WHAT TO EXPECT WHEN PRETRAINING AN LLM

Note: This information is based solely on the pretraining of a model in this repo:

https://github.com/logic-OT/Decoder-Only-LLM

The model size is about 37M parameters

1. Number of epochs: 20 to 30 epochs should be enough for your model to capture semantic relationships between tokens. For the model above, i used 24 epochs.

2.Learning rate: A leaning rate of 10^{-4} (0.0001) is okay. But you ca start with 10^{-3} and move to 10^{-4} around 10 epochs. Better still, use a shedule to dynamically adjust the learning rate as the model trains.

3. Predictions: At the beginning of the training, you would realise that the model predicts the most common or recurring token. So you would see tokens like "the" and "and" alot. But by the 10th epoch, it would start understanding language semantics.

4. Batch size: I mean batch size depends on your compute. But for P100 the max is 50. Also, it makes sense to start with a large batch size and reduce the batch size at later epochs.

5. Training Time: Training on P100 is really slow, it take about 4 hours to finish 1 epoch. So be patient. You can 2 epochs a day.