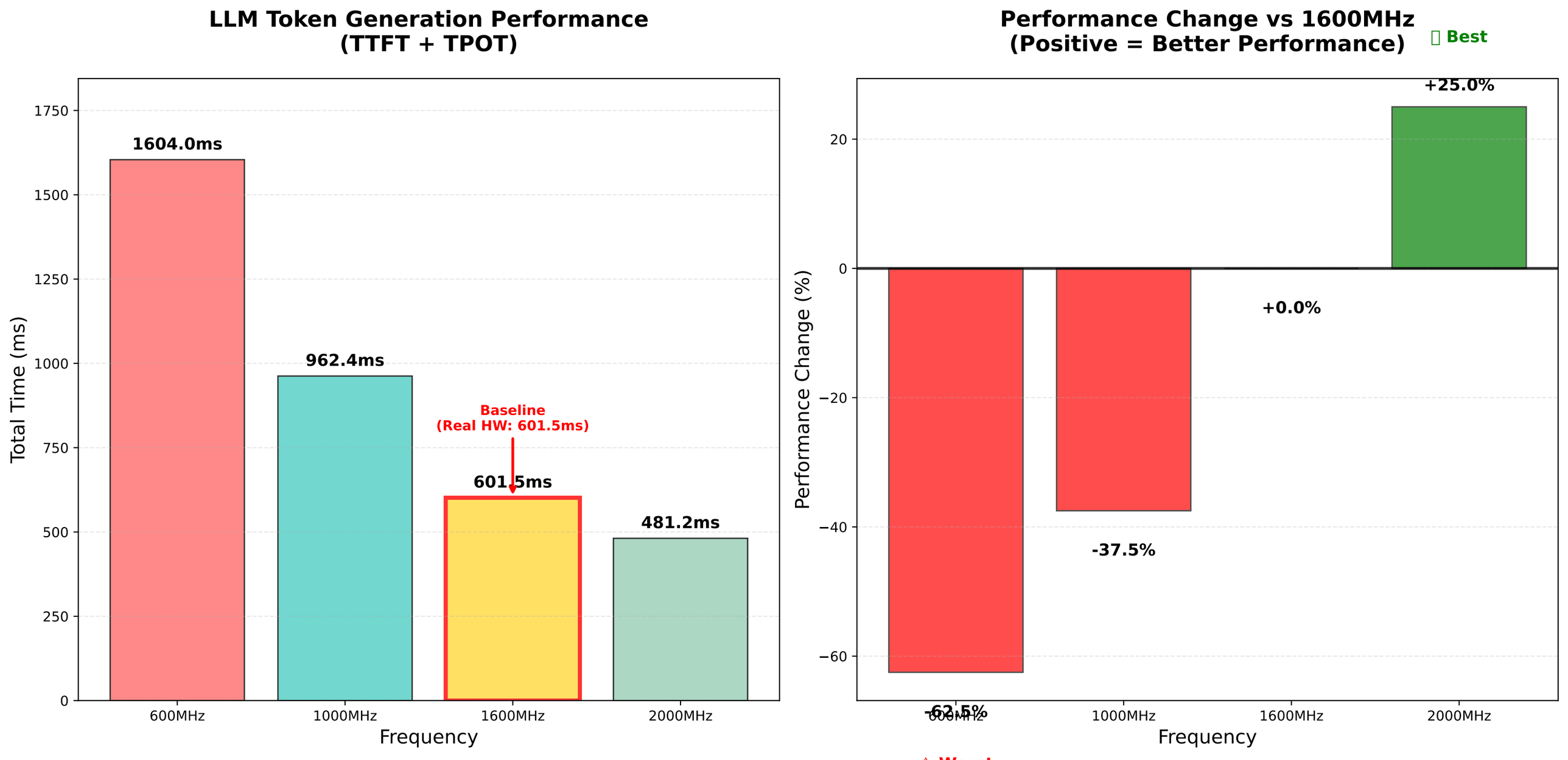
LLM Token Generation Performance
(TTFT + TPOT)
Total Time (ms)
1750
1500
1250
1000
750
500
250
0
1604.0ms
962.4ms
Baseline
(Real HW: 601.5ms)
601.5ms
481.2ms
600MHz
1000MHz
1600MHz
2000MHz
Frequency
Performance Change vs 1600MHz
(Positive = Better Performance)
Best
Performance Change (%)
20
0
−20
−40
−60
+25.0%
+0.0%
-37.5%
-62.5%
1000MHz
1600MHz
2000MHz
Frequency
Worst