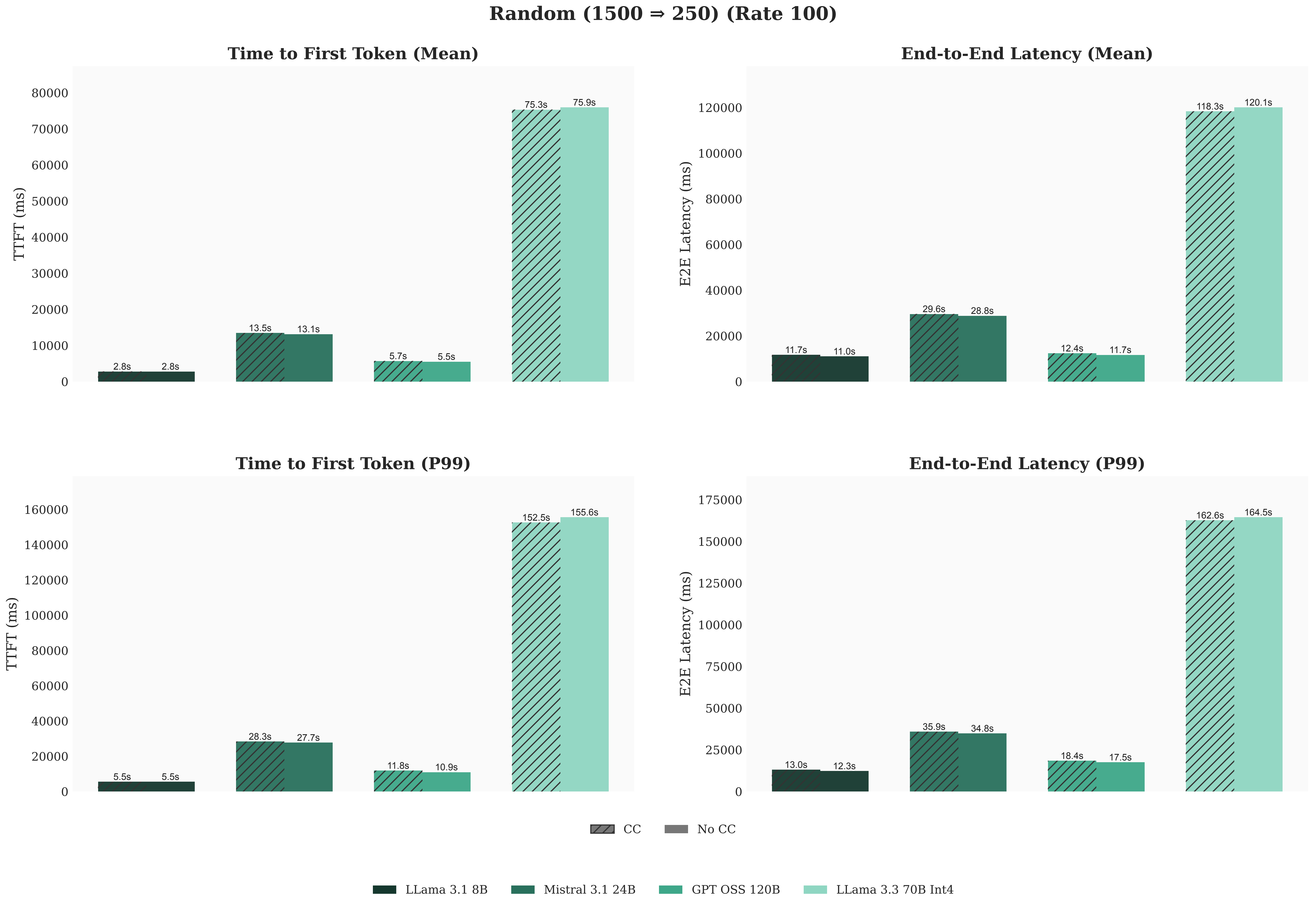
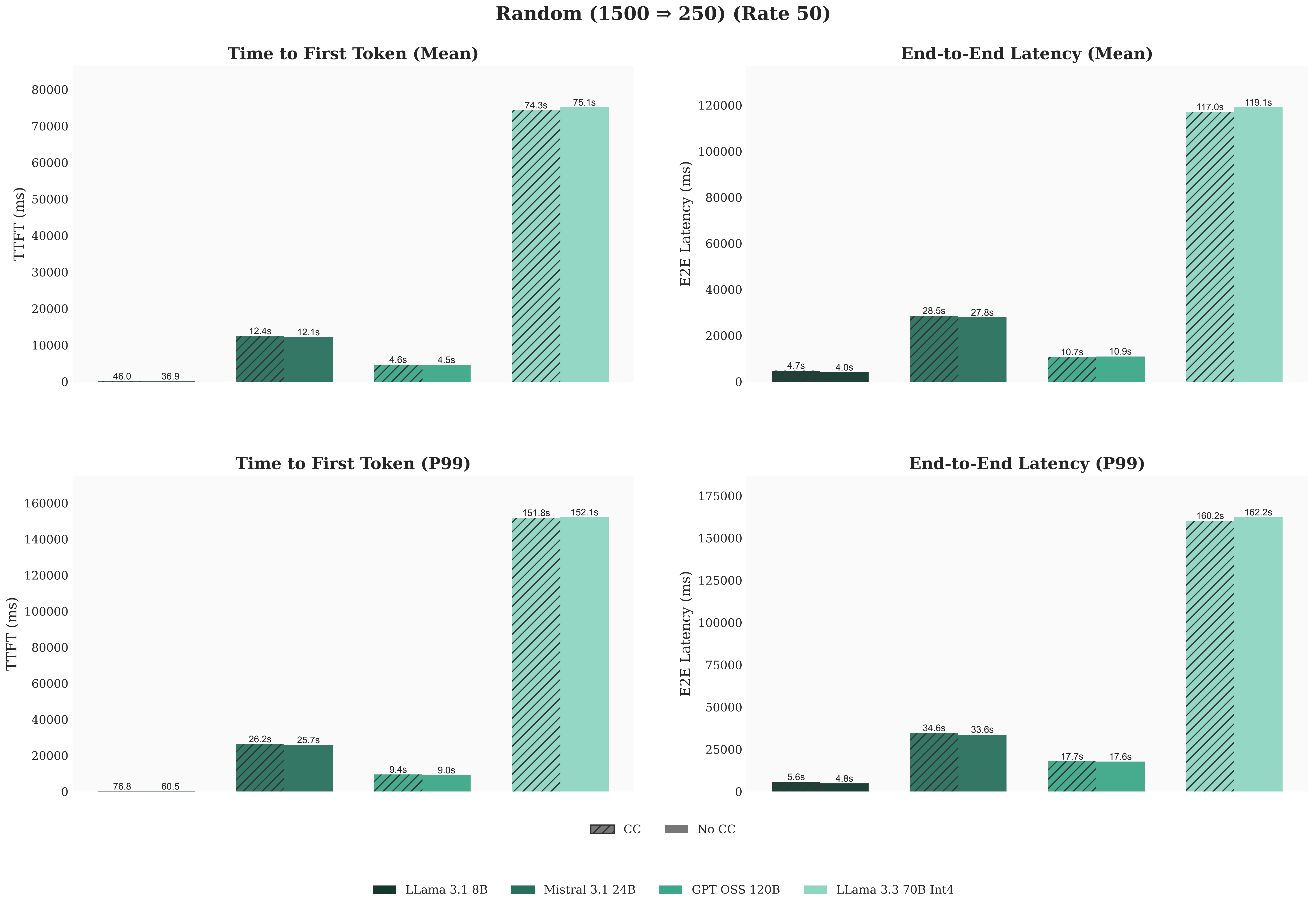

Random (1500 ⇒ 250) (Rate 100)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
60000
70000
80000
2.8s
2.8s
13.5s
13.1s
5.7s
5.5s
75.3s
75.9s
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
11.7s
11.0s
29.6s
28.8s
12.4s
11.7s
118.3s
120.1s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
160000
5.5s
5.5s
28.3s
27.7s
11.8s
10.9s
152.5s
155.6s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
13.0s
12.3s
35.9s
34.8s
18.4s
17.5s
162.6s
164.5s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4



Random (1500 ⇒ 250) (Rate 50)

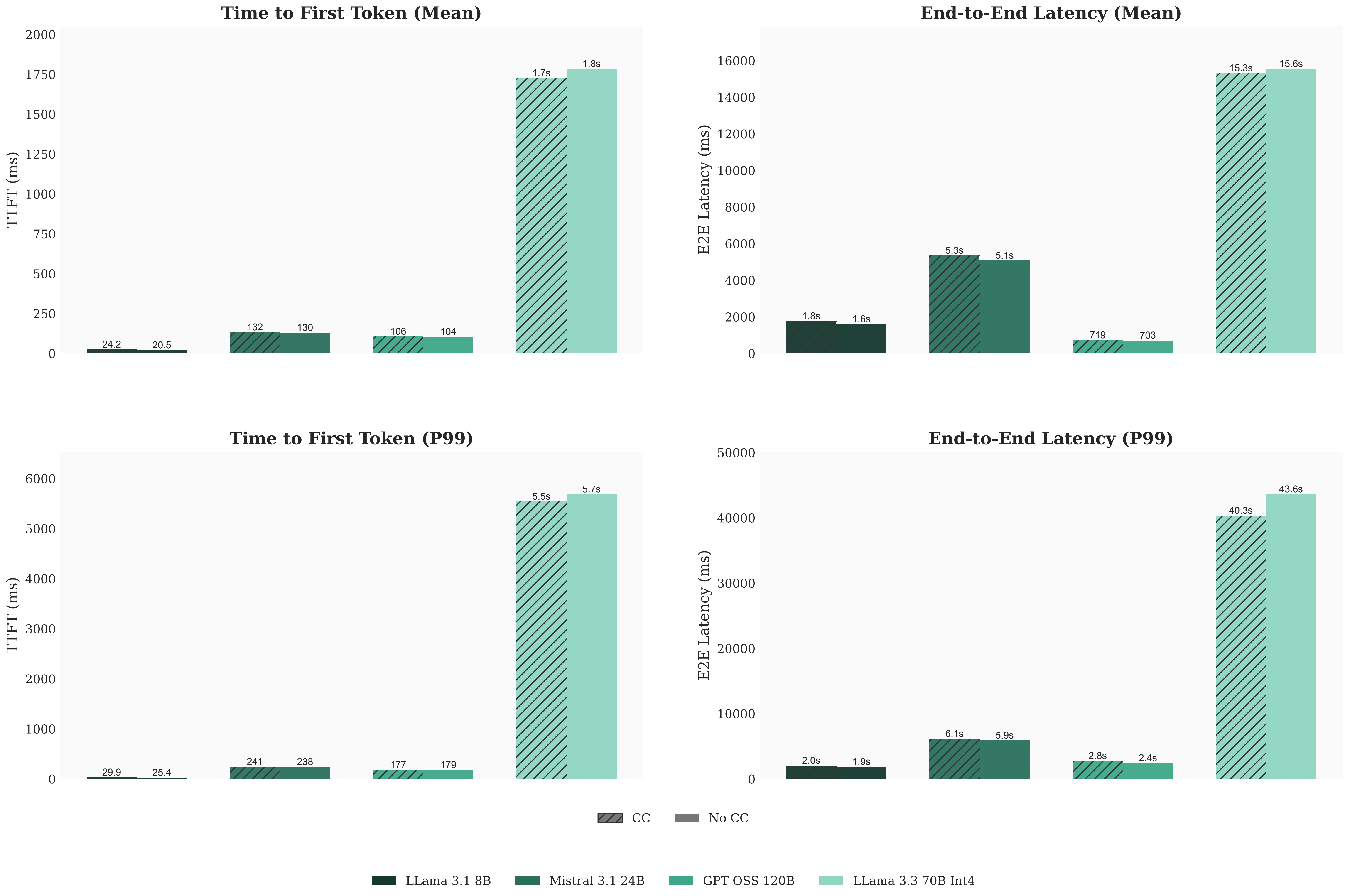
Time to First Token (Mean)

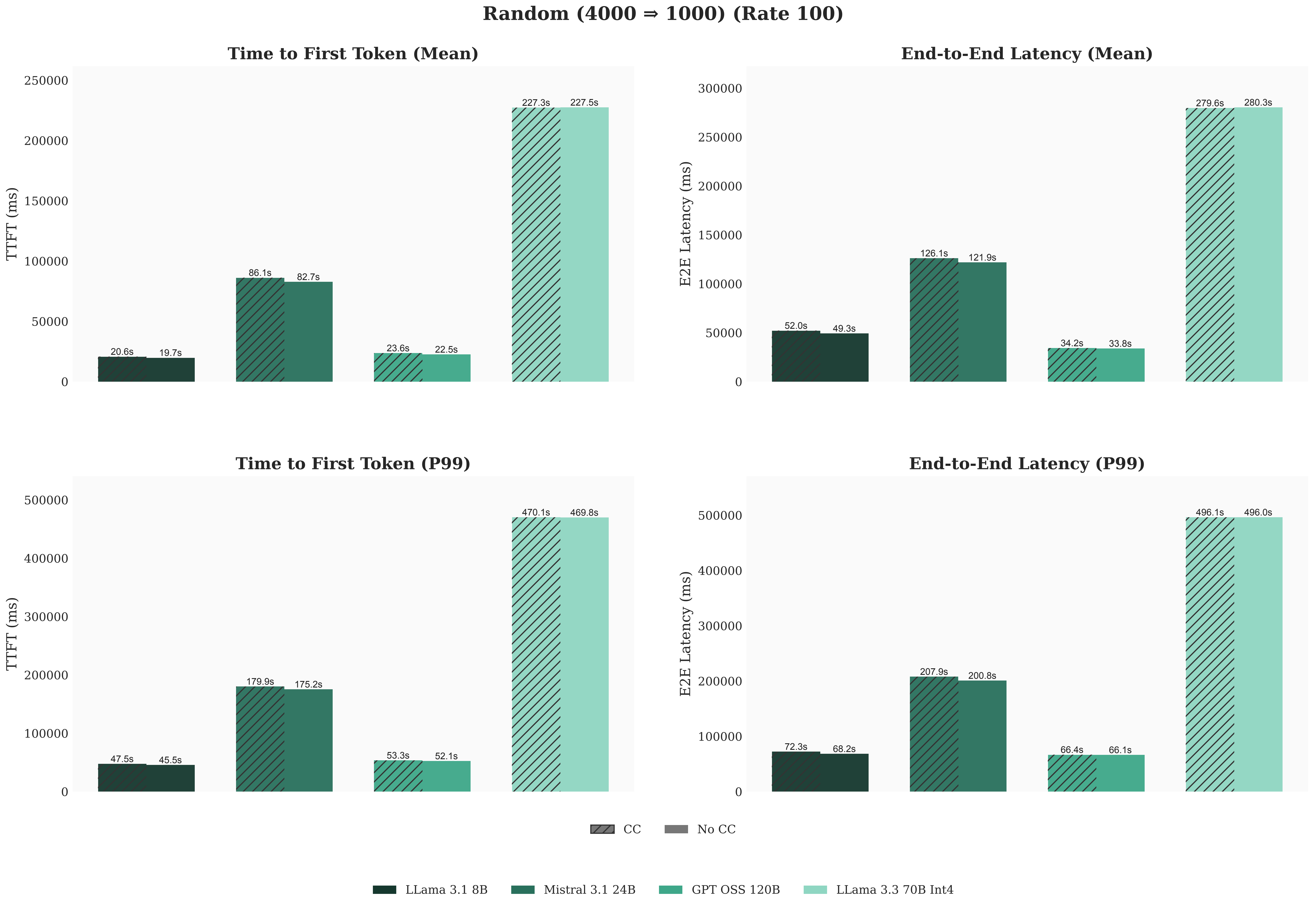
End-to-End Latency (Mean)

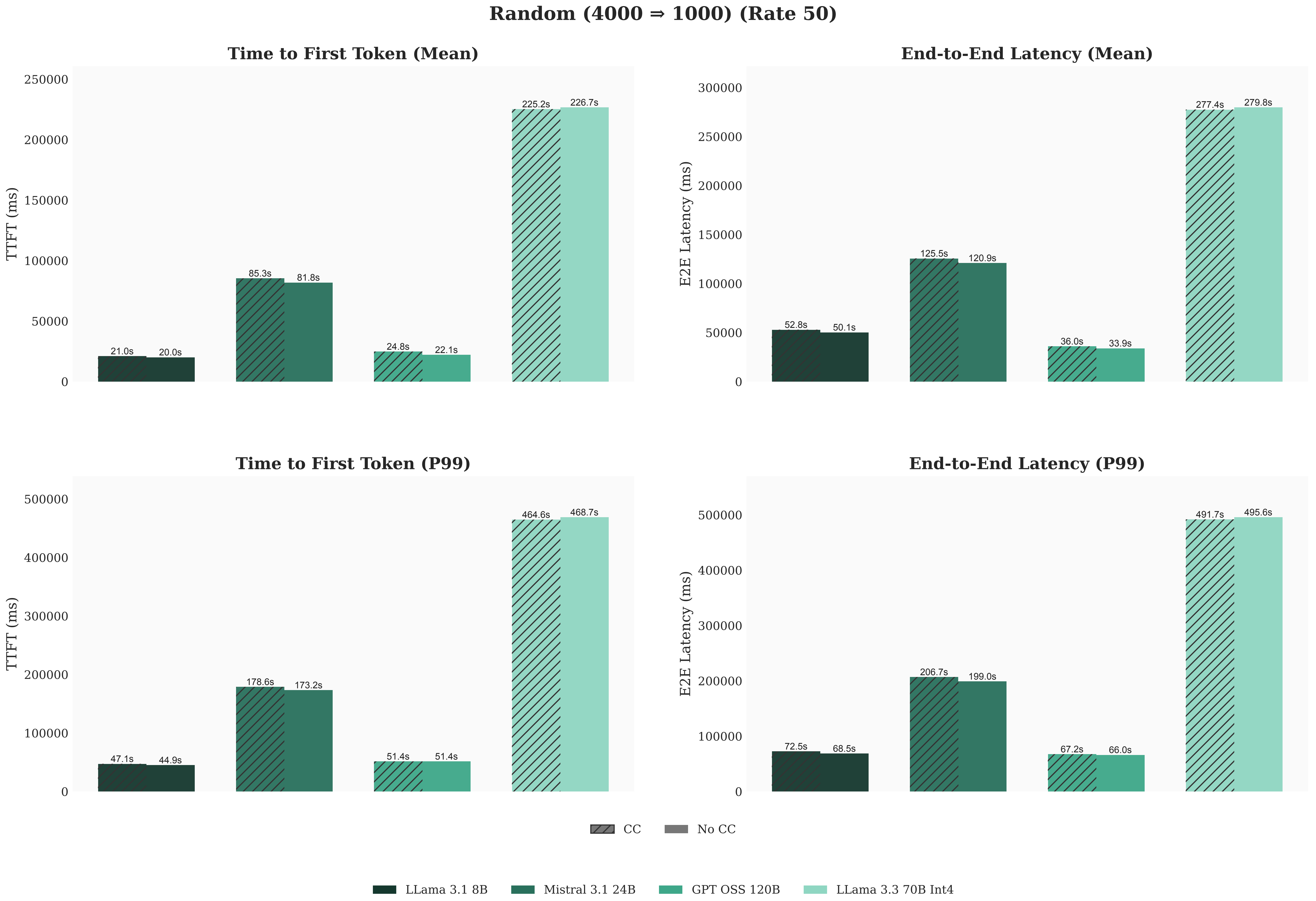
Time to First Token (P99)

End-to-End Latency (P99)

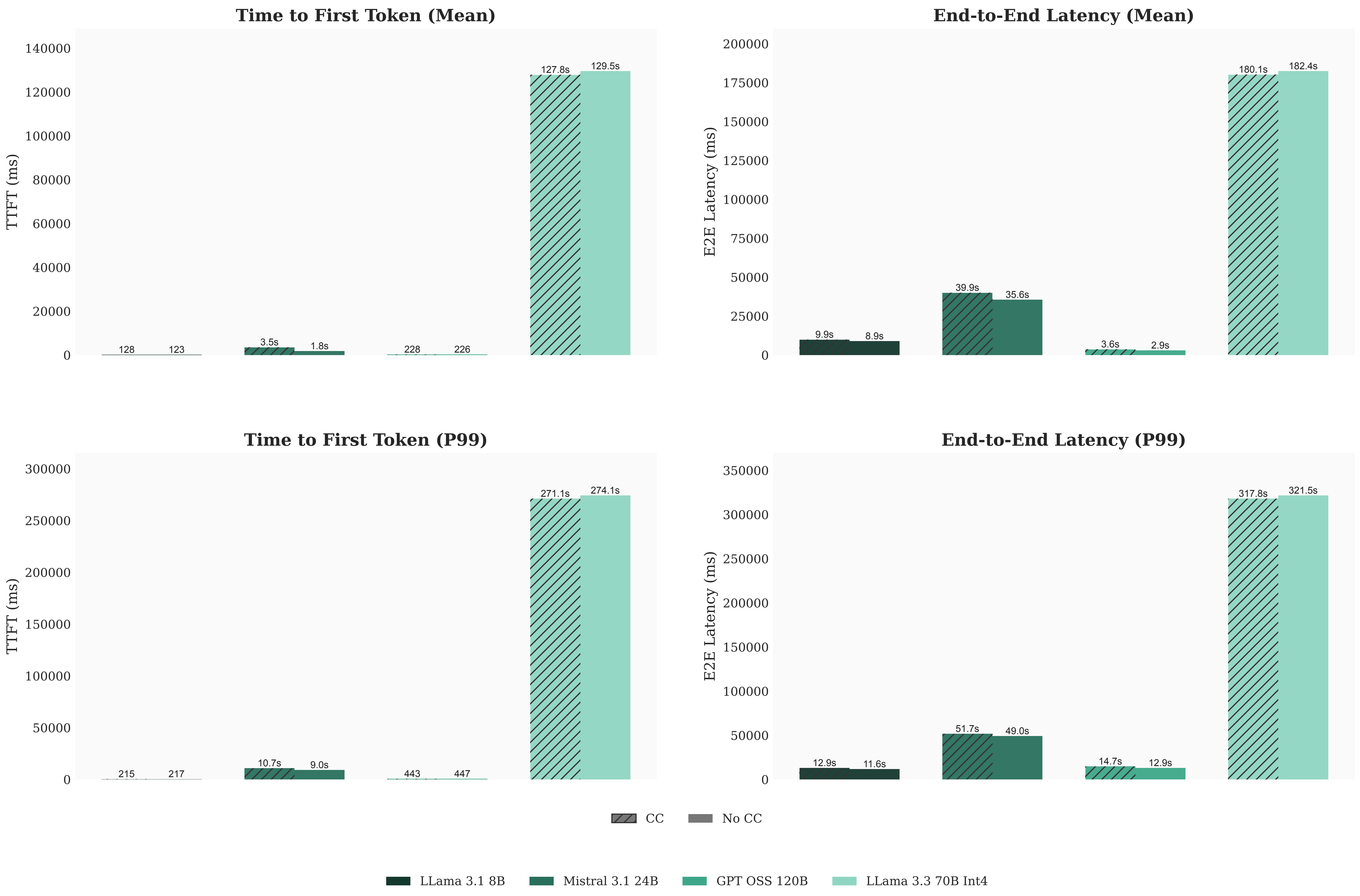
Random (1500 ⇒ 250) (Rate 1)




Random (4000 ⇒ 1000) (Rate 100)
Time to First Token (Mean)
TTFT (ms)
250000
200000
150000
100000
50000
0
20.6s
19.7s
86.1s
82.7s
23.6s
22.5s
227.3s
227.5s
End-to-End Latency (Mean)
E2E Latency (ms)
300000
250000
200000
150000
100000
50000
0
52.0s
49.3s
126.1s
121.9s
34.2s
33.8s
279.6s
280.3s
Time to First Token (P99)
TTFT (ms)
500000
400000
300000
200000
100000
0
47.5s
45.5s
179.9s
175.2s
53.3s
52.1s
470.1s
469.8s
End-to-End Latency (P99)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
72.3s
68.2s
207.9s
200.8s
66.4s
66.1s
496.1s
496.0s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Random (4000 ⇒ 1000) (Rate 50)
Time to First Token (Mean)
TTFT (ms)
0
50000
100000
150000
200000
250000
21.0s
20.0s
85.3s
81.8s
24.8s
22.1s
225.2s
226.7s
End-to-End Latency (Mean)
E2E Latency (ms)
0
50000
100000
150000
200000
250000
300000
52.8s
50.1s
125.5s
120.9s
36.0s
33.9s
277.4s
279.8s
Time to First Token (P99)
TTFT (ms)
0
100000
200000
300000
400000
500000
47.1s
44.9s
178.6s
173.2s
51.4s
51.4s
464.6s
468.7s
End-to-End Latency (P99)
E2E Latency (ms)
0
100000
200000
300000
400000
500000
72.5s
68.5s
206.7s
199.0s
67.2s
66.0s
491.7s
495.6s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

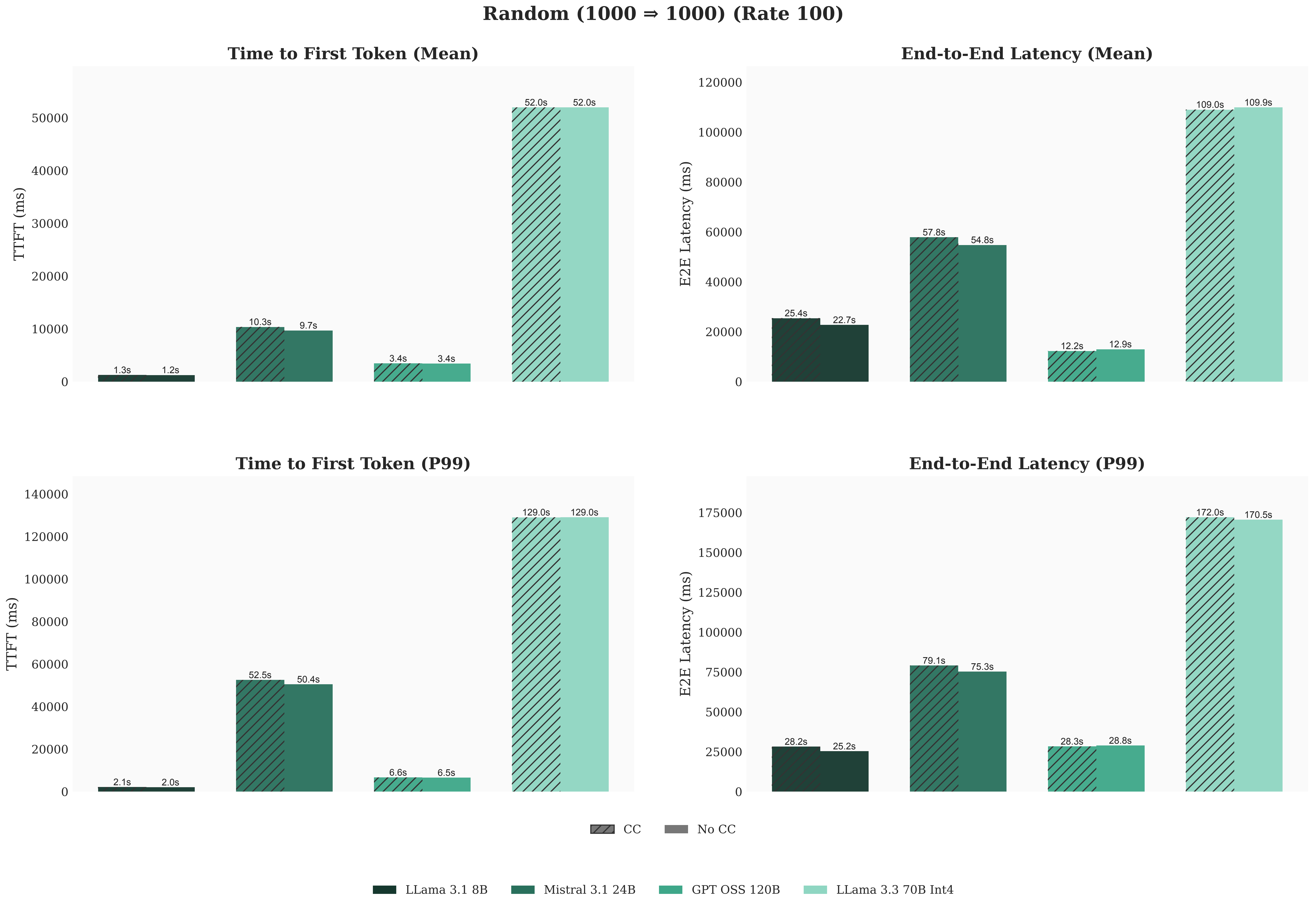
Random (4000 ⇒ 1000) (Rate 1)

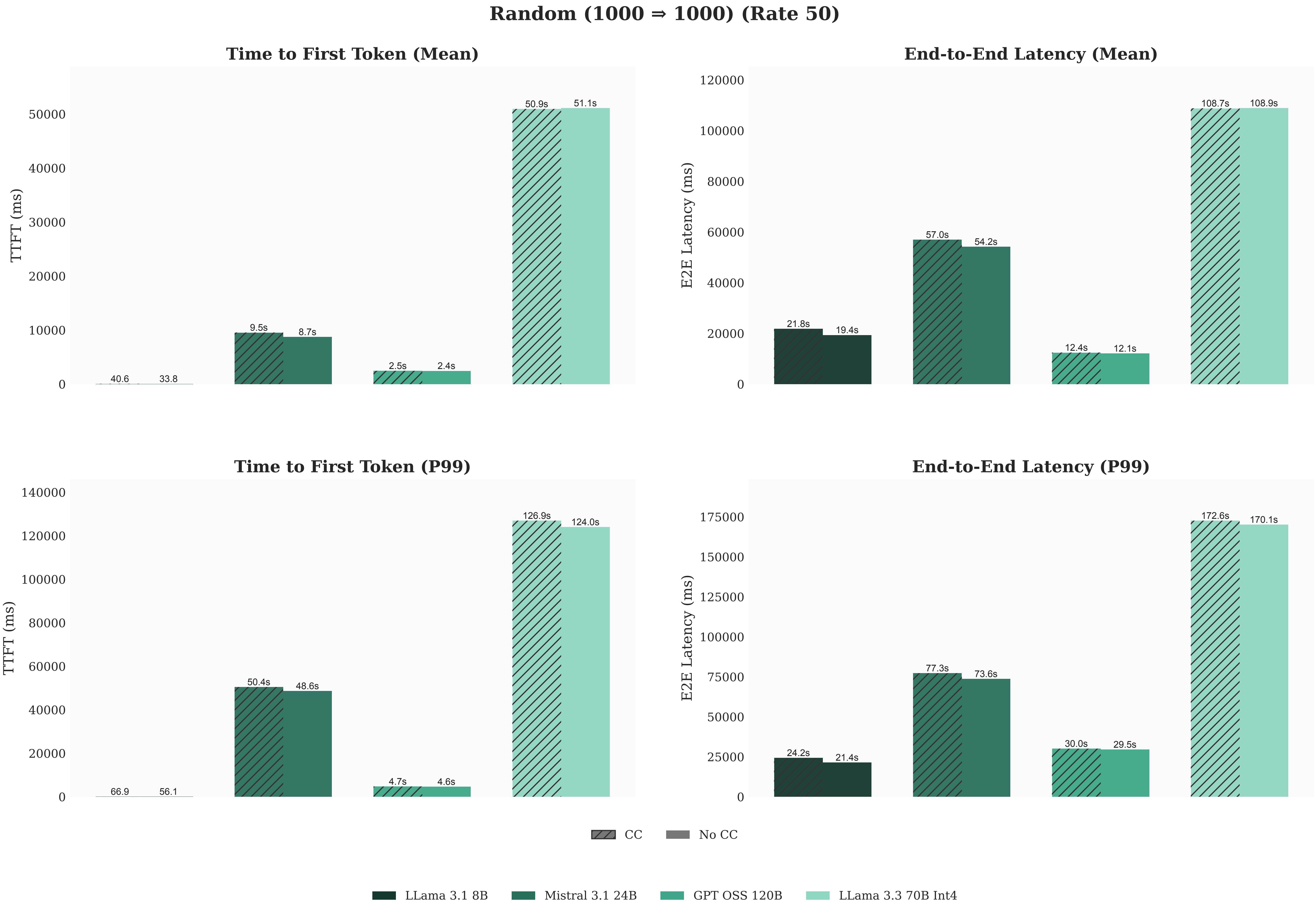

Time to First Token (Mean)

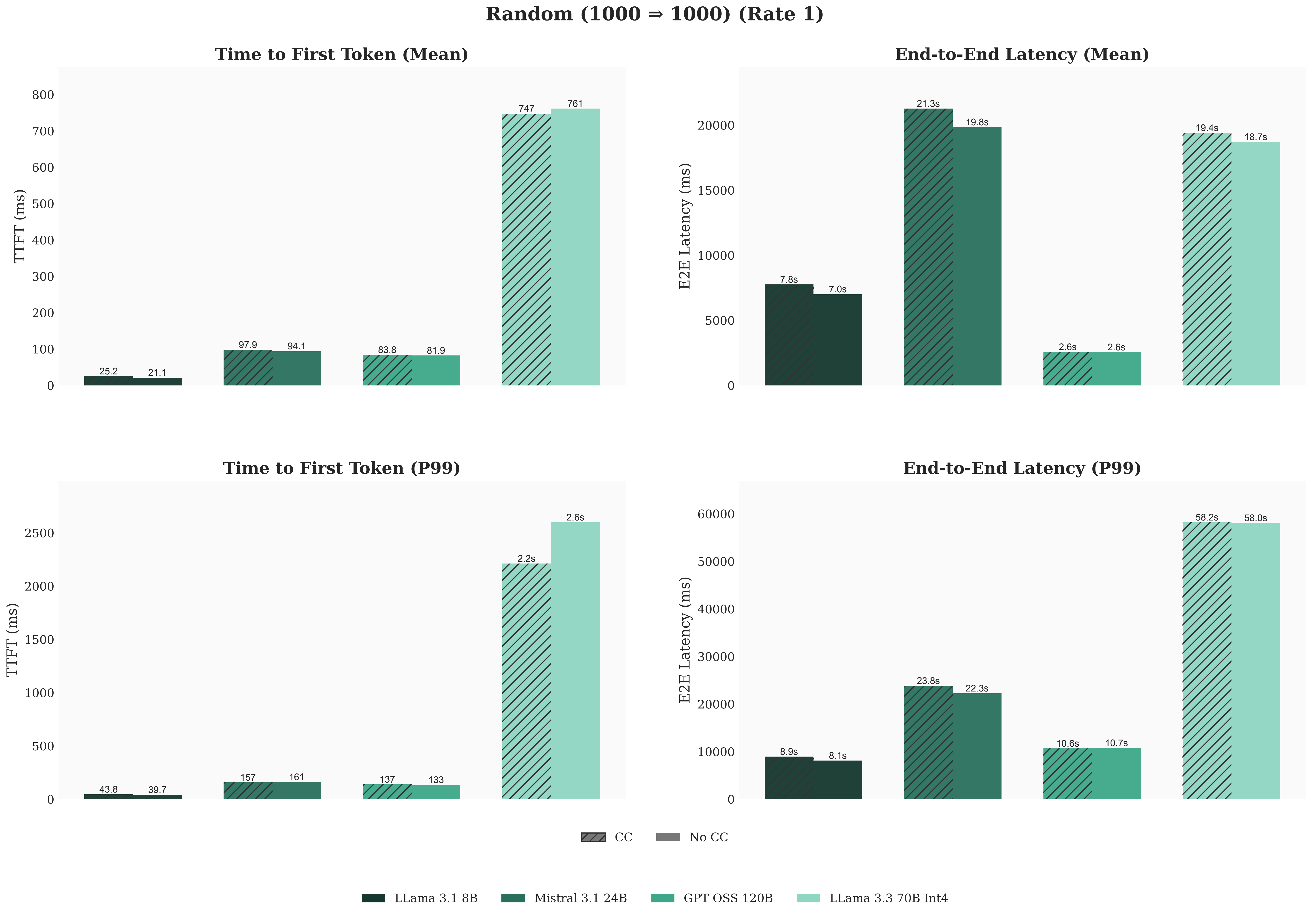
End-to-End Latency (Mean)

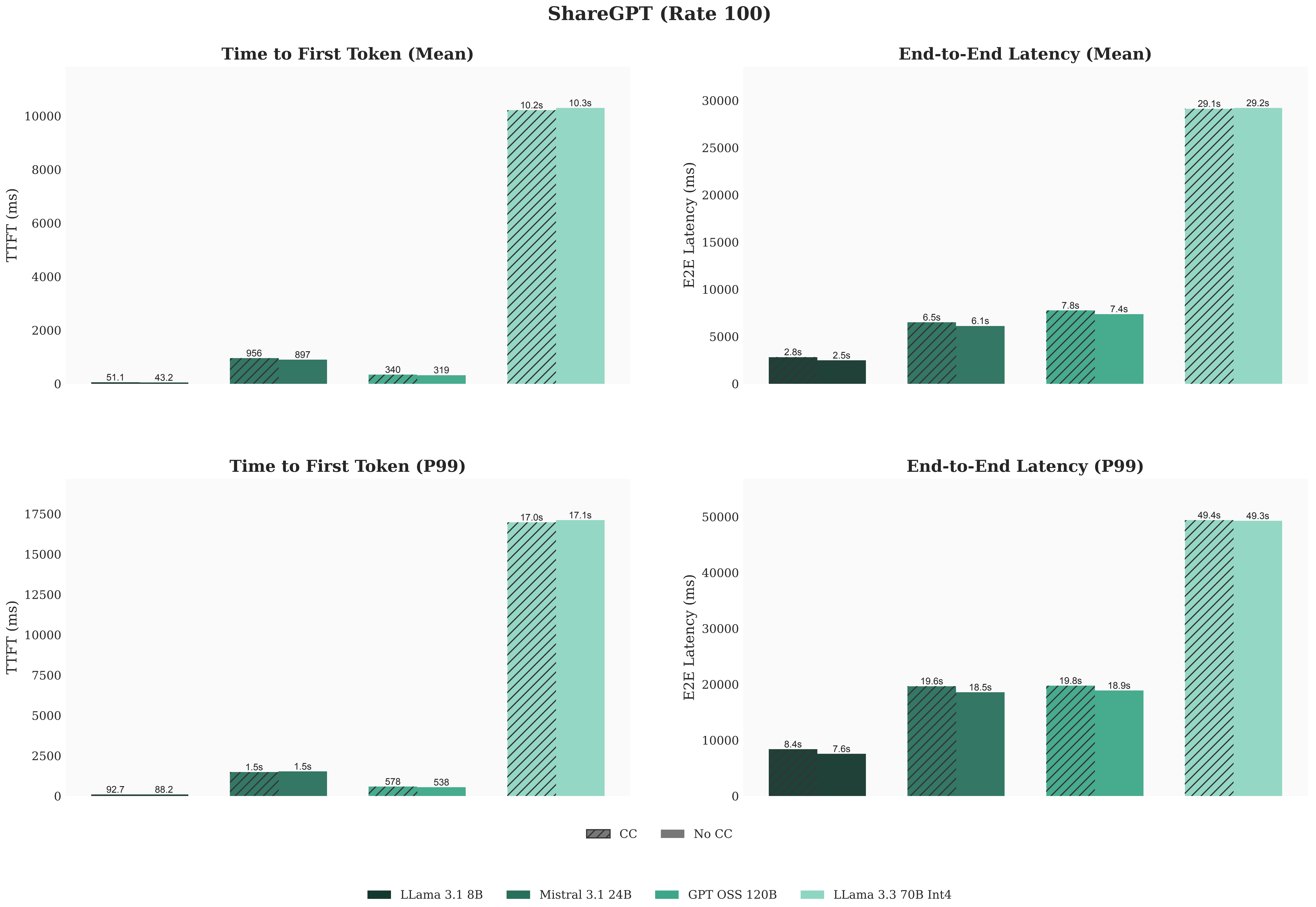
Time to First Token (P99)

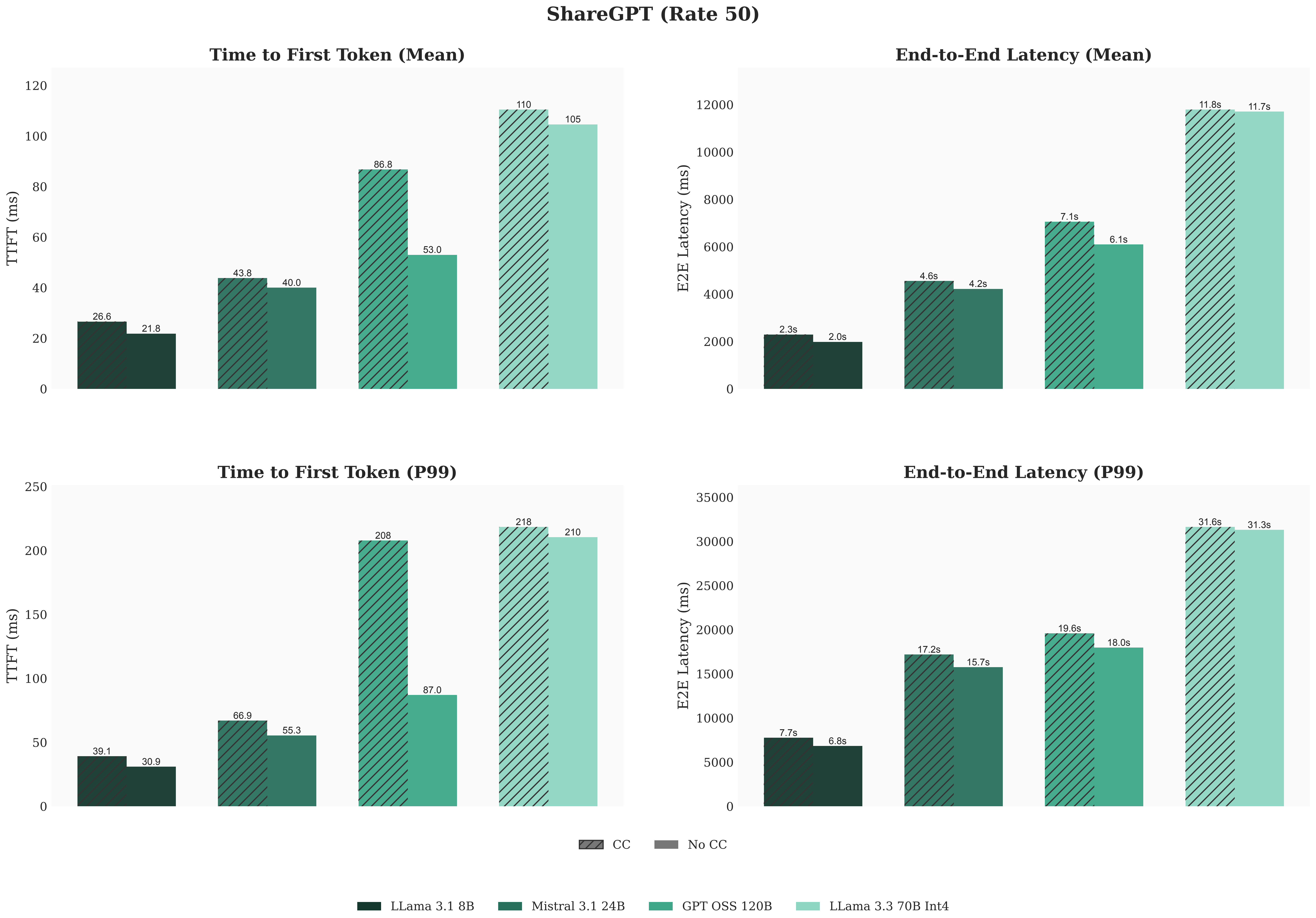
End-to-End Latency (P99)

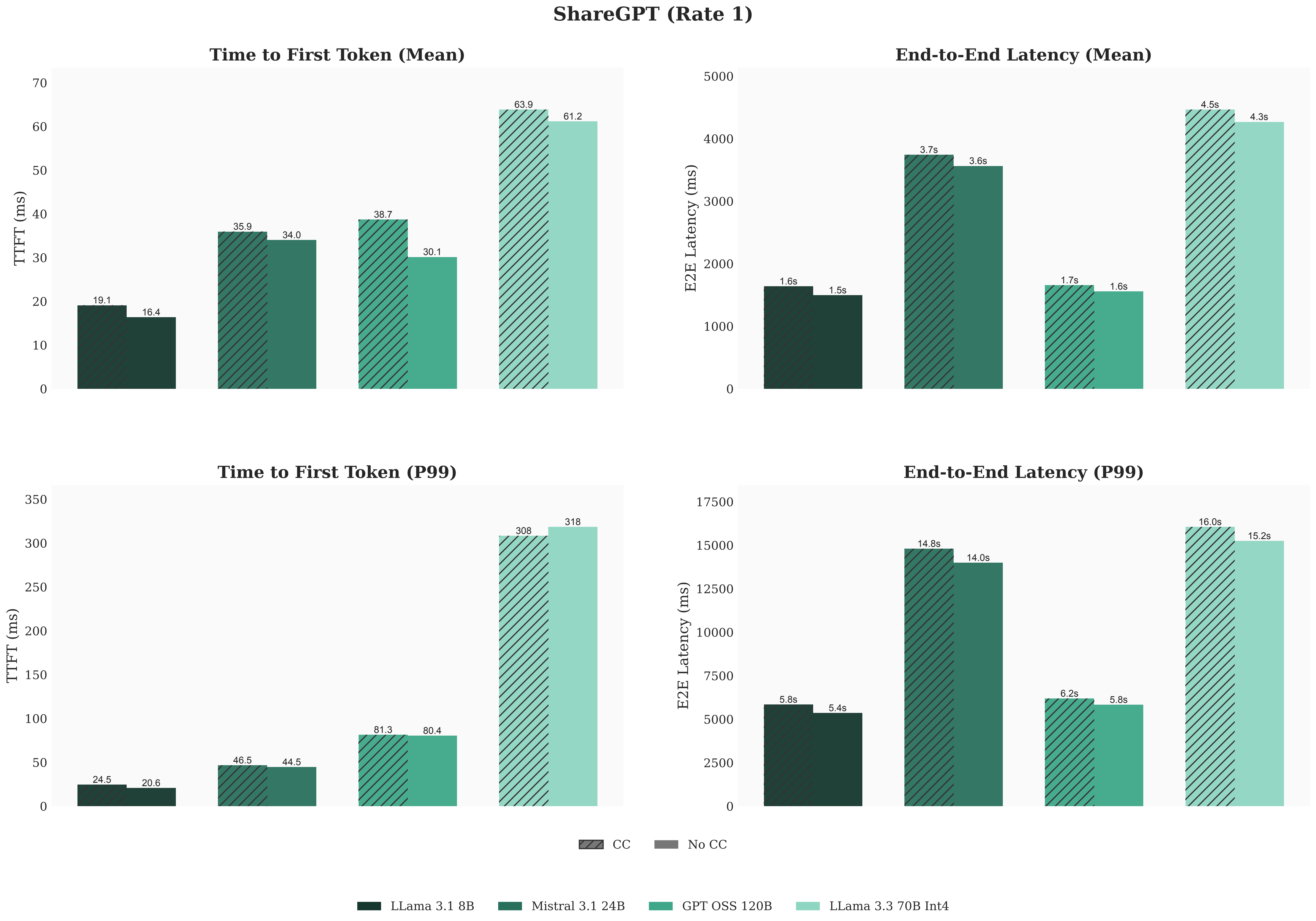

Random (1000 ⇒ 1000) (Rate 100)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
1.3s
1.2s
10.3s
9.7s
3.4s
3.4s
52.0s
52.0s
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
25.4s
22.7s
57.8s
54.8s
12.2s
12.9s
109.0s
109.9s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
2.1s
2.0s
52.5s
50.4s
6.6s
6.5s
129.0s
129.0s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
28.2s
25.2s
79.1s
75.3s
28.3s
28.8s
172.0s
170.5s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

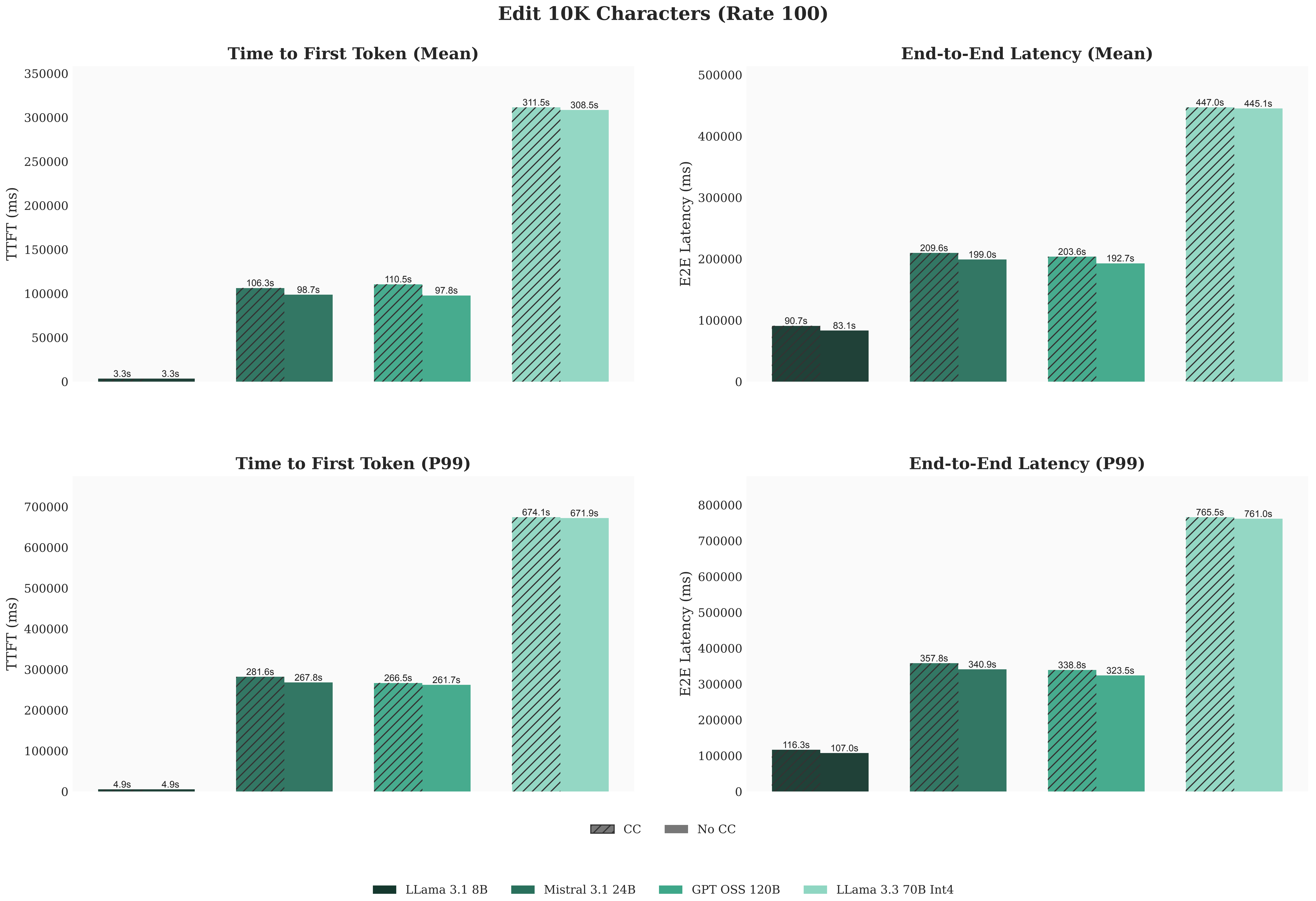

Random (1000 ⇒ 1000) (Rate 50)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
40.6
33.8
9.5s
8.7s
2.5s
2.4s
50.9s
51.1s
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
21.8s
19.4s
57.0s
54.2s
12.4s
12.1s
108.7s
108.9s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
66.9
56.1
50.4s
48.6s
4.7s
4.6s
126.9s
124.0s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
24.2s
21.4s
77.3s
73.6s
30.0s
29.5s
172.6s
170.1s
▨ CC
■ No CC
■ LLama 3.1 8B
■ Mistral 3.1 24B
■ GPT OSS 120B
■ LLama 3.3 70B Int4

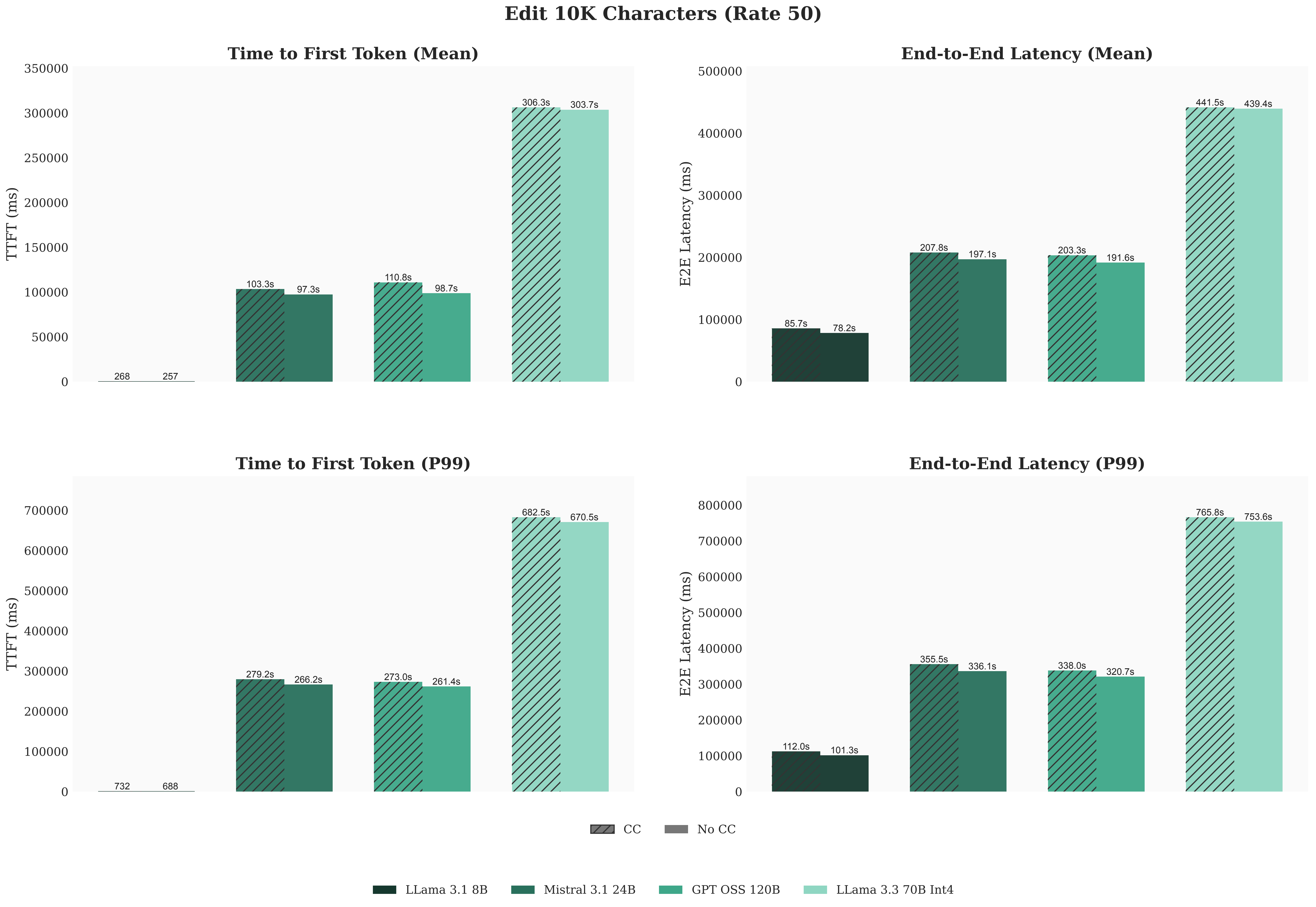

Random (1000 ⇒ 1000) (Rate 1)
Time to First Token (Mean)
TTFT (ms)
0
100
200
300
400
500
600
700
800
25.2
21.1
97.9
94.1
83.8
81.9
747
761
End-to-End Latency (Mean)
E2E Latency (ms)
0
5000
10000
15000
20000
7.8s
7.0s
21.3s
19.8s
2.6s
2.6s
19.4s
18.7s
Time to First Token (P99)
TTFT (ms)
0
500
1000
1500
2000
2500
43.8
39.7
157
161
137
133
2.2s
2.6s
End-to-End Latency (P99)
E2E Latency (ms)
0
10000
20000
30000
40000
50000
60000
8.9s
8.1s
23.8s
22.3s
10.6s
10.7s
58.2s
58.0s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

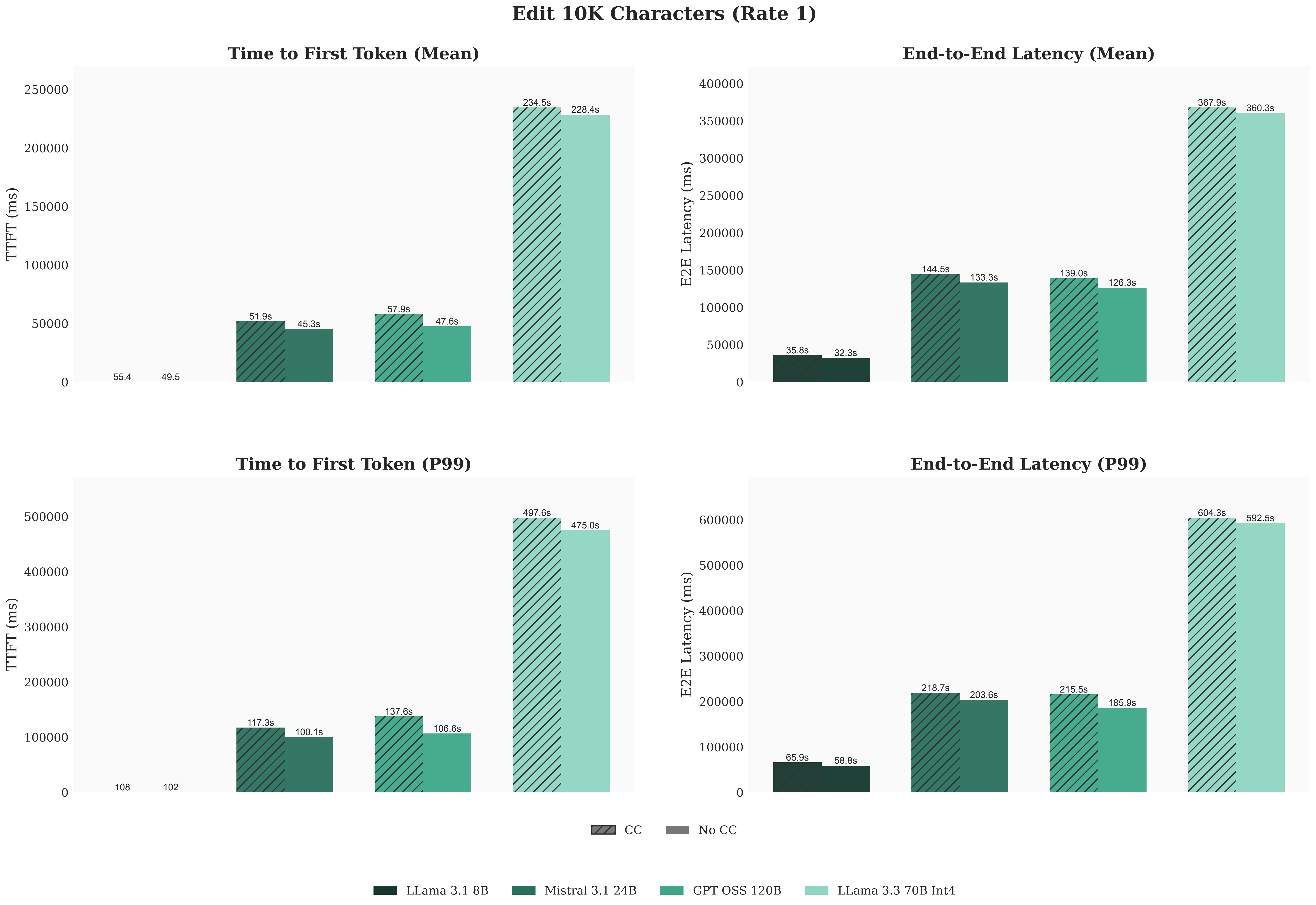

ShareGPT (Rate 100)
Time to First Token (Mean)
TTFT (ms)
0
2000
4000
6000
8000
10000
51.1
43.2
956
897
340
319
10.2s
10.3s
End-to-End Latency (Mean)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
2.8s
2.5s
6.5s
6.1s
7.8s
7.4s
29.1s
29.2s
Time to First Token (P99)
TTFT (ms)
0
2500
5000
7500
10000
12500
15000
17500
92.7
88.2
1.5s
1.5s
578
538
17.0s
17.1s
End-to-End Latency (P99)
E2E Latency (ms)
0
10000
20000
30000
40000
50000
8.4s
7.6s
19.6s
18.5s
19.8s
18.9s
49.4s
49.3s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

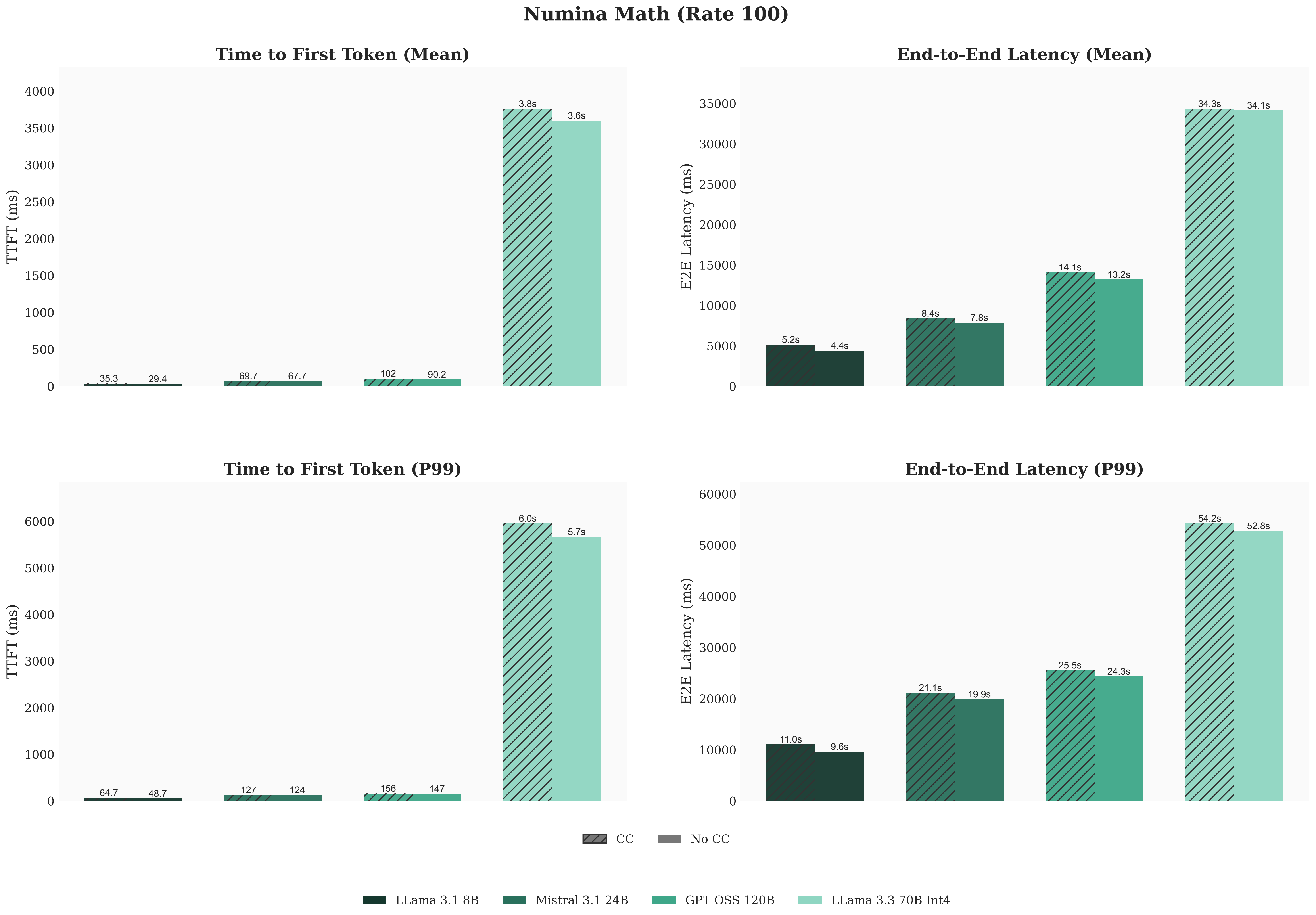

ShareGPT (Rate 50)
Time to First Token (Mean)
TTFT (ms)
0
20
40
60
80
100
120
26.6
21.8
43.8
40.0
86.8
53.0
110
105
End-to-End Latency (Mean)
E2E Latency (ms)
0
2000
4000
6000
8000
10000
12000
2.3s
2.0s
4.6s
4.2s
7.1s
6.1s
11.8s
11.7s
Time to First Token (P99)
TTFT (ms)
0
50
100
150
200
250
39.1
30.9
66.9
55.3
208
87.0
218
210
End-to-End Latency (P99)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
35000
7.7s
6.8s
17.2s
15.7s
19.6s
18.0s
31.6s
31.3s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

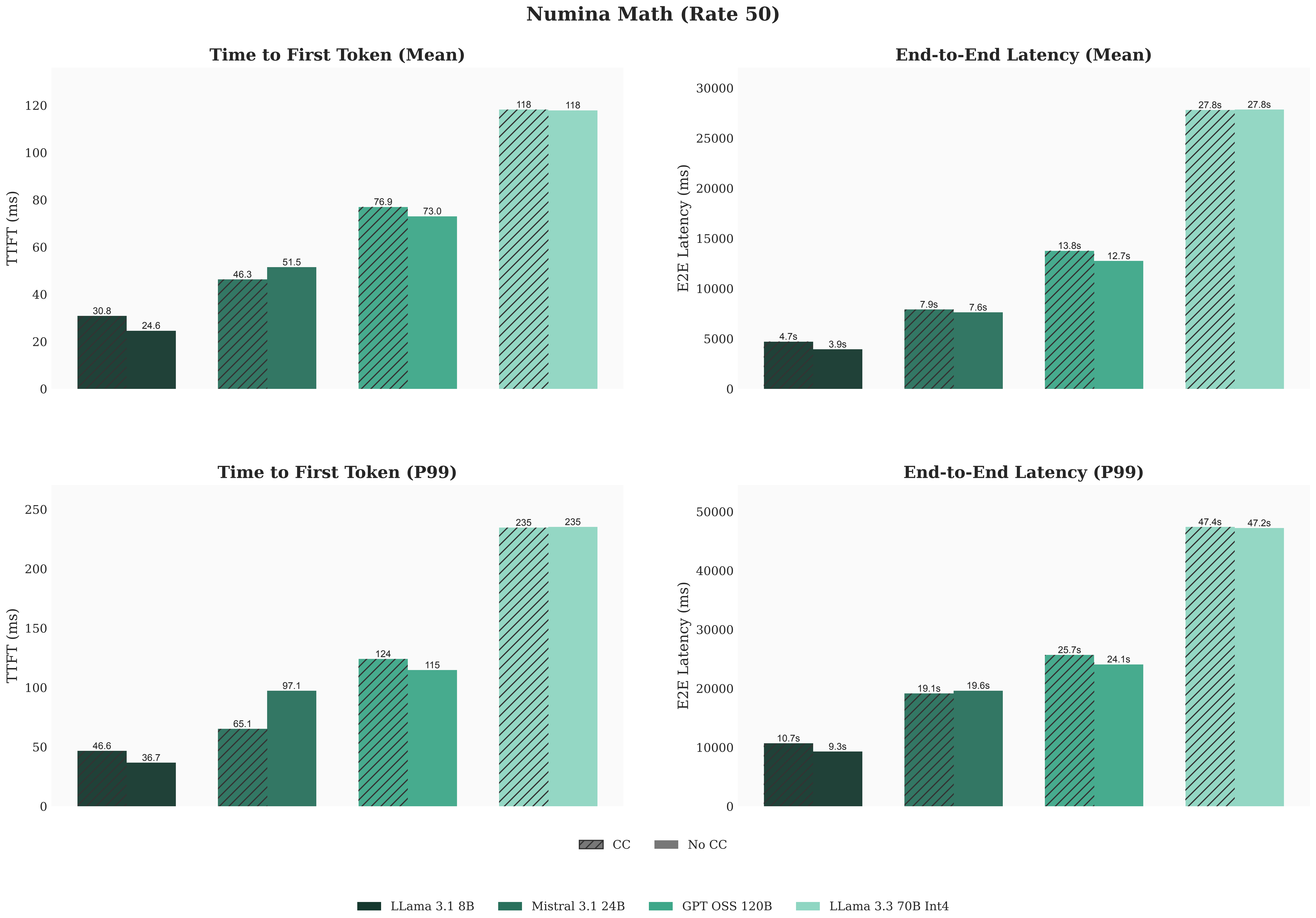

ShareGPT (Rate 1)
Time to First Token (Mean)
TTFT (ms)
0
10
20
30
40
50
60
70
19.1
16.4
35.9
34.0
38.7
30.1
63.9
61.2
End-to-End Latency (Mean)
E2E Latency (ms)
0
1000
2000
3000
4000
5000
1.6s
1.5s
3.7s
3.6s
1.7s
1.6s
4.5s
4.3s
Time to First Token (P99)
TTFT (ms)
0
50
100
150
200
250
300
350
24.5
20.6
46.5
44.5
81.3
80.4
308
318
End-to-End Latency (P99)
E2E Latency (ms)
0
2500
5000
7500
10000
12500
15000
17500
5.8s
5.4s
14.8s
14.0s
6.2s
5.8s
16.0s
15.2s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

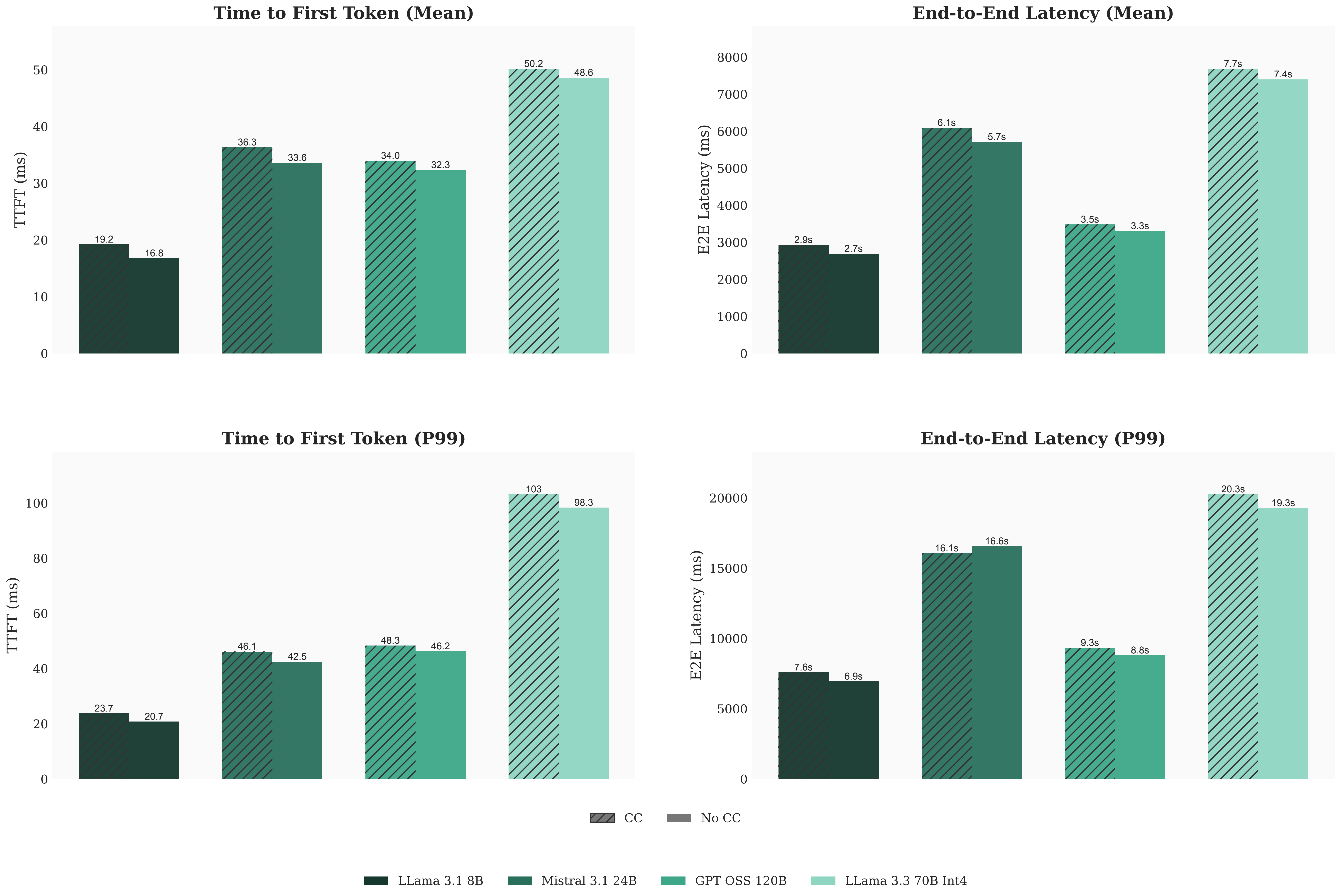

Edit 10K Characters (Rate 100)
Time to First Token (Mean)
TTFT (ms)
350000
300000
250000
200000
150000
100000
50000
0
3.3s
3.3s
106.3s
98.7s
110.5s
97.8s
311.5s
308.5s
End-to-End Latency (Mean)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
90.7s
83.1s
209.6s
199.0s
203.6s
192.7s
447.0s
445.1s
Time to First Token (P99)
TTFT (ms)
700000
600000
500000
400000
300000
200000
100000
0
4.9s
4.9s
281.6s
267.8s
266.5s
261.7s
674.1s
671.9s
End-to-End Latency (P99)
E2E Latency (ms)
800000
700000
600000
500000
400000
300000
200000
100000
0
116.3s
107.0s
357.8s
340.9s
338.8s
323.5s
765.5s
761.0s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Edit 10K Characters (Rate 50)
Time to First Token (Mean)
TTFT (ms)
350000
300000
250000
200000
150000
100000
50000
0
268
257
103.3s
97.3s
110.8s
98.7s
306.3s
303.7s
End-to-End Latency (Mean)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
85.7s
78.2s
207.8s
197.1s
203.3s
191.6s
441.5s
439.4s
Time to First Token (P99)
TTFT (ms)
700000
600000
500000
400000
300000
200000
100000
0
732
688
279.2s
266.2s
273.0s
261.4s
682.5s
670.5s
End-to-End Latency (P99)
E2E Latency (ms)
800000
700000
600000
500000
400000
300000
200000
100000
0
112.0s
101.3s
355.5s
336.1s
338.0s
320.7s
765.8s
753.6s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Edit 10K Characters (Rate 1)
Time to First Token (Mean)
TTFT (ms)
0
50000
100000
150000
200000
250000
55.4
49.5
51.9s
45.3s
57.9s
47.6s
234.5s
228.4s
End-to-End Latency (Mean)
E2E Latency (ms)
0
50000
100000
150000
200000
250000
300000
350000
400000
35.8s
32.3s
144.5s
133.3s
139.0s
126.3s
367.9s
360.3s
Time to First Token (P99)
TTFT (ms)
0
100000
200000
300000
400000
500000
108
102
117.3s
100.1s
137.6s
106.6s
497.6s
475.0s
End-to-End Latency (P99)
E2E Latency (ms)
0
100000
200000
300000
400000
500000
600000
65.9s
58.8s
218.7s
203.6s
215.5s
185.9s
604.3s
592.5s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (Rate 100)
Time to First Token (Mean)
TTFT (ms)
0
500
1000
1500
2000
2500
3000
3500
4000
35.3
29.4
69.7
67.7
102
90.2
3.8s
3.6s
End-to-End Latency (Mean)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
35000
5.2s
4.4s
8.4s
7.8s
14.1s
13.2s
34.3s
34.1s
Time to First Token (P99)
TTFT (ms)
0
1000
2000
3000
4000
5000
6000
64.7
48.7
127
124
156
147
6.0s
5.7s
End-to-End Latency (P99)
E2E Latency (ms)
0
10000
20000
30000
40000
50000
60000
11.0s
9.6s
21.1s
19.9s
25.5s
24.3s
54.2s
52.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (Rate 50)
Time to First Token (Mean)
TTFT (ms)
0
20
40
60
80
100
120
30.8
24.6
46.3
51.5
76.9
73.0
118
118
End-to-End Latency (Mean)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
4.7s
3.9s
7.9s
7.6s
13.8s
12.7s
27.8s
27.8s
Time to First Token (P99)
TTFT (ms)
0
50
100
150
200
250
46.6
36.7
65.1
97.1
124
115
235
235
End-to-End Latency (P99)
E2E Latency (ms)
0
10000
20000
30000
40000
50000
10.7s
9.3s
19.1s
19.6s
25.7s
24.1s
47.4s
47.2s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4



Numina Math (Rate 1)