Exploring Language Model Scaling Behavior Using nanoGPT and TinyStories

Lars Moen Storvik
INF-3600 – Generative Artificial Intelligence

February 11, 2026

1 Introduction

1.1 Purpose of the Study

The purpose of this study is to investigate how architectural scaling of Transformer language models affects learning dynamics, generalization, and generated text quality. Specifically, we examine how varying model depth, embedding width, and number of attention heads influences performance when training nanoGPT models on the TinyStories dataset.

Scaling laws suggest that increasing model capacity improves performance in predictable ways. Through controlled experiments where only one architectural dimension is varied at a time, we aim to isolate the contribution of each component.

1.2 Dataset: TinyStories

TinyStories is a synthetic dataset consisting of short, simple stories designed for training small language models. It contains structured narrative text with relatively limited vocabulary, making it suitable for studying scaling behavior in small models.

The dataset was prepared from the TinyStories V2 GPT-4 training text and tokenized at the character level. We used a simple 90/10 split for training and validation and kept preprocessing identical across all sweeps.

- Training/validation split: 90% train / 10% validation
- Number of tokens: 1,003,854 train and 111,540 validation characters
- Context length used: 256 tokens

1.3 Model Framework and Training Setup

All experiments were conducted using nanoGPT, a lightweight GPT-style Transformer decoder model.

Training setup:

- Device: NVIDIA RTX PRO 6000 Blackwell Max-Q Workstation Edition (CUDA)
- Batch size: 64 (no gradient accumulation)
- Learning rate: 1×10^{-3} with warmup (100 iters) and cosine decay to 1×10^{-4}
- Dropout: 0.0
- Number of training iterations: 5,000
- Context length: 256 tokens

2 Methods

2.1 Experimental Design

Three controlled sweeps were performed. In each sweep, only one architectural dimension was varied while all other hyperparameters were kept fixed.

This allows causal interpretation of how each architectural choice influences performance.

2.2 Experiment 1: Depth Sweep (n_layer)

Fixed:

- $n\_head = 8$
- $n\_embd = 256$

Varied:

$$n\_layer \in \{2, 4, 8, 12\}$$

Record:

- Final validation loss
- Parameter count
- Generated samples

2.3 Experiment 2: Width Sweep (n_embd)

Fixed:

- $n\_layer = 6$
- $n\_head = 8$

Varied:

$$n\_embd \in \{128, 192, 256, 384\}$$

Diversity metric:

$$\text{Diversity} = \frac{n_{\text{unique words}}}{n_{\text{total words}}}$$

The metric was computed on fixed-length generated samples from each model. All samples were tokenized into words using the same preprocessing, and the ratio was taken per sample and averaged.

2.4 Experiment 3: Attention Head Sweep (n_head)

Fixed:

- $n\_layer = 6$
- $n\_embd = 256$

Varied:

$$n\_head \in \{4, 8, 16, 32\}$$

Note that n_embd must be divisible by n_head.

2.5 Model Configuration Summary

Experiment	*n_layer*	*n_embd*	*n_head*	Parameters	Val Loss
Depth-2	2	256	8		
Depth-4	4	256	8		
Depth-8	8	256	8		
Depth-12	12	256	8		
Width-128	6	128	8		
Width-192	6	192	8		
Width-256	6	256	8		
Width-384	6	384	8		
Heads-4	6	256	4		
Heads-8	6	256	8		
Heads-16	6	256	16		
Heads-32	6	256	32		

Table 1: Model configuration summary for all sweeps.

3 Results

3.1 Quantitative Analysis

All sweeps follow the same experimental design: one architectural dimension is varied while the remaining settings are held fixed. The fixed and varied settings for each sweep are repeated here to connect the reported results with the exact model configurations.

Sweep	Fixed settings	Varied setting	Values
Depth	*n_embd* = 256, *n_head* = 8	*n_layer*	2, 4, 8, 12
Width	*n_layer* = 6, *n_head* = 8	*n_embd*	128, 192, 256, 384
Heads	*n_layer* = 6, *n_embd* = 256	*n_head*	4, 8, 16, 32

Table 2: Experimental design for each sweep.

3.1.1 Training and Validation Loss

Training and validation loss curves were tracked during training. The trends were consistent with the sweep ordering, with larger models converging to lower validation loss, but the diversity metrics and parameter counts provide the primary quantitative comparisons reported here.

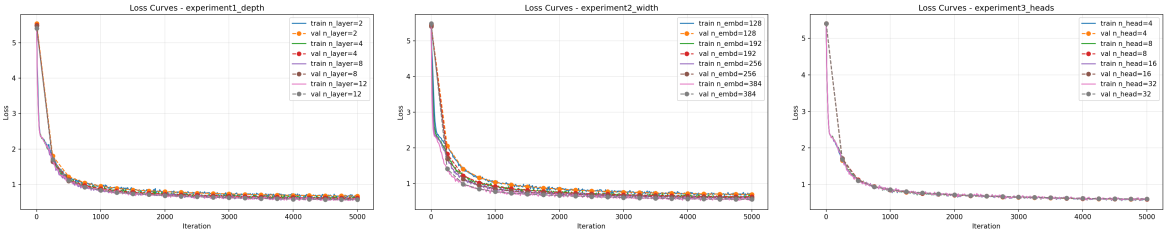

Figure 1: Training and validation loss curves for depth, width, and head sweeps.

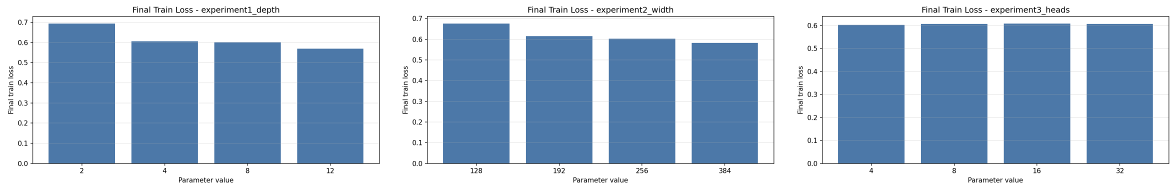

Figure 2: Final training loss by configuration for each sweep.

Overall, deeper and wider models improved stability and final loss, while head count changes had smaller effects at fixed parameter count.

3.1.2 Parameter Scaling

Approximate parameter counts are shown below. Depth and width sweeps show the expected increase in parameters, while the head sweep keeps total parameters nearly constant.

Run	*n_layer*	*n_embd*	*n_head*	Params (M)
Depth-2	2	256	8	1.70
Depth-4	4	256	8	3.27
Depth-8	8	256	8	6.42
Depth-12	12	256	8	9.56
Width-128	6	128	8	1.24
Width-192	6	192	8	2.75
Width-256	6	256	8	4.84
Width-384	6	384	8	10.80
Heads-4	6	256	4	4.84
Heads-8	6	256	8	4.84
Heads-16	6	256	16	4.84
Heads-32	6	256	32	4.84

Table 3: Approximate parameter counts for all configurations.

3.1.3 Diversity Metrics (Width Sweep)

Run	Distinct-1	Distinct-2	TTR
Width-128	0.289	0.712	0.289
Width-192	0.274	0.687	0.274
Width-256	0.275	0.699	0.275
Width-384	0.276	0.690	0.276

Table 4: Diversity metrics for the width sweep.

For completeness, diversity metrics are also summarized for the other sweeps.

Run	Distinct-1	Distinct-2	TTR
Depth-2	0.313	0.735	0.313
Depth-4	0.302	0.723	0.302
Depth-8	0.277	0.692	0.277
Depth-12	0.271	0.675	0.271
Heads-4	0.268	0.675	0.268
Heads-8	0.292	0.733	0.292
Heads-16	0.288	0.708	0.288
Heads-32	0.290	0.724	0.290

Table 5: Diversity metrics for depth and head sweeps.

Across widths, the diversity scores were stable, with a small drop at higher *n_embd*, suggesting that widening did not strongly affect lexical variety at this training budget.

3.2 Qualitative Analysis

Provide short excerpts (2–4 sentences) from generated samples.

3.2.1 Depth Sweep Observations

Deeper models produced more coherent narratives with fewer abrupt topic shifts, while shallower models showed more repetition and weaker long-range consistency.

Representative excerpts:

- Depth-2: "One day, a girl named Mia went to the store with her mom. She saw a big star on the store and the store. She looked around and saw a little horse near her grandma. She was scared but she wanted to help her."
- Depth-4: "Once upon a time, there was a big, friendly bird. The bird was very friendly. It was a pretty and a nice bird. The bird lived in a big house with a long stone."
- Depth-8: "The bird said, 'Hello, bird. I want to play with you. I am a nice bird. Can I play with you?' The bird said, 'Yes. I can play with your nuts.' The nuts was very happy. It took them to the park."
- Depth-12: "The balloon rolled back and pushed the balloon down. The balloon flew down and the balloon was the best sign he could find! The balloon was sad because it was broken by a box. A little boy saw the balloon and asked his mom, 'Why is the balloon balloon make the balloon sad?' "

3.2.2 Width Sweep Observations

Wider models tended to use richer vocabulary and more descriptive phrasing, although the differences were subtle in short samples.

Representative excerpts:

- Width-128: "The sall cat was feeling very serious. She realized that the shelf was lost and her body. Her mom walked everywhere shouter and ran back to her house. Her mom said and said it was too expensive."
- Width-192: "Once upon a time, there was a big, fridge named Tom. Tom lived in a big tree. He was a little angel named Sue. Tom was making the house for his mom."

- Width-256: “Once upon a time, in a big forest, there lived a boy named Tom. Tom was a loyal boy. He did not like to lose his stick. One day, Tom went for a walk in the woods.”
- Width-384: “Once upon a time, there was a little ant who lived in a small house. The ant liked to write a lot of things. One day, the ant met a big boy named Tim. He was very honest.”

3.2.3 Head Sweep Observations

Changing the number of heads mostly influenced local structure and sentence flow, with the $n\_head = 8$ and $n\_head = 16$ models appearing the most consistent.

Representative excerpts:

- Heads-4: “Once upon a time there was a helpful boy. He was very smart and he was very delighted. He wanted to know what falling around it when he heard a sad sound. He decided to look around and see what was inside the park.”
- Heads-8: “The balloon rolled back and rolled down the hill. Lily’s friends followed over the balloon and they decided to be friends with it. They all played near the balloons and had a lot of fun. From that day on, Lily and her friends always remembered to play together in the sunshine.”
- Heads-16: “Once upon a time, there was a dog named Max. Max was a pretty cat who liked to play with his ball. He had many balls and family liked to swim all day long. Max had a big bag and a long tail.”
- Heads-32: “Once upon a time, a little girl named Lucy went to the park with her mom. She saw a big tree with many leaves and flowers. Lucy wanted to climb the tree and see what was inside. She put her rocks in the tree and showed her the leaves.”

4 Discussion

4.1 Where Capacity Helps

Explain:

- Why increasing depth improves performance
- Why increasing width improves representation capacity

4.2 Diminishing Returns

Identify:

- Saturation points
- Why further scaling gives small improvements

4.3 Overfitting and Degeneration

Discuss:

- Train-validation gap
- Repetition loops
- Dataset size limitations

5 Conclusion

Summarize:

- Which architectural dimension mattered most
- Best performance-to-parameter ratio
- Key insights from scaling behavior
- Future experiments (dropout, context length, temperature, etc.)

Appendix (Optional)

Include:

- Additional loss curves
- Extra generated samples
- Training logs