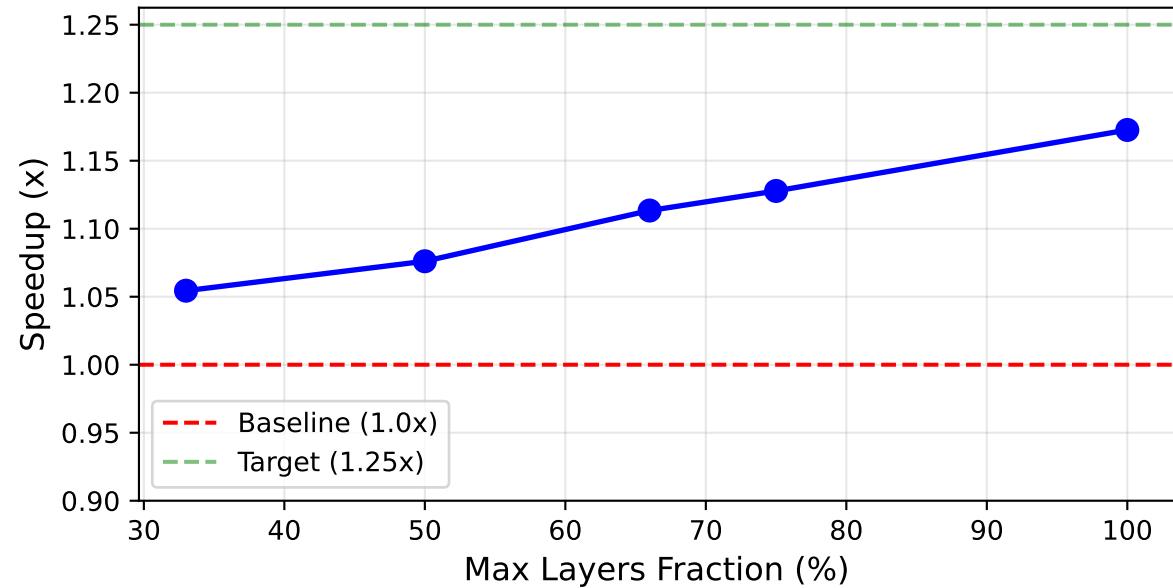
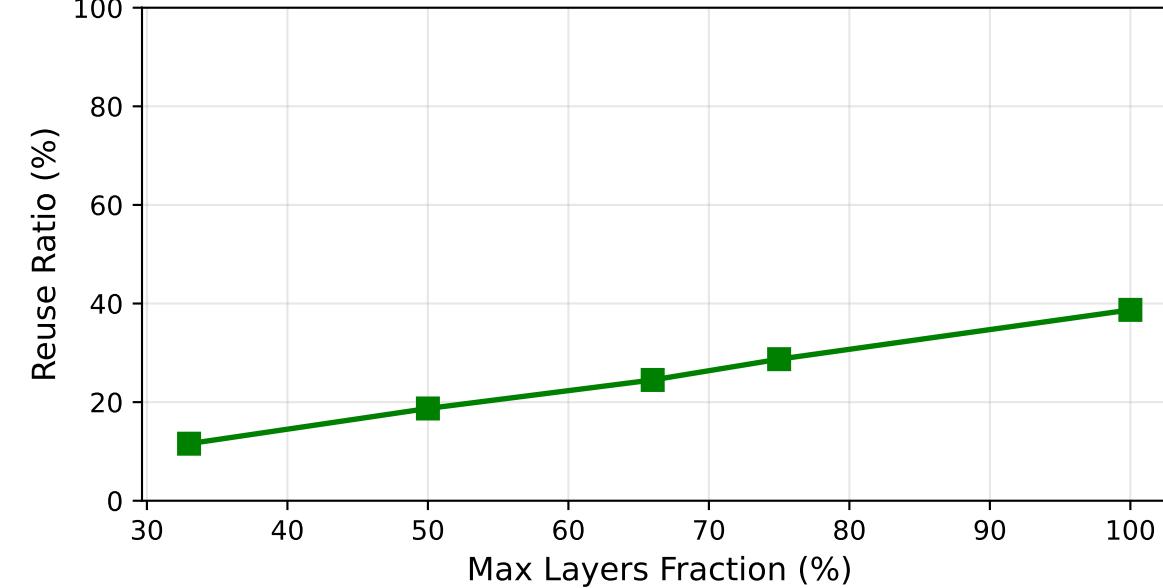


DQAR Layer Fraction Sweep Results (warmup=20%)

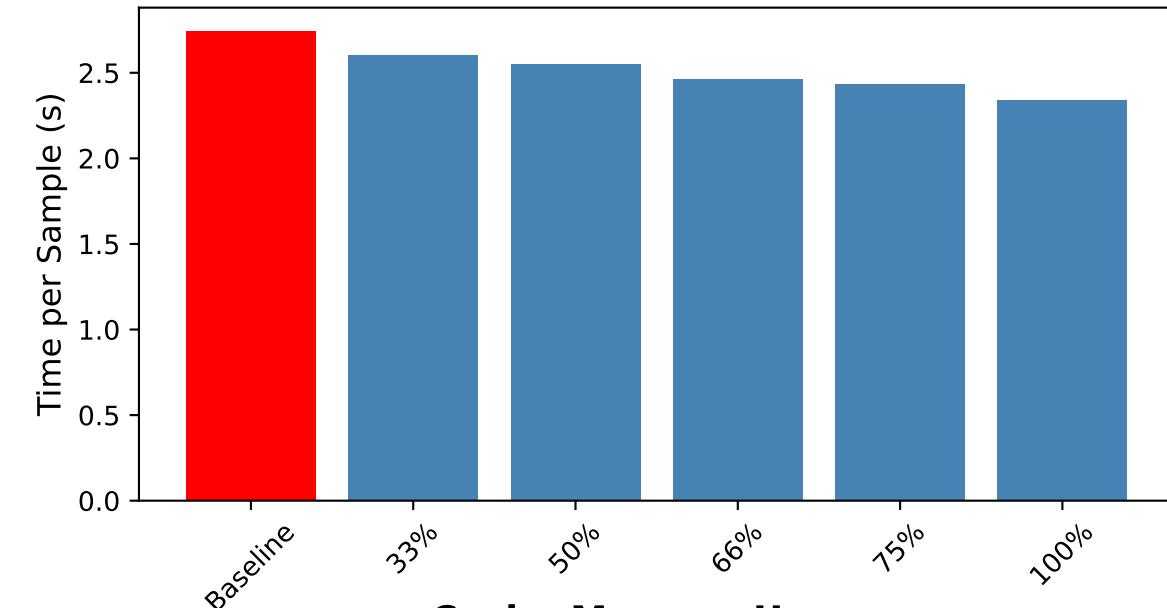
Speedup vs Layer Fraction (warmup=20%)



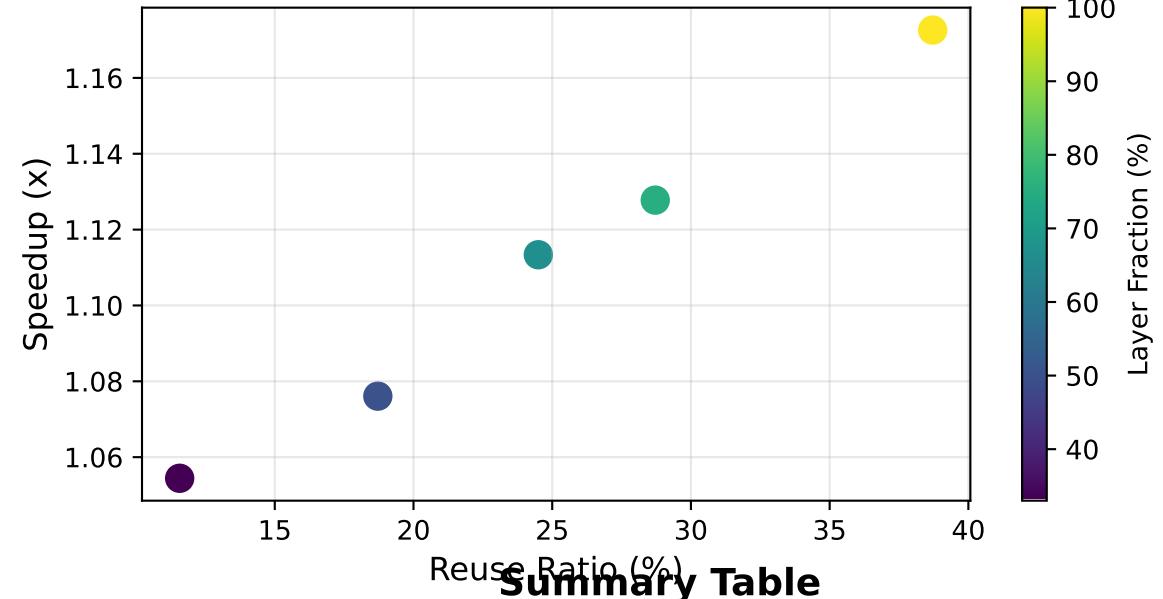
Attention Reuse vs Layer Fraction



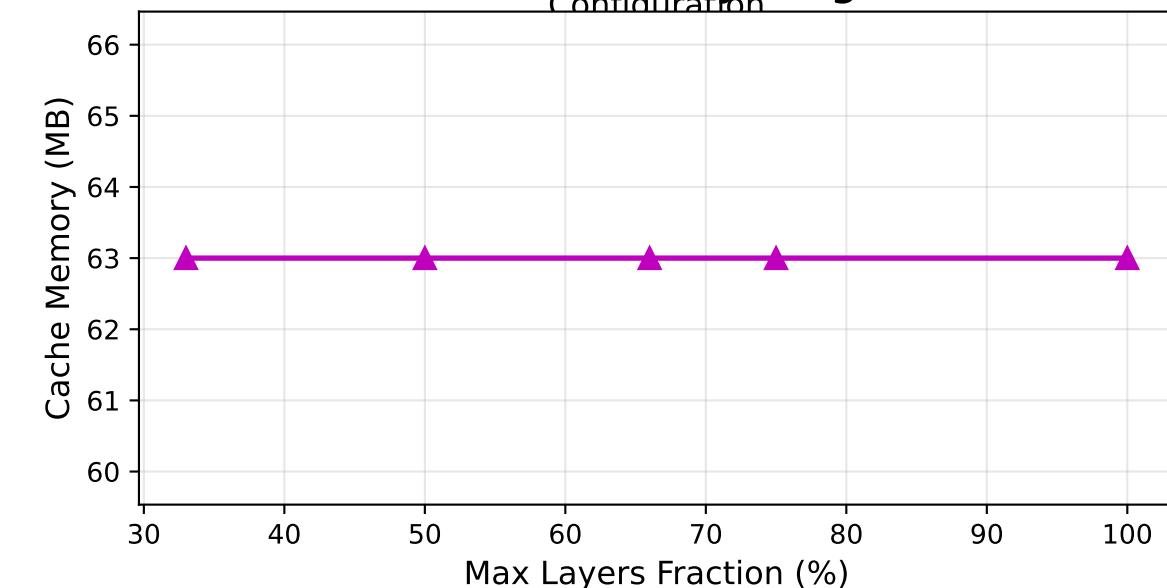
Inference Time Comparison



Speed-Quality Trade-off



Cache Memory Usage



Config	Layers	Speedup	Reuse %	Time (s)
Baseline	-	1.00x	0%	2.74
layers_33pct	33%	1.05x	11.6%	2.60
layers_50pct	50%	1.08x	18.7%	2.55
layers_66pct	66%	1.11x	24.5%	2.46
layers_75pct	75%	1.13x	28.7%	2.43
layers_100pct	100%	1.17x	38.7%	2.34