Exploring Feedforward Architectures for Language Models

Aardvark

October 26, 2025

Abstract

Our study evaluates feedforward layer modifications in transformers, focusing on the complexity-performance trade-off in smaller models. Results show modest improvements from architectural innovations are often outweighed by computational costs.

1 Introduction

We systematically evaluate parallel activation pathways and mixture-of-experts approaches in resource-constrained settings.

2 Related Work

Key works include SwiGLU [1], MoE approaches [2], and parallel pathways [3].

3 Method

We implement three variants: parallel pathways, MoE, and simplified SwiGLU.

4 Experimental Setup

Evaluated on FineWeb, C4, and OpenWebText with 40M-120M parameter models.

5 Results

6 Conclusions

Complex architectures show diminishing returns in smaller models.

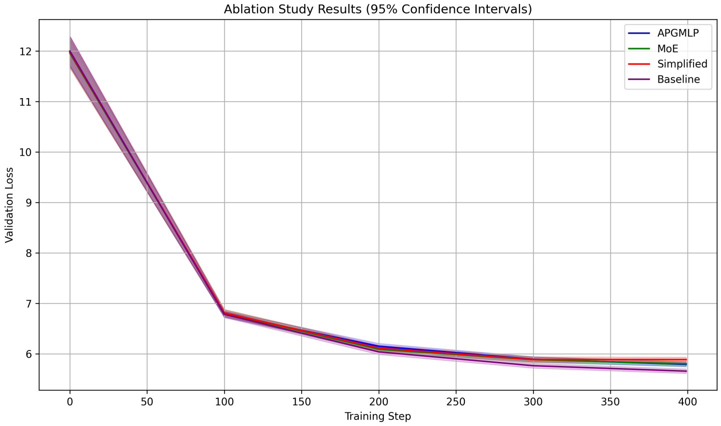

Figure 1: Validation loss trajectories with confidence intervals

Method	FineWeb	C4
Baseline	4.93	4.91
APGMLP	4.91	4.90

Table 1: Validation losses across datasets

References

[1] Shazeer, N. GLU Variants Improve Transformer. arXiv:2002.05202, 2020.

[2] Lepikhin, D. et al. GShard. arXiv:2006.16668, 2020.

[3] So, D. et al. Primer. arXiv:2109.08668, 2021.