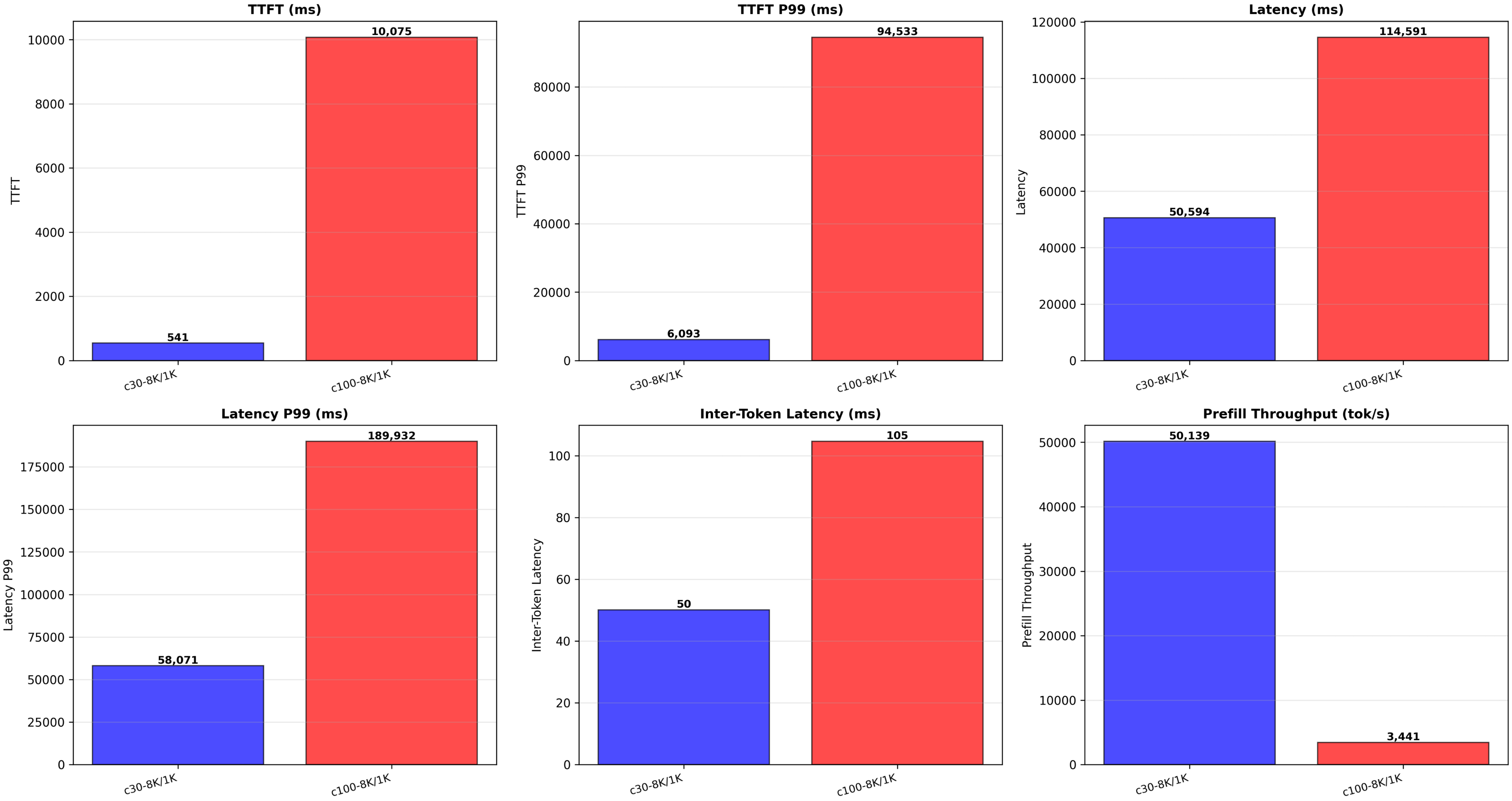
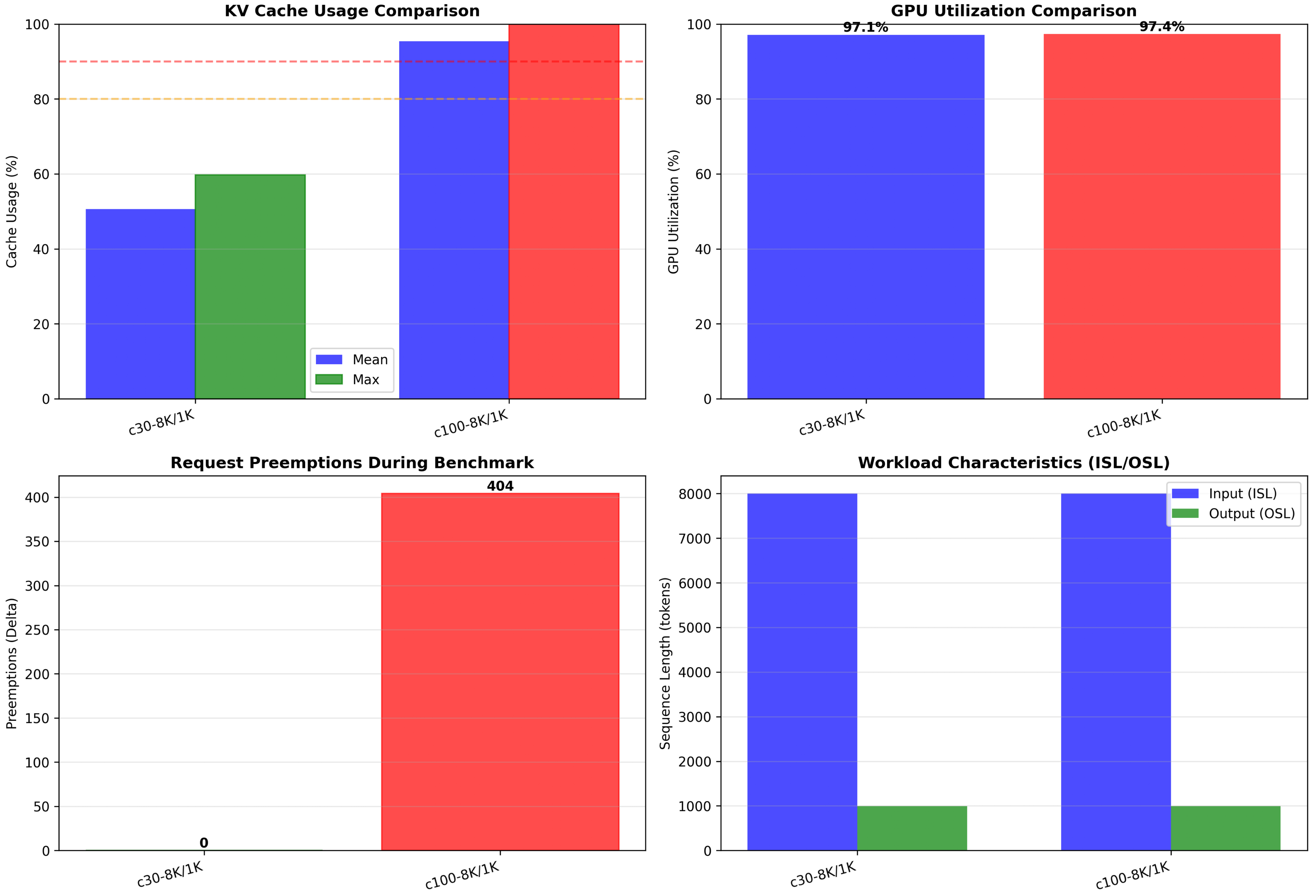
Performance Comparison Across Runs



Resource Utilization Comparison

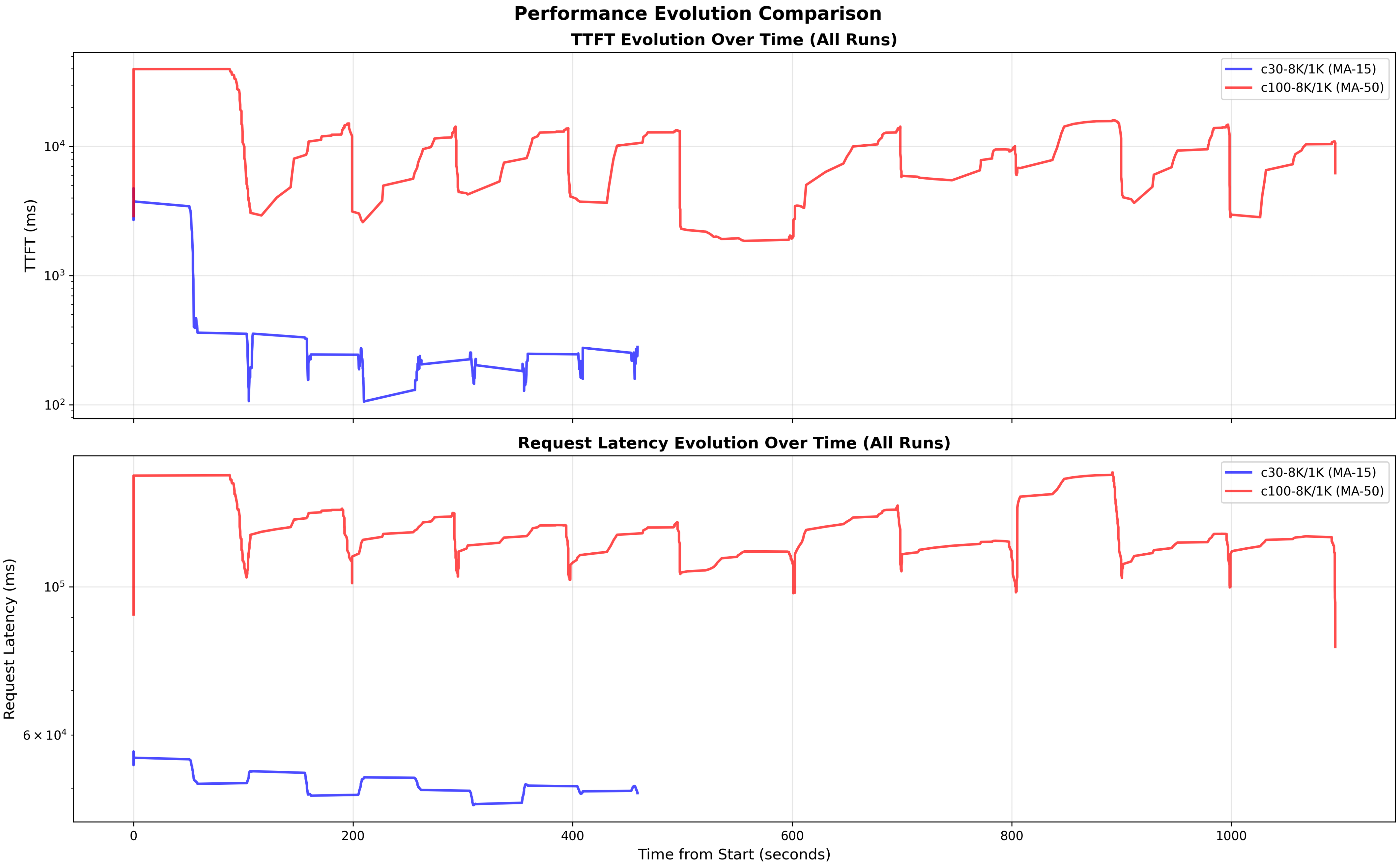


KV Cache Usage Comparison

GPU Utilization Comparison

Request Preemptions During Benchmark

Workload Characteristics (ISL/OSL)

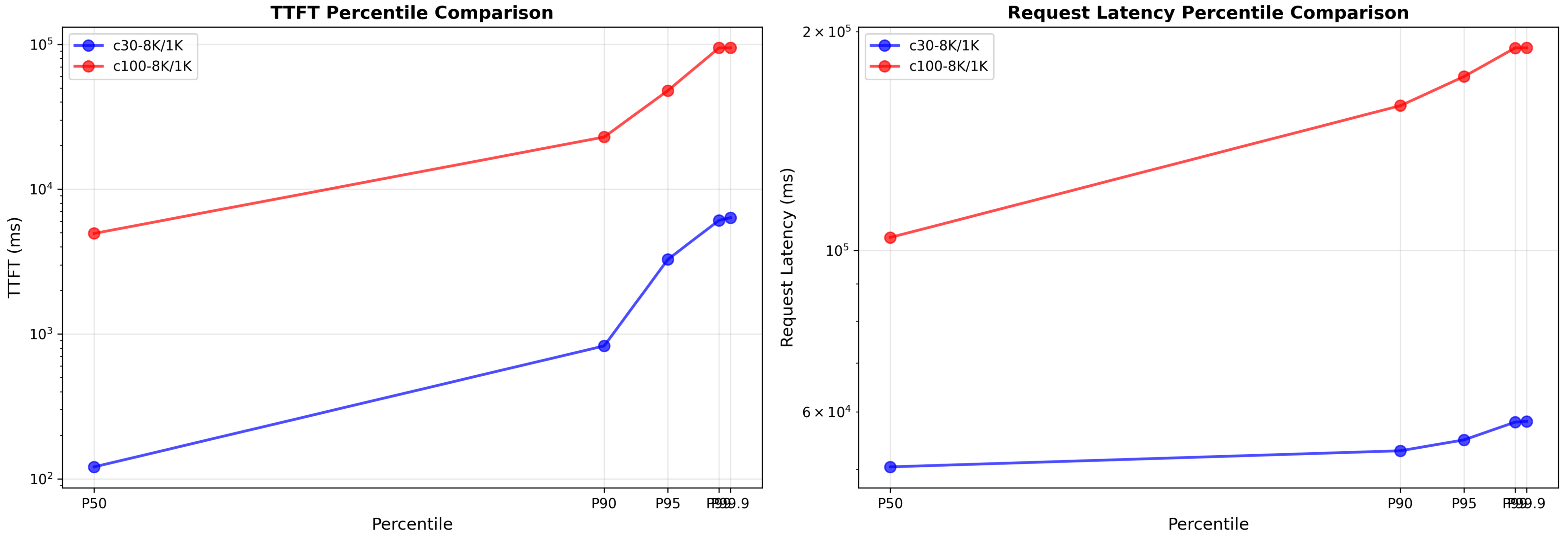


Performance Evolution Comparison

TTFT Evolution Over Time (All Runs)

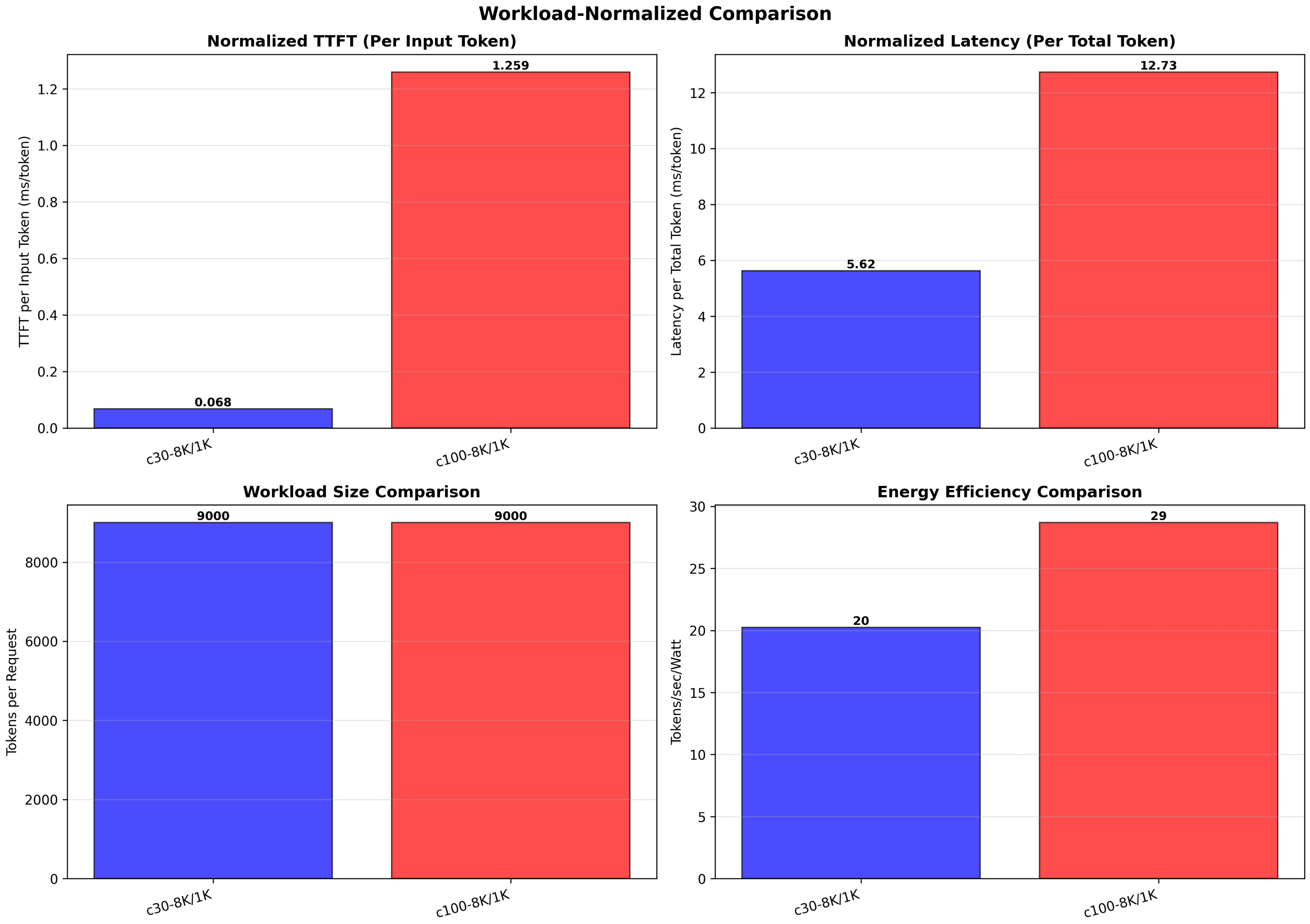
Request Latency Evolution Over Time (All Runs)

Percentile Distribution Comparison



TTFT Percentile Comparison

Request Latency Percentile Comparison



Workload-Normalized Comparison

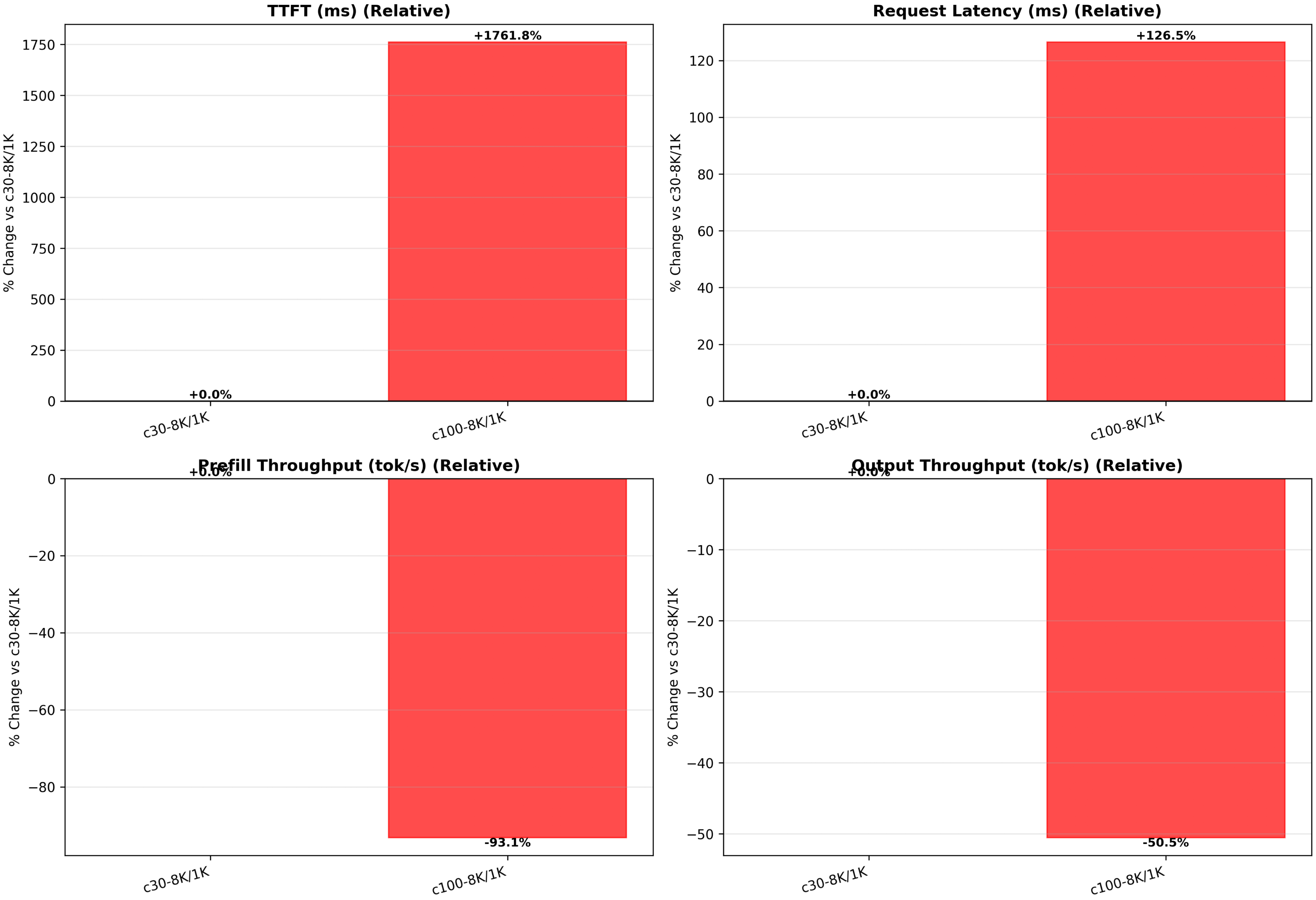
Normalized TTFT (Per Input Token)

Normalized Latency (Per Total Token)

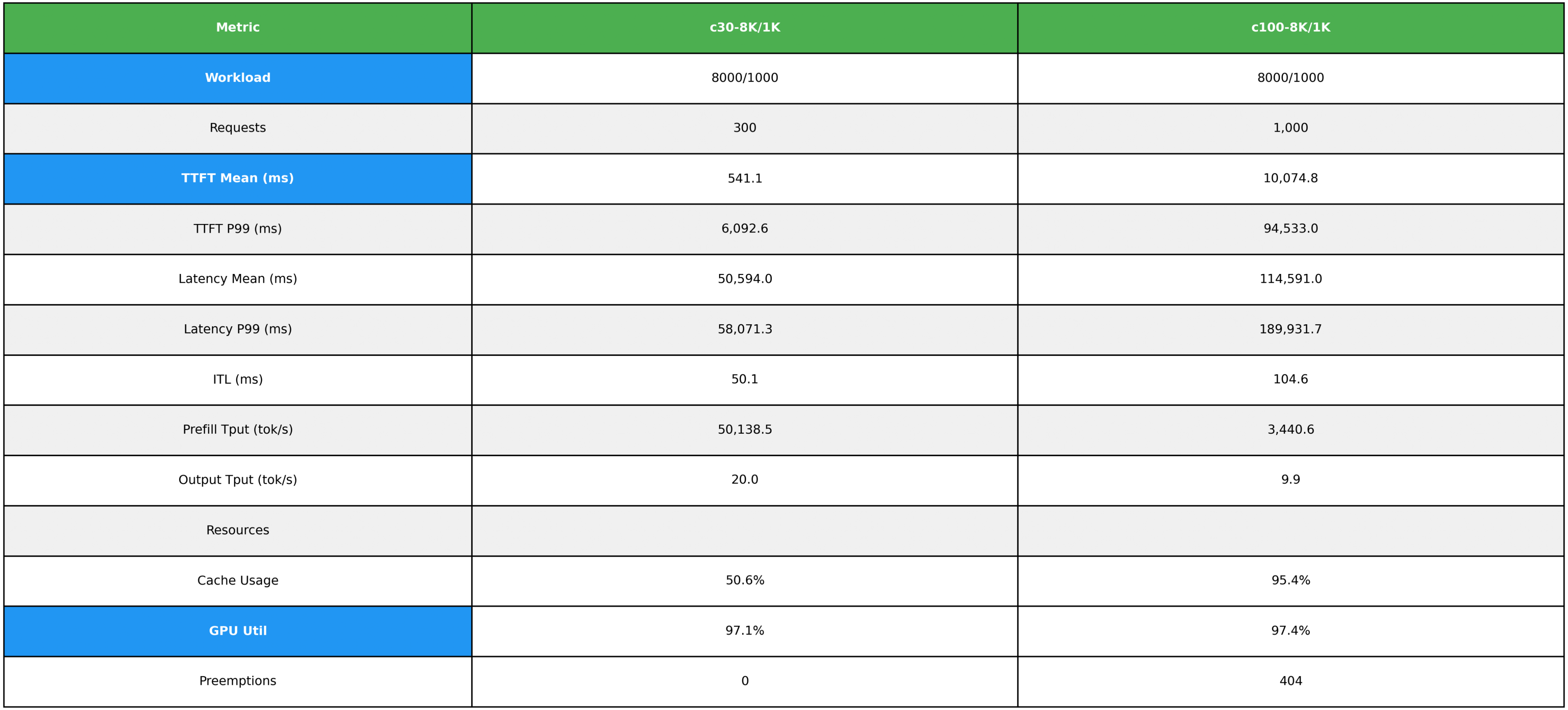
Workload Size Comparison

Energy Efficiency Comparison

Relative Performance vs Baseline (c30-8K/1K)



Comprehensive Run Comparison - Summary Table



Metric	c30-8K/1K	c100-8K/1K
Workload	8000/1000	8000/1000
Requests	300	1,000
TTFT Mean (ms)	541.1	10,074.8
TTFT P99 (ms)	6,092.6	94,533.0
Latency Mean (ms)	50,594.0	114,591.0
Latency P99 (ms)	58,071.3	189,931.7
ITL (ms)	50.1	104.6
Prefill Tput (tok/s)	50,138.5	3,440.6
Output Tput (tok/s)	20.0	9.9
Resources		
Cache Usage	50.6%	95.4%
GPU Util	97.1%	97.4%
Preemptions	0	404