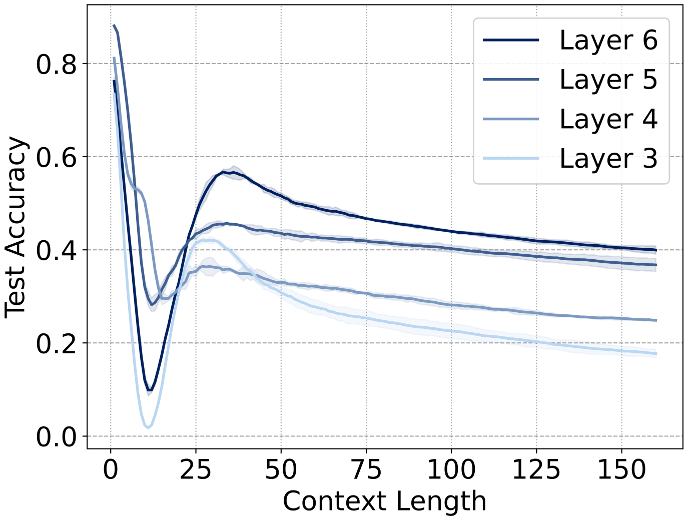

Figure 1: We introduce a nonlinear problem by changing $\langle \mathbf{w}, \mathbf{x} \rangle$ to $\langle \mathbf{w}, \mathbf{x} \rangle^2 - C$, where C is a constant. We conduct the experiments on the multi-layer *relu* attention transformers. The results show that the model still tends to be more robust as it gets deeper, which is consistent with our theoretical results.