# Readme

A close-up of a chart

Description automatically generated

## A diagram of a network Description automatically generated

## Pruning Mistral 7B and LLama 7B Models

python main2.py \

--model mistralai/Mistral-7B-Instruct-v0.1 \

--prune\_method wanda \

--sparsity\_ratio 0.75 \

--sparsity\_type unstructured \

--save out/llama\_7b/unstructured/wanda/ \

--save\_model /mnt/parscratch/users/lip23pdr/mistral\_75\_pruned

python main2.py \

--model "meta-llama/Llama-2-7b-hf" \

--prune\_method wanda \

--sparsity\_ratio 0.25 \

--sparsity\_type unstructured \

--save out/llama\_7b/unstructured/wanda/ \

--save\_model /mnt/parscratch/users/lip23pdr/llama2\_25\_pruned

* The above commands were used to prune the Mistral 7B and LLama 7B models.
* Change the sparsity accordingly to get the desired pruned models.
* Models were then saved which were then further used to evaluate on CommonsenseQA datasets and ARC Challenge Dataset.
* You can find the .ipynb files of every sparsity ratio, model and dataset attached as well.
* Pruned models where not commited in this github it takes more than 200GB of Data Storage Github doesn’t allow.
* Use the above commands and tweak the paramaters and add tokenisers according while using the model from huggingface.