Scaling Behavior of Sparse Attention Mechanisms in Transformer Language Models

Your Name Co-Author Name

November 13, 2026

Abstract

We study how sparse attention patterns affect the scaling laws of transformer language models across model sizes from 125M to 2.7B parameters. We find that sparsity ratio interacts with model scale: larger models tolerate more aggressive sparsification with negligible loss in validation perplexity.

Keywords: scaling laws, sparse attention, transformer architectures, language modeling

1 Introduction

Scaling laws describe how model loss decreases predictably with parameters, data, and compute. Whether these laws hold when attention is sparsified remains underexplored. We train a family of models with varying sparsity ratios to characterize this interaction.

2 Materials and Methods

We train six model sizes (125M–2.7B parameters) at four sparsity ratios (0%, 50%, 75%, 87.5%) on a 300B-token corpus, holding compute budget fixed via the Chinchilla-optimal token-to-parameter ratio. Sparsity is implemented as fixed strided attention with block size 64.

3 Results

Table 1 reports validation loss at convergence for each configuration.

Params	Dense	75% sparse	87.5% sparse
125M	3.21	3.24	3.31
1.3B	2.74	2.75	2.79
2.7B	2.58	2.58	2.61

4 Discussion

The gap between dense and sparse configurations narrows as model size increases, suggesting that larger models redistribute representational capacity to compensate for restricted attention patterns. This has direct implications for the compute-optimal design of long-context models.

5 Conclusion

Sparsity ratio should be treated as a scale-dependent hyperparameter rather than a fixed architectural choice.

Acknowledgments

We thank the systems team for compute infrastructure support and colleagues for feedback on early drafts.

References

[1] Kaplan, J. et al. Scaling Laws for Neural Language Models. 2020.

[2] Hoffmann, J. et al. Training Compute-Optimal Large Language Models. 2022.