods. PostShare, on the other hand, shows an impressive performance on keeping LLM's capabilities, needing to compromise on time efficiency. Last, we hope our work inspires researchers to explore better fine-grained weight sharing techniques for memory-efficient LLMs.

Limitations

This paper primarily focuses on the head-wise weight sharing in Multi-Head Attention (MHA) block, inspired by the attention map similarity across heads. However, the Feed-Forward Network (FFN) block has more parameters compared to the MHA block. To further reduce the memory usage for LLMs, there is necessary to investigate the feasibility of applying weight sharing to FFN block. Subsequently, similar to exploration in MHA block, we should determine whether layer-wise weight sharing in FFN block is enough, otherwise fine-grained shared modules are needed to keep more performance. We leave it as future work.

Furthermore, the computing resources limited our ability to conduct experiments on LLMs with a model size of more than 13B. Although we hypothesize that our approach can still work in larger models, which proves to have redundant parameters (Frantar and Alistarh, 2023), it is crucial to validate this hypothesis with further exploration.

References

Haoli Bai, Wei Zhang, Lu Hou, Lifeng Shang, Jing Jin, Xin Jiang, Qun Liu, Michael Lyu, and Irwin King. 2020. Binarybert: Pushing the limit of bert quantization. *arXiv preprint arXiv:2012.15701*.

Baichuan. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Srinadh Bhojanapalli, Ayan Chakrabarti, Andreas Veit, Michal Lukasik, Himanshu Jain, Frederick Liu, Yin-Wen Chang, and Sanjiv Kumar. 2021. Leveraging redundancy in attention with reuse transformers. *arXiv preprint arXiv:2110.06821*.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language.

Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Ale s Tamchyna. 2014. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Yen-Hao Chen, Shang-Wen Li, and Hung-yi Lee. 2021. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 344–350. IEEE.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL*.

OpenCompass Contributors. 2023. Opencompass: A universal evaluation platform for foundation models. `https://github.com/open-compass/opencompass`.

Raj Dabre and Atsushi Fujita. 2019. Recurrent stacking of layers for compact neural machine translation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6292–6299.

Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. 2018. Universal transformers. *arXiv preprint arXiv:1807.03819*.

Tim Dettmers, Ruslan Svirschevski, Vage Egiazarian, Denis Kuznedelev, Elias Frantar, Saleh Ashkboos, Alexander Borzunov, Torsten Hoefler, and Dan Alistarh. 2023. Spqr: A sparse-quantized representation for near-lossless llm weight compression. *arXiv preprint arXiv:2306.03078*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Wikimedia Foundation. Wikimedia downloads.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Ze-Feng Gao, Kun Zhou, Peiyu Liu, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Small pre-trained language models can be fine-tuned as large models via over-parameterization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3819–3834.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *arXiv preprint arXiv:2305.08322*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Hector Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Citeseer.

Yudong Li, Yuqing Zhang, Zhe Zhao, Linlin Shen, Weijie Liu, Weiquan Mao, and Hui Zhang. 2022. Csl: A large-scale chinese scientific literature dataset. *arXiv preprint arXiv:2209.05034*.

Zhuohan Li, Eric Wallace, Sheng Shen, Kevin Lin, Kurt Keutzer, Dan Klein, and Joey Gonzalez. 2020. Train big, then compress: Rethinking model size for efficient training and inference of transformers. In *International Conference on machine learning*, pages 5958–5968. PMLR.

Peiyu Liu, Ze-Feng Gao, Yushuo Chen, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Enhancing scalability of pre-trained language models via efficient parameter sharing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13771–13785.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xiuqing Lv, Peng Zhang, Sunzhu Li, Guobing Gan, and Yueheng Sun. 2023. Lightformer: Light-weight transformer using svd-based weight transfer and parameter sharing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10323–10335.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. *arXiv preprint arXiv:2305.11627*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *EMNLP*.

Machel Reid, Edison Marrese-Taylor, and Yutaka Matsuo. 2021. Subformer: Exploring weight sharing for parameter efficiency in generative transformers. *arXiv preprint arXiv:2101.00234*.

Kyuhong Shim, Jungwook Choi, and Wonyong Sung. 2023. Exploring attention map reuse for efficient transformer neural networks. *arXiv preprint arXiv:2301.12444*.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2023. A simple and effective pruning approach for large language models. *arXiv preprint arXiv:2306.11695*.

Sho Takase and Shun Kiyono. 2021. Lessons on parameter sharing across layers in transformers. *arXiv preprint arXiv:2104.06022*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Shicheng Tan, Weng Lam Tam, Yuanchun Wang, Wenwen Gong, Yang Yang, Hongyin Tang, Keqing He, Jiahao Liu, Jingang Wang, Shu Zhao, et al. 2023. Gkd: A general knowledge distillation framework for large-scale pre-trained language model. *arXiv preprint arXiv:2306.06629*.

Chaofan Tao, Lu Hou, Haoli Bai, Jiansheng Wei, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2023. Structured pruning for efficient generative pre-trained language models. In *Findings of the Association for*

Computational Linguistics: ACL 2023, pages 10880–10895.

Chaofan Tao, Lu Hou, Wei Zhang, Lifeng Shang, Xin Jiang, Qun Liu, Ping Luo, and Ngai Wong. 2022. Compression of generative pre-trained language models via quantization. *arXiv preprint arXiv:2203.10705*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. *arXiv preprint arXiv:1906.05714*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Siyue Wu, Hongzhan Chen, Xiaojun Quan, Qifan Wang, and Rui Wang. 2023. Ad-kd: Attribution-driven knowledge distillation for language model compression. *arXiv preprint arXiv:2305.10010*.

Yingce Xia, Tianyu He, Xu Tan, Fei Tian, Di He, and Tao Qin. 2019. Tied transformers: Neural machine translation with shared encoder and decoder. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5466–5473.

Tong Xiao, Yinqiao Li, Jingbo Zhu, Zhengtao Yu, and Tongran Liu. 2019. Sharing attention weights for fast transformer. *arXiv preprint arXiv:1906.11024*.

Canwen Xu and Julian McAuley. 2023. A survey on model compression and acceleration for pretrained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10566–10575.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, et al. 2020. Clue: A chinese language understanding evaluation benchmark. *arXiv preprint arXiv:2004.05986*.

Ziqing Yang, Yiming Cui, Xin Yao, and Shijin Wang. 2022. Gradient-based intra-attention pruning on pre-trained language models. *arXiv preprint arXiv:2212.07634*.

Chengxuan Ying, Guolin Ke, Di He, and Tie-Yan Liu. 2021. Lazyformer: Self attention with lazy update. *arXiv preprint arXiv:2102.12702*.

Jinnian Zhang, Houwen Peng, Kan Wu, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. 2022. Minivit: Compressing vision transformers with weight multiplexing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12145–12154.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Daquan Zhou, Bingyi Kang, Xiaojie Jin, Linjie Yang, Xiaochen Lian, Zihang Jiang, Qibin Hou, and Jiashi Feng. 2021. Deepvit: Towards deeper vision transformer. *arXiv preprint arXiv:2103.11886*.

Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*.

A Implementation Details

In this section, we will provide additional information about our experimental implementation.

A.1 For the Baseline

To our knowledge, there is no existing baseline for our methods, due to the absence of prior research on fine-grained weight sharing for LLMs. To provide a comprehensive demonstration of the effectiveness of our **DirectShare**, we can only choose another important memory-efficient method of a different category for comparison. Here, we select two model pruning methods applied in LLMs: one classical model pruning method **Magnitude Pruning** and one state-of-the-art structured pruning method **LLM-Pruner**. We do not consider unstructured pruning methods in this paper since they can not achieve real memory reduction without specialized hardware or software.

Based on the results presented in Table 7, 8, 9, it is evident that our **DirectShare** performs on par with one of the prior best structured pruning methods regarding the overall performance, superior to the standard magnitude pruning. Consequently, we claim that designing such a fine-grained (i.e., head-wise) weight sharing strategy with a specific focus on LLMs is indeed simple but effective and this would be a good direction for future work.

A.2 For the Post-training

For carrying out the post-training process, we employ the code framework from LLaMA-Factory repository[1] with DeepSpeed ZeRO-1[2]. The Adam optimizer with a learning rate of 5e-5 is used in our experiment and the parameter values assigned during training are $\beta_1 = 0.9$ and $\beta_2 = 0.95$. For Llama 2-7B model, we set the batch size to 32. While for Llama 2-13B model, the batch size of training is only 8 subject to the limited computational resources. Besides, the maximum context size and γ are set to 4096 and 0.5, respectively.

B Experimental Results based on Baichuan 2 Models

We re-implement **Magnitude Pruning** and **LLM-Pruner** with their public code to accommodate Baichuan 2 models.

B.1 Logical and Common Sense Reasoning

Table 7 presents a comparison on five datasets about reasoning abilities for three memory-efficient methods performed on the Baichuan 2 models.

Our results show that compared to NLU and knowledge-related abilities (listed in Table 8,9), **DirectShare** can indeed maintain its reasoning abilities to a large extent. Specifically, at 30% ratio, DirectShare remains competitive with LLM-Pruner.

Ratio	Method	CMNLI	OCNLI	AX-b	AX-g	RTE
0%	**Baichuan 2-7B**	33.37	41.88	51.90	50.28	57.40
10%	Magnitude	33.11	33.12	**55.62**	50.84	55.96
	LLM-Pruner	**37.31**	40.62	49.18	50.00	60.65
	DirectShare	33.00	**41.25**	49.55	**51.12**	60.29
30%	Magnitude	32.97	31.25	48.28	**51.97**	46.57
	LLM-Pruner	**34.20**	**34.38**	47.55	50.84	**51.26**
	DirectShare	32.97	30.63	**54.71**	51.69	49.82
0%	**Baichuan 2-13B**	33.21	40.62	59.69	50.59	44.77
10%	Magnitude	33.21	31.25	55.62	48.60	46.93
	LLM-Pruner	**33.66**	36.88	**58.51**	49.72	47.65
	DirectShare	33.23	**40.00**	53.71	**53.37**	**53.07**
30%	Magnitude	**33.21**	30.00	50.91	48.03	43.32
	LLM-Pruner	33.04	**36.88**	**55.71**	**50.28**	44.04
	DirectShare	33.11	30.00	54.98	50.00	**45.13**

Table 7: Evaluation Results on Reasoning of the Memory-efficient Baichuan 2-7B & Baichuan 2-13B.

B.2 Natural Language Understanding

Table 8 presents the performance for each NLU task discussed in Section 5.2.1 when applying **DirectShare** to Baichuan 2 models. Consistent with the experiments on Llama 2-7B and Llama 2-13B models, similar performance drop exists. Thus, at the cost of post-training time, our PostShare can narrow the gap observed across the majority of datasets. With regard to individual datasets, it remains to be seen if the gap can be largely recovered given the best training epoch[3].

Ratio	Method	RACE-middle	RACE-high	OBQA	CSL	TNEWS
0%	**Baichuan 2-7B**	51.04	52.63	32.20	66.25	28.60
10%	Magnitude	24.37	28.13	30.20	57.50	**27.60**
	LLM-Pruner	25.42	35.36	**32.60**	61.25	26.05
	DirectShare	**50.49**	**48.46**	28.20	**63.75**	26.23
30%	Magnitude	21.80	21.67	**27.60**	**57.50**	13.66
	LLM-Pruner	22.56	22.67	27.40	53.12	**21.31**
	DirectShare	**25.14**	**23.44**	**27.60**	52.50	18.40
0%	**Baichuan 2-13B**	68.94	67.27	42.20	63.12	28.96
10%	Magnitude	25.56	26.33	26.20	45.62	11.38
	LLM-Pruner	41.71	46.80	**32.40**	62.50	**29.23**
	DirectShare	**47.56**	**49.34**	31.20	**64.38**	22.22
30%	Magnitude	**24.58**	**24.58**	25.40	50.62	6.65
	LLM-Pruner	22.63	21.81	**26.80**	**55.00**	**24.13**
	DirectShare	22.14	23.99	26.60	53.13	17.58

Table 8: NLU Abilities of the Memory-efficient Models.

B.3 Knowledge-related Tasks

The results of Baichuan 2 models on knowledge-related tasks are shown in Table 9. Similar decline appears in Llama 2-7B and Llama 2-13B models as well.

Ratio	Method	WinoGrande	BoolQ	C-Eval	MMLU
0%	**Baichuan 2-7B**	54.04	63.30	56.19	54.65
10%	Magnitude	50.18	57.06	34.70	45.47
	LLM-Pruner	50.53	**59.30**	48.14	**51.78**
	DirectShare	**51.58**	58.01	**50.41**	49.96
30%	Magnitude	49.12	**55.41**	**23.91**	24.36
	LLM-Pruner	51.23	48.93	22.11	**25.62**
	DirectShare	**51.58**	51.53	21.86	24.05
0%	**Baichuan 2-13B**	56.14	67.00	59.21	59.58
10%	Magnitude	50.53	40.55	25.22	25.55
	LLM-Pruner	51.23	**65.87**	49.60	51.49
	DirectShare	**53.33**	61.04	**53.65**	**52.60**
30%	Magnitude	50.18	50.09	**25.35**	24.66
	LLM-Pruner	**50.53**	**59.42**	21.09	**24.95**
	DirectShare	48.77	40.83	23.25	24.82

Table 9: Results on Knowledge-related Tasks of the Memory-efficient Models.

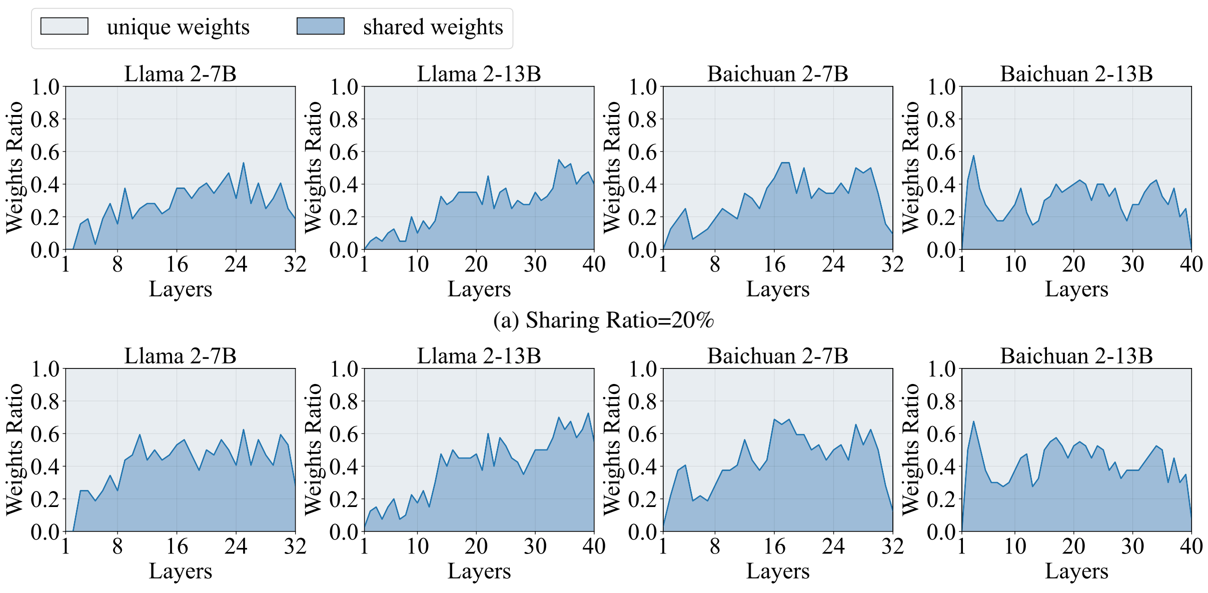
[1] https://github.com/hiyouga/LLaMA-Factory

[2] Because of our designed special loss function in the post-training stage, only DeepSpeed ZeRO-1 can work.

[3] We speculate that it may be attributed to overfitting issue. Furthermore, as the model size increases, it becomes increasingly difficult to determine the optimal training epoch for effectively mitigating overfitting.

Ratio	Method	WinoGrande	BoolQ	C-Eval	MMLU	RACE-middle	RACE-high	OBQA	OBQA-fact
0%	**Llama 2-13B**	55.44	71.50	40.17	55.81	60.24	58.03	42.40	60.00
30%	DirectShare	50.18	59.36	22.30	30.79	26.53	27.53	27.40	27.80
	PostShare*	53.68 ↑ 3.50	71.25 ↑ 11.89	25.80 ↑ 3.50	33.90 ↑ 3.11	32.03 ↑ 3.30	29.07 ↑ 1.54	33.60 ↑ 6.20	38.80 ↑ 11.00

Table 10: Performance of the Memory-efficient Llama 2-13B via **PostShare**. * means choosing relatively good performance across different training steps.



(a) Sharing Ratio=20%

(b) Sharing Ratio=30%

Figure 6: Ratios of Weight Sharing across the MHA Layers in LLaMA2-7B/13B & Baichuan2-7B/13B.

C PostShare on Llama 2-13B Model

In addition to Llama 2-7B, we also experiment with Llama 2-13B to evaluate **PostShare** (See Table 10). Compared to Llama 2-7B, the best training epoch on Llama 2-13B is much smaller: approximately hundreds of training steps is enough, otherwise it may suffer from overfitting issue. However, the overfitting problem seems to be obvious as model size increases, resulting in the challenge with regard to choosing the best training epoch.

D More Analysis

D.1 Overfitting Phenomenon in PostShare

Figure 7 shows the performance curves on different kinds of datasets across various post-training steps. Remarkably, our **PostShare** requires no more than 1 epoch that can push the selected weights closer for sharing while keeping the performance. However, we observe the slight overfitting phenomenon in **PostShare**, i.e., the capabilities initially improve and then experience a slight decline. Besides, it is clear that the turning point about performance varies with datasets. Detailed statistical data are provided in Table 11.

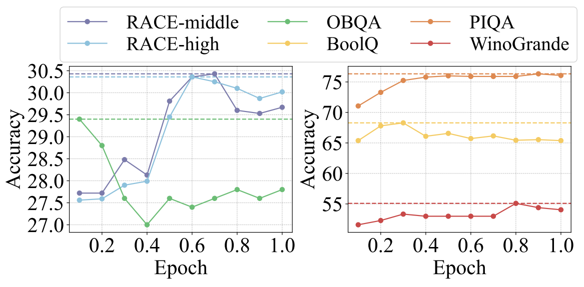


Figure 7: Accuracy across Different Training Steps during **PostShare**.

Epoch	RACE-middle	RACE-high	OBQA	BoolQ	PIQA	Wino-Grande
0.10	27.72	27.56	**29.40**	65.38	71.06	51.58
0.20	27.72	27.59	28.80	67.80	73.29	52.28
0.30	28.48	27.90	27.60	**68.29**	75.24	53.33
0.40	28.13	27.99	27.00	66.09	75.79	52.98
0.50	29.81	29.45	27.60	66.57	76.00	52.98
0.60	30.36	**30.36**	27.40	65.72	75.90	52.98
0.70	**30.43**	30.25	27.60	66.15	75.90	52.98
0.80	29.60	30.10	27.80	65.44	75.90	**55.09**
0.90	29.53	29.87	27.60	65.54	**76.33**	54.39
1.00	29.67	30.02	27.80	65.38	76.06	54.04

Table 11: Accuracy across Different Training Steps during **PostShare**.

Sharing Ratio	5%		10%		15%		20%		25%		30%		35%	40%
Dataset	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	OBQA	OBQA
W^q	74.92	29.2	74.97	27.5	73.29	27.8	70.89	27.7	64.64	27.5	58.43	25.5	24.4	25.7
W^k	74.92	28.7	74.27	27.6	71.71	27.7	70.35	27.6	68.77	**27.6**	64.36	27.2	27.6	26.9
W^v	74.92	28.1	74.48	27.7	73.29	26.7	70.46	**28.5**	68.39	25.6	60.17	23.1	23.9	22.5
W^q, W^k, W^v	71.71	27.6	63.55	27.8	54.03	26.8	50.16	24.5	51.41	25.5	51.09	25.5	**29.0**	25.3
$W^q\|\|W^k\|\|W^v$	74.59	**34.7**	74.59	**30.0**	73.45	**30.3**	70.73	28.2	66.59	**27.6**	63.33	27.6	27.1	25.0
$W^q\|\|W^k(Ours)$	**75.84**	33.9	**75.30**	28.2	**74.54**	27.5	**73.01**	27.3	**69.37**	27.5	**65.56**	**28.0**	27.6	**28.6**

Table 12: Results on PIQA and OBQA with Different Head-wise Matching Functions for Baichuan 2-7B model.

D.2 Impact of Different Head-wise Matching Functions

The selection of shared heads plays a crucial role in weight sharing. An ablation experiment for this is shown in Table 12.

D.3 Visualization Study on the Shared Weights

As depicted in Figure 6, the distribution of ratios of shared weights across attention heads is similar regardless of the sharing ratio.

E Generations from Models after Weight Sharing

Table 13, 14 and 15 show more examples of the models after head-wise weight sharing. We present the generation results of both the shared model via DirectShare and PostShare.

Model	**Generated Responses**
Llama 2-7B w/ DirectShare	The future is going to be one that presents many challenges, but also a great deal of opportunity. Everybody knows this, and it's why we have been working so hard on the newest version of the future. The problem with the future is that it's not always easy to see what you are doing wrong. It can be difficult to find
	Neural Sequential Model, especially transformers, are a powerful tool for the analysis of time-series data. surely, they can be used to predict the future. Their use is not limited to the financial industry. In this article, we will discuss the benefits and limitations of the
Llama 2-7B w/ PostShare	The future is going to be one that presents many challenges, and the world will need a new generation of leaders. The United Nations has identified the following as key issues for the future: Climate change, Food security, Energy security, Water scarcity, Global health, Human rights, International law, Peacekeeping, Sustainable
	Neural Sequential Model, especially transformers, is a type of artificial neural network that uses a sequence of vectors to represent the input and output. The model was introduced by Geoffrey Hinton in 1987. It has been used for speech recognition, handwriting recognition,

Table 13: Responses from Llama 2-7B (Sharing Ratio=30%).

Model	**Generated Responses**
Baichuan 2-7B w/ DirectShare (English)	There are many reasons to support why perseverance is of great importance. The first and foremost, it's the only way you can get better at what ever your doing! It also gives us an opportunity to learn from our mistakes or
	Her eyes shone with happiness. "I'm so happy to see you," She said, "and I am very proud of the work that we have done together in this project and it is a great honor for me as an artist because there are many people who has been working on these projects over the years but not all of them were able to do what was needed from us at any given time when things got out of hand or if they had lost their way"
Baichuan 2-7B w/ DirectShare (Chinese)	如何健康的生活 1. 注意休息,保证睡眠。2.合理饮食:多吃水果、蔬菜等。3.适当运动(如散步) 4.保持乐观心态5.多饮水6.避免劳累7.不要熬夜8.劳逸结合9.注意安全10.尊重生命
	想要成功需要把握什么?1. 首先,要有一个好的心态。2.其次,要有一定的经济基础3.接着就是自己的实力了4. 最后一个条件是:自己一定要努力!

Table 14: Responses from Baichuan 2-7B (Sharing Ratio=30%).

Model	**Generated Responses**	**Reference**
GPT2-small w/ PostShare	In the letter, Mr Snowden said he believed the support of the international community could persuade the US government to abandon criminal charges against him.<to-fr> Dans la lettre, M. le député a dit que l'appui de la communauté internationale pourrait convaincre le gouvernement américain de abandonner la charge pénale contre lui.	Dans sa lettre, M. Snowden a écrit qu'il pensait que le soutien de la communauté internationale pourrait persuader le gouvernement américain d'abandonner les charges pénales retenues contre lui.
	The letter extends an offer to cooperate with German authorities "when the difficulties of this humanitarian situation have been resolved." <to-fr> La lettre élargit une invitation à coopérer avec les autorités allemands lorsque les difficultés de cette situation humanitaire sont réglées.	La lettre avance une offre de coopération avec les autorités allemandes « lorsque les difficultés de cette situation humanitaire auront été résolues ».
	The first test plane was unveiled in March and took flight for the first time in September after months of delays.<to-fr> Le premier étudiant a été démontré en mars et a fait l'avion pour la première fois après des mois de retard.	Le premier avion d'essai a été dévoilé en mars et s'est envolé pour la première fois en septembre après des mois de retard.

Table 15: Responses from GPT2-small (Sharing Ratio=30%).