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

Student Angom Umakanta Singh
Entry No.: 2025EEZ8518
Course: ELL8299

Experiments completed

1. Layer Wise static pruning
2. Percentage based static pruning

1 Layer Wise static pruning

Qwen/Qwen3-0.6B is used as a model to perform the pruning. The model has 28 transformer layers. After pruning each layer, the pruned model is evaluated on 10% of the MMLU and GSM8K dataset. The result of the experiment is as given in Table 1.

The experiment was carried out with python 3.10.16 and Nvidia RTX5000 Quadro with 7 CPUs, 32 GB RAM and 16 GB VRAM.

1.1 Experimental Results

1.1.1 Plots

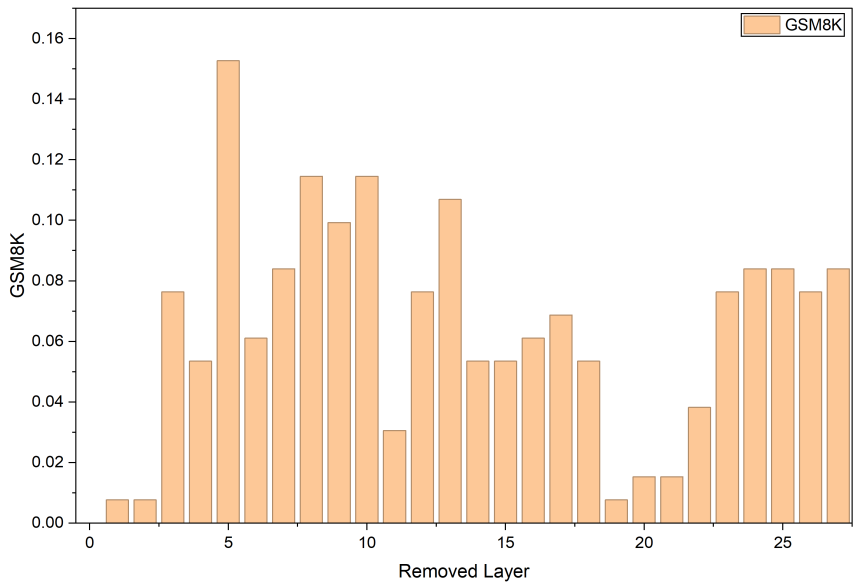

Figure 1: GSM8K performance plot

Removed Layer	MMLU Accuracy	GSM8K Accuracy
0	0	0.00
1	0	0.01
2	0	0.01
3	0	0.08
4	0	0.05
5	0	0.15
6	0	0.06
7	0	0.08
8	0	0.11
9	0	0.10
10	0	0.11
11	0	0.03
12	0	0.08
13	0	0.08
14	0	0.11
15	0	0.05
16	0	0.06
17	0	0.07
18	0	0.05
19	0	0.01
20	0	0.02
21	0	0.02
22	0	0.04
23	0	0.08
24	0	0.08
25	0	0.08
26	0	0.08
27	0	0.08

Table 1: Layerwise pruning performance

2 Percentage based model static pruning

Qwen/Qwen3-0.6B is used as a model to perform the pruning. The model weight is pruned in ascending order of magnitude. The pruned model is evaluated on 10% of the MMLU and GSM8K data set. The result of the experiment is as given in Table 2.

The experiment was carried out with python 3.10.16 and Nvidia RTX5000 Quadro with 7 CPUs, 32 GB RAM and 16 GB VRAM.

2.1 Experimental Results

Pruning percentage	MMLU Accuracy	GSM8K Accuracy
0	0	0.08
10	0	0.08
20	0	0.05
30	0	0.00
40	0	0.03
50	0	0.00
60	0	0.01
70	0	0.01
80	0	0.01
90	0	0.01

Table 2: Result of Percentage based model static pruning

2.1.1 Plots

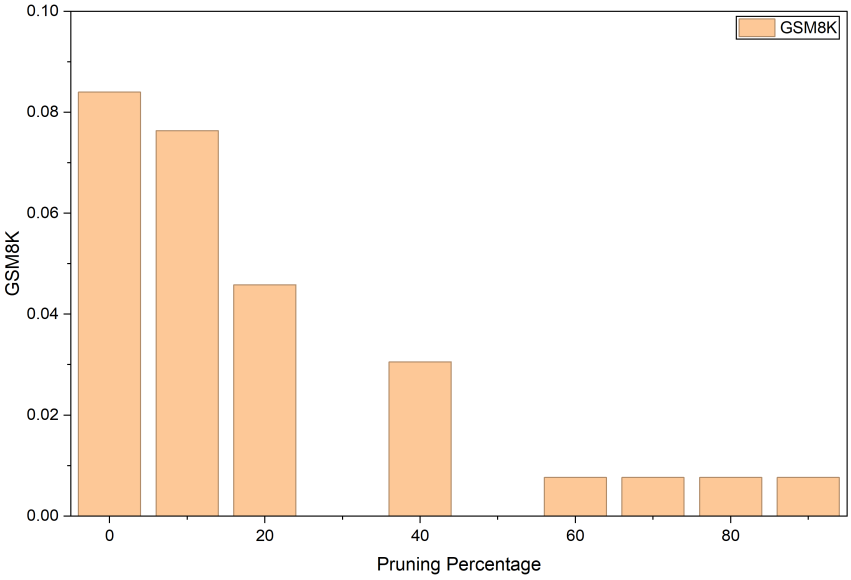

Figure 2: GSM8K performance plot

Observation

Layerwise pruning

It is observed as shown in the accuracy plot 1 that the performance is worst for layer 0 pruning followed by layers 1,2 and 19. Hence, the said layers may be considered as important for performing mathematical reasoning task.

Magnitude pruning

It is observed as shown in the accuracy plot 2 that the performance is worst at 30% and 50% followed by 60% to 90%. Hence, even when most of the weights are pruned, the model can still perform some mathematical reasoning task.