Jonathan Mandl 211399175
Danielle Shrem

Part 6

Model setup with k=10

We repurposed our tagging architecture into a character-level language model.
The model takes as input a fixed-length window of k = 10 characters and predicts the next character using a simple feedforward neural network.

Each character is embedded into a 64-dimensional vector, concatenated across the 10-character context, and passed through a hidden layer of 128 ReLU units followed by a softmax output layer.

We trained the model on the eng.txt corpus, containing the complete works of Shakespeare (~1.1M characters), for :

- Learning rate: 0.003
- Epochs: 10
- Batch size: 256

Results for k=10

As seen in the loss curve (Figure: part6_k10_loss.png), the validation loss steadily decreased from 1.67 to 1.51 over 10 epochs. This suggests that the model learned useful representations of character-level sequences.

We also generated samples after each epoch (see appendix). Initially the model outputs gibberish, but from epoch 5 onward, we start seeing recognizable English words, Shakespearean names, and punctuation patterns. By epoch 10, the output is mostly syntactically plausible and stylistically close to the training corpus.

Sample after epoch 10:

The 'tis tears in hath so prephesprord, in them not the come to my wife himself dilious case thwardly co

Model setup with k=5

We trained the same character-level language model with a shorter context window of k = 5 characters.
The architecture remains identical: each character is embedded into a 64-dimensional vector, concatenated across the 5-character context, and passed through a hidden ReLU layer with 128 units, followed by a softmax layer predicting the next character.

The model was trained on the eng.txt Shakespeare corpus for 10 epochs using the same training configuration as before.

Results for k=5

The validation loss improved from 1.6532 to 1.5398, showing stable and consistent learning. The rate of improvement was slightly slower compared to k=10, but the final loss was competitive.

The generated samples also showed progress. Early epochs produced mostly random strings, but by epoch 10, we see a stronger presence of real words, names, and recognizable dialogue patterns. However, sentences tended to be shorter and less coherent compared to those generated with k=10.

Sample after epoch 10:

The cad!

Clown: to yearn,
Mind sirraifferty to true,
And by say, and it
Thy resely; knew can are staon,

Model setup with k=20

To explore the effect of longer context, we trained the character-level language model with a window of **k = 20** characters.
The model and training parameters were identical to previous experiments: character embeddings of size 64, a fully connected ReLU layer of 128 units, softmax output, and Adam optimizer.

The model was trained on the same English corpus (eng.txt) over 10 epochs with a batch size of 256 and learning rate of 0.003.

Results for k=20

The validation loss improved from **1.7162** to **1.5273**, showing the best final performance of the three configurations.
While the model started out with a higher loss than both k=5 and k=10, the longer context enabled it to converge to a better local minimum.

Generated text at epoch 10 appears more stylistically coherent, and the model often captures complex patterns such as speaker names (MERCUTIO, BIANCA, MENENIUS) and sentence structures with greater consistency. However, it also seems to overfit character patterns (e.g., repetitive colons or pseudo-names like CIRCAI:: and MIMIMIM:), which may suggest excessive reliance on memorization for longer k.

Sample after epoch 10:

The WISNI:wra her my hadower most truliged:
And here I passitious
like the now to thou so God, be mene s

Comparison of different context lengths (k)

k	Train Loss (Epoch 10)	Validation Loss (Epoch 10)	Text Quality	Observations
5	1.5522	1.5398	Basic, short, mostly valid words	Fast convergence, but limited context
10	1.5412	1.5134	Balanced, fluent, Shakespearean style	Best overall coherence and style
20	1.5655	1.5273	Complex, consistent, sometimes odd	Captures structure well, but tends to repeat

Summary and conclusion

Our experiments show that the context size k has a significant impact on both model performance and the quality of generated text.
A smaller value like k=5 enables fast training but provides only limited contextual understanding, leading to fragmented output.
The medium setting k=10 offered the best tradeoff: it produced coherent, stylistically appropriate text while maintaining stable convergence.
The longest context, k=20, led to the lowest validation loss, indicating deeper pattern learning. However, it also introduced overfitting artifacts such as pseudo-repetitive speaker names and less diversity.

In conclusion, **k=10** appears to be the most effective choice, balancing structure, fluency, and generalization.

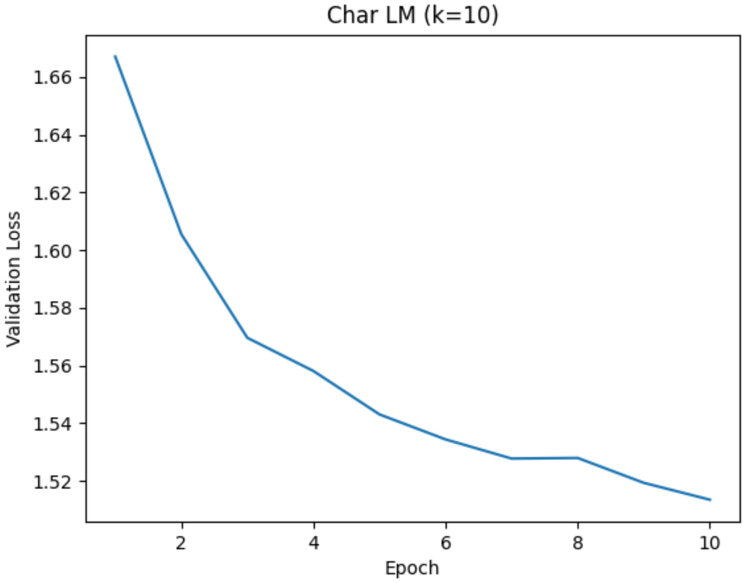
Sampling without prefix (k=10)

We also evaluated the model's generative ability without providing any initial context. Interestingly, the model was able to generate reasonably structured Shakespearean-style text even when starting from an empty prefix.
While the coherence was somewhat reduced, many of the outputs included character names (e.g., BENVOLIO) and plausible sentence openings.

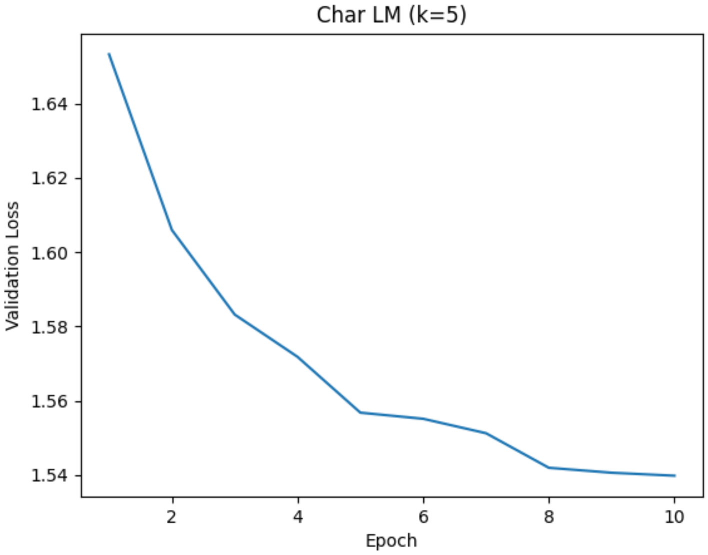
For example: *POPSO: Of conter's so thee starrance?*
BENVOLIO: Welcome that throughs sundriver his taloucwertly co

Figures

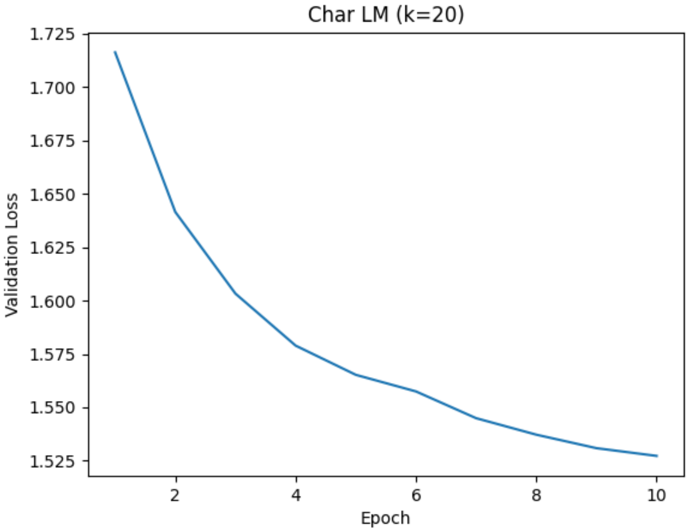
k=10



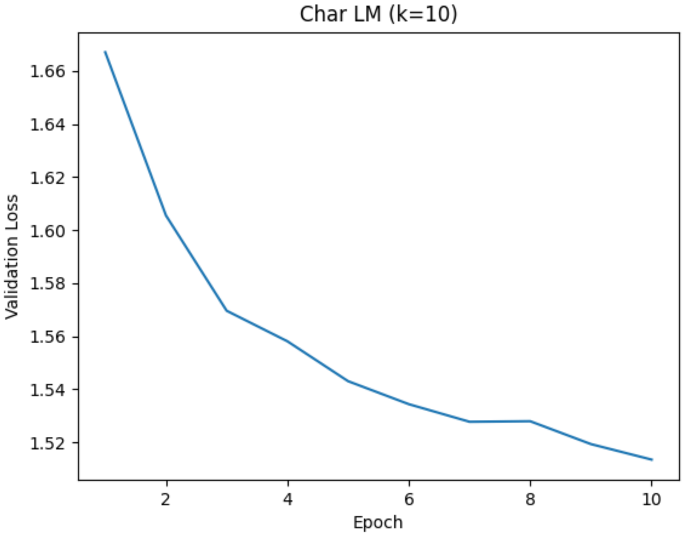
k=5



k=20



k=10 without prefix


Char LM (k=10)
Validation Loss
Epoch