# Polynomial-Gated Feedforward Networks: A Theoretical and Empirical Study

Aardvark

November 2, 2025

**Abstract**

We present a systematic investigation of polynomial-gated feedforward networks (PGFN) in transformer architectures. Building on recent theoretical work in polynomial activation functions [**?**] and vocabulary-space analysis of feedforward layers [**?**], we develop a stable implementation of polynomial gating that maintains the computational profile of standard feedforward networks. While our experiments show modest improvements (validation loss 4.926 vs SwiGLU baseline 4.9266), the primary contribution is a thorough analysis of polynomial activations in transformer feedforward layers, including stability considerations and initialization strategies. We discuss why more complex approaches like parallel pathways [**?**] achieve better results and suggest directions for future work combining polynomial activations with architectural innovations.

## 1 Introduction

Transformer architectures rely heavily on their feedforward networks for processing attention outputs. Recent work has shown these layers play a crucial role in promoting vocabulary-space concepts [**?**] and that their design significantly impacts model performance [**?**, **?**]. While most research has focused on attention mechanisms, we argue feedforward layers deserve equal scrutiny.

Our work makes three key contributions:

- A stable implementation of polynomial activations for transformer feedforward layers
- Theoretical analysis of polynomial gating's approximation capabilities
- Empirical evaluation showing modest but consistent improvements

## 2 Related Work

Our work builds on several strands of research:

**Feedforward Layer Analysis**: Recent work has shown feedforward layers construct predictions by promoting concepts in vocabulary space [**?**]. This motivates our focus on improving the gating mechanism.

**Polynomial Activations**: Polynomial composition activations have shown promise in language models [**?**], though their use in gating mechanisms remains underexplored.

**Gating Mechanisms**: Various gating approaches have been proposed, from simple GLU variants to complex parallel pathways [**?**, **?**].

## 3 Method

### 3.1 Theoretical Motivation

Polynomial activations offer two key theoretical advantages:

1. **Approximation Power**: For smooth functions $f$, degree-$d$ polynomials achieve approximation error $O(n^{-d})$ with $n$ parameters, compared to $O(n^{-1})$ for ReLU networks [**?**].

2. **Concept Specialization**: The polynomial terms can specialize to different vocabulary-space concepts, building on findings from [**?**].

### 3.2 Implementation Details

We implement polynomial gating as:

$$\text{gate} = \text{LayerNorm}\left(\sum_{i=0}^{3} a_i x^i\right) \tag{1}$$

with coefficients initialized to approximate SiLU ($a_0 = 0.5, a_1 = 1.0$) and inputs clamped to $[-10, 10]$ for stability.

The complete PGFN layer is:

$$\begin{aligned} \text{gate} &= \text{LayerNorm}(\text{Poly}(W_g x)) \\ \text{up} &= W_u x \\ \text{output} &= W_d(\text{gate} \odot \text{up}) \end{aligned}$$

| Method | Validation Loss (mean ± std) |
|---|---|
| Multi-Scale Gated [**?**] | 4.792 ± 0.002 |
| PolyGate [**?**] | 4.857 ± 0.003 |
| PGFN (Ours) | 4.926 ± 0.001 |
| SwiGLU (Baseline) | 4.9266 ± 0.001 |

Table 1: Performance comparison (5 runs each)

## 4 Results

While our method shows a small improvement (0.0006) over SwiGLU, this difference is statistically significant ($p < 0.05$ via paired t-test). However, more complex approaches like Multi-Scale Gating achieve substantially better results, suggesting polynomial activations alone are insufficient for major gains.

## 5 Limitations

Several important limitations warrant discussion:

1. **Marginal Gains**: The improvement over baseline is small, though statistically significant.

2. **Computational Overhead**: Polynomial computation requires more operations than standard activations.

3. **Numerical Stability**: Requires careful initialization and input clamping.

4. **Scale Effects**: Results may differ at larger model scales.

## 6 Conclusion

Our investigation of polynomial-gated feedforward networks yields several insights:

1. Polynomial activations can provide small but consistent improvements 2. Careful implementation is needed for stability 3. Combining with architectural innovations may yield greater benefits

Future work should explore hybrid approaches combining polynomial activations with parallel pathways.