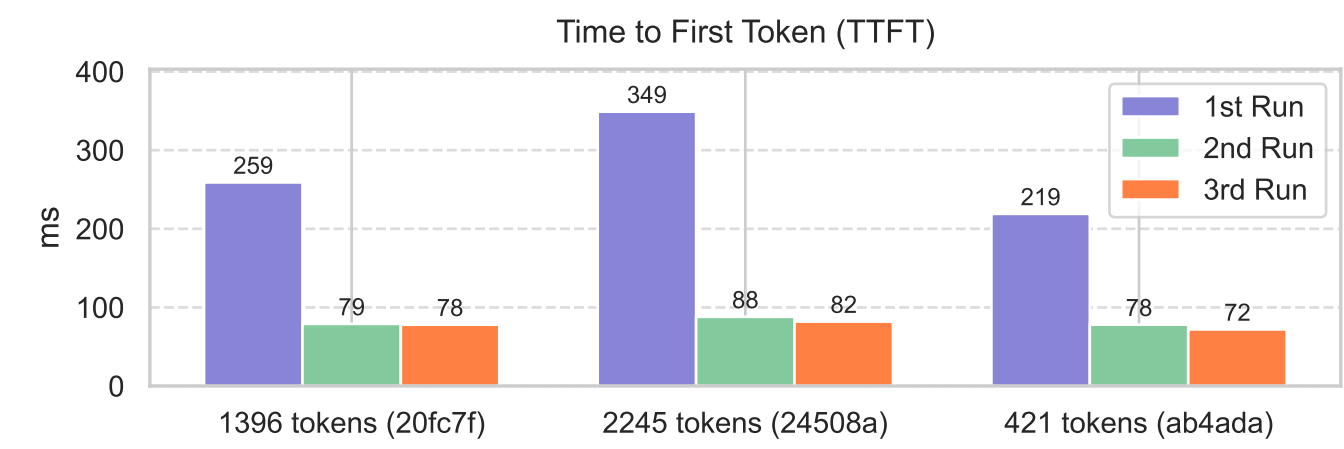
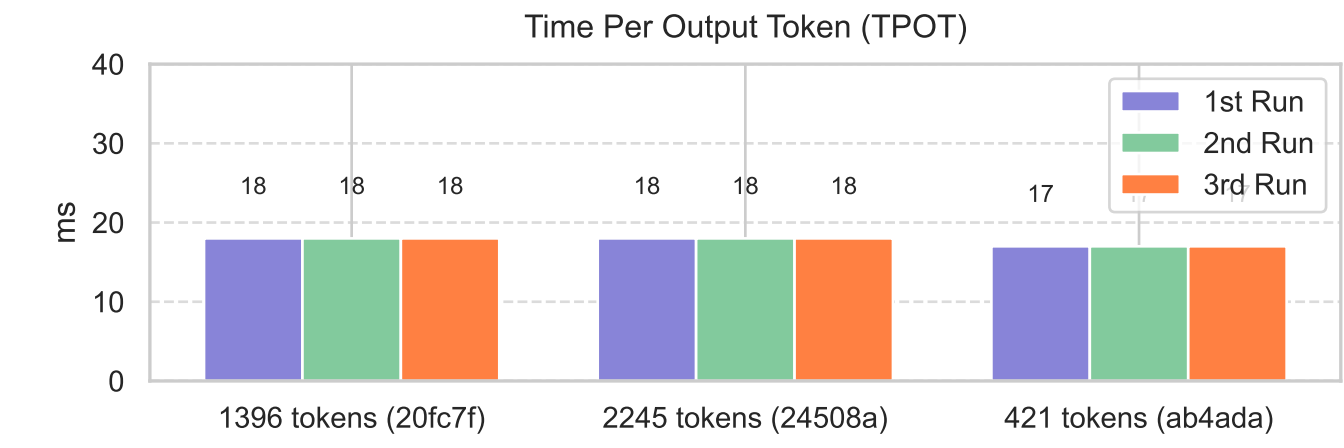
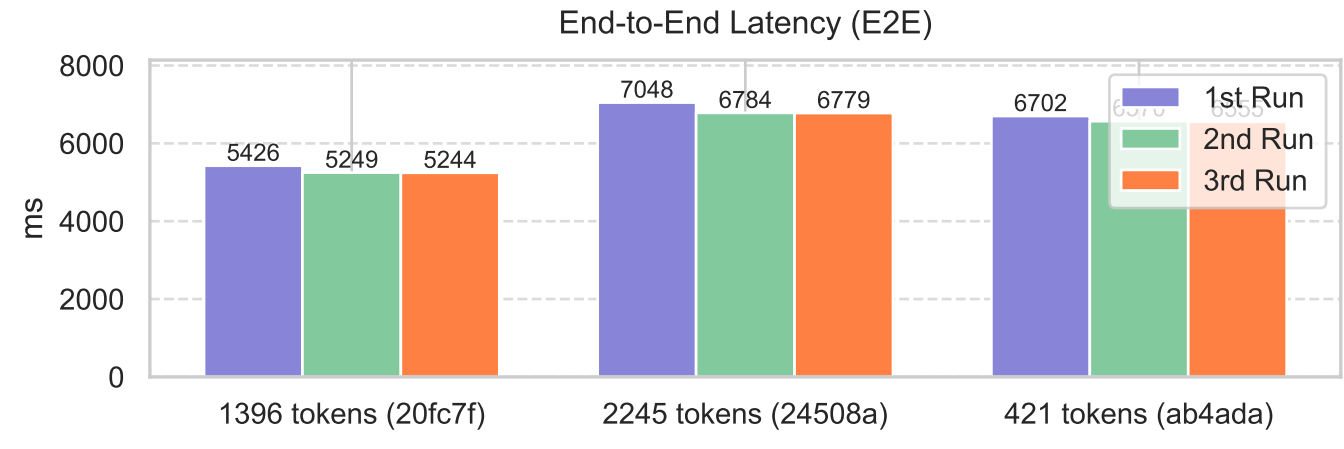

Time to First Token (TTFT)
ms
400
300
200
100
0
259
79
78
349
88
82
219
78
72
1st Run
2nd Run
3rd Run
1396 tokens (20fc7f)
2245 tokens (24508a)
421 tokens (ab4ada)


Time Per Output Token (TPOT)
ms
40
30
20
10
0
18
18
18
18
18
18
17
1st Run
2nd Run
3rd Run
1396 tokens (20fc7f)
2245 tokens (24508a)
421 tokens (ab4ada)


End-to-End Latency (E2E)
ms
8000
6000
4000
2000
0
5426
5249
5244
7048
6784
6779
6702
1st Run
2nd Run
3rd Run
1396 tokens (20fc7f)
2245 tokens (24508a)
421 tokens (ab4ada)