1. Dataset & Preprocessing

For this assignment, I used *An Inquiry into the Nature and Causes of the Wealth of Nations* by Adam Smith, sourced from Project Gutenberg. This text was cleaned to remove license text, normalize spacing, and standardize punctuation. The processed corpus contained **441,551 tokens** and **10,359 unique words**.

I applied **word-level tokenization** and introduced four special tokens: <pad>, <unk>, <bos>, and <eos> for padding, unknown words, and sequence boundaries. The text was split into **80% train, 10% validation, and 10% test**, preserving order to reflect real sequential prediction.

In early experiments, I restricted the token count to reduce runtime, but this led to unmeaningful ablation results because the reduced dataset lacked sufficient variation for robust parameter evaluation. After that, I trained on the full dataset, enabling **more reliable insights** into model performance and stability.

2. Models & Training Setup

I implemented two architectures: a **vanilla RNN**, and an **LSTM** representing different stages of sequence modeling.

- **RNN**: Processes tokens sequentially with a hidden state. It is simple and fast but struggles with long-term dependencies due to vanishing gradients.
- **LSTM**: Enhances RNNs with memory cells and gating (input, forget, output), enabling longer-range modeling and mitigating gradient issues.

Training setup: Models were trained in PyTorch with mini-batch optimization, cross-entropy loss, and perplexity as the evaluation metric. Input sequences had length 128, and the training used the following hyperparameters:

- Batch size: 64
- Embedding size: 128
- Hidden size: 256
- Layers: 2
- Dropout: 0.2
- Learning rate: 0.001 (AdamW)
- Gradient clipping: 1.0
- Max epochs: 10 (with early stopping, patience=2)

Early Stopping:
Early stopping was applied based on validation perplexity with a patience of 2 epochs. This prevented overfitting and improved efficiency. I decided to use early stopping after observing that both RNN and LSTM tended to diverge when trained for the full 10 epochs: although the training loss continued decreasing, validation perplexity increased exponentially. This indicated overfitting and vanishing gradient issues, making early stopping a necessary safeguard.

```
rnn | epoch 01 | val PPL 383.95 | time 201.7s
rnn | epoch 10 | val PPL 1513.54 | time 199.0s
lstm | epoch 01 | val PPL 2610.04 | time 303.5s
lstm | epoch 10 | val PPL 59763.24 | time 303.1s
```

All experiments were run on GPU (Tesla T4) in Google Colab, with runtime tracked per model for comparison.

3. Results & Evaluation

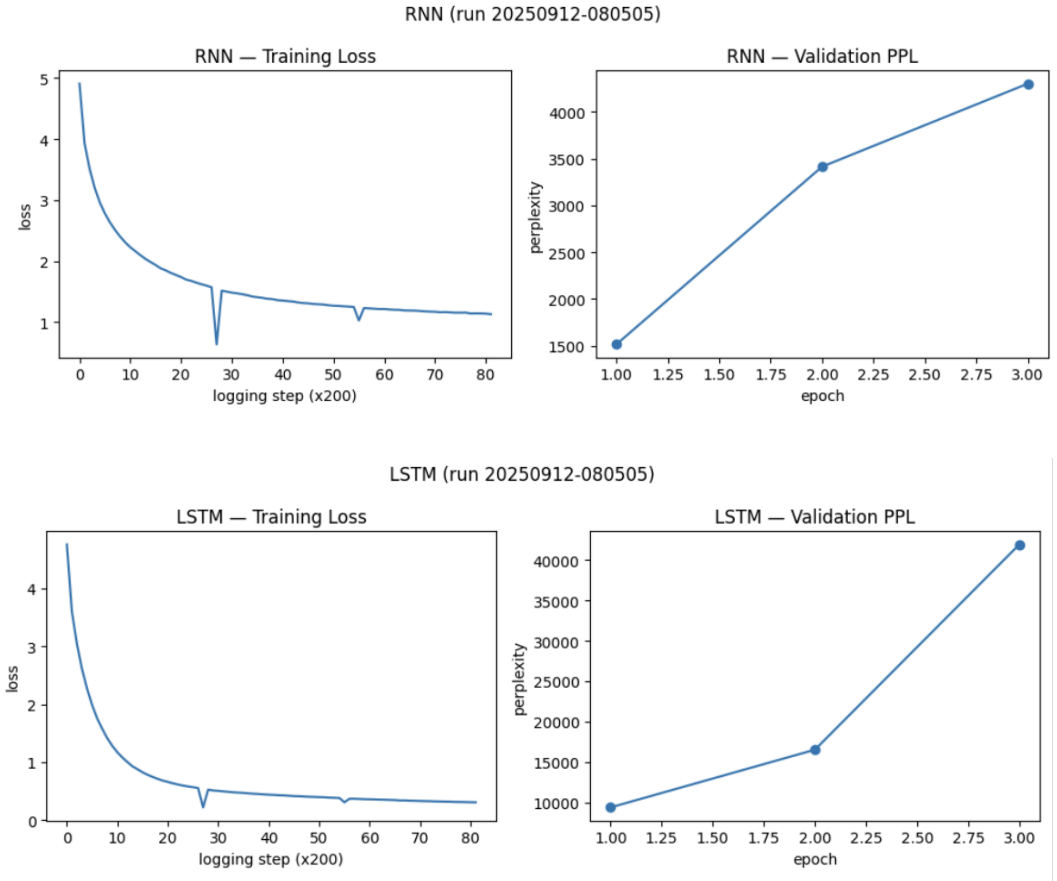
3.1 Validation and Test Perplexity

To evaluate the quantitative performance of the models, I measured validation and test perplexity (PPL). Perplexity is a standard metric for language modeling, reflecting how well a model predicts the next token in a sequence. Lower perplexity indicates better predictive performance and greater confidence in the model's output distribution.

- *RNN | val PPL: 1515.10 | test PPL: 1065.74*
- *LSTM | val PPL: 9376.26 | test PPL: 6937.49*

Surprisingly, the RNN outperformed the LSTM. Although LSTMs usually handle long-range dependencies better, **here the dataset is relatively small and comes from a single book**. The added complexity of the LSTM likely caused overfitting and unstable convergence, while the simpler RNN generalized more effectively.

3.2 **Training and Validation Curves**



From 2 figures, both models learned the training data quickly: their losses dropped sharply and approached low (0-1) within a few epochs. However, validation perplexity told a different story.

- **RNN**: Training loss fell smoothly, but validation perplexity climbed from ~1,500 to above 4,000 by epoch 3. This shows the model memorized training patterns but failed to generalize, a clear case of overfitting.
- **LSTM**: Loss decreased even faster, yet validation perplexity exploded from ~9,000 to over 40,000 within 3 epochs. The larger capacity and gating made it prone to overfitting on this small, single-book dataset, leading to instability.

Interpretation: Despite its simplicity, the RNN was more stable than the LSTM here. The curves highlight why validation metrics—not training loss—must guide early stopping. In this data regime, the lighter RNN generalized better, while the LSTM's extra complexity backfired.

3.3 Training Time Summary

The training times for both models are summarized below:

- *RNN | run=20250912-080505 | epochs=3 | total time: 16.11 min | per-epoch avg: 322.1s*
- *LSTM | run=20250912-080505 | epochs=3 | total time: 21.19 min | per-epoch avg: 423.8s*

The LSTM took longer per epoch due to its extra gates and parameters. In theory, this added complexity should capture longer patterns, but here it only made training slower without improving performance. Under this setup, the RNN was both **faster and more effective**, showing that simplicity can be an advantage when data is limited.

4. Ablation Studies

4.1 Dropout Ablation

To investigate the effect of dropout regularization, I trained both RNN and LSTM models with dropout values of 0.0 and 0.2 while keeping all other hyperparameters fixed.

Results:

- *RNN | drop=0.0 | val PPL=49965.69 | test PPL=158282.01 | time=10.67 min*
- *RNN | drop=0.2 | val PPL=1565.06 | test PPL=2468.63 | time=10.72 min*
- *LSTM | drop=0.0 | val PPL=74691.81 | test PPL=87013.92 | time=14.03 min*
- *LSTM | drop=0.2 | val PPL=10212.43 | test PPL=13671.51 | time=14.12 min*

Analysis:
Without dropout, both models massively overfit, with validation and test perplexities exploding. Introducing dropout (0.2) made a huge difference: the RNN's test perplexity dropped by more than 30x, and the LSTM's by about 7x. Training time remained nearly identical across settings, meaning the improvements came at no extra computational cost. The RNN benefited the most from dropout, achieving substantially lower perplexities than the LSTM, which still lagged in performance even with regularization.

Conclusion:
Dropout proved essential for this task. It greatly improved stability and generalization, especially for the RNN, which with dropout (0.2) achieved the best balance of efficiency and accuracy. While the LSTM also improved with dropout, it consistently underperformed compared to the RNN. This highlights how proper regularization can allow simpler recurrent models to perform competitively on language modeling tasks.

4.2 Context Length Ablation

I tested sequence lengths of 128 and 256 for both RNN and LSTM:

Results:

- *RNN | seq_len=128 | val PPL=1466.81 | test PPL=2393.76 | time=10.59 min*
- *RNN | seq_len=256 | val PPL=4508.12 | test PPL=7497.54 | time=20.79 min*
- *LSTM | seq_len=128 | val PPL=9852.43 | test PPL=13462.89 | time=14.00 min*
- *LSTM | seq_len=256 | val PPL=35031.13 | test PPL=41908.94 | time=27.22 min*

Analysis:

Extending the sequence length from 128 to 256 tokens consistently worsen performance for both models. The RNN saw perplexity roughly triple, while the LSTM degraded even more severely, with test perplexity exceeding 40k. Training time also doubled with longer contexts, but accuracy dropped sharply, suggesting that these models struggled

to exploit extended dependencies effectively. A likely reason is that both RNNs and LSTMs suffer from vanishing or exploding gradients when handling longer sequences, and their limited capacity prevents them from capturing long-range structure in the data. By contrast, shorter contexts allowed the models to focus on more local patterns and train more stably.

Conclusion:
On this dataset, **shorter context windows (128) worked much better**. The RNN with seq_len=128 achieved the best trade-off between efficiency and generalization, while the LSTM underperformed across both settings. These results confirm that simply increasing context is not always beneficial—especially with small recurrent models—since optimization issues can outweigh potential gains.

5. Generation Samples

For the prompt: “What are the fundamental principles that lead to a nation's wealth?”

=== RNN ===

[T = 0.7] what are the fundamental principles that lead to a <unk> wealth ? the money price of corn , and the rent of land is always a greater proportion , than with that of butcher ’ s meat . the whole value of the annual produce , the quantity of gold and employed in the british manufactures of those particular , far from tending to make anybody gradually loss how much . though the loss of this waste and unimproved east india , the greater part of it was found to be so . before that period , it was found , that the one species of food , clothing ,

[T = 1.0] what are the fundamental principles that lead to a <unk> wealth ? the number of a particular commodity , over and above the number of other people are , the more stock than any one of them must have been the merchants of several nations ; the trade would lie much cheaper , and go to one half . the rent above the new town , he had been accommodated to a certain quantity of corn , or at least retire in the government for the market of his corn , and must as it be in foreign markets , can be no more than sufficient to supply the effectual demand .

[T = 1.3] what are the fundamental principles that lead to a <unk> wealth ? upon a inland university foreign cattle , of which the money naturally sailed from ireland , or the smallest quantity of money , it is said , will soon bring down value it . a particular set of roads are by no means one of the simplest operations ; and in those of france , the extension is the means of one man . a simplest revenue , which might , in many respects , be maintained every cause , have left the industry for their own . the commerce , besides , have no better in the condition ,

=== LSTM ===

[T = 0.7] what are the fundamental principles that lead to a <unk> wealth ? the legal rate of interest seems to render it ten or fifteen years worth of goods , the whole extent of his dress is very far . from the growing of the french , gives is an unwholesome encouragement , to a trade in general . if the monopoly has not been very properly , so , by the famous authors of indigent people , and who want those profitable productions of the earth , or for 20s . and for the year , therefore , it had been higher than it would otherwise have risen

[T = 1.0] what are the fundamental principles that lead to a <unk> wealth ? the commerce of purchasing in gold and silver , is , in this case , equally unproductive . i have never been thought of proposed . such are the causes of the same causes as which are circumstances naturally suit the same privilege . the frequent who established the great lords - spirited - trading upon a particular branch of commerce ; and the produce of the industry of the country easily replaces , by every such operation of the two , that industry is requisite to condition for the wines of france ,

[T = 1.3] what are the fundamental principles that lead to a <unk> wealth ? upon the books of a man who has supposed , a few people to suffer the business , though the greater part of them is greater or less , because they could not afford to pay the old farther . besides , there would purchase , three guilders a great part of the necessary subsistence of the labourer , or , on account of the share and which that part of it consists always ultimately altogether . should this superiority of profit , therefore , in extending or single those poor consume his inhabitants of a few years ; and

Analysis

Both models produced text that was grammatically plausible but often incoherent, which is expected given the small dataset.

- **RNN**: Outputs were shorter and stayed closer to the theme of economics (e.g., trade, corn, merchants). At low temperature (T=0.7) the text was repetitive but on-topic; higher temperatures increased diversity but quickly reduced coherence.
- **LSTM**: Generated longer and more complex sentences, but they drifted into word salad and showed clear signs of overfitting, echoing memorized fragments from the book.

Overall, the generations from both the RNN and LSTM models did not successfully capture Adam Smith's fundamental principles of national wealth which are 6 categories. Instead of identifying key ideas such as division of labour, free markets, free trade, capital accumulation, and the limited role of government, the outputs drifted into fragmented references to corn, rents, interest rates, and gold. The RNN results were especially incoherent and repetitive, while the LSTM produced slightly longer, more fluent sentences but still failed to convey the core concepts. Overall, neither model provided a clear or complete answer to the prompt. However, **the models show some partial awareness of economic themes**.