From Layers to Latents: Pruning and Aligning LLMs for Efficiency and Safety

Kushal Pal Singh
2025BSZ8215
bsz258215@iitd.ac.in

Abstract

This is project project report for project 3, titled "From Layers to Latents: Pruning and Aligning LLMs for Efficiency and Safety". This project's objective is a. compression of models using static and dynamic pruning and b. alignment of models to safety alignment by identifying and mitigating harmful latent directions inside the model's internal representation space.

For pruning, entire transformer layers are removed systematically and also magnitude-based weight pruning is applied to Qwen3-0.6B model. The analyses includes study of how model accuracy and perplexity degrade as pruning increases. Dynamic pruning is also implemented for attention-head pruning via a learned router network with a load-balancing loss, based on MoH. This experiment is done on Llama1.1 model.

For alignment, a mixed dataset of neutral, biased, and toxic prompts is created from 2 differnt sources and is used to train layer-wise probes on GPT-2 Large. Probe accuracy reveals the "most harmful" layer. The harmful direction from activations is extracted and activation editing at inference time is applied to suppress toxic behaviour while preserving performance.

1 Introduction

Though the Transformer-based language models exhibit strong performance but are computationally expensive and amplify harmful biases. The project explores two directions to handle this:

a. Efficiency - Static Pruning using layer removal (structured) & weights removal (unstructured) and Dynamic Pruning of Attention Heads by investigating the importance of layers, weights and attention heads to decide the candidates for pruning.

b. Safety - aligning a model towards safety by identifying and reducing biases present in a language model and aligning it towards safer behavior.

The detailed report on each task is given in the following sections.

2 Task 1: Investigating the Importance of Layers and Attention Heads

This task involved 3 activities:

a. Static Pruning using layer removal (structured),

b. Weights removal (unstructed), and,

c. Dynamic Pruning of Attention Heads in MoE style.

2.1 Dataset

All the experiments for this task have been performed on MMLU[(Dan Hendrycks and Steinhardt, 2021)] and GSM8K [(Cobbe et al., 2021)] datasets. a subset of both these datasets is used for evaluation specifically 451 examples from MMLU and 131 examples from GSM8K have been taken.

2.2 Layer Removal

This experiment has been performed on Qwen3-0.6B [(Qwen, 2025)] model. Each transformer layer of the model is removed one by and its impact on model accuracy and perplexity is observed. The layer is removed by simply passing the input to the layer as output to the next layer. The impact of removal of different layers on accuracy and perplexity is shown in Fig. 1. It is observed that removal of initial layers has catastrophic effect both on accuracy and perplexity. The accuracy of model drops significantly on both the dataset. Same effect is seen on perplexity also. However, there is not much impact on performance and perplexity if later layers (4 onwards) are removed. In fact, as expected there is a positive impact of layer removal and the performance of the model improves.

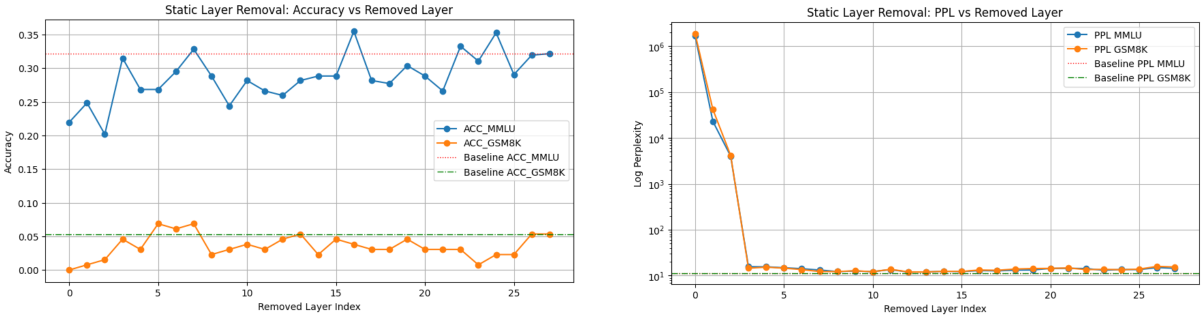

Figure 1: Effect of removing different layers from the model on (a) accuracy and (b) perplexity

2.3 Weight Removal

This experiment has also been performed on Qwen3-0.6B [(Qwen, 2025)] model. In this, rather than removing the complete layer, individual weights are removed. Given a parameter p, the percentage of parametes to be removed, the top $p\%$ minimum magnitude weights are selected and pruned by multiplying them with *zero*. In this experiment, the effect of different levels of pruning (i.e. different values of p) is studied. The experiment is done with $p = 5\%$, 10%, 20%, 30%, 40% and 50%. The effect of removal of diffrent percentage of weights is shown in Fig. 2. The experiments show that a small percentage (upto 20%) weight removal has positive or close to neutral effect on the accuracy of the model, keeping perplexity close to intact. But as the percentage of pruned weights is increased, there is adverse effect, both on accuracy and perplexity.

2.4 Dynamic Pruning

This experiment has been performed on TinyLlama-1.1B-v1.1 [(Zhang et al., 2024)] model as with experiments on Qwen3-0.6B model was giving NaN loss always. Importance weights are shown in Fig. 3.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Steven Basart Andy Zou Mantas Mazeika Dawn Song Dan Hendrycks, Collin Burns and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*, pages 1–2.

Qwen. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024. Tinyllama: An open-source small language model. *ArXiv*, abs/2401.02385.

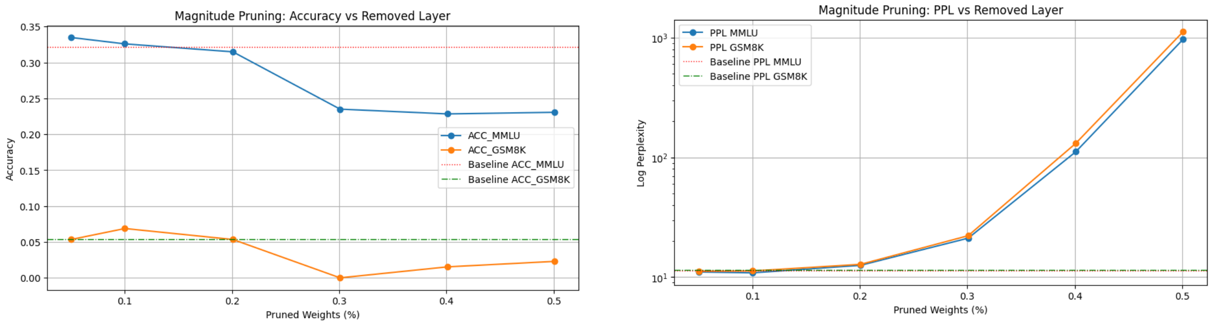

Figure 2: Effect of removing least significant weights from the model on (a) accuracy and (b) perplexity

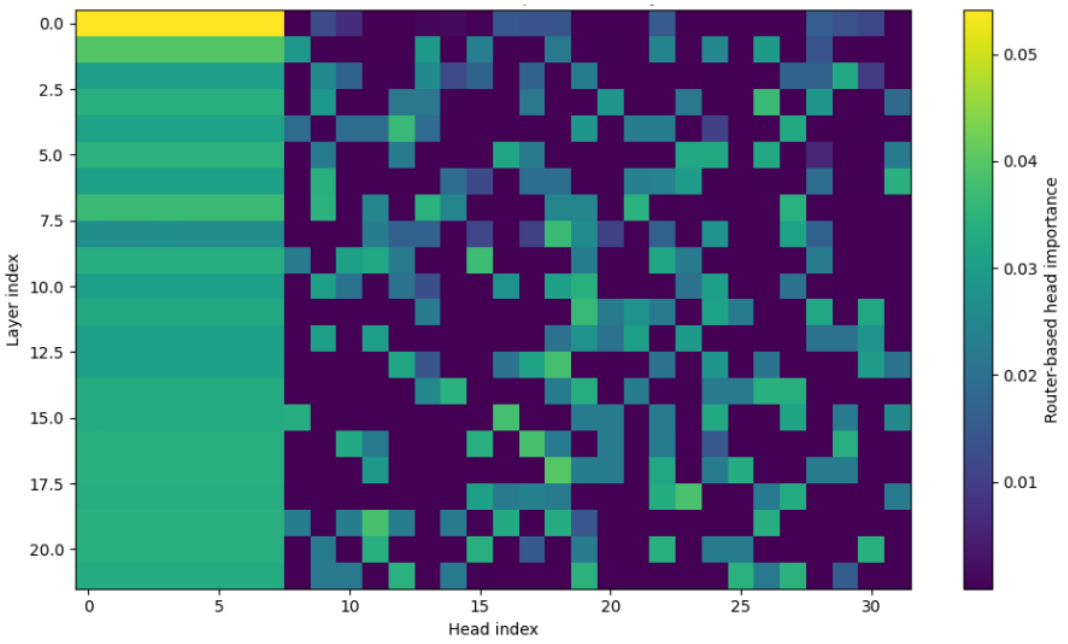

Figure 3: Importance of different heads of different layers.