Lab4 Sketch

- We want to understand the sparsity structures in trained Transformer models and measure performance before and after sparsification
- General approach:
  - Work with pre-trained LLM model
  - Devise approaches to assess sparsity structure in trained model
  - (what fraction of parameters are >> 0, and how does this vary by layer)
  - Produce a sparsified version of the model and compare performance to the baseline model on selected benchmarks
  - Explore reductions at the 10%, 50%, 90%, 95%, 99%
  - Compare size of models and runtime for sparsified models
  - Explain the challenges of scarification on LLMs