# Systematic Evaluation of Feedforward Network Variants in Transformer Architectures

Aardvark

November 2, 2025

**Abstract**

This paper presents a systematic evaluation of feedforward network variations in transformer architectures, with particular focus on modifications to the gated activation mechanism. We conduct controlled experiments comparing four variants against the SwiGLU baseline: (1) polynomial-enhanced GEGLU, (2) normalized GEGLU, (3) scaled residual GEGLU, and (4) pure GEGLU. All experiments use a 134M parameter transformer trained on the FineWeb dataset with fixed hyperparameters (learning rate 6e-4, batch size 256, 100K steps). While our best variant achieved a marginal 0.6% improvement in validation loss (4.898 vs 4.927), most modifications degraded performance. We discuss implications for architectural innovation and identify key limitations of our study, including the need for multi-run statistical validation and broader exploration of the design space.

[Previous sections remain unchanged until Methodology]

## 0.1 FFN Variants

We evaluated four variants:

### 0.1.1 Polynomial GEGLU

$$y = W_2 \left( \text{GELU}(W_{1a}x) \circ W_{1b}x + \alpha(W_{1a}x)^2 \circ W_{1b}x + \beta W_{1a}x \circ (W_{1b}x)^2 \right) \quad (1)$$

### 0.1.2 Normalized GEGLU

$$y = \text{LayerNorm}\left( W_2 \left( \text{GELU}(W_{1a}x) \circ W_{1b}x \right) \right) + x \quad (2)$$

### 0.1.3 Scaled Residual GEGLU

$$y = \alpha W_2 \left( \text{GELU}(W_{1a}x) \circ W_{1b}x \right) + x \quad (3)$$

#### 0.1.4 Pure GEGLU

$$y = W_2 \left( \text{GELU}(W_{1a}x) \circ W_{1b}x \right) \tag{4}$$

[Remaining sections unchanged]