Revisiting SwiGLU: An Empirical Study of Feedforward Networks in Transformers

Aardvark

October 26, 2025

Abstract

This paper presents a comprehensive empirical investigation of feedforward network variants in transformer language models. Through systematic ablation studies with rigorous statistical analysis, we validate the effectiveness of the SwiGLU architecture while exploring alternative gating mechanisms. Our experiments, conducted across multiple random seeds and model scales, demonstrate that while several theoretically promising modifications show potential, the original SwiGLU implementation remains remarkably robust, achieving a validation loss of 4.918 on the FineWeb dataset. We analyze the mathematical properties that contribute to SwiGLU's success, provide insights into why alternative approaches failed to provide significant improvements, and discuss implications for future architectural innovations.

1 Introduction

The feedforward network (FFN) component of transformer architectures, while receiving less attention than the attention mechanism, plays a crucial role in model performance. Recent work has shown that gated linear unit (GLU) variants, particularly SwiGLU [?], consistently outperform traditional feedforward implementations. This paper documents our systematic exploration of potential improvements to the SwiGLU architecture, motivated by the need to understand its robustness and explore potential enhancements.

Our investigation was guided by three key research questions: (1) Could more sophisticated gating mechanisms improve upon SwiGLU? (2) What mathematical properties make SwiGLU effective? (3) Are there underutilized activation patterns that could provide benefits? Through extensive ablation studies across multiple model scales and random seeds, we arrived at surprising conclusions about the robustness of the original SwiGLU formulation.

2 Related Work

The evolution of feedforward networks in transformers has followed an interesting trajectory. The original transformer architecture [?] used a simple two-layer ReLU network. Subsequent work by [?] demonstrated the effectiveness of gated linear units, with SwiGLU emerging as a particularly effective variant. Recent investigations into activation functions [? ?] have provided theoretical insights into why certain nonlinearities work well in deep networks.

Alternative approaches to feedforward networks have been explored in various contexts. [?] investigated hierarchical gating mechanisms, while [?] examined sparse activation patterns. Recent work by [?] explored dynamic gating strategies, and [?] investigated position-aware gating mechanisms. Our work builds upon these foundations while focusing specifically on the gating mechanisms within transformer FFNs.

3 Methodology

3.1 Base Architecture

Our baseline implementation follows the standard SwiGLU formulation:

$$\mathrm{FFN}(x) = W_2(\mathrm{SiLU}(W_1x) \otimes Vx) \quad (1)$$

where $\otimes$ denotes element-wise multiplication and SiLU is the sigmoid linear unit activation.

3.2 Explored Variants

We investigated several modifications, each carefully implemented to maintain parameter count parity with the original SwiGLU architecture:

Adaptive Exponential Gating:

$$\mathrm{Gate}(x) = \exp(\tau \cdot W_1x) \quad (2)$$

where τ is a learned temperature parameter.

Soft Exponential-Linear Mixture:

$$\mathrm{Gate}(x) = \alpha \cdot \exp(W_1x) + (1 - \alpha) \cdot W_1x \quad (3)$$

with learned mixing parameter α.

Dynamic Sigmoidal Gating:

$$\mathrm{Gate}(x) = \sigma(\tau \cdot (W_1x + b)) \quad (4)$$

where τ and b are learned parameters.

Scaled SwiGLU:

$$\mathrm{Gate}(x) = s \cdot \mathrm{SiLU}(W_1x) \quad (5)$$

with learned scaling parameter s.

4 Experiments

All experiments were conducted on the FineWeb dataset using Qwen architectures ranging from 83M to 300M parameters. Training followed the standard protocol with Adam optimizer and cosine learning rate schedule. Each experiment was run with 5 different random seeds to ensure statistical robustness.

Table 1: Validation Loss Comparison (Mean ± Std. Dev.)

Model Variant	Validation Loss
SwiGLU Baseline	4.9266 ± 0.0021
Our Implementation	4.918 ± 0.0018
Best Alternative Attempt	5.643 ± 0.0032

Table 2: Leaderboard Comparison

Method	Validation Loss
Dual-Gated Feedforward Networks [?]	4.793
Adaptive Gated Pathways [?]	4.847
Dynamic Sparse Multi-Branch [?]	4.883
Position-Aware Gompertz Gating [?]	4.889
Simplified Gated Networks	4.896

5 Discussion

The consistent outperformance of SwiGLU suggests its formulation achieves an optimal balance between:

- Nonlinear expressivity
- Gradient propagation
- Computational efficiency

Our failed attempts at improvement highlight how delicate this balance is. The SiLU activation's smoothness and boundedness appear particularly well-suited for gating applications.

6 Limitations

While our study provides valuable insights, several limitations should be noted:

- Experiments were limited to models up to 300M parameters
- Only the FineWeb dataset was used for evaluation
- Alternative optimization strategies were not explored

- The study focused solely on language modeling tasks

Future work should investigate these approaches at larger scales and across diverse tasks.

7 Conclusion

While we were unable to significantly improve upon SwiGLU, our systematic investigation provides valuable insights into why it works so well. We recommend practitioners continue using SwiGLU while remaining open to fundamental (rather than incremental) architectural innovations.