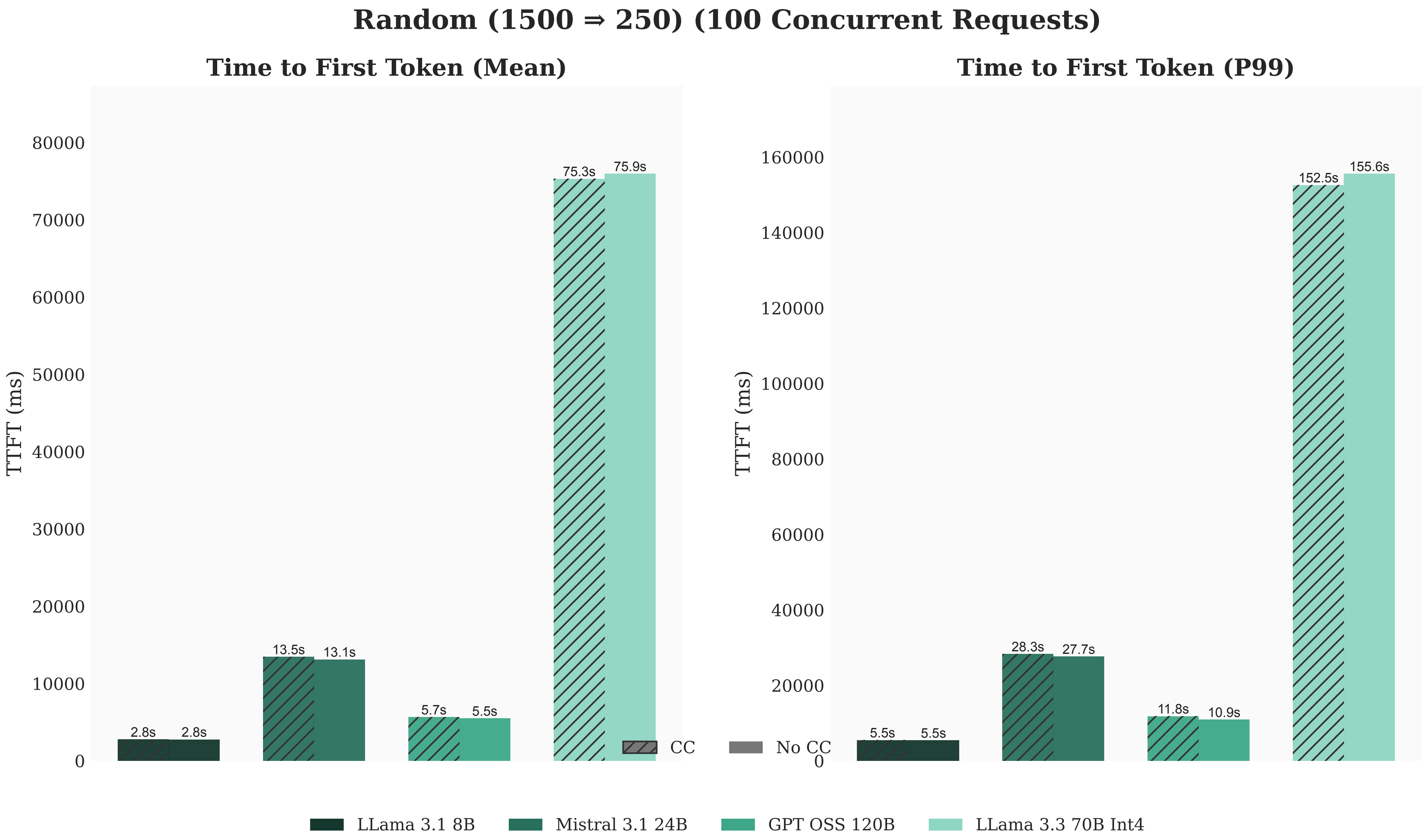
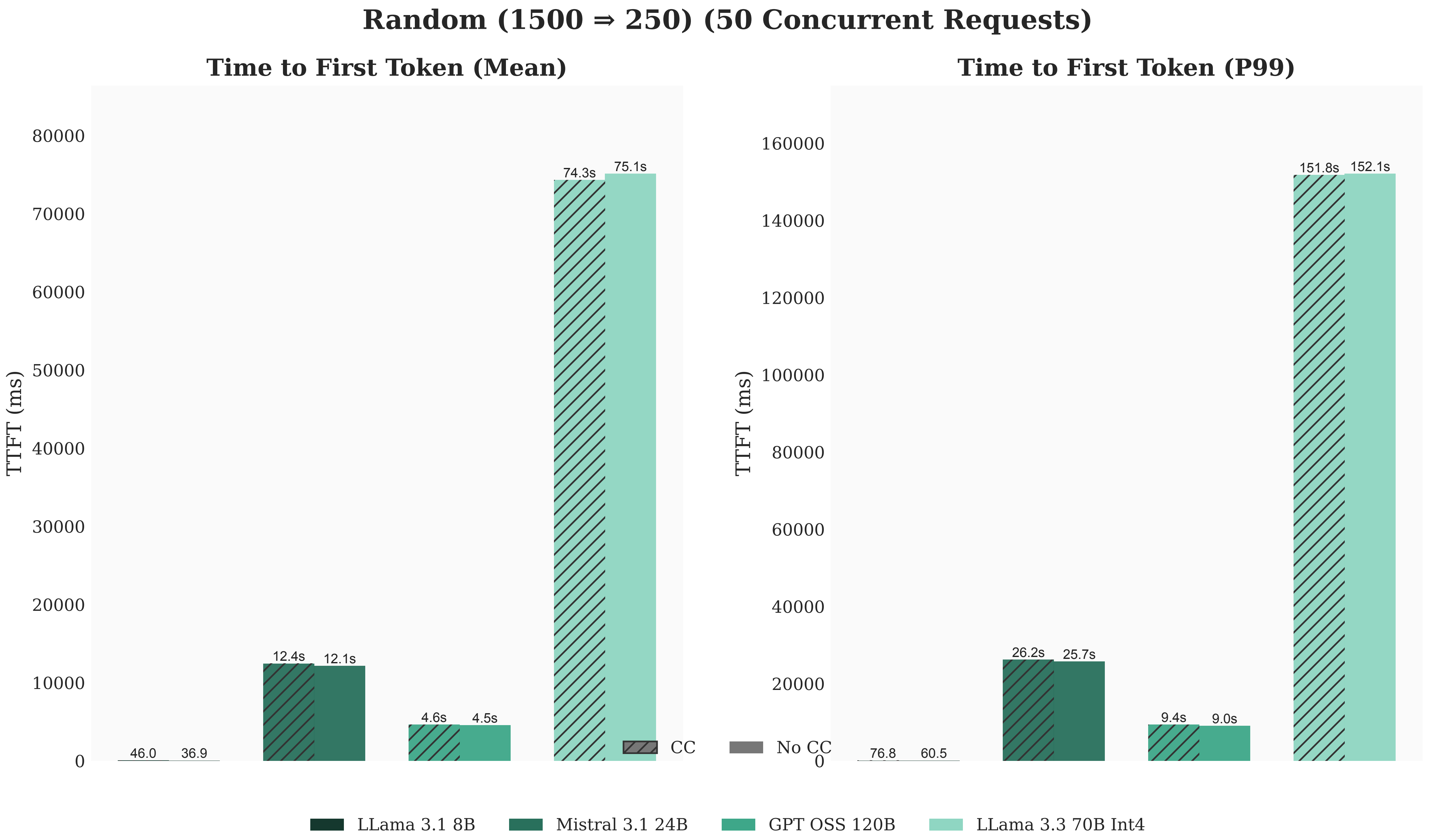
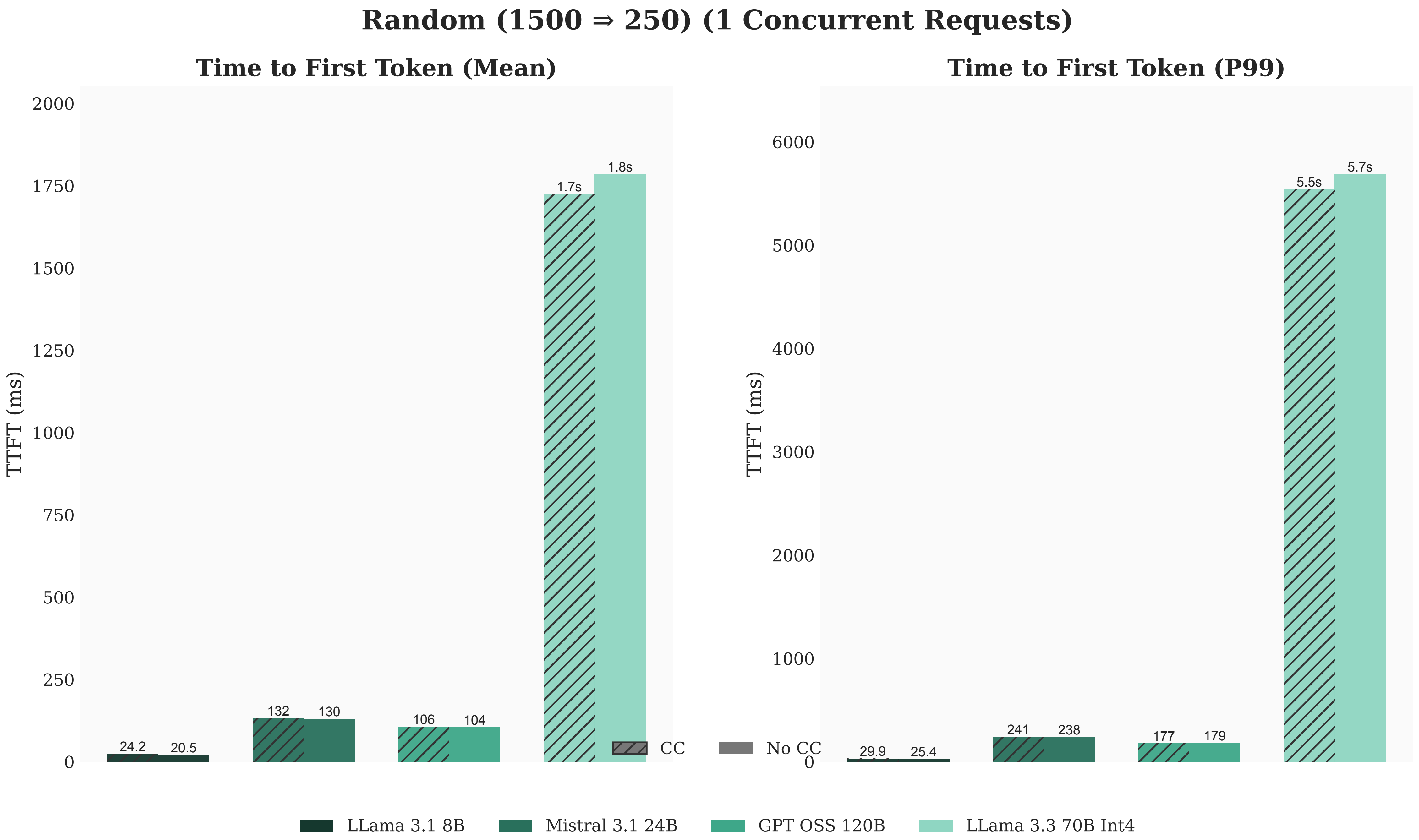
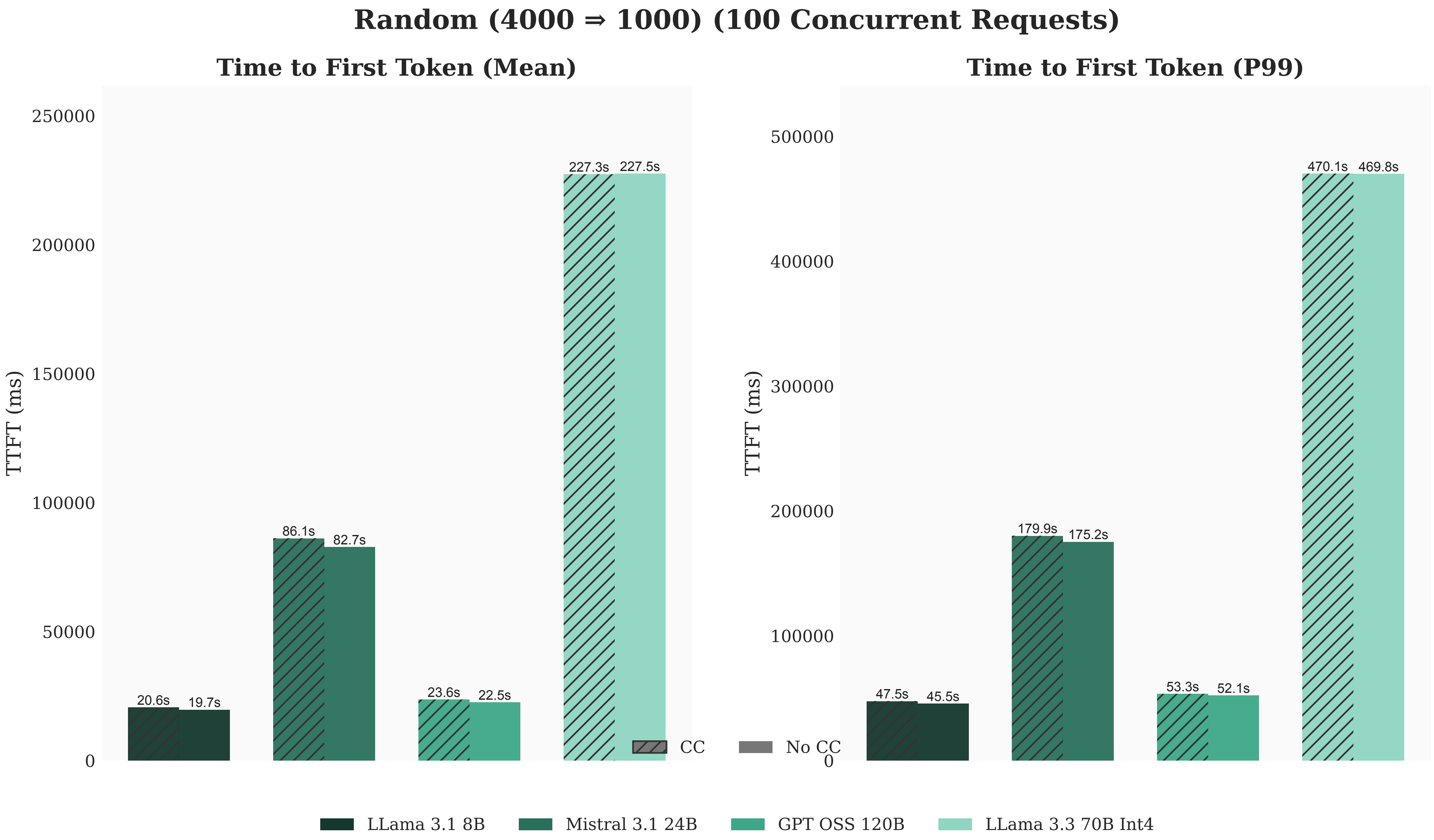
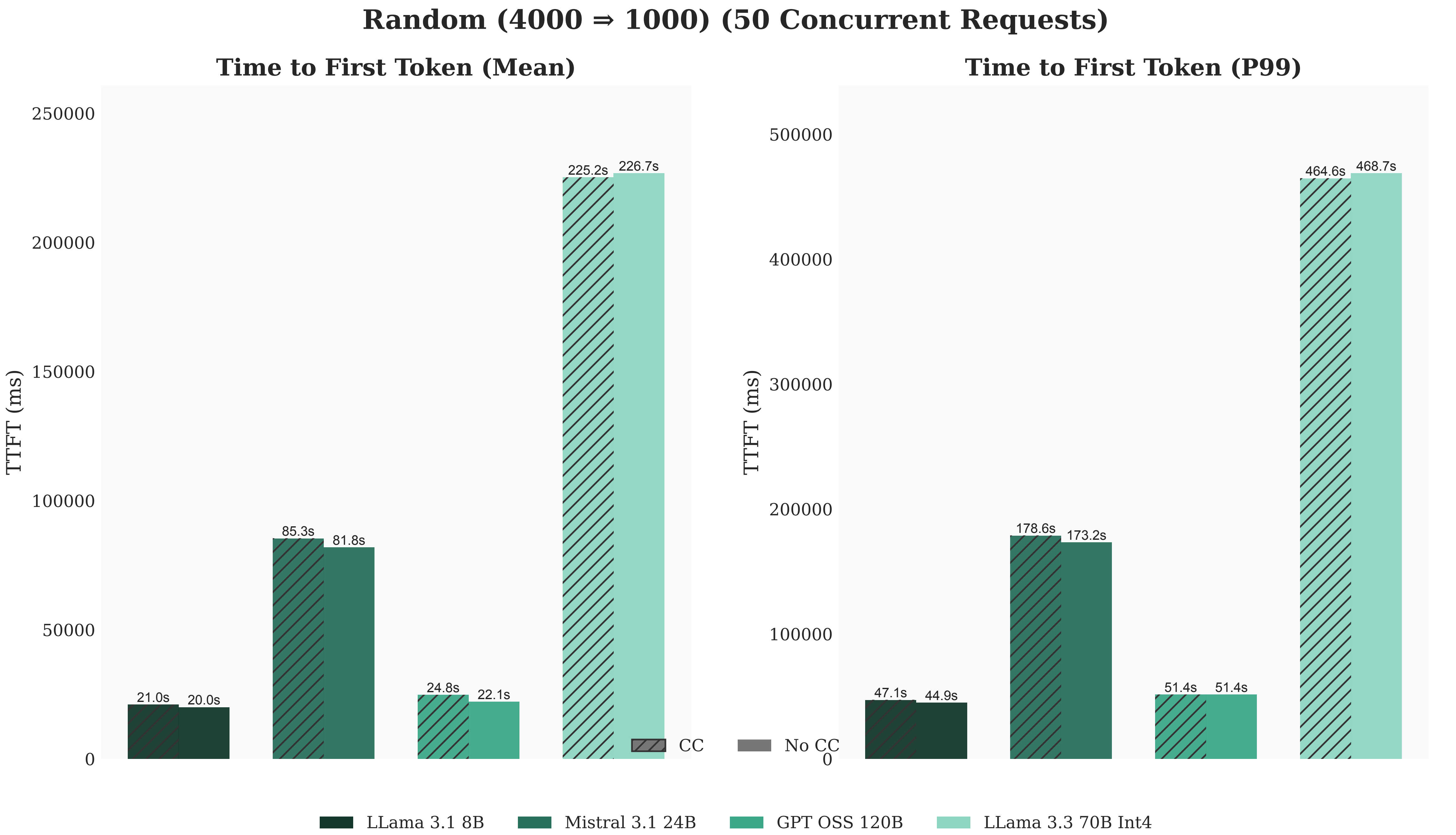
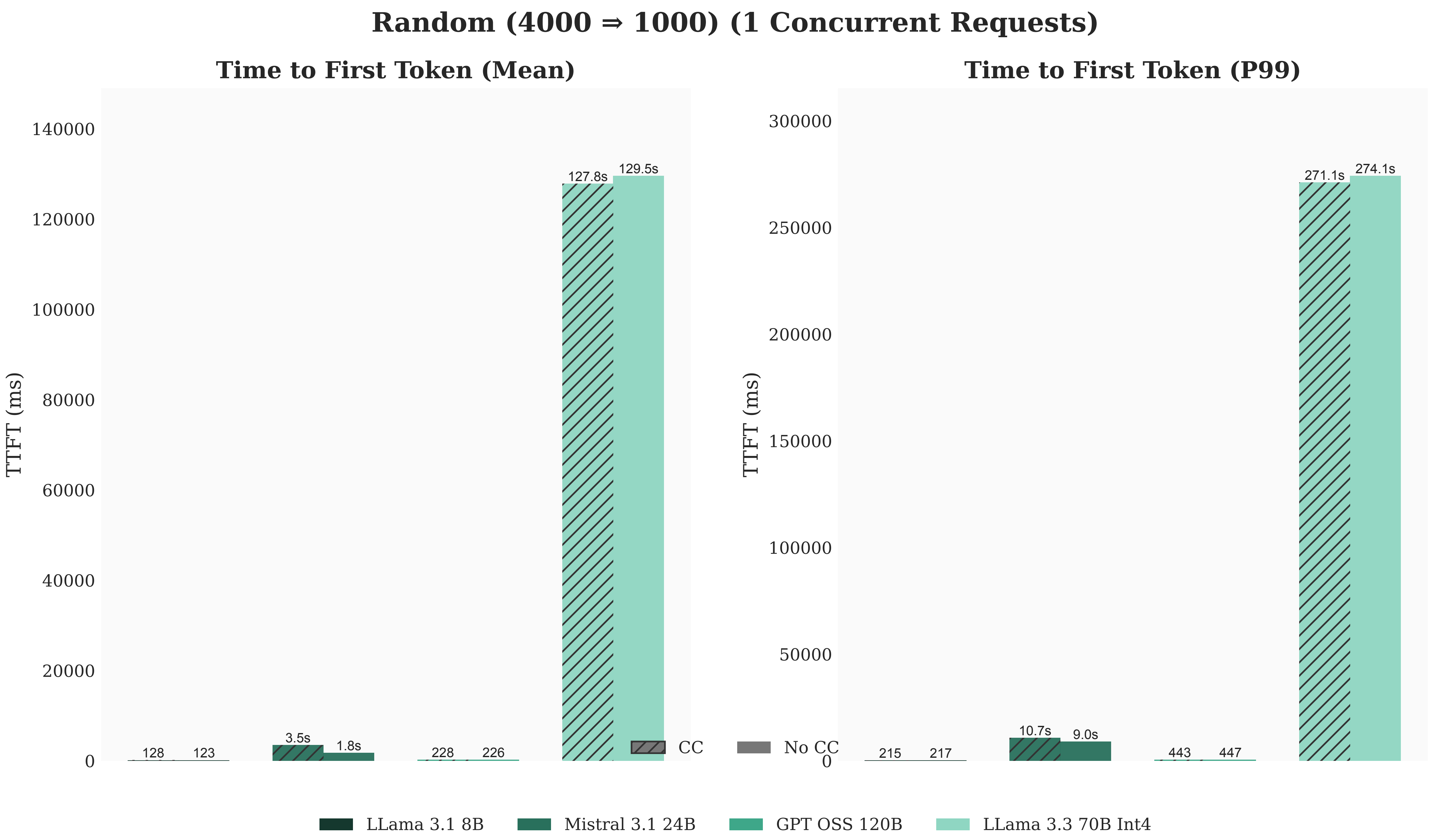

Random (1500 ⇒ 250) (100 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
60000
70000
80000
2.8s
2.8s
13.5s
13.1s
5.7s
5.5s
75.3s
75.9s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
160000
5.5s
5.5s
28.3s
27.7s
11.8s
10.9s
152.5s
155.6s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

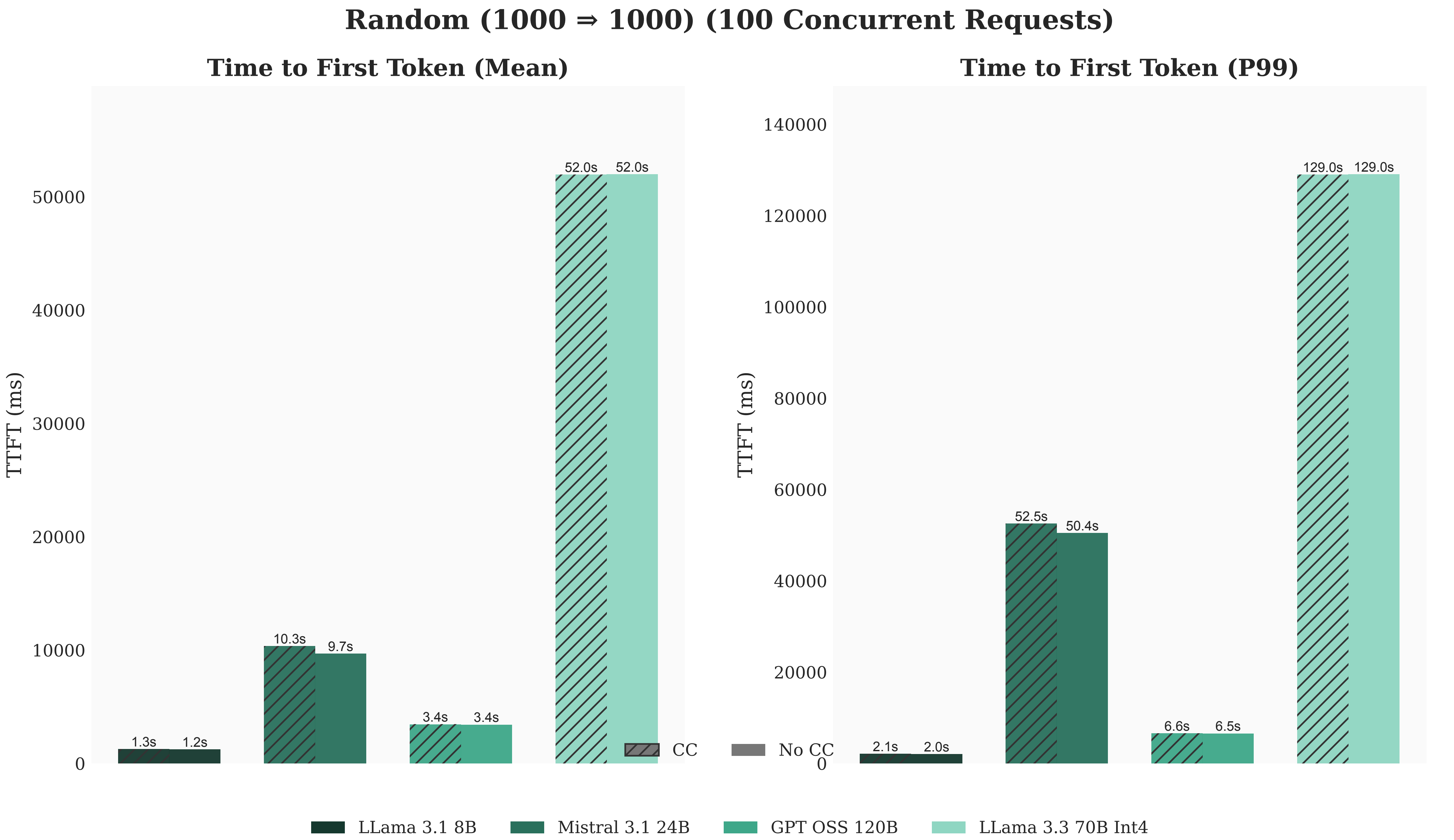

Random (1500 ⇒ 250) (50 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
60000
70000
80000
46.0
36.9
12.4s
12.1s
4.6s
4.5s
74.3s
75.1s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
160000
76.8
60.5
26.2s
25.7s
9.4s
9.0s
151.8s
152.1s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

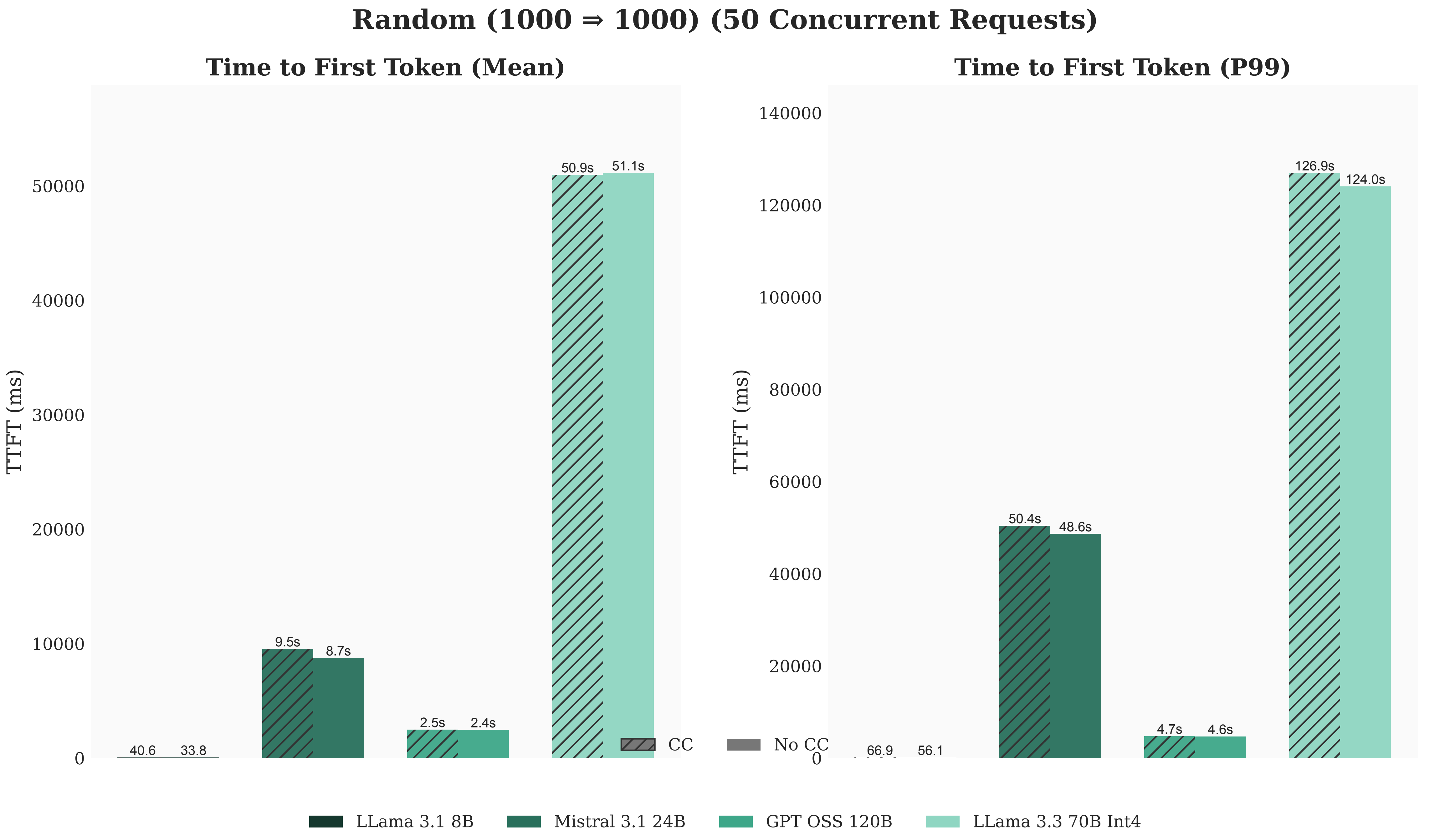

Random (1500 ⇒ 250) (1 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
0
250
500
750
1000
1250
1500
1750
2000
24.2
20.5
132
130
106
104
1.7s
1.8s
Time to First Token (P99)
TTFT (ms)
0
1000
2000
3000
4000
5000
6000
29.9
25.4
241
238
177
179
5.5s
5.7s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

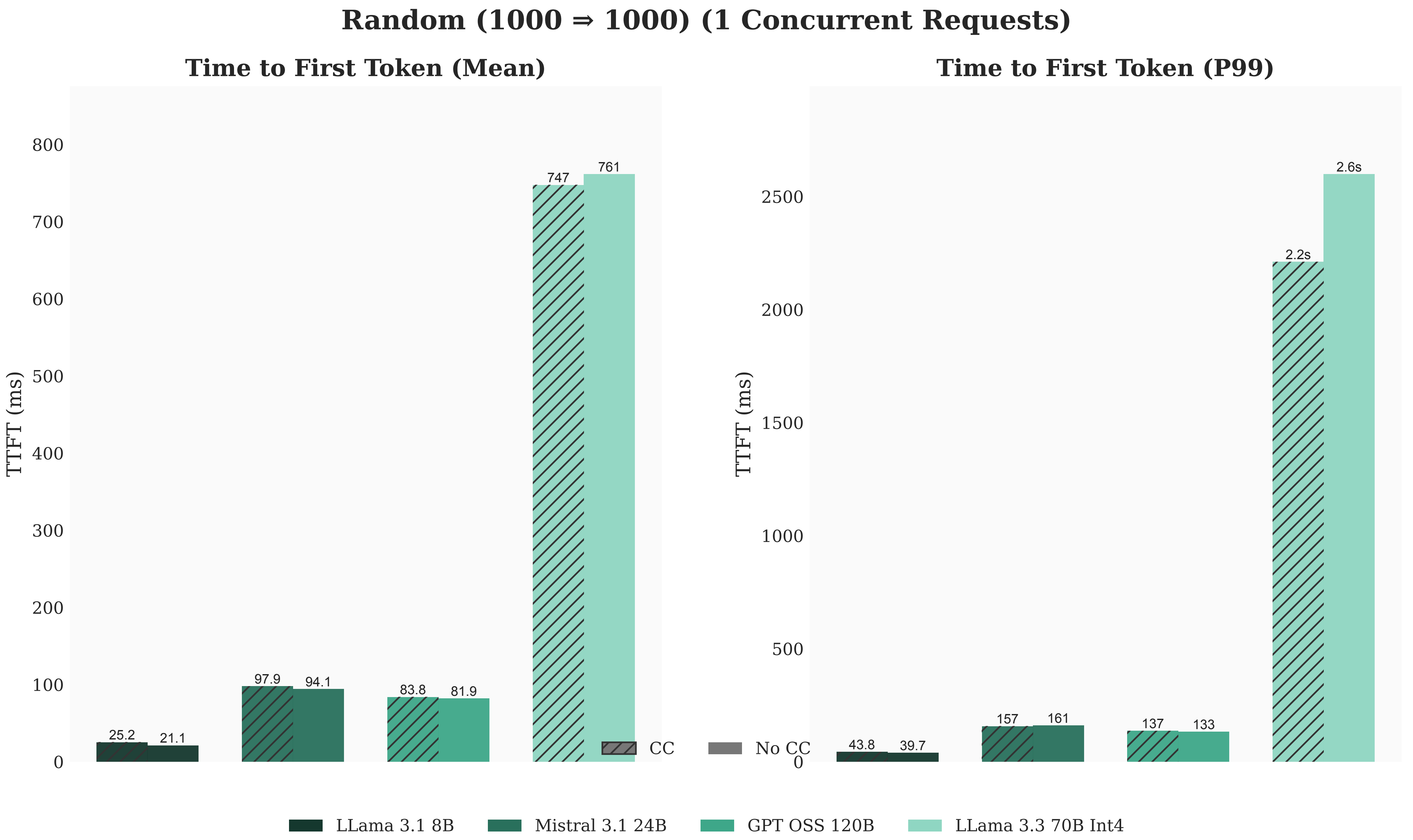

Random (4000 ⇒ 1000) (100 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
250000
200000
150000
100000
50000
0
20.6s
19.7s
86.1s
82.7s
23.6s
22.5s
227.3s
227.5s
Time to First Token (P99)
TTFT (ms)
500000
400000
300000
200000
100000
0
47.5s
45.5s
179.9s
175.2s
53.3s
52.1s
470.1s
469.8s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

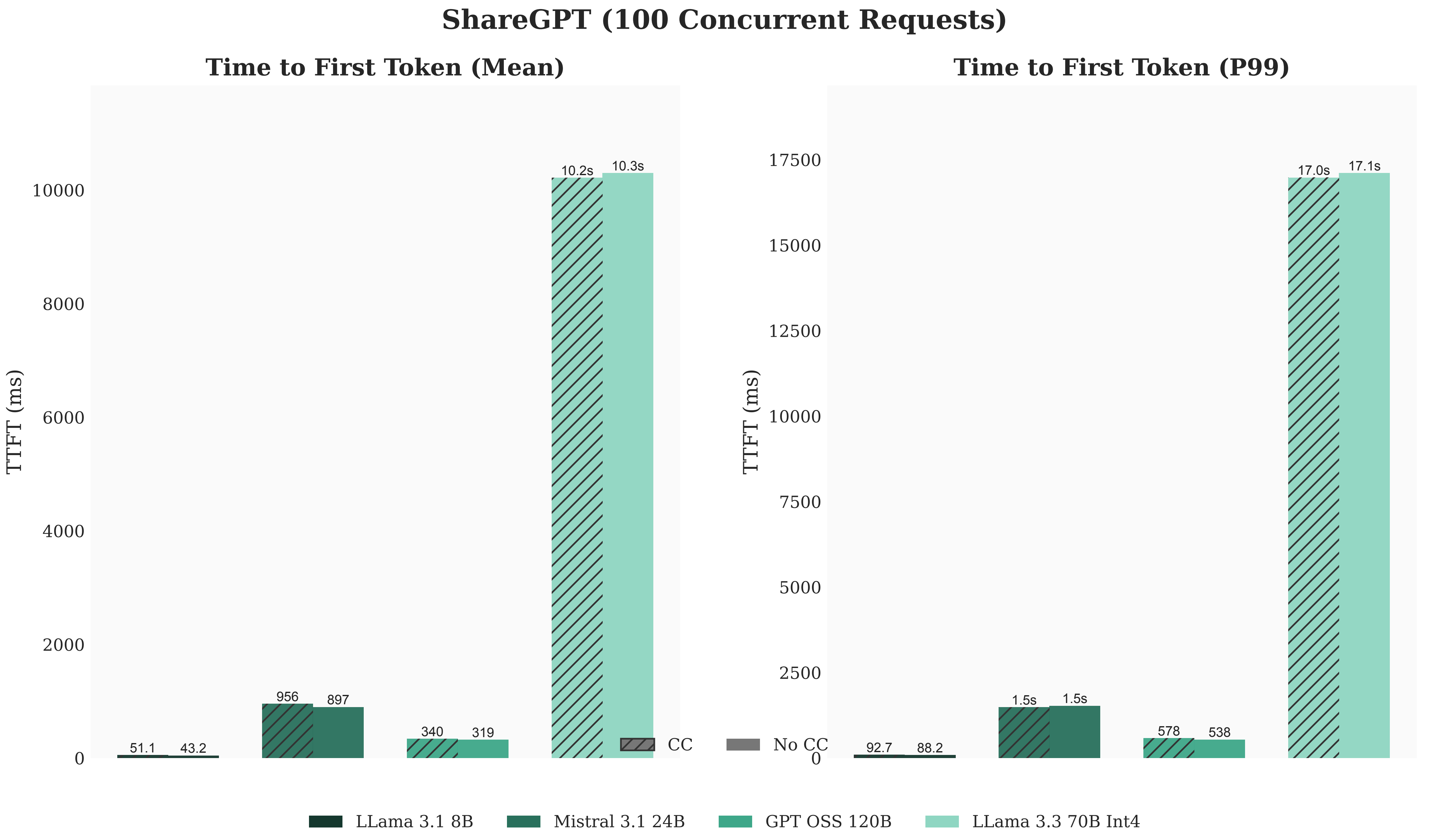

Random (4000 ⇒ 1000) (50 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
250000
200000
150000
100000
50000
0
21.0s
20.0s
85.3s
81.8s
24.8s
22.1s
225.2s
226.7s
Time to First Token (P99)
TTFT (ms)
500000
400000
300000
200000
100000
0
47.1s
44.9s
178.6s
173.2s
51.4s
51.4s
464.6s
468.7s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

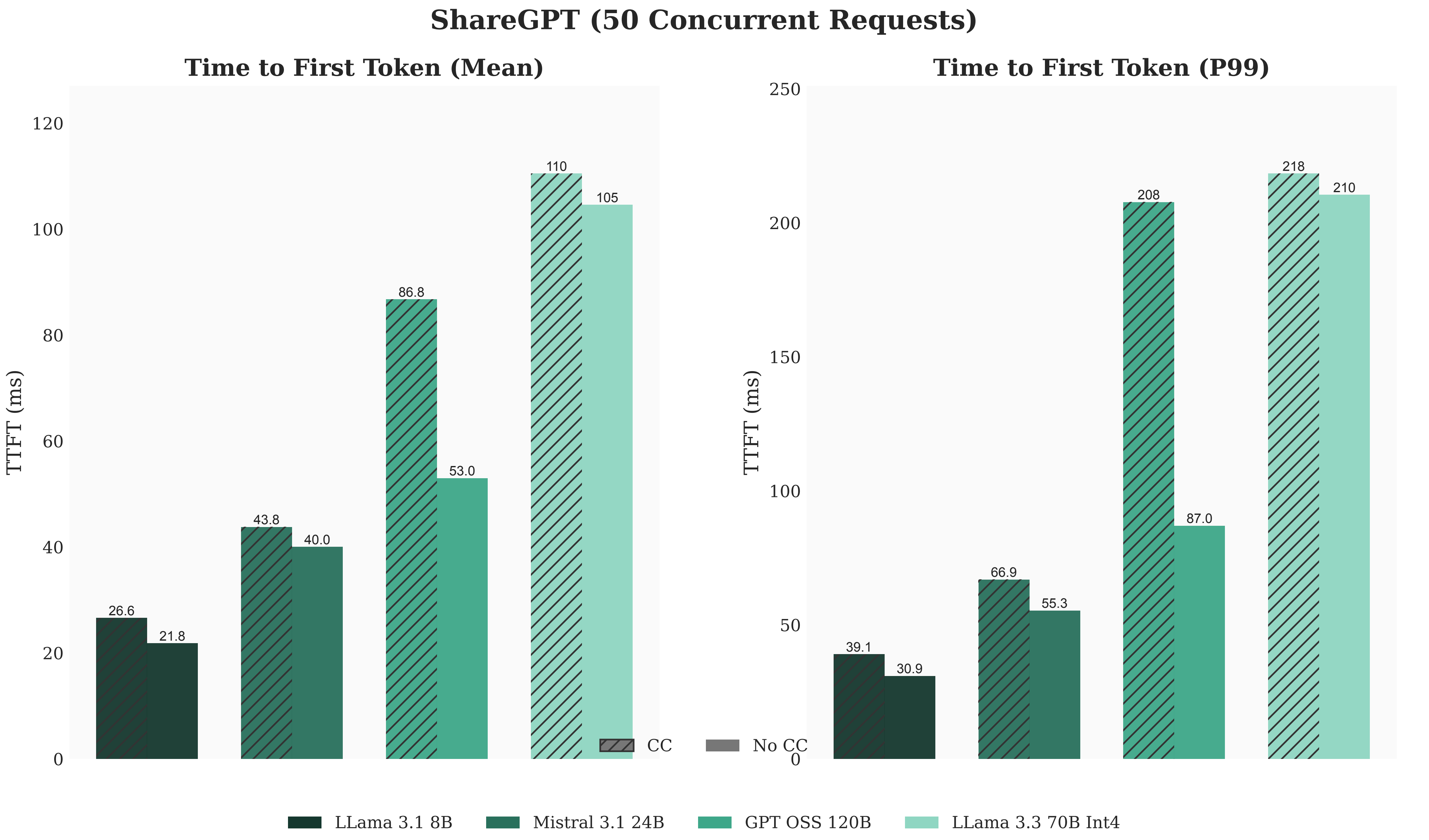

Random (4000 ⇒ 1000) (1 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
128
123
3.5s
1.8s
228
226
127.8s
129.5s
Time to First Token (P99)
TTFT (ms)
0
50000
100000
150000
200000
250000
300000
215
217
10.7s
9.0s
443
447
271.1s
274.1s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

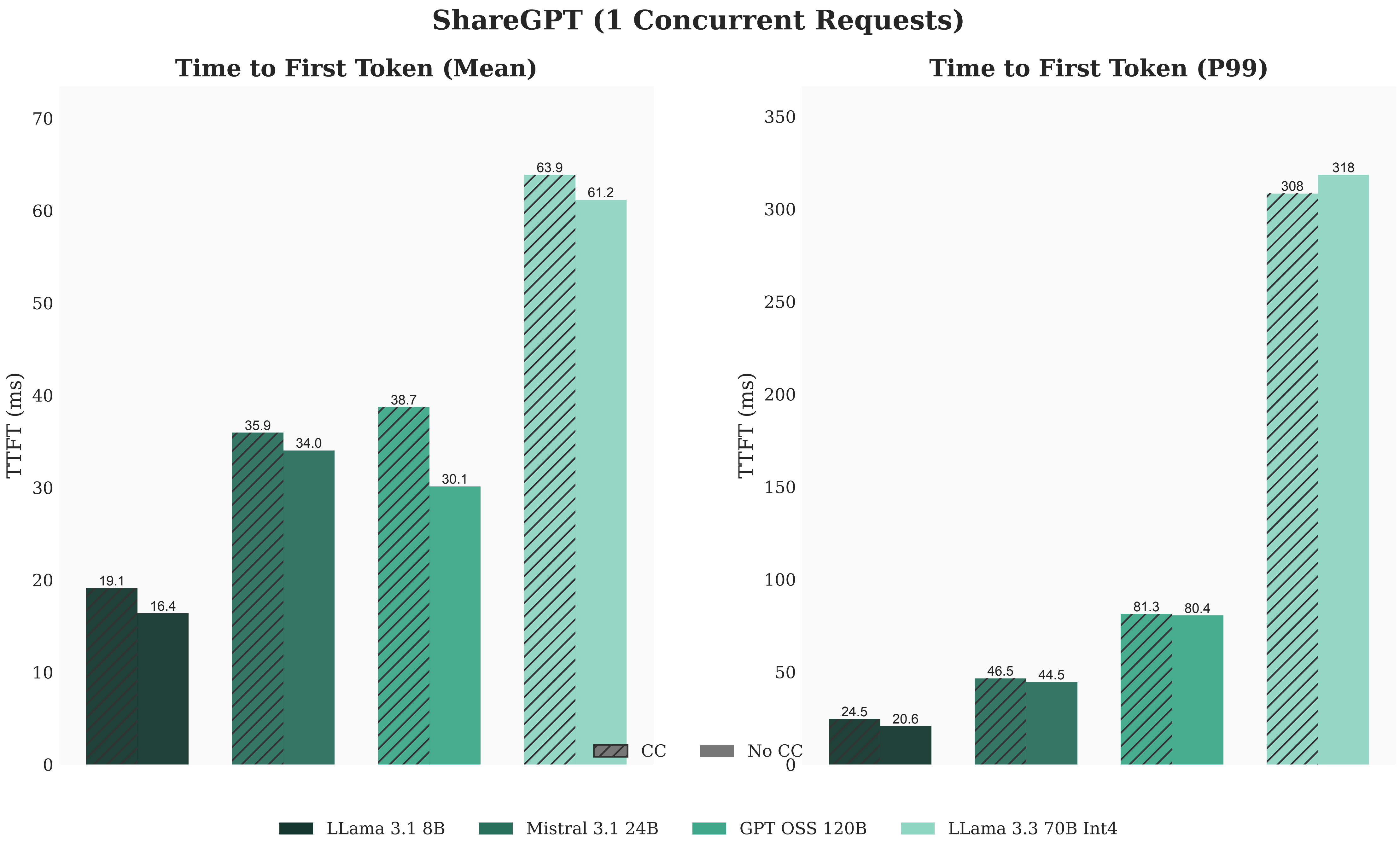

Random (1000 ⇒ 1000) (100 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
1.3s
1.2s
10.3s
9.7s
3.4s
3.4s
52.0s
52.0s
CC
No CC
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
2.1s
2.0s
52.5s
50.4s
6.6s
6.5s
129.0s
129.0s
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

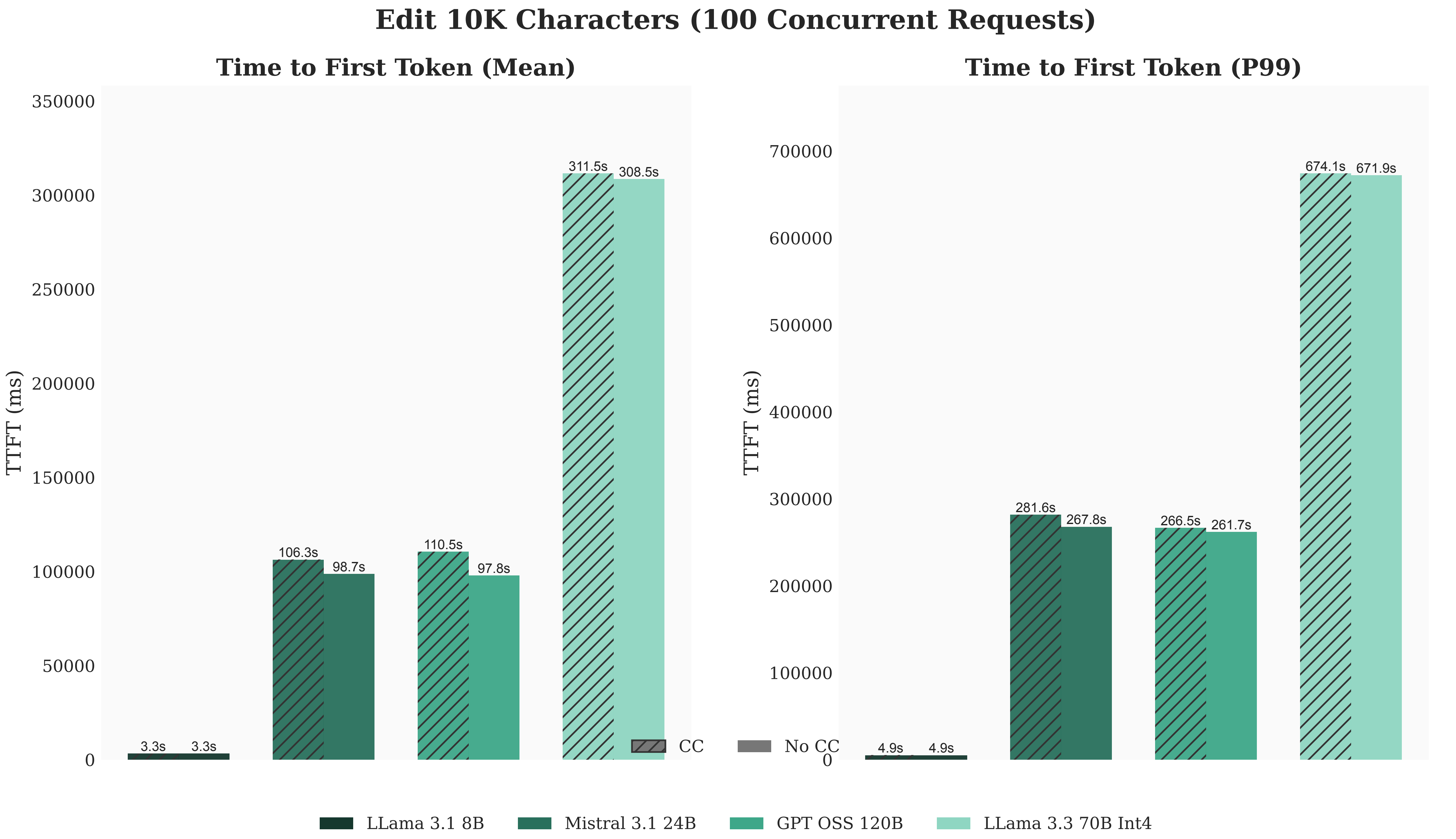

Random (1000 ⇒ 1000) (50 Concurrent Requests)
Time to First Token (Mean)
Time to First Token (P99)
TTFT (ms)
0
10000
20000
30000
40000
50000
40.6
33.8
9.5s
8.7s
2.5s
2.4s
50.9s
51.1s
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
66.9
56.1
50.4s
48.6s
4.7s
4.6s
126.9s
124.0s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (1 Concurrent Requests)
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
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
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


ShareGPT (100 Concurrent Requests)
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
CC
No CC
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
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


ShareGPT (50 Concurrent Requests)
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
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
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

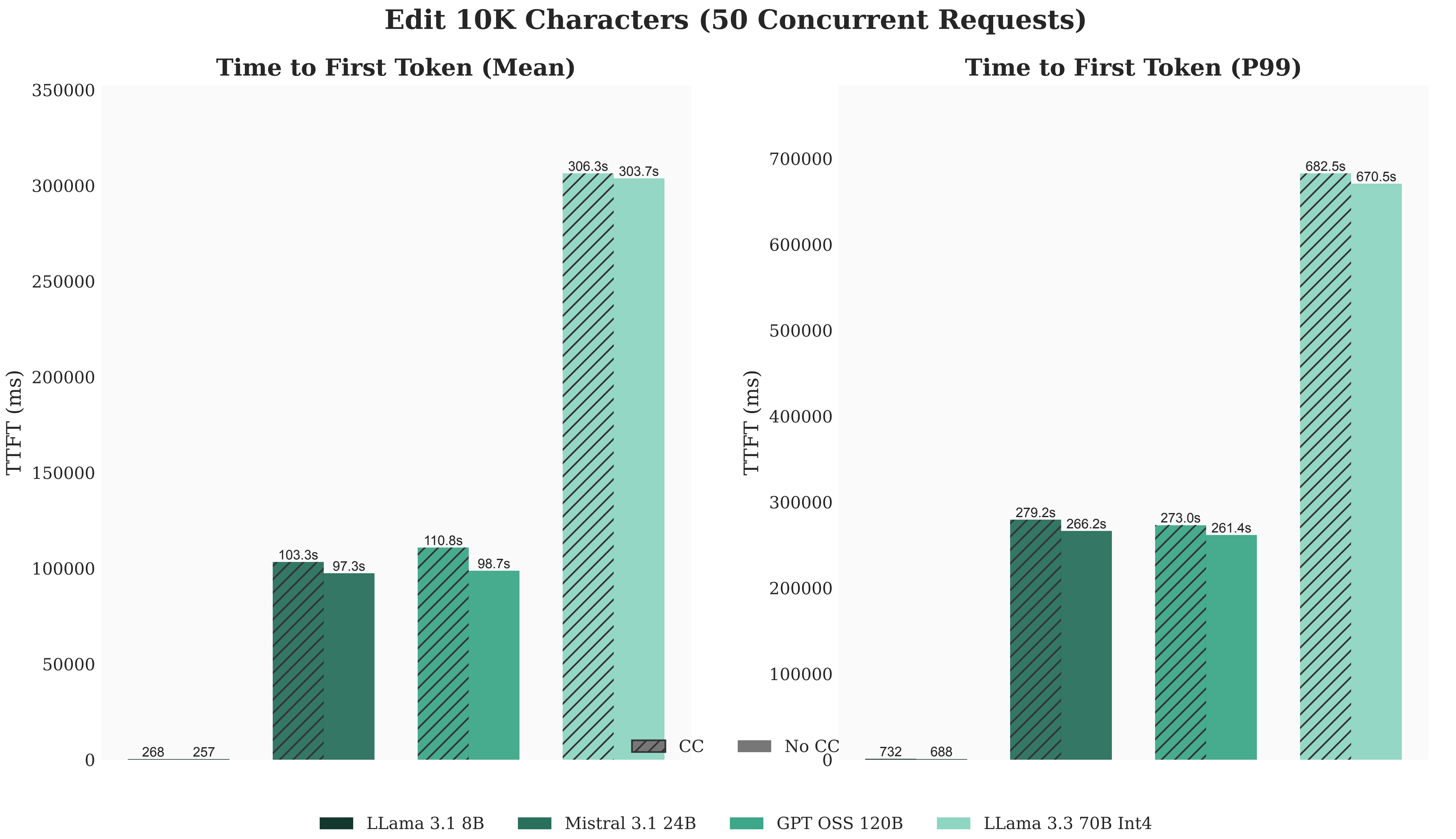

ShareGPT (1 Concurrent Requests)
Time to First Token (Mean)
Time to First Token (P99)
TTFT (ms)
0
10
20
30
40
50
60
70
19.1
16.4
35.9
34.0
38.7
30.1
63.9
61.2
0
50
100
150
200
250
300
350
24.5
20.6
46.5
44.5
81.3
80.4
308
318
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

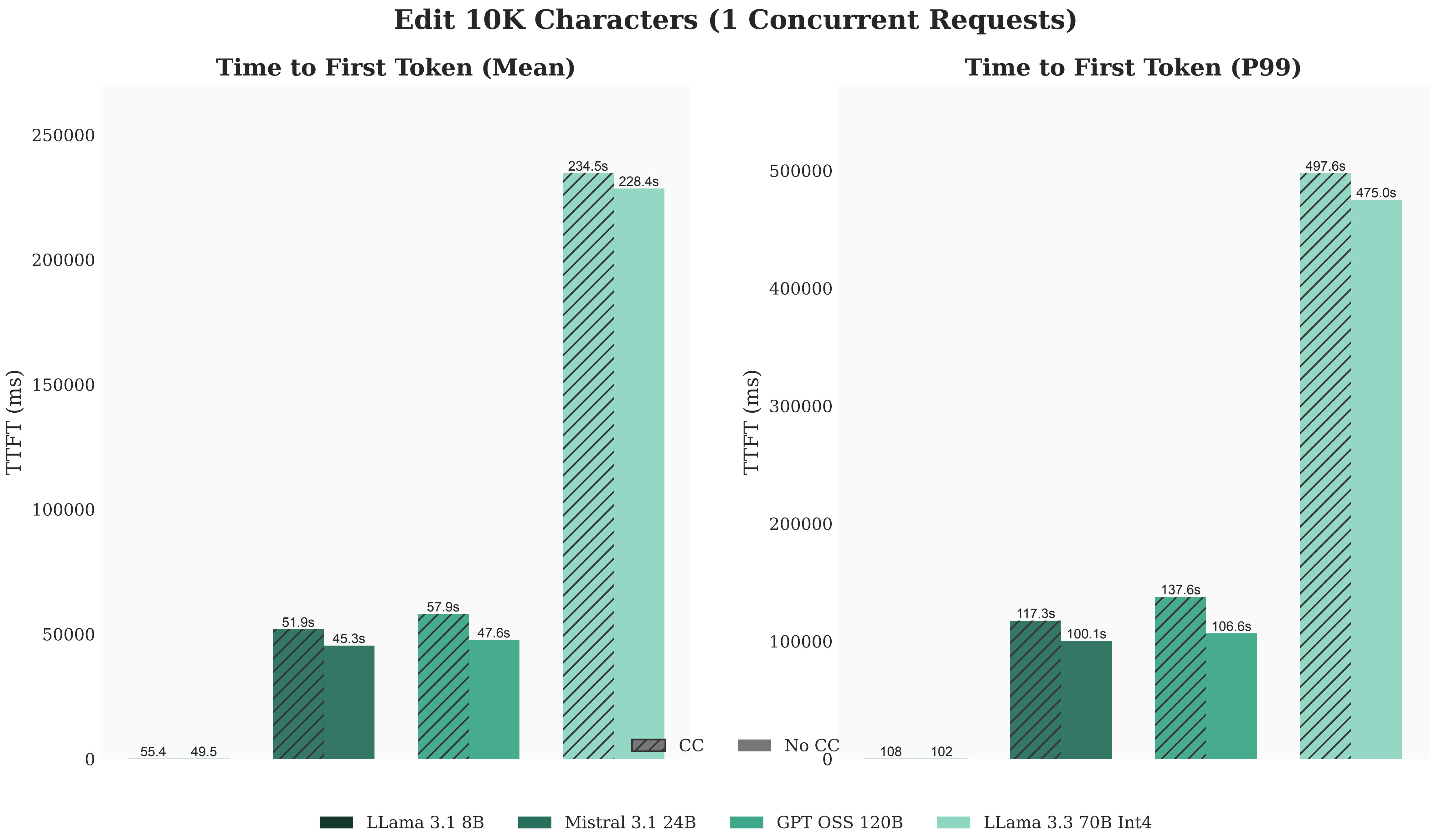


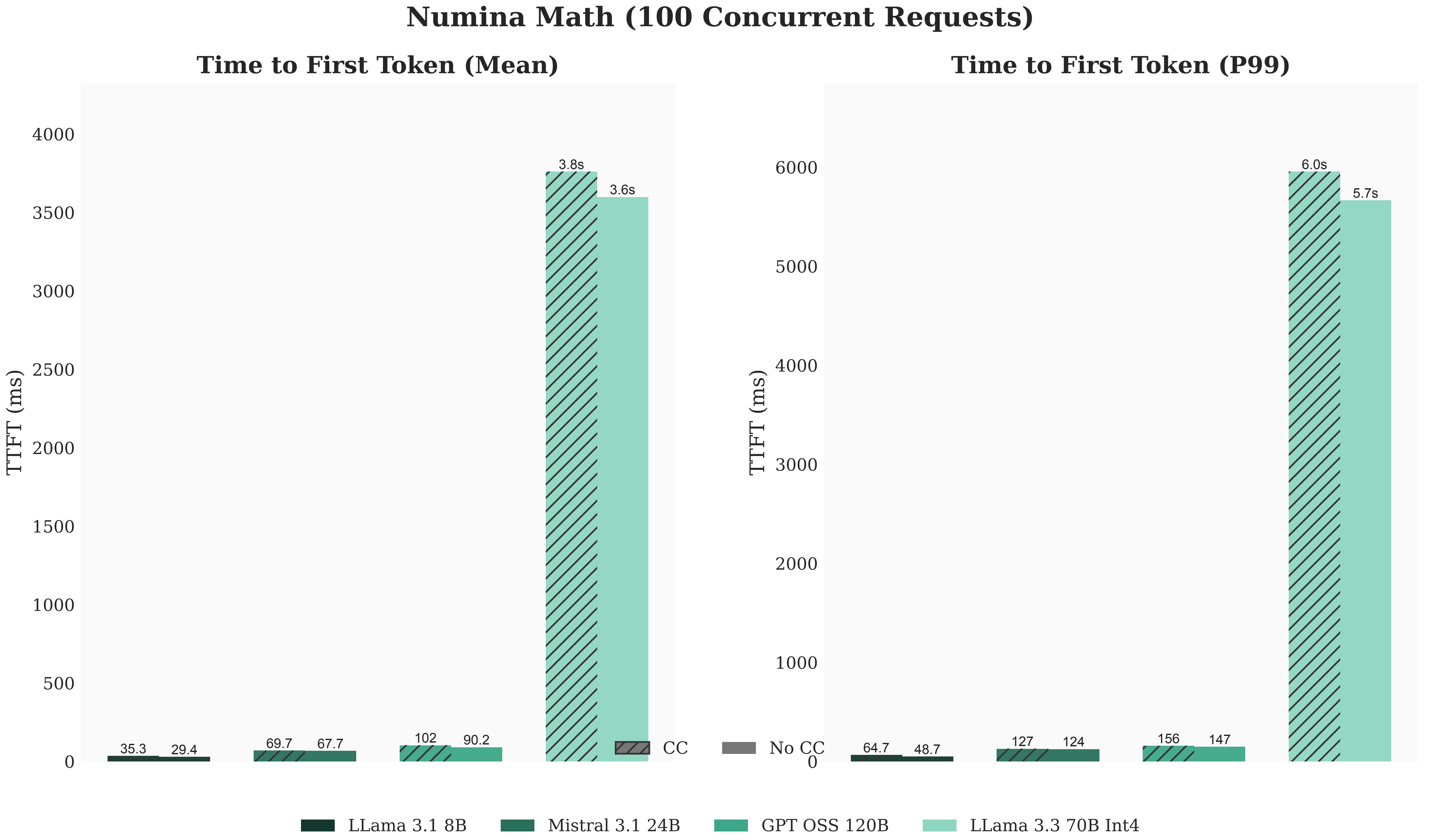
Edit 10K Characters (100 Concurrent Requests)

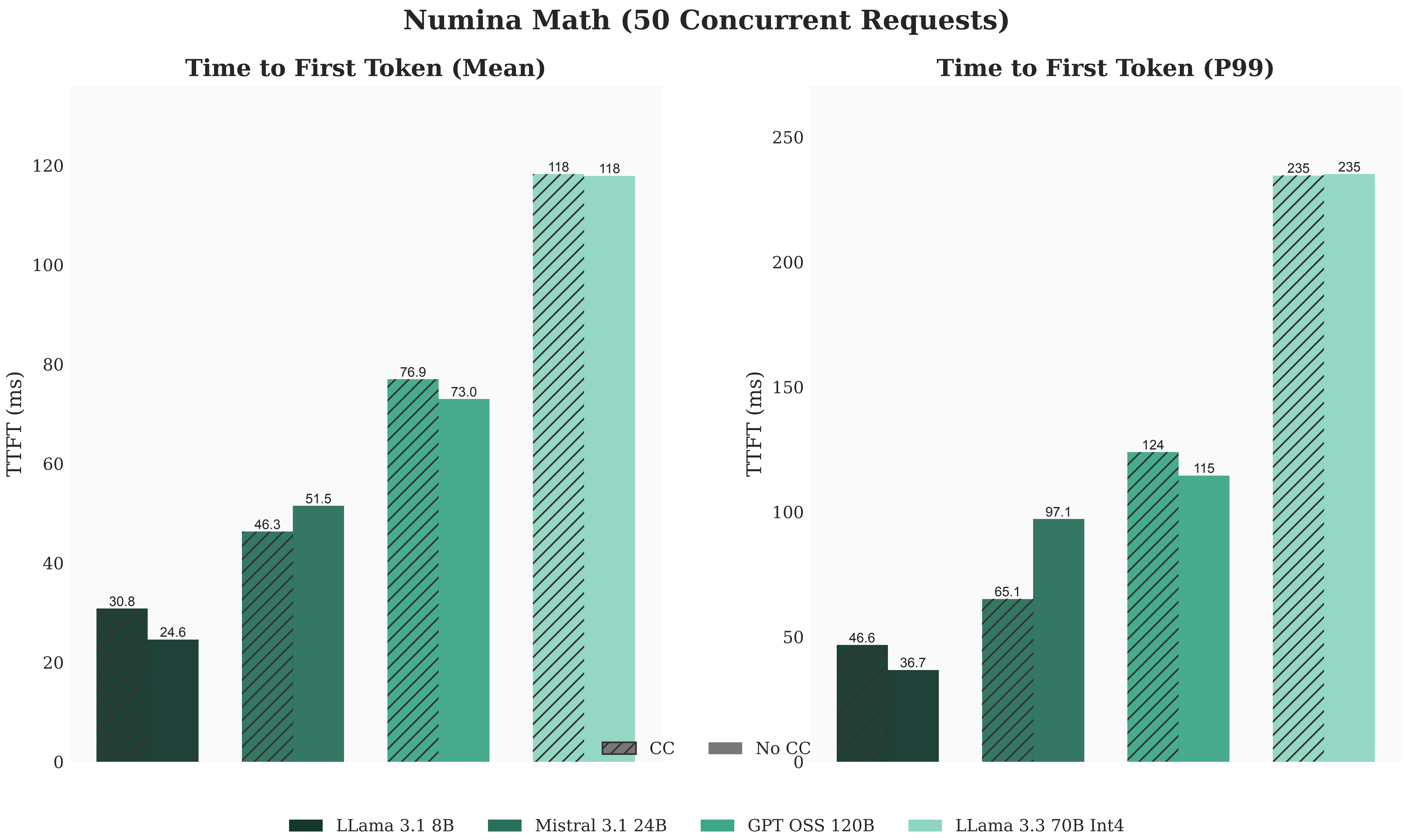
Time to First Token (Mean)

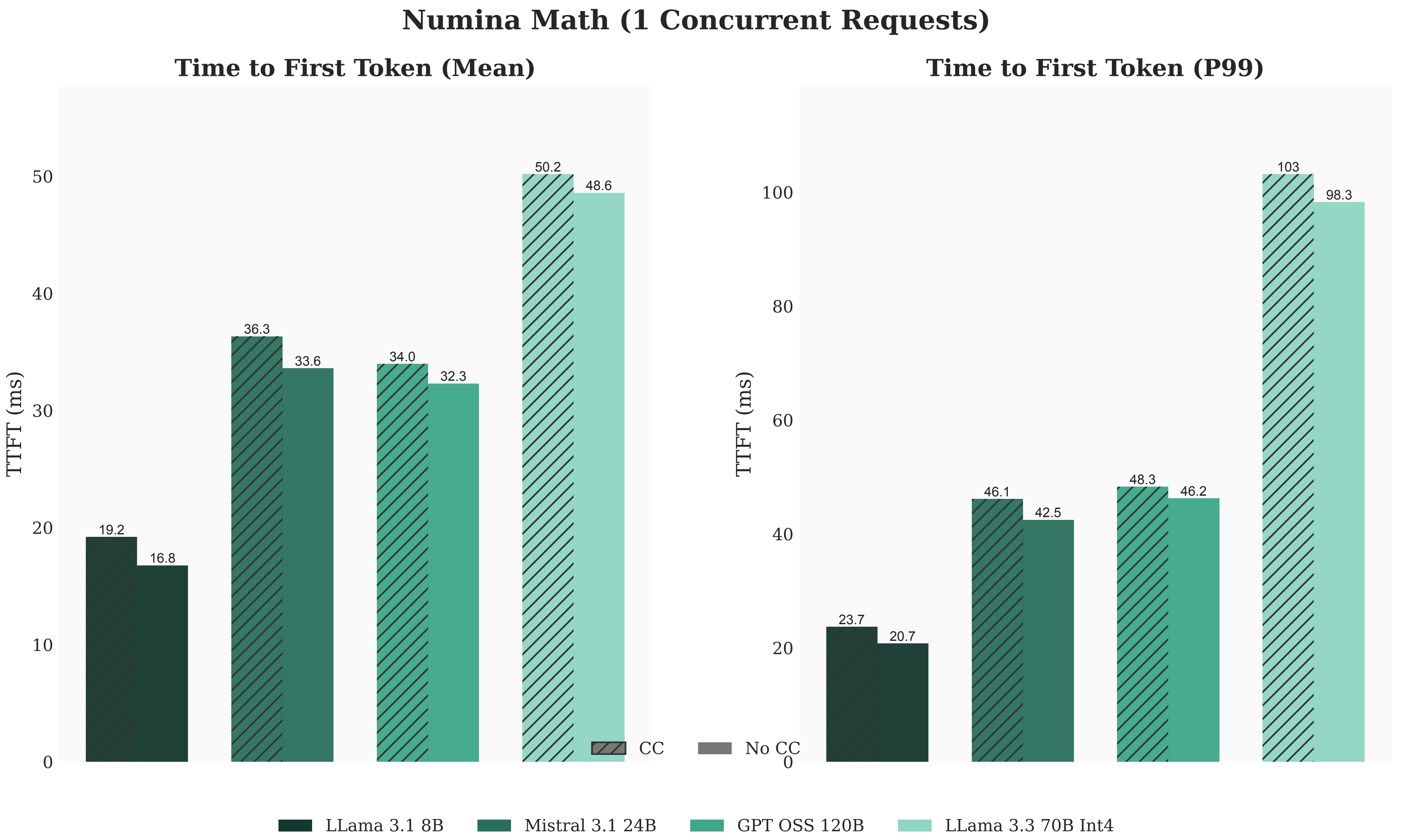
Time to First Token (P99)


Edit 10K Characters (50 Concurrent Requests)
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
CC
No CC
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
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Edit 10K Characters (1 Concurrent Requests)
Time to First Token (Mean)
TTFT (ms)
250000
200000
150000
100000
50000
0
55.4
49.5
51.9s
45.3s
57.9s
47.6s
234.5s
228.4s
Time to First Token (P99)
500000
400000
300000
200000
100000
0
108
102
117.3s
100.1s
137.6s
106.6s
497.6s
475.0s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (100 Concurrent Requests)
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
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
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (50 Concurrent Requests)
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
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
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (1 Concurrent Requests)
Time to First Token (Mean)
Time to First Token (P99)
TTFT (ms)
0
10
20
30
40
50
19.2
16.8
36.3
33.6
34.0
32.3
50.2
48.6
TTFT (ms)
0
20
40
60
80
100
23.7
20.7
46.1
42.5
48.3
46.2
103
98.3
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4