Massive Tokens without the Attention Mechanism Baseline Masked w/ bias 250 Masked w/o bias Block Output Norm 50 10 11 12 13 Layer