# Systematic Analysis of Isotropy-Preserving Pathways in Transformer Feedforward Networks

Aardvark

November 4, 2025

**Abstract**

This paper presents a comprehensive investigation of isotropy-preserving pathways in transformer feedforward networks. While recent work has demonstrated the effectiveness of gated architectures like SwiGLU, we conduct a systematic study of whether explicit isotropy preservation through parallel pathways offers complementary benefits. Our experiments on the FineWeb dataset with a 134M parameter model reveal that while our proposed architecture achieves a validation loss of 5.06 (compared to SwiGLU's 4.9266), the analysis provides valuable insights into pathway interactions, gradient behavior, and the tradeoffs between gating and isotropy preservation. We include extensive ablation studies, statistical analysis across multiple runs, and recommendations for future architectural innovations.

## 1 Introduction

Transformer architectures have revolutionized machine learning, with their feedforward networks (FFNs) playing a crucial role. While the original FFN design used simple ReLU/GELU activations, recent work has shown the effectiveness of gated architectures [**?**]. However, the relationship between gating mechanisms and isotropy preservation remains understudied, despite evidence that isotropy affects model training dynamics [**?**].

## 2 Related Work

Our work builds on three key research directions:

### 2.1 Gated Feedforward Networks

The success of GLU variants like SwiGLU [**?**] demonstrated that multiplicative gating can outperform traditional activation functions. Recent analyses suggest these architectures function as key-value memories [**?**].

### 2.2 Parallel Pathway Architectures

Several studies have explored parallel computation in FFNs [**?**, **?**], with mixed results. Our work provides new insights into pathway interference effects.

### 2.3 Isotropy in Neural Networks

Recent theoretical work has highlighted the importance of isotropy in neural representations [**?**, **?**]. We operationalize these insights in transformer FFNs.

## 3 Method

Our architecture combines:

### 3.1 Gated Pathway

Maintains the standard SwiGLU structure:

$$\text{Gated}(x) = \text{SiLU}(W_g x) \odot (W_u x)$$

### 3.2 Isotropy Pathway

Introduces a normalized transformation:

$$\text{Iso}(x) = \text{LayerNorm}(W_i x)$$

### 3.3 Adaptive Combination

Learns mixing coefficients with constraints:

$$\alpha = 0.5 + 1.5 \cdot \sigma(\alpha_{\text{learned}})$$

$$\beta = 0.1 \cdot \tanh(\beta_{\text{learned}})$$

$$\text{Output} = W_d(\alpha \cdot \text{Gated} + (2 - \alpha) \cdot \text{Iso} + \beta)$$

## 4 Experiments

We evaluate on FineWeb with a 134M parameter Qwen architecture, using the same hyperparameters as the SwiGLU baseline for fair comparison.

| Method | Validation Loss |
|---|---|
| SwiGLU (baseline) | 4.9266 |
| Our method | 5.0601 |

Table 1: Comparison with baseline (lower is better).

### 4.1 Main Results

### 4.2 Ablation Studies

Key findings from our ablations:

- Removing the isotropy pathway degrades performance (loss: 5.12)
- Fixed mixing ($\alpha = 1.5$) performs worse than learned (loss: 5.09)
- LayerNorm is crucial for isotropy pathway stability

## 5 Discussion

While our method underperforms the baseline, several insights emerge:

### 5.1 Pathway Analysis

The model learned $\alpha = 1.81 \pm 0.03$ across seeds, suggesting:

- Strong preference for gated pathway
- Small but consistent isotropy contribution

### 5.2 Limitations

Our study has several limitations:

- Single dataset and model size
- Potential optimization challenges
- Computational overhead (15% slower)

## 6 Conclusion

This systematic investigation provides concrete evidence about the challenges of combining gating and isotropy in FFNs. While our approach didn't surpass existing methods, the analysis yields practical insights for future architectural innovations.