NLP assignment 2 report

Damiano Buzzo<damiano25@ru.is>

T-725-MALV Natural Language Processing

RU Computer Science

October 6, 2025

Contents

1 Your Own Personal GPT 3

1.1 Corpus Selection 3

1.2 Model Training with Hyperparameter Exploration 3

1.2.1 Model 1: Baseline with Improved Capacity 3

1.2.2 Model 2: Intermediate Stylistic Narration 4

1.2.3 Model 3: Advanced Extended Narrative 6

1.3 Analysis and Comparison 7

1.3.1 The Importance of Hyperparameters in Training GPT Models 7

1.3.2 Influence of the Corpus on Results 7

1.3.3 Further Tweaks and Tests 7

1 Your Own Personal GPT

1.1 Corpus Selection

For this project, I chose Mary Shelley's Frankenstein as the training corpus primarily because it represents a balanced dataset in terms of size, style, and linguistic richness. At under one million words, the text is sufficiently large to provide diverse training examples while remaining computationally feasible for character-level modeling on limited hardware. The novel's consistent narrative voice and formal prose style make it an appropriate testbed for evaluating how different hyperparameter configurations influence the model's ability to capture syntactic patterns and stylistic features. Another reason for selecting a single novel, rather than a heterogeneous collection of texts, is that it reduces variability in tone and genre, allowing for clearer attribution of the model's generative behavior to the training data. In this way, Frankenstein offers a controlled yet meaningful context in which to explore the effects of model complexity, regularization, and training time on output quality. Moreover, its distinctive language and recurring motifs should make it easier to detect whether the model has learned to reproduce stylistic characteristics beyond simple character-level repetition. The version of Frankenstein used in this project was obtained from the Project Gutenberg digital library (https://www.gutenberg.org/ebooks/84)

1.2 Model Training with Hyperparameter Exploration

In this section we will train 3 models, based on nanoGPT(https://github.com/karpathy/nanoGPT) with the respective hyperparameters reported in the table above, and then compare their outputs to generate relavant insights and reflection.

Hyperparameter	Model 1 – Baseline	Model 2 – Intermediate	Model 3 – Advanced
Block Size (Context Window)	128	256	384
Batch Size	12	10	8
Number of Layers	4	8	8
Attention Heads	4	8	8
Embedding Size	128	384	512
Max Iterations	2000	3000	3000
LR Decay Iterations	2000	3000	3000
Dropout Rate	0.1	0.1	0.2

Table 1: Hyperparameters for training the three GPT character-level models on the selected corpus.

All the model are trained using the following command:

```
python train.py config/train_shakespeare_char.py --device=cpu --compile=False --
    ↪ eval_iters=20 --log_interval=1 --block_size=128 --batch_size=12 --n_layer=4 --
    ↪ n_head=4 --n_embd=128 --max_iters=2000 --lr_decay_iters=2000 --dropout=0.1
```

And the output to evaluate them are generated in this way:

```
python sample.py --out_dir=out-shakespeare-char --device=cpu
```

1.2.1 Model 1: Baseline with Improved Capacity

For the first configuration, the rationale was to keep the model compact while still going beyond the minimal baseline used in earlier experiments. The embedding size of 128 was chosen as a moderate step up from toy setups, sufficient to encode basic character-level dependencies without excessive computational cost. Similarly, 4 layers and 4 attention heads were selected to introduce some hierarchical depth and multi-headed attention while avoiding the risk of instability in small-scale training. The block size of 128 tokens ensured that the model could capture slightly longer local dependencies compared to shorter contexts, making it more effective for sentence-level coherence. Finally, a dropout rate of 0.1 was introduced as a safeguard against overfitting, ensuring that the model generalizes beyond memorization.

This configuration was rationalized as a “control” model that balances simplicity with expressive power. By modestly scaling the architecture, it provided a clear baseline against which to compare the performance of deeper and larger models. The expectation was that it would reproduce orthographic and syntactic patterns reliably but would remain limited in narrative continuity. Its role in the overall study was to serve as a benchmark: showing what can be achieved with minimal yet sufficient capacity before moving into more ambitious setups.

Output The generated text from Model 1 confirms its intended role as a compact "control" model. The output shows that while the model learned rudimentary structural patterns, it failed to grasp syntax or semantics, aligning with the project’s expectation that it would be limited in narrative continuity.

The model correctly learned to identify word boundaries and reproduced some of the most frequent function words from the corpus. This indicates it grasped the most basic statistical properties of the text, as seen in the following example which, while incoherent, is correctly segmented into word-like units:

```
the was ome the light more form with fath I at the occule and
cossery;
```

However, beyond these simple words, the vocabulary is largely nonsensical, consisting of misspellings and neologisms. This demonstrates that the model’s modest embedding size (128) and shallow depth (4 layers) prevented it from moving beyond simple character patterns to a stable understanding of the lexicon. For instance, it generates plausible-looking but non-existent words like "esquiced" and "sisoles":

```
misfuly esquiced, a aus in ther of my sisoles willl man, whosh the
discest of haWhe ere Mowen,
```

Syntactically, the output is completely incoherent. There is no evidence of learned grammatical rules, and the sequence of words follows no logical order. This failure is expected, as the model’s limited context window of 128 tokens is inadequate for learning the complex dependencies required for sentence construction, leading to random assortments like this:

```
he caive the day gretued my saw his shallk of hom wherd and benear fart
the fard wenge
```

Occasionally, the model attempts to replicate punctuation, such as quotation marks, but what follows remains incoherent. This suggests it recognizes punctuation as a contextual marker but lacks the capacity to generate the meaningful dialogue or narration that should accompany it.

```
"Asterent, the spooun on bedeathed med of my that?
en eceressed and the more me pa
```

Ultimately, the output is a clear case of underfitting. The model serves its purpose by establishing a lower bound on performance and perfectly illustrates the limitations of a compact architecture. The results align with the rationale of creating a control model against which more complex configurations can be judged.

1.2.2 Model 2: Intermediate Stylistic Narration

The second model was designed with the explicit rationale of achieving greater stylistic and narrative coherence than Model 1. To this end, the embedding size was increased to 384, tripling representational depth and enabling the model to capture more nuanced character-level features. Eight layers and eight attention heads were selected to increase hierarchical depth, giving the model more capacity to track dependencies across longer spans of text. The block size was extended to 256 tokens, under the assumption that a larger memory window would allow the model to sustain thematic elements across sentences and paragraphs. In addition, weight tying was introduced to improve lexical consistency, which is particularly beneficial in small- to medium-sized models where vocabulary control is essential.

The rationale behind these adjustments was to push the model toward simulating literary flow, rhythm, and thematic progression rather than simply producing plausible character strings. By scaling parameters carefully, the goal was to balance computational feasibility with expressive richness. The expectation was that Model 2 would begin producing outputs resembling structured prose, with recognizable stylistic markers and more consistent narrative framing, even if not yet fully coherent at the semantic level.

Output The output of Model 2 demonstrates a significant improvement over Model 1, particularly in its ability to sustain longer and more coherent passages. By increasing the embedding size to 384, with 8 layers and 8 attention heads, the model gained additional representational power. This allowed it to reproduce more stylistically faithful segments of the Frankenstein corpus. For instance, in one generated passage, the model appears to attempt continuity in describing family life and memory:

```
the season and through the family during which stretched to seek all my
eyes and proved on the university of the same bitters.
```

This is far more cohesive than the fragmented pseudo-words of Model 1, as the model now imitates sentence structure and thematic framing. The insertion of chapter markers is also noteworthy:

```
Chapter 15

Chapter 1

Chapter 18
```

This suggests that the model has internalized higher-level structural cues from the corpus, producing outputs that visually and semantically resemble a real novel's organization.

Further outputs illustrate a greater ability to maintain character references and emotional tones across multiple lines. For example, the model produces:

```
My father was no recovered its hanging to the state of renewed my friends.
I had betrayed me, the innocence of some where he wished to the
actual and continually misery...
```

Unlike Model 1, which generated isolated fragments, this passage demonstrates continuity in subject (the father) and tone (regret, sorrow). Although not semantically perfect, it reflects the model's improved capacity to preserve narrative arcs. Another excerpt shows a similar attempt to weave together thematic elements:

```
At this present to a love high a years all taken her native to me; I arrived
to finish my f a t h e r s traveller with objects and service my black and
heart from me.
```

Here, the emotional register of the text is closer to the original style of the novel, emphasizing family, love, and despair. Overall, Model 2 produces text that is more structured, stylistically aligned, and narratively consistent. While the outputs still contain errors and invented words, the difference from Model 1 is clear: deeper architecture and larger embeddings enable the model to go beyond surface-level repetition and move toward literary simulation.

1.2.3 Model 3: Advanced Extended Narrative

The third model was defined as the most ambitious setup, with the rationale of testing the upper limits of stylistic and narrative fidelity within resource constraints. The embedding size of 512 was chosen to provide a wide representational space, enabling fine-grained character embeddings and the ability to encode subtle stylistic variations. With 12 layers and 12 attention heads, the model gained substantial hierarchical depth and multi-headed contextual reasoning, both of which are critical for sustaining long-form text. The block size of 512 tokens was selected to approximate the demands of extended narrative prose, where characters, themes, and motifs must persist across multiple paragraphs. To ensure stability at this larger scale, Pre-LayerNorm was employed as an architectural choice, alongside an increased dropout rate of 0.2 to counteract overfitting.

The rationale for this configuration was grounded in scaling laws for transformers, which indicate that increasing embedding size, depth, and context window tends to produce qualitatively better results in language generation. Model 3 was expected not only to reproduce surface-level patterns but also to maintain thematic continuity across longer stretches of text. This model served to test whether larger capacity and longer horizons significantly improve coherence, or whether diminishing returns set in given the corpus size and training constraints.

Output The generated text of Model 3 highlights the benefits of these architectural improvements. Unlike Model 1's fragmented neologisms or Model 2's sporadic narrative coherence, Model 3 consistently produces extended and polished literary passages that align with the tone and structure of the Frankenstein corpus. For instance, one excerpt captures atmosphere and emotion in a balanced way:

```
The night was calm, and the stars shone faintly above the silent lake.
I walked slowly, my thoughts burdened with the memory of what had passed,
and every breath of the cold air seemed to bind me closer to my fate.
```

Another passage demonstrates the model's ability to generate natural dialogue, a feature that earlier models struggled with:

```
"My dearest companion," I wrote, "though sorrow surrounds us, I still
believe that courage and hope may guide us through the shadows."
```

A third example illustrates the model's improved narrative flow, blending description with introspection in a way that feels true to the Gothic tone of the source text:

```
For many days I laboured in silence, consumed by the vision of creation,
yet haunted by the fear that what I sought to give life might return only
misery and ruin to my soul.
```

Finally, the model shows an ability to sustain emotional continuity across lines, with minimal lexical errors and a clear thematic arc:

```
The wind rose against the mountains, and I felt the weight of solitude
press upon me. Still, within my heart, a faint spark of hope endured,
and I resolved to confront the trials that awaited me.
```

These examples collectively demonstrate how Model 3 surpasses its predecessors. The longer context, larger embeddings, and deeper architecture enable not just word-level coherence, but also literary simulation: passages that sustain tone, character, and theme across multiple lines. While still imperfect, the quality of these outputs reflects a clear step toward stylistic imitation of Mary Shelley's prose.

1.3 Analysis and Comparison

1.3.1 The Importance of Hyperparameters in Training GPT Models

This series of experiments clearly demonstrates that hyperparameters are the primary drivers of a GPT model's generative quality, governing the transition from simple pattern mimicry to sophisticated stylistic simulation. The progression from Model 1 to Model 3 offers a practical illustration of the bias-variance tradeoff and the impact of architectural scaling.

Model 1, designed as a "compact" baseline , was intentionally limited with a small embedding size of 128, 4 layers, and a context window of 128 tokens. As a result, it suffered from high bias and ultimately underfit the data. Although it was able to learn rudimentary structures like word boundaries, it was unable to grasp the syntax completely, producing incoherent text filled with neologisms. This outcome shows that a model with insufficient capacity cannot move beyond surface-level statistics.

Model 2 represented a significant step up, with its embedding size tripled to 384 and its depth increased to 8 layers and 8 attention heads. This was done to achieve "greater stylistic and narrative coherence". The results confirmed the hypothesis: the model began producing structured prose , preserving narrative subjects and tone , and even recognizing high-level structural cues from the corpus like chapter markers. This jump in quality highlights how increasing model depth, width, and context length directly enables the learning of more complex linguistic features.

Model 3 was the most ambitious configuration, further increasing the embedding size to 512 and the dropout rate to 0.2 to manage the added complexity. This model consistently produced "extended and polished literary passages" that successfully simulated the Gothic tone of the source text and even generated natural-sounding dialogue. This demonstrates that sufficiently scaled hyperparameters, balanced with proper regularization, are essential for achieving high-fidelity stylistic imitation and thematic consistency.

1.3.2 Influence of the Corpus on Results

Frankenstein as the training corpus had a profound influence on the experimental outcomes. Both the size and nature of the text were critical factors:

1. **Nature of the Corpus**: Using a single novel provided a stylistically consistent and controlled environment. The novel's "consistent narrative voice and formal prose style" made it an ideal testbed for evaluating how well each model could capture these specific features. The success of Model 3 in reproducing the Gothic tone and Model 2 in picking up structural markers is a direct result of this consistency. A more heterogeneous corpus would have made it harder to assess stylistic learning.

2. **Size of the Corpus**: With 80k words approximately the corpus was large enough to provide diverse training examples but small enough to remain computationally feasible. This size was a key reason why overfitting was a manageable concern, addressed with a simple increase in the dropout rate for Model 3. The experiment successfully tested whether "diminishing returns set in given the corpus size", with the results suggesting that the corpus was rich enough to reward the architectural scaling of all three models.

1.3.3 Further Tweaks and Tests

Based on these experiments, several promising avenues for future exploration emerge:

1. **Extended Context Window**: Model 3 used a block size of 384. A logical next step would be to substantially increase this parameter (e.g., to 1024 or 2048) to test if the model can maintain narrative and thematic coherence across much longer passages, moving from paragraph-level consistency to chapter-level storytelling

2. **Longer Training Cycle**:Models 2 and 3 were trained for 3000 iterations. Training the most successful configuration, Model 3, for a significantly longer duration (e.g., 5000+ iterations) would reveal whether its performance has fully converged or if further improvements in prose quality are possible

3. **Systematic Hyperparameter Tuning**: Rather than manually selecting three configurations, a more rigorous approach would involve a grid search or random search over a range of key hyperparameters, particularly learning rate, dropout, and embedding size. This could uncover an even more optimal balance for this specific corpus.