**Abstract**

This paper investigates dynamic activation weighting in transformer feedforward networks. We evaluate a dual-pathway architecture combining SiLU and GELU activations with learned weights. Experiments on an 83M parameter model show our approach achieves 5.124 validation loss, underperforming the SwiGLU baseline (4.927) while using more memory. The results suggest current implementations of dynamic weighting may not outperform simpler approaches.

# 1 Introduction

Transformer architectures rely heavily on their feedforward components. We examine whether dynamically weighting between activation functions can improve performance compared to fixed approaches.

# 2 Method

Our architecture processes inputs through parallel SiLU and GELU pathways, combined using learned weights from a softmax layer. The model was trained on FineWeb with standard transformer optimization settings.

# 3 Results

Key results:

- Validation loss: 5.124 (SwiGLU baseline: 4.927)
- Memory usage: 39.5GB (vs 31.5GB for baseline)
- Training time: comparable to baseline

# 4 Discussion

The results indicate that:

- Dynamic weighting adds computational overhead
- The benefits may require larger model scales
- Simple fixed activations remain competitive

# 5 Conclusion

While dynamic activation weighting shows theoretical promise, our implementation did not outperform established baselines. Future work should explore alternative gating mechanisms and larger scales.