NanoGPT-Z: Measuring Catastrophic Forgetting in Style Fine-tuning of Small Language Models

Nikolaos Fourlis
School of Electrical and Computer Engineering
National Technical University of Athens (NTUA)
Machine Learning Engineer - Sophea AI
Athens, Greece
hi@nikosfourlis.com

Abstract

Fine-tuning language models on domain-specific data is known to degrade performance on general language tasks—a phenomenon termed catastrophic forgetting. While extensively studied in large-scale models, its dynamics at the small model scale remain underexplored. We train a 17M parameter GPT from scratch on 1B tokens of standard English, then fine-tune it on Gen Z slang across eight data regimes ranging from 1k to 1,000k tokens, measuring WikiText-103 perplexity and slang token prevalence at every 200 fine-tuning steps. We find that the onset of forgetting is immediate and severe: WikiText perplexity doubles within the first 200 optimization steps. By the end of the 5,000-step fine-tuning run on just 1k tokens, perplexity increases 47x (from 258 to 12,611). Meanwhile, style shift plateaus at 18–20% across all subset sizes. These results suggest that at small model scales, style acquisition capacity is rigidly bounded independently of fine-tuning data volume, while the degradation of pre-trained knowledge scales disastrously with extended training.

1 Introduction

The ability to adapt a pretrained language model to new tasks or domains via fine-tuning is one of the central practical techniques in modern NLP. However, fine-tuning comes at a cost: as the model adapts to new data, it tends to overwrite the representations learned during pretraining—a phenomenon known as catastrophic forgetting [1]. This trade-off between plasticity and stability is well-documented in the continual learning literature [2, 4], but most empirical studies focus on large models ($>$1B parameters) or task-switching scenarios rather than style adaptation in small models [3].

Style adaptation, teaching a model to generate text in a specific register, dialect, or persona, is a practically important use case. Yet the forgetting dynamics of style fine-tuning remain poorly understood, particularly at the small model scale relevant for on-device and edge deployment.

In this work, we study catastrophic forgetting through the lens of style acquisition. We train a 17M parameter GPT on 1B tokens of standard English text, then fine-tune it on Gen Z slang across eight fine-tuning corpus sizes. We track two metrics throughout fine-tuning: WikiText-103 perplexity as a proxy for retained English fluency, and slang token prevalence as a proxy for acquired style.

Our main findings are: (1) the onset of forgetting is immediate, with baseline perplexity doubling within the first 200 steps regardless of corpus size; (2) style shift plateaus at approximately 18–20% across all data regimes, suggesting a capacity ceiling for style acquisition at this model scale; and (3) larger fine-tuning corpora cause disproportionately more forgetting without proportionally more style

Preprint.

gain. To support deployment, we built a custom C++ inference engine compiled to WASM; code is available at `github.com/fourlhs/nano-gpt-z`.

2 Method

2.1 Model and Architecture

We train a decoder-only transformer [7] with the following architecture: 6 layers, 8 attention heads, 256 embedding dimension, context window of 64 tokens, and weight-tied input/output embeddings. This yields approximately 17M parameters. We use the GPT-2 tokenizer (tiktoken, vocabulary size 50,257). Attention is implemented using `F.scaled_dot_product_attention` with causal masking. The model is trained from scratch with GPT-2 weight initialization ($\sigma = 0.02$).

2.2 Pretraining

The model is pretrained on 1B tokens sampled from FineWeb-Edu [6], a high-quality filtered web text dataset. Training uses AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay = 0.1), batch size 512, peak learning rate 1e-3 with linear warmup over 500 steps followed by cosine decay to 6e-5, and gradient clipping at 1.0. Training runs for approximately 30k steps on a single NVIDIA RTX 4090, reaching a final validation loss of 4.24.

2.3 Fine-tuning Data and Procedure

Gen Z fine-tuning data is drawn from two public datasets: `Sam-genz-omni` and `genz_brainrot_dataset`. Combined, these yield approximately 1M tokens. We construct eight fine-tuning subsets by taking the first N tokens: 1k, 5k, 20k, 50k, 100k, 200k, 500k, and 1000k. Each subset run starts from the same pretrained checkpoint. Fine-tuning uses AdamW with learning rate 3e-5, batch size 32, cosine decay, and gradient clipping at 1.0. Each run trains for 5,000 steps.

2.4 Evaluation Metrics

WikiText-103 perplexity measures retained English fluency. We evaluate on the WikiText-103 validation split every 200 fine-tuning steps. Perplexity is computed as $\exp(\text{mean cross-entropy loss})$ over 100 random batches.

Style shift measures acquired Gen Z register. We compute the proportion of generated tokens belonging to a slang vocabulary—the top-200 tokens ranked by normalized frequency ratio between the Gen Z corpus and WikiText-103.

3 Results

3.1 Forgetting Curves

Figure 1 shows WikiText-103 perplexity over fine-tuning steps for all eight subset sizes. The onset of forgetting is immediate and steeply continuous. Within the first 200 optimization steps, the baseline perplexity of 258 doubles (reaching approximately 537) across all subset sizes. By the end of the 5,000-step runs, the degradation becomes catastrophic: the 1k and 5k subsets peak around 12,000–13,000 perplexity, while the 500k and 1000k subsets suffer extreme collapse, continuing to rise beyond 40,000.

3.2 Style Acquisition and Trade-offs

Table 1 shows final WikiText perplexity and style shift for each subset. Style shift increases from a baseline of 1.4% to approximately 18–20% across all subsets. Notably, this plateau is reached by the 1k subset and does not meaningfully increase with more data—the 1000k subset achieves 19.9% style shift compared to 18.1% for the 1k subset, despite 1000x more fine-tuning data. The 500k subset is an exception, showing the highest perplexity increase despite the lowest style shift, which we discuss in Section 4.

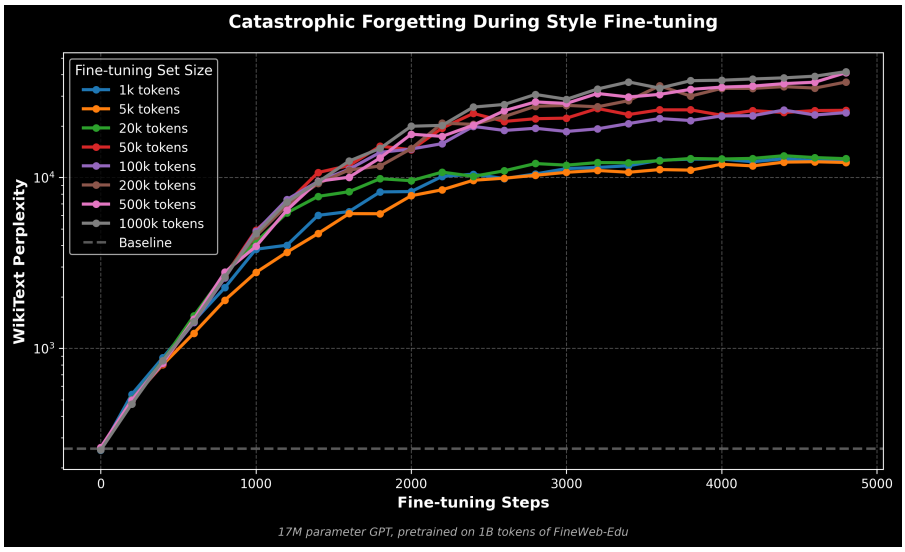

Figure 1: Catastrophic Forgetting During Style Fine-tuning. The onset is immediate, with perplexity doubling in the first 200 steps and degrading continuously thereafter.

Table 1: Final WikiText-103 perplexity and style shift after 5,000 fine-tuning steps.

Fine-tune tokens	WikiText Perplexity	Δ Perplexity	Style Shift
Baseline (none)	258	—	1.4%
1k	12,611	+12,353	18.1%
5k	12,339	+12,081	19.0%
20k	13,009	+12,750	17.7%
50k	24,988	+24,730	17.9%
100k	23,736	+23,478	18.7%
200k	34,129	+33,871	20.1%
500k	40,518	+40,260	16.5%
1000k	39,535	+39,277	19.9%

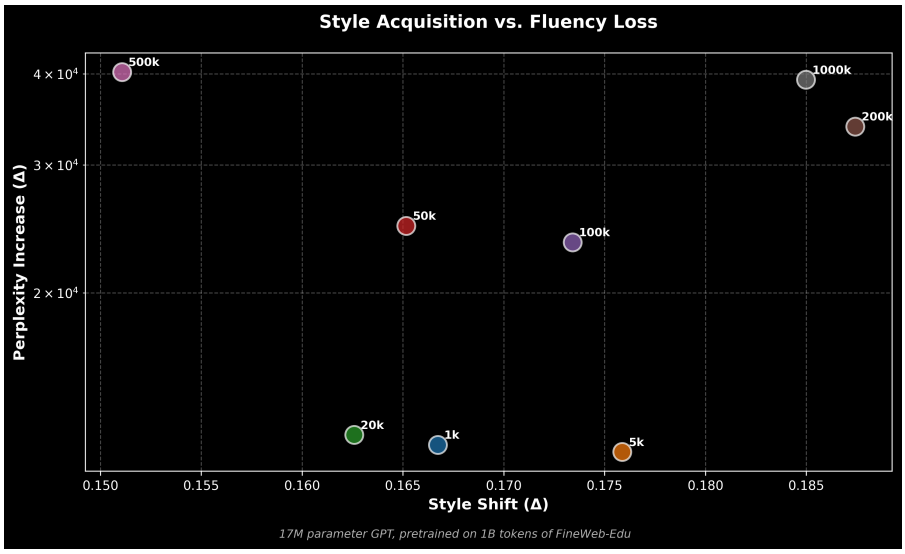

Figure 2: Style Acquisition vs. Fluency Loss. Perplexity increase vs. style shift delta across eight fine-tuning corpus sizes.

4 Discussion

The Onset vs. The Collapse. The steepness of the forgetting curves reveals a critical vulnerability in small models. While the ultimate collapse (reaching >12,000 perplexity) takes several thousand steps, the initial onset is instantaneous. The fact that general language representations double their perplexity in just 200 steps suggests that a 17M parameter model lacks the capacity to hold pre-trained representations in superposition with new stylistic gradients.

Bounded style acquisition. The plateau at 18–20% style shift across all corpus sizes suggests that this model has a hard capacity limit for style adoption. This may reflect the limited context window (64 tokens), the size of the slang vocabulary, or the stylistic distance between standard English and Gen Z slang. Feeding more data strictly penalizes general fluency without yielding stylistic gains.

Limitations. Several limitations should be noted. First, the baseline WikiText perplexity of 258 is high relative to larger models (GPT-2 small achieves $\sim$29), reflecting the intentionally small scale of our model. Second, the style shift metric relies on a simple token frequency ratio, which is sensitive to corpus composition at intermediate subset sizes, as evidenced by the 500k outlier. Third, our experiments use full fine-tuning. Parameter-efficient methods such as LoRA [5] likely reduce forgetting and represent a natural direction for future work.

5 Conclusion

We present NanoGPT-Z, a controlled study of catastrophic forgetting during style fine-tuning in a 17M parameter language model. Our results demonstrate that the onset of forgetting is immediate—doubling perplexity in 200 steps—and culminates in a severe collapse of general language capabilities over 5,000 steps. Furthermore, style acquisition plateaus rapidly regardless of fine-tuning corpus size. These findings suggest that small language models are particularly vulnerable to catastrophic forgetting, highlighting critical constraints for on-device personalization and edge deployment scenarios.

References

[1] McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24, 109–165.

[2] Kirkpatrick, J., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521–3526.

[3] Luo, Y., et al. (2023). An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *arXiv preprint arXiv:2308.08747*.

[4] Goodfellow, I. J., et al. (2013). An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*.

[5] Hu, E. J., et al. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[6] Penedo, G., et al. (2024). The FineWeb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*.

[7] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.