AI and the Enterprise

WISCONSIN SCHOOL OF BUSINESS

Assignment 2: GPT Parameter Tuning

The purpose of this assignment is to adjust parameters of the provided language model and train to observe the impact on **training loss** and **validation loss**.

Tasks

1. Parameter Tuning for Model Training

- Adjust parameters in the *LLM tuning.ipynb* code to loop and compare tuning performance.
- Track the **training loss** and **validation loss** at regular intervals.

Metric Visualization

- Use the *MetricsTracker* class to store and visualize training and validation losses.
- Compare the impact of each parameter across runs with the training time.

Parameters

- **Batch size**: 24, 64, 256
- **Block size**: 16, 64, 128
- **Learning rate**: 1e-1, 1e-3, 1e-4
- **Layers**: 2, 4, 8
- **Heads**: 2, 8, 16
- **Dropout**: 0.1, 0.2, 0.3

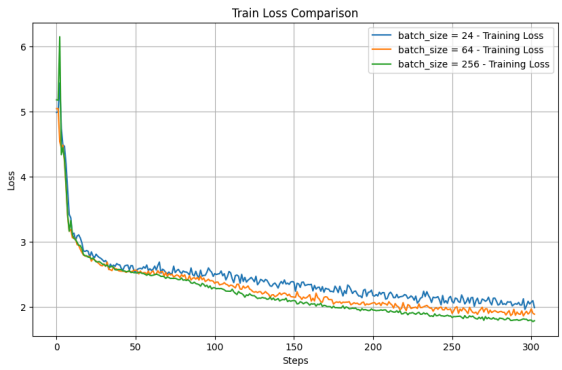
Deliverable

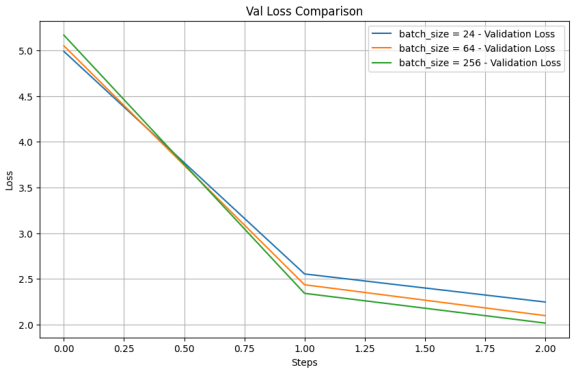
1. **Loss Comparison**: Create line charts to show training and validation loss across different hyperparameter settings.
2. **Performance Summary**: Describe the impact of changing each hyperparameter on model performance. Identify which configurations resulted in the **lowest training and validation loss**.
3. What were the key observations about how each parameter influenced the loss?
4. Which combination of parameters worked best?
5. What challenges or patterns did you notice during the experiments?

Batch Size: 24 | 64 | 256

Small (24) noisiest loss; medium (64) balances stability and speed; large (256) minimizes loss with slowest training time.

Training Time
Batch Size 24: 21.93 seconds
Batch Size 64: 25.93 seconds
Batch Size 256: 83.51 seconds

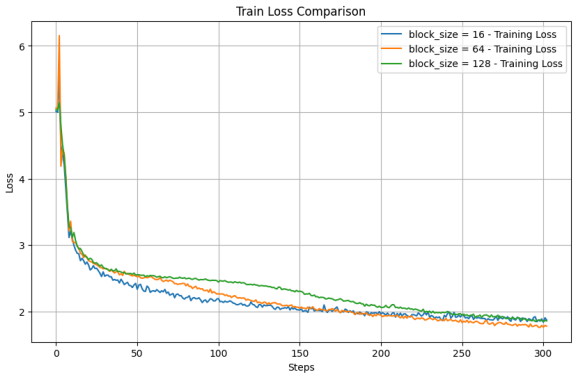


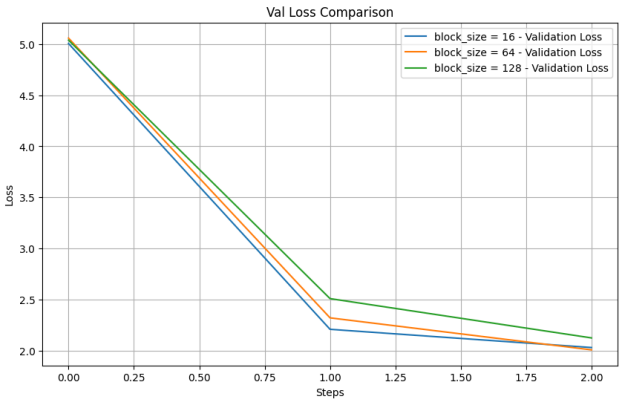


Block Size: 16 | 64 | 128

Medium (64) provides lowest loss over time with training efficiency. Small (16) is fastest and yields early performance but degrades over time. Large (128) increases cost without corresponding benefit.

Training Time
Block Size 16: 21.31 seconds
Block Size 64: 37.05 seconds
Block Size 128: 73.88 seconds



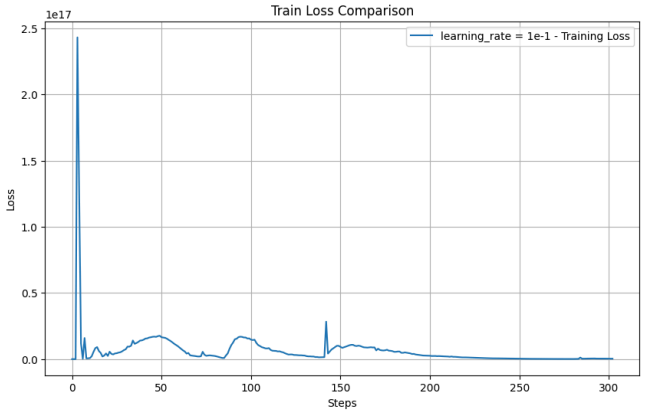


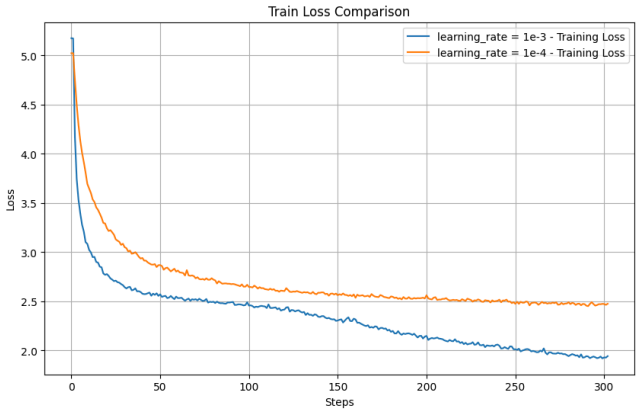
Learning Rate: 1e-1 | 1e-3 | 1e-4:

Higher learning rate (1e-4) significantly outperforms other options and has the lowest cost. Low learning rate (1e-1) causes instability and possible early stoppage.

Training Time

1e-1 Learning Rate: 72.56 seconds
1e-3 Learning Rate: 73.54 seconds
1e-4 Learning Rate: 73.77 seconds



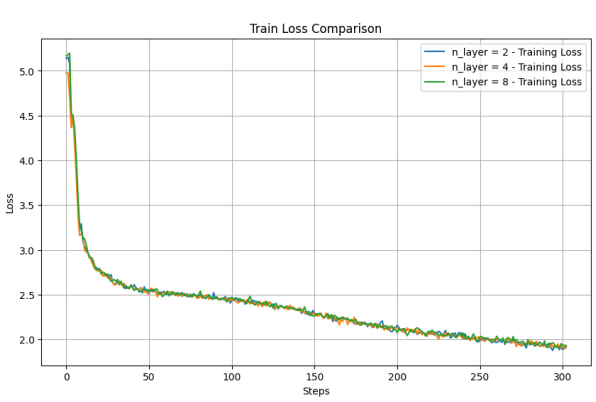


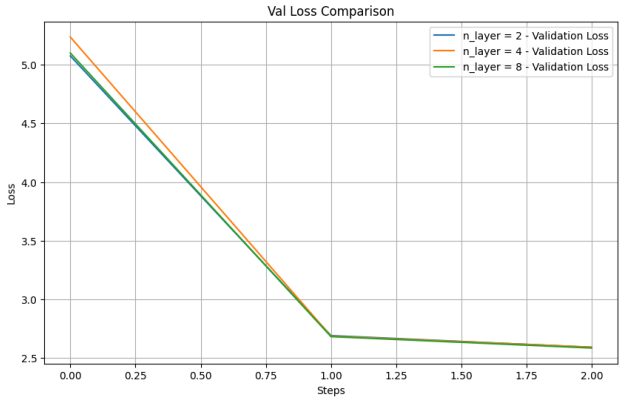
Layers: 2 | 4 | 8

Training Time

2 Layers: 134.80 seconds
4 Layers: 134.65 seconds
8 Layers: 134.76 seconds

Not much difference

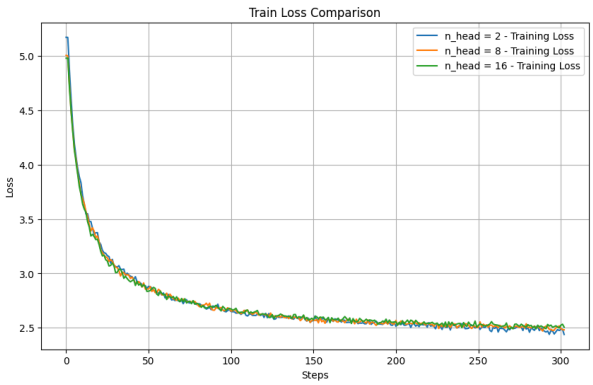


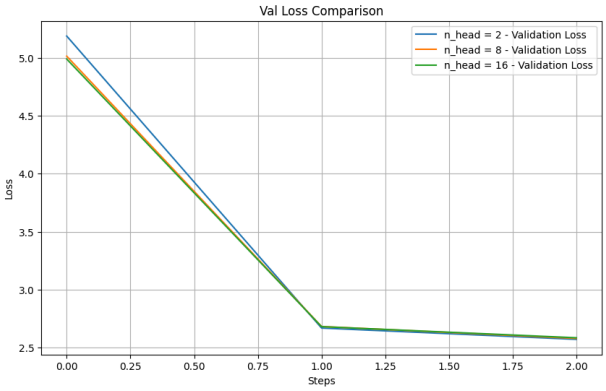


Heads: 2 | 8 | 16

Training Time
2 Heads: 65.82 seconds
8 Heads: 90.40 seconds
16 Heads: 134.43 seconds

All options have similar loss outcomes with 2 heads having lowest compute time.





Dropout: 0.1 | 0.2 | 0.3

Training Time
2 Layers: 135.05 seconds
4 Layers: 134.50 seconds
8 Layers: 134.62 seconds

Not much difference in training or validation loss