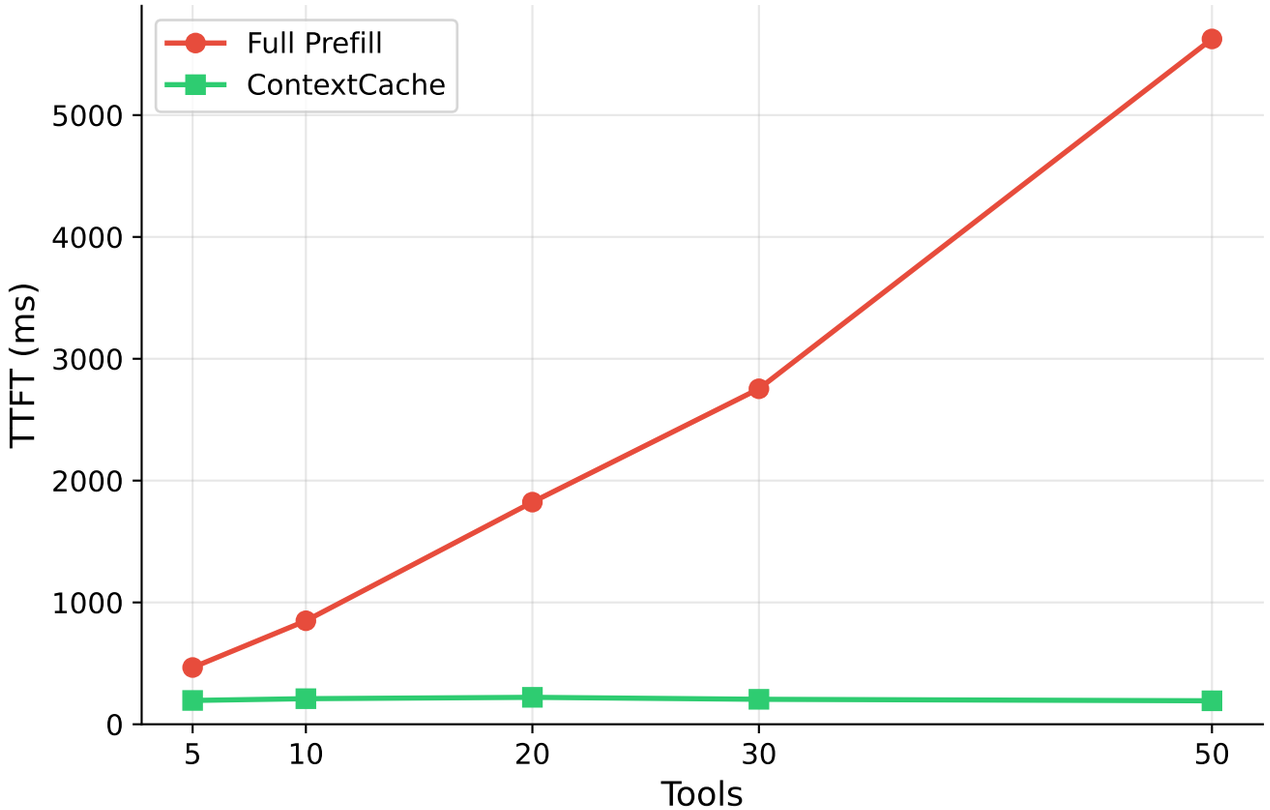
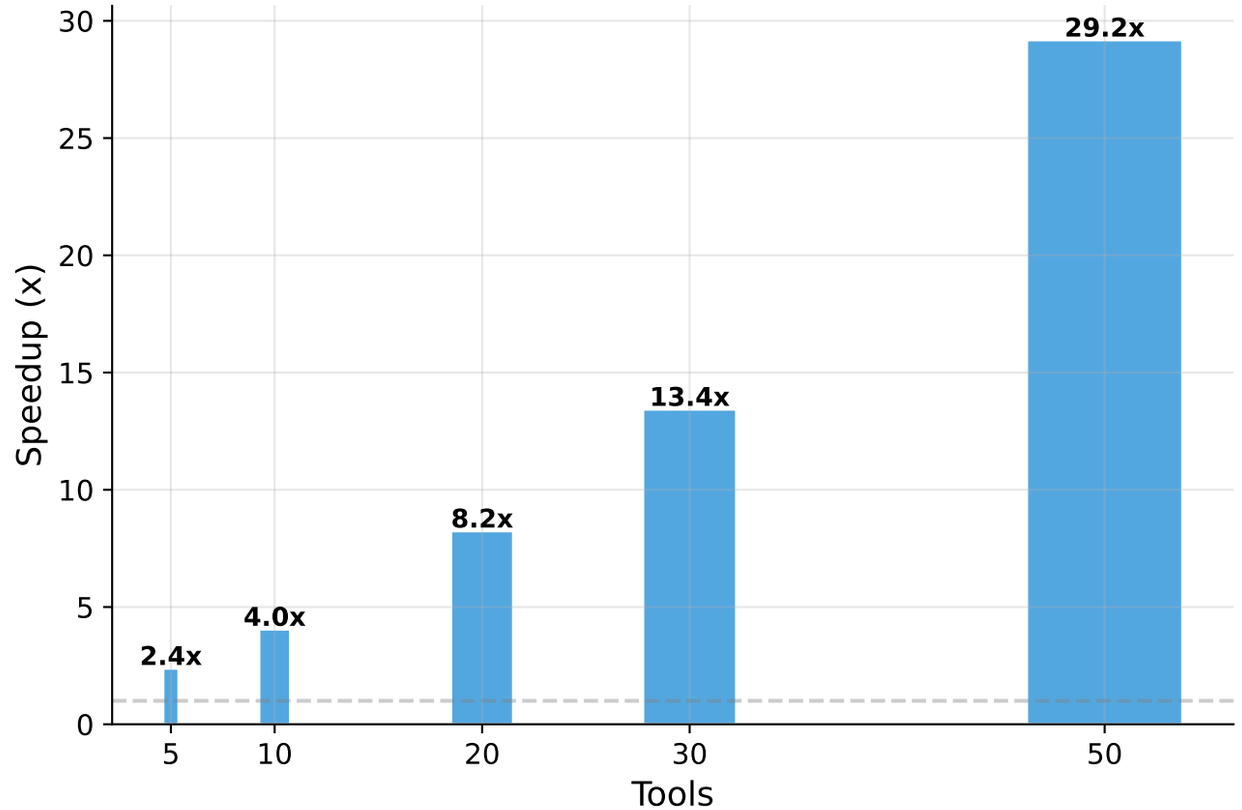
ContextCache Scaling: 5 to 100 Tools (Qwen3-8B, 4-bit NF4, RTX 3090 Ti)

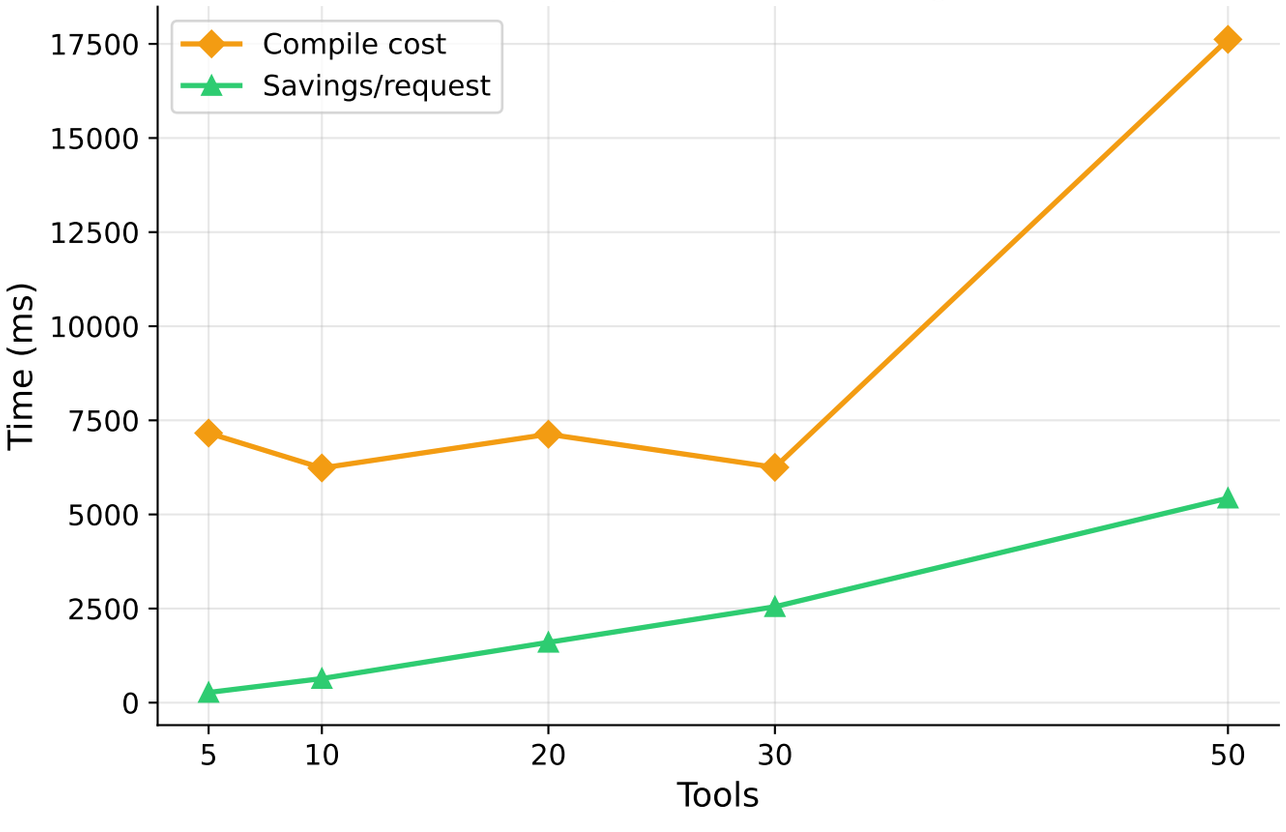
(a) Time to First Token



(b) TTFT Speedup



(c) Compile Cost vs Savings



(d) GPU Cache Size