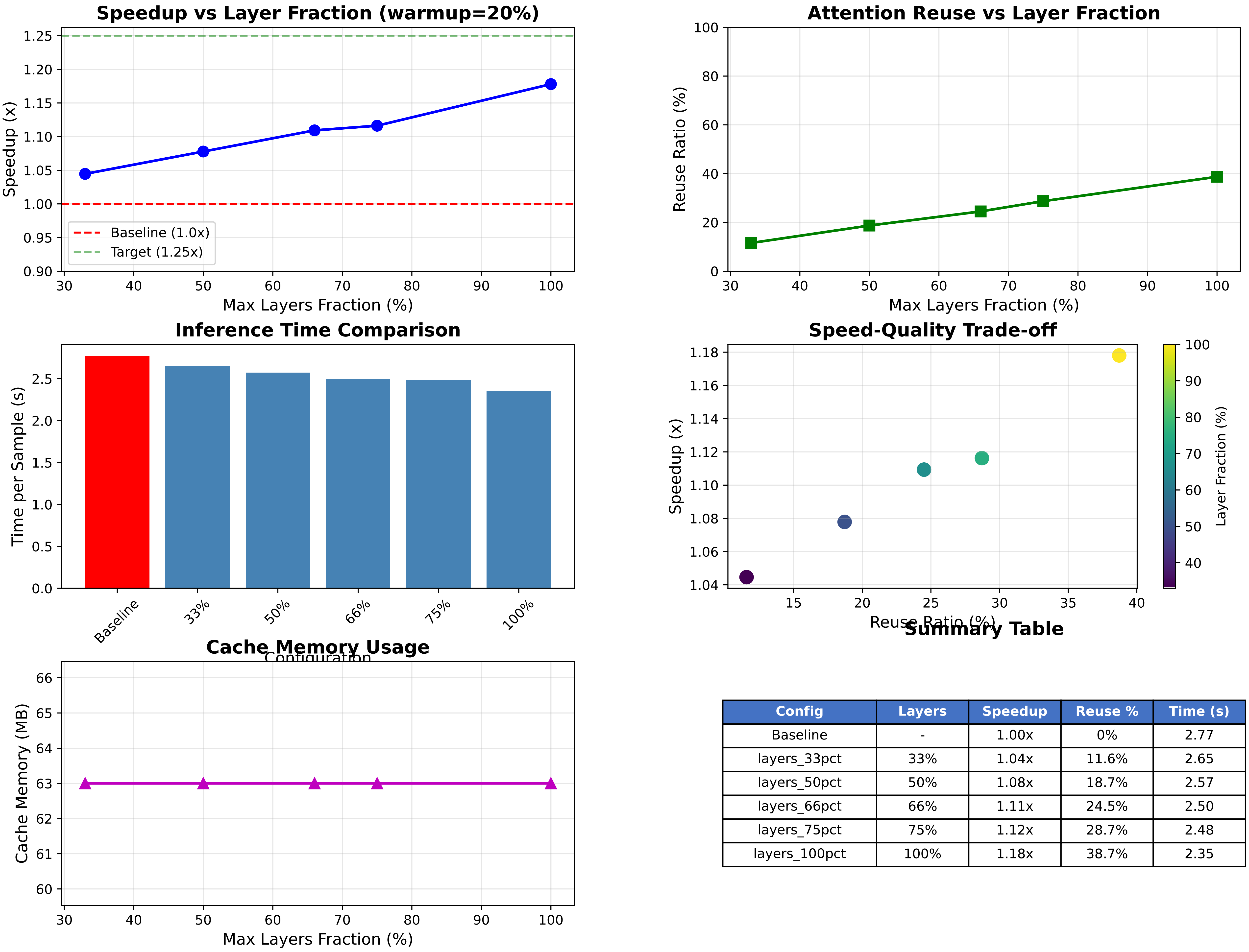
DQAR Layer Fraction Sweep Results (warmup=20%)



Speedup vs Layer Fraction (warmup=20%)

Attention Reuse vs Layer Fraction

Inference Time Comparison

Speed-Quality Trade-off

Cache Memory Usage

Summary Table

Config	Layers	Speedup	Reuse %	Time (s)
Baseline	-	1.00x	0%	2.77
layers_33pct	33%	1.04x	11.6%	2.65
layers_50pct	50%	1.08x	18.7%	2.57
layers_66pct	66%	1.11x	24.5%	2.50
layers_75pct	75%	1.12x	28.7%	2.48
layers_100pct	100%	1.18x	38.7%	2.35