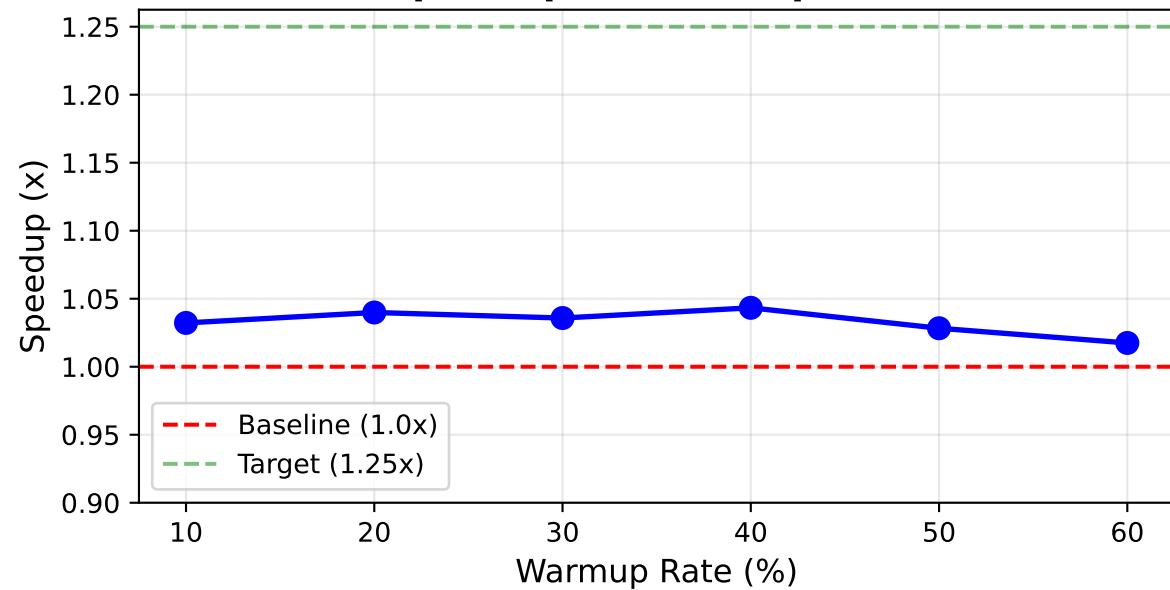
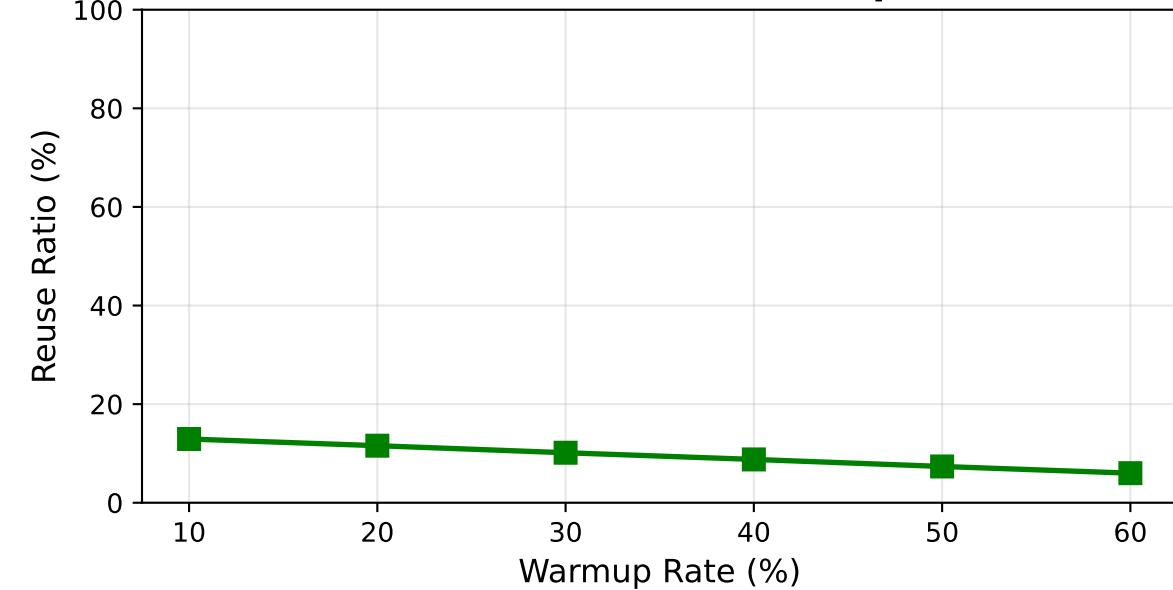


# DQAR Warmup Rate Sweep Results

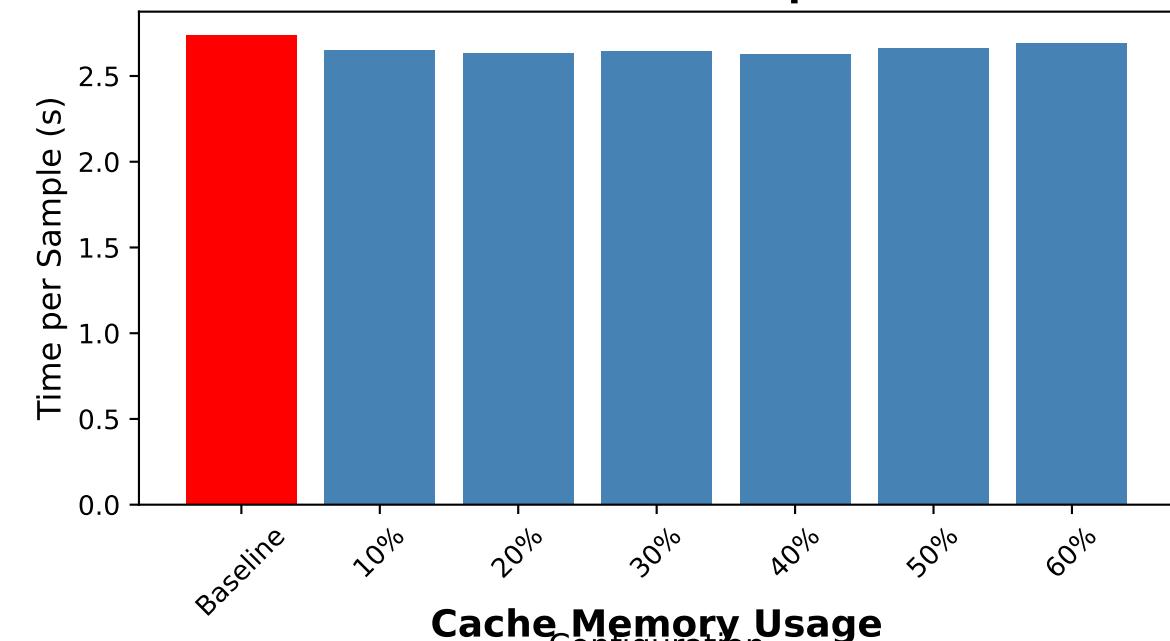
## Speedup vs Warmup Rate



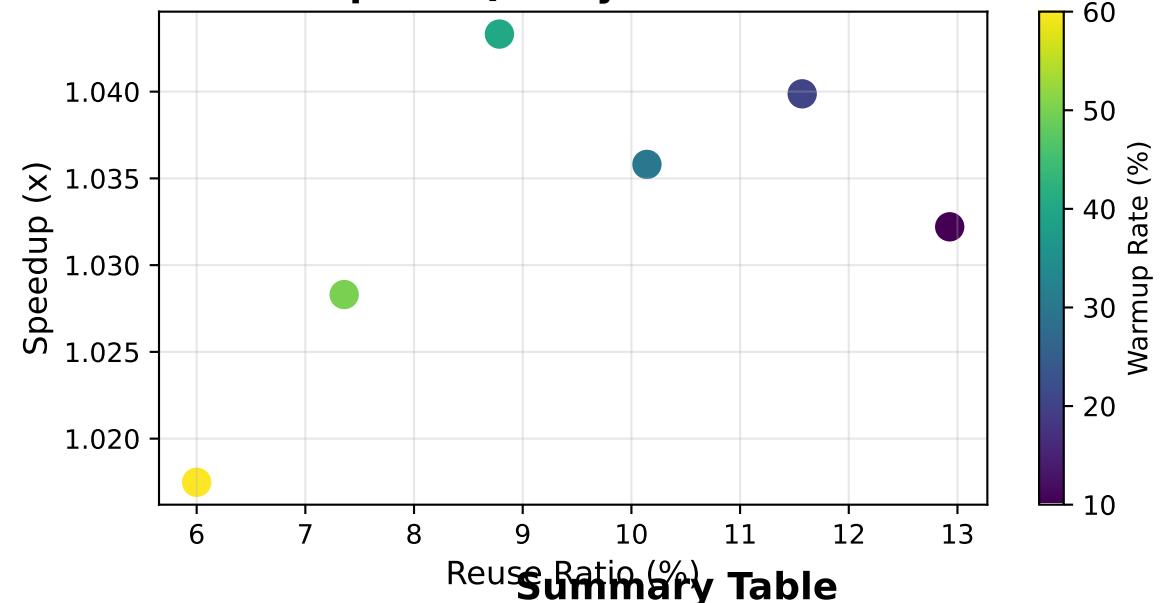
## Attention Reuse vs Warmup Rate



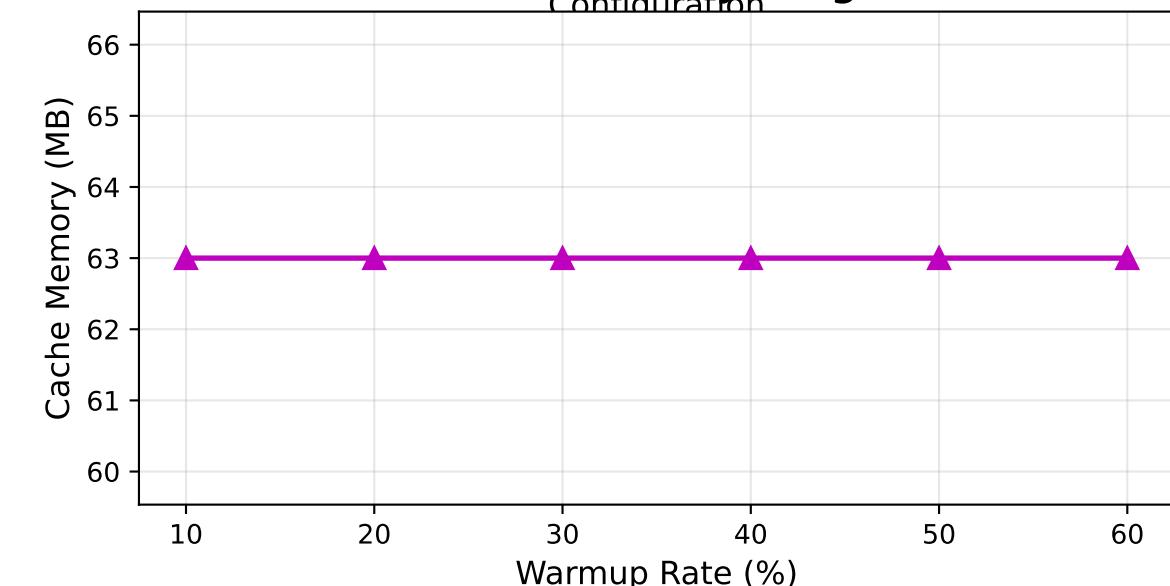
## Inference Time Comparison



## Speed-Quality Trade-off



## Cache Memory Usage



Config	Warmup	Speedup	Reuse %	Time (s)
Baseline	-	1.00x	0%	2.74
warmup_10pct	10%	1.03x	12.9%	2.65
warmup_20pct	20%	1.04x	11.6%	2.63
warmup_30pct	30%	1.04x	10.1%	2.64
warmup_40pct	40%	1.04x	8.8%	2.62
warmup_50pct	50%	1.03x	7.4%	2.66
warmup_60pct	60%	1.02x	6.0%	2.69