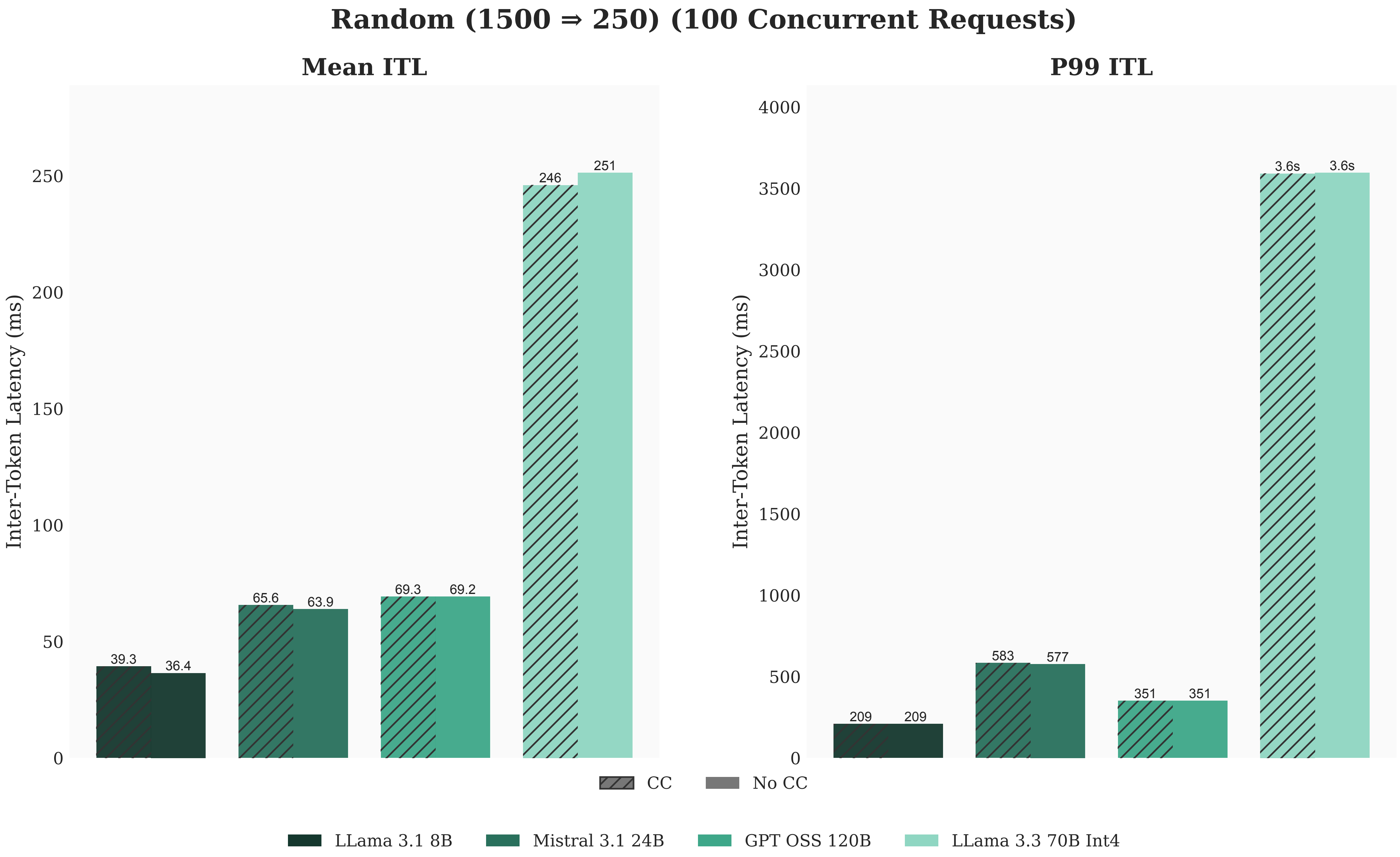
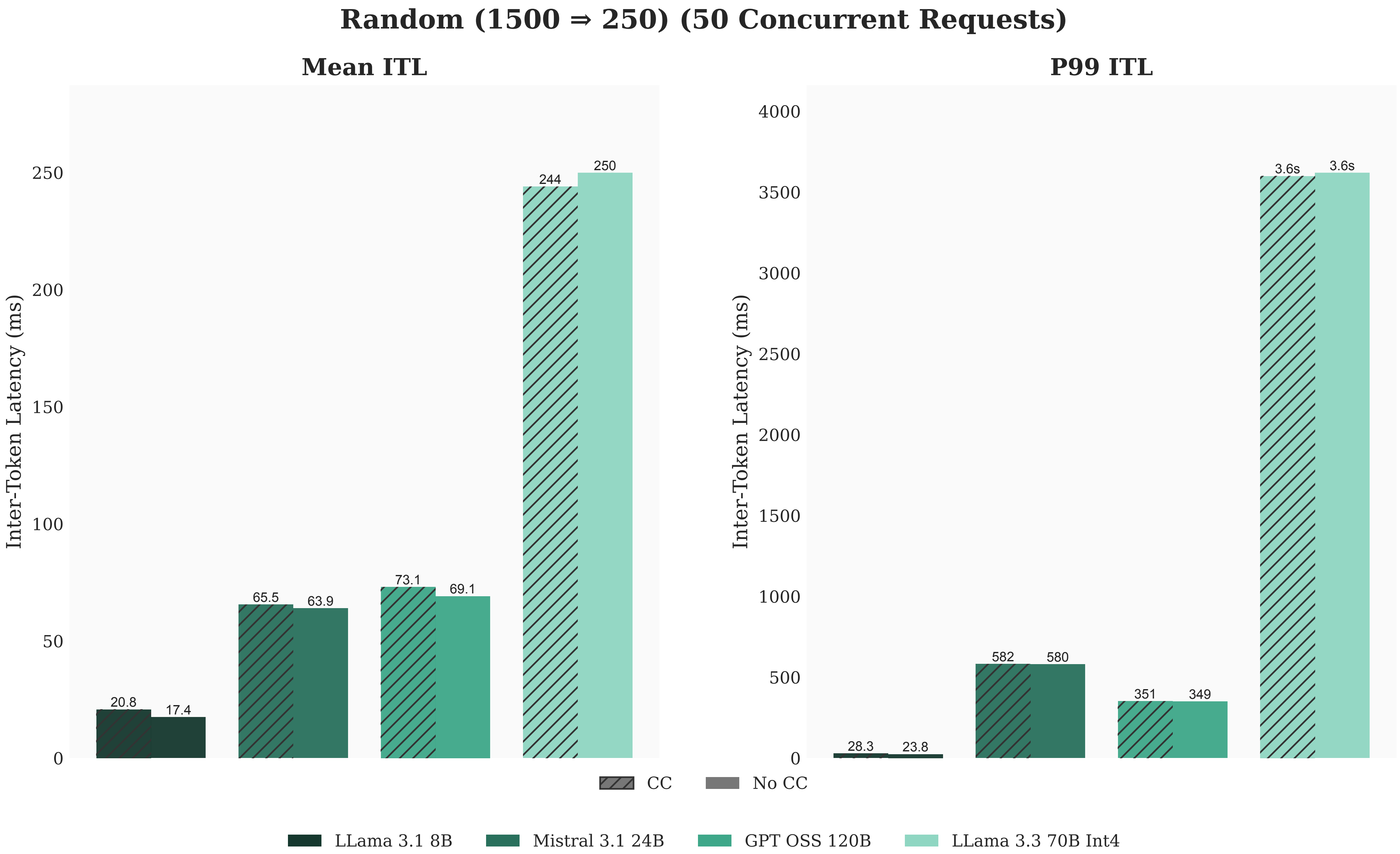
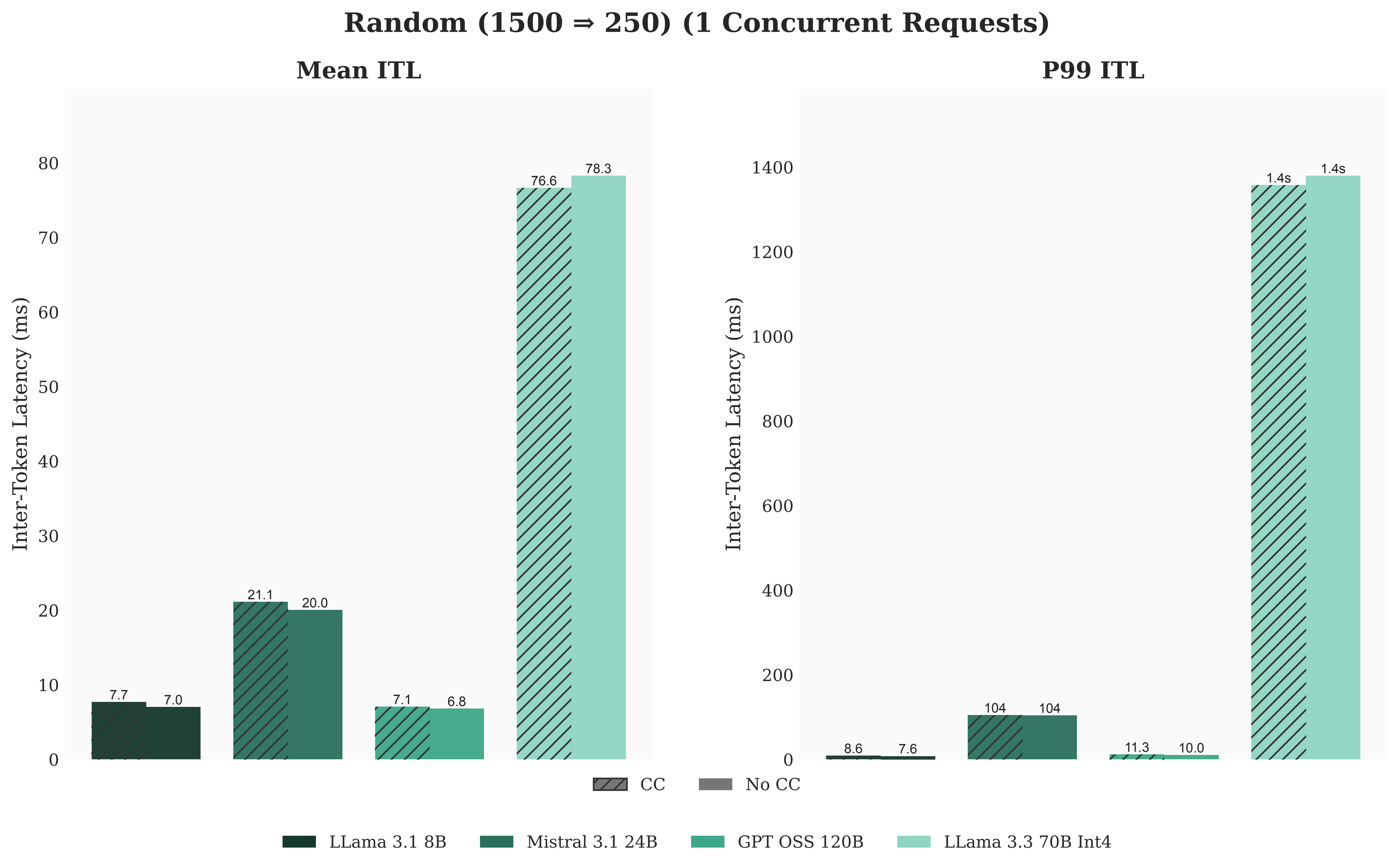
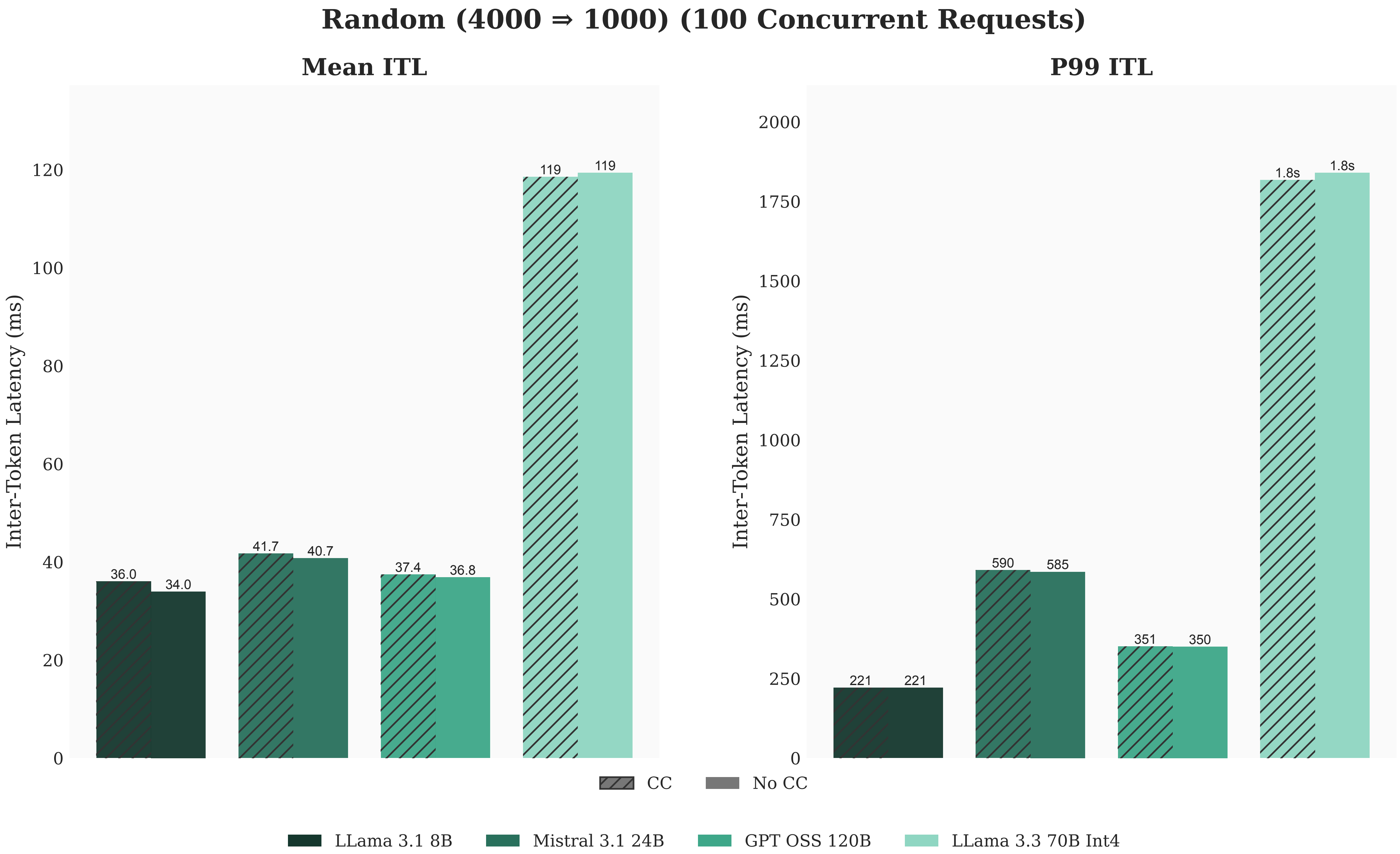
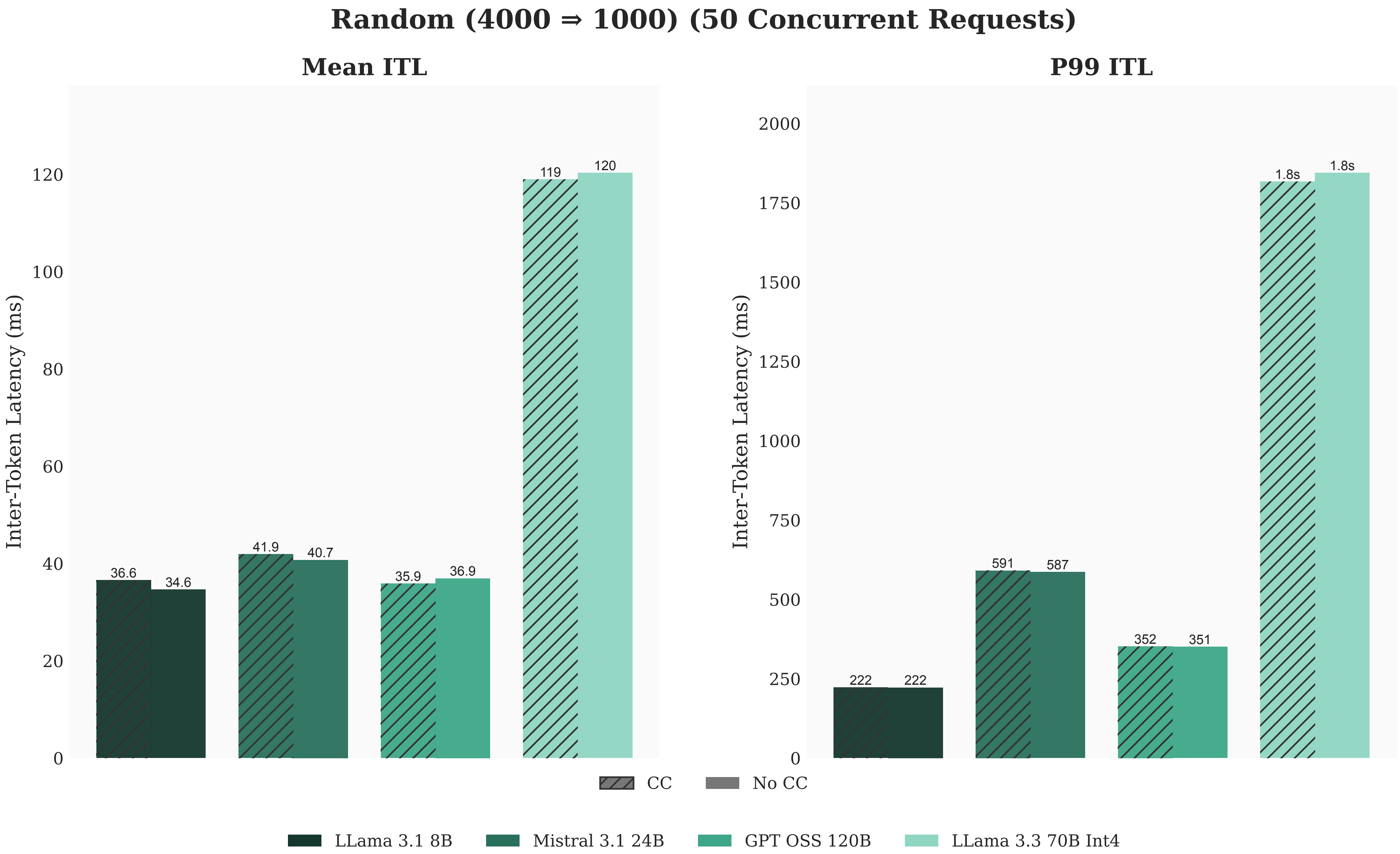
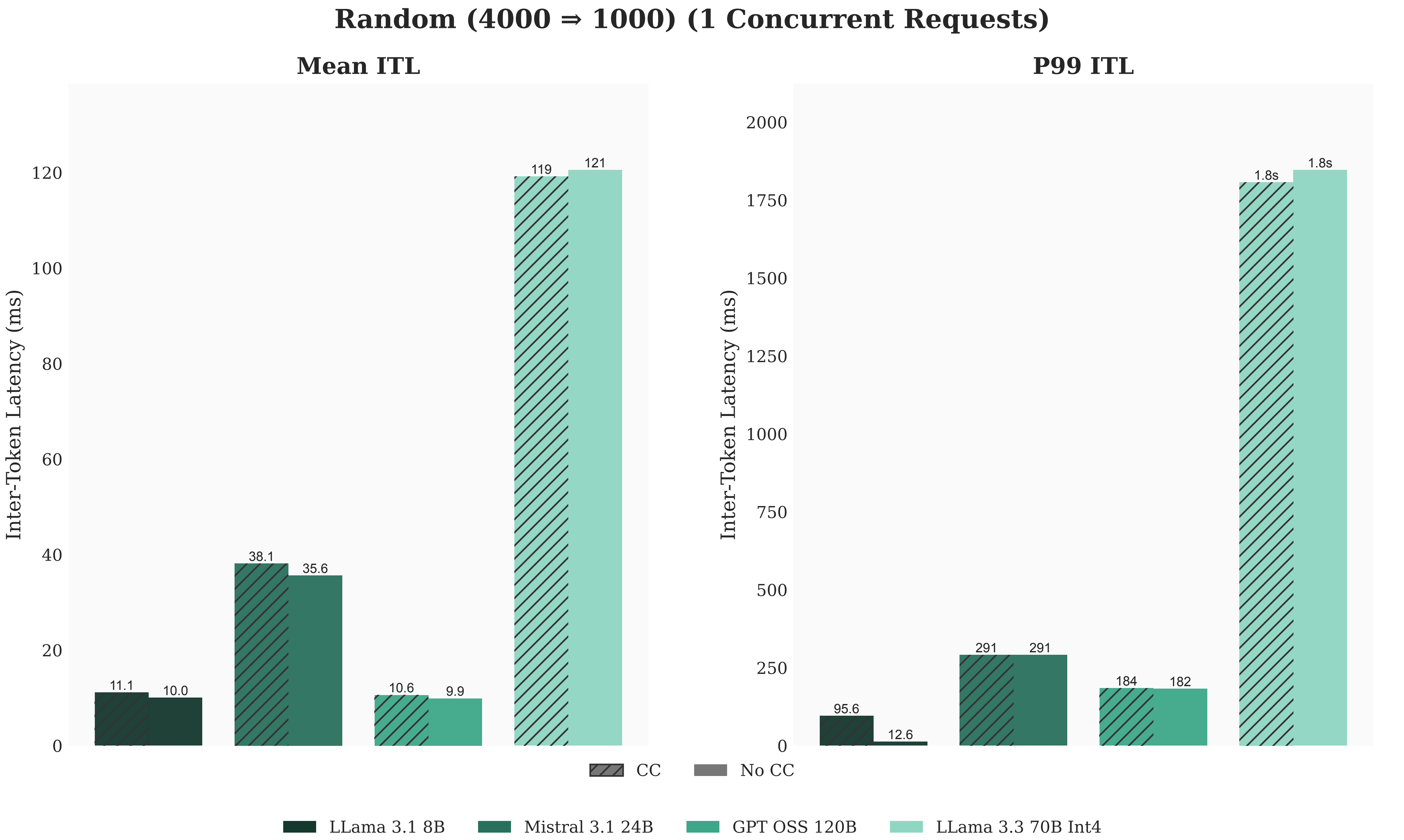

Random (1500 ⇒ 250) (100 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
50
100
150
200
250
39.3
36.4
65.6
63.9
69.3
69.2
246
251
Inter-Token Latency (ms)
0
500
1000
1500
2000
2500
3000
3500
4000
209
209
583
577
351
351
3.6s
3.6s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Random (1500 ⇒ 250) (50 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
50
100
150
200
250
20.8
17.4
65.5
63.9
73.1
69.1
244
250
Inter-Token Latency (ms)
0
500
1000
1500
2000
2500
3000
3500
4000
28.3
23.8
582
580
351
349
3.6s
3.6s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

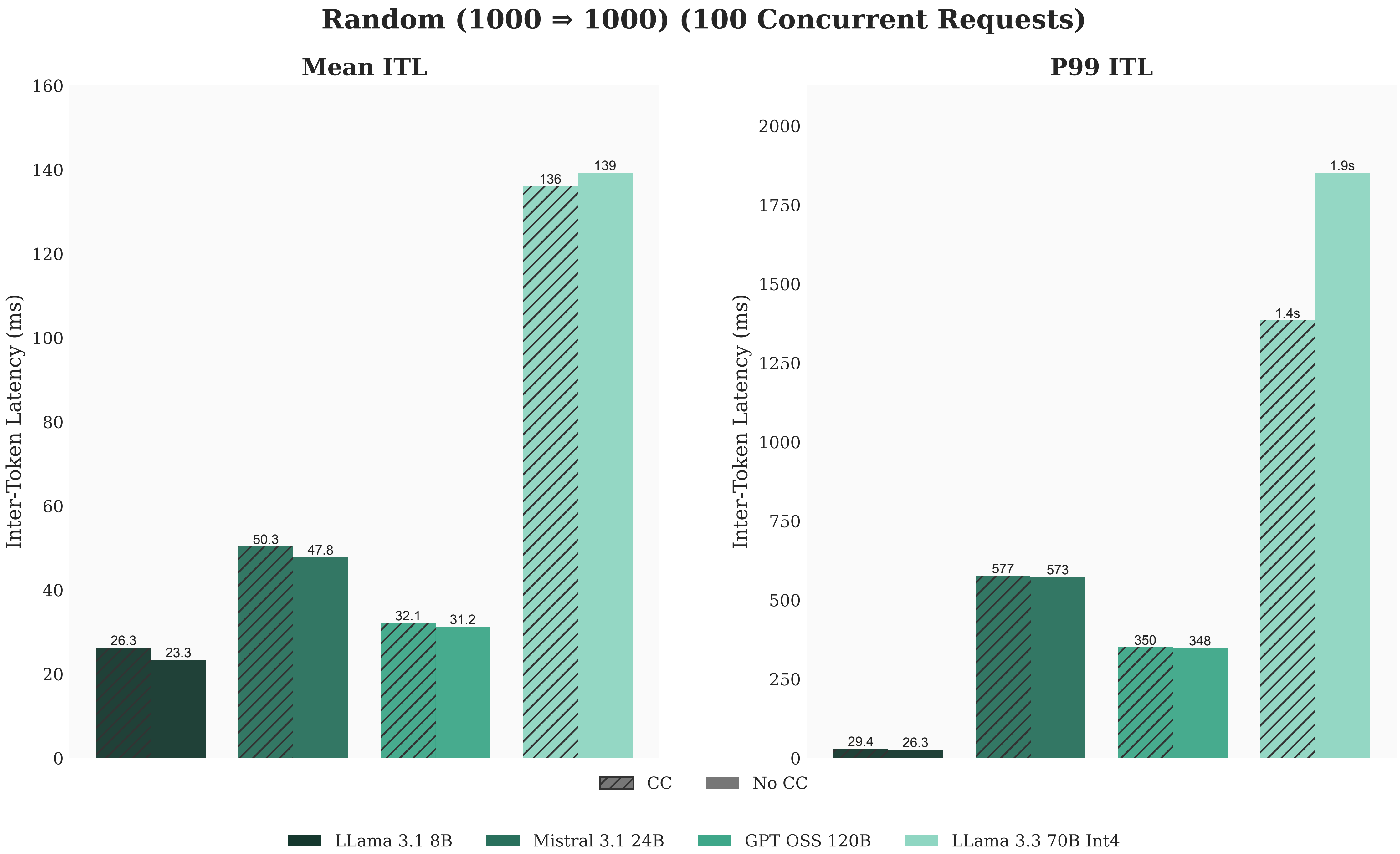

Random (1500 ⇒ 250) (1 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
60
70
80
7.7
7.0
21.1
20.0
7.1
6.8
76.6
78.3
Inter-Token Latency (ms)
0
200
400
600
800
1000
1200
1400
8.6
7.6
104
104
11.3
10.0
1.4s
1.4s
▨ CC
■ No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

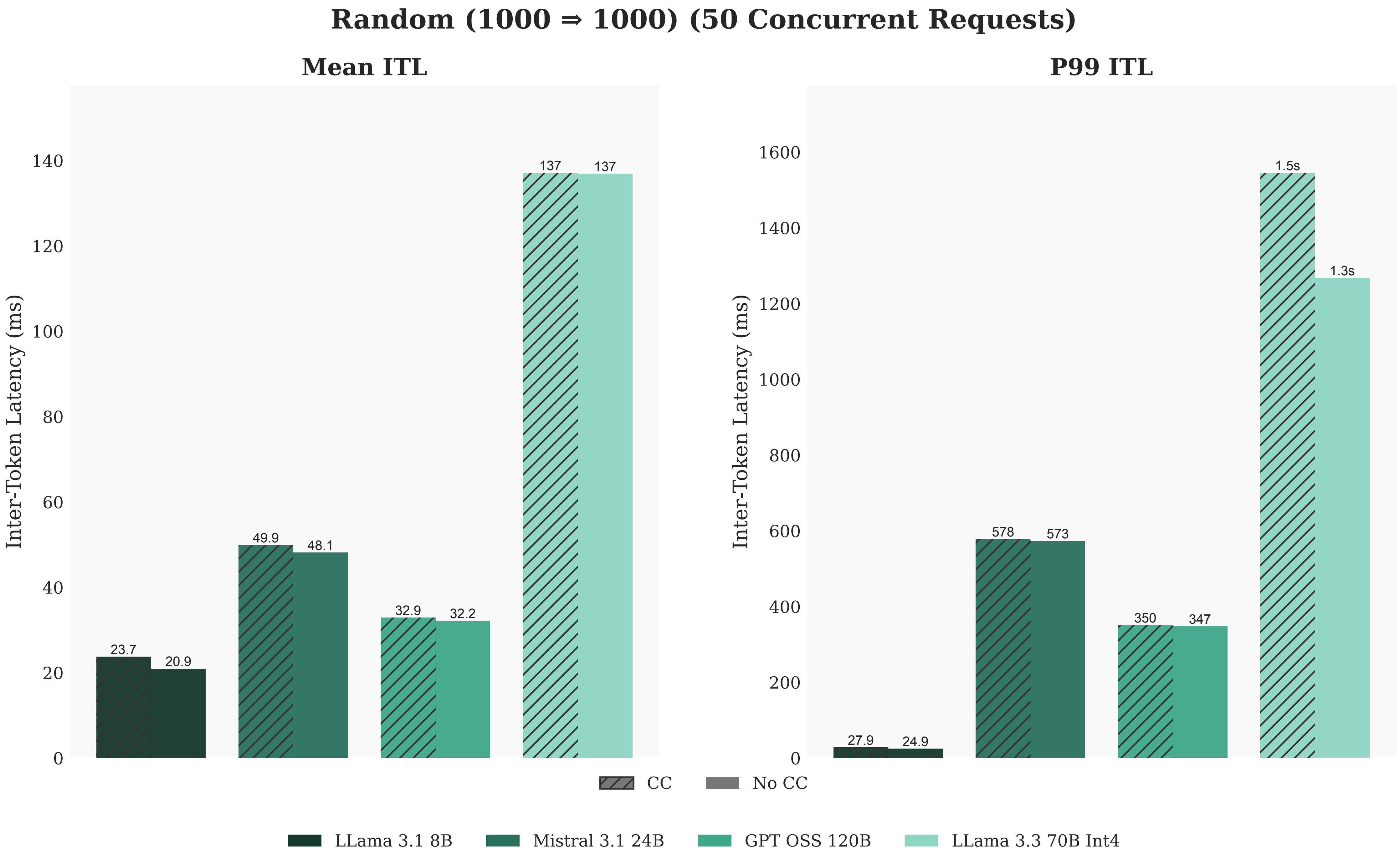

Random (4000 ⇒ 1000) (100 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
36.0
34.0
41.7
40.7
37.4
36.8
119
119
Inter-Token Latency (ms)
0
250
500
750
1000
1250
1500
1750
2000
221
221
590
585
351
350
1.8s
1.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

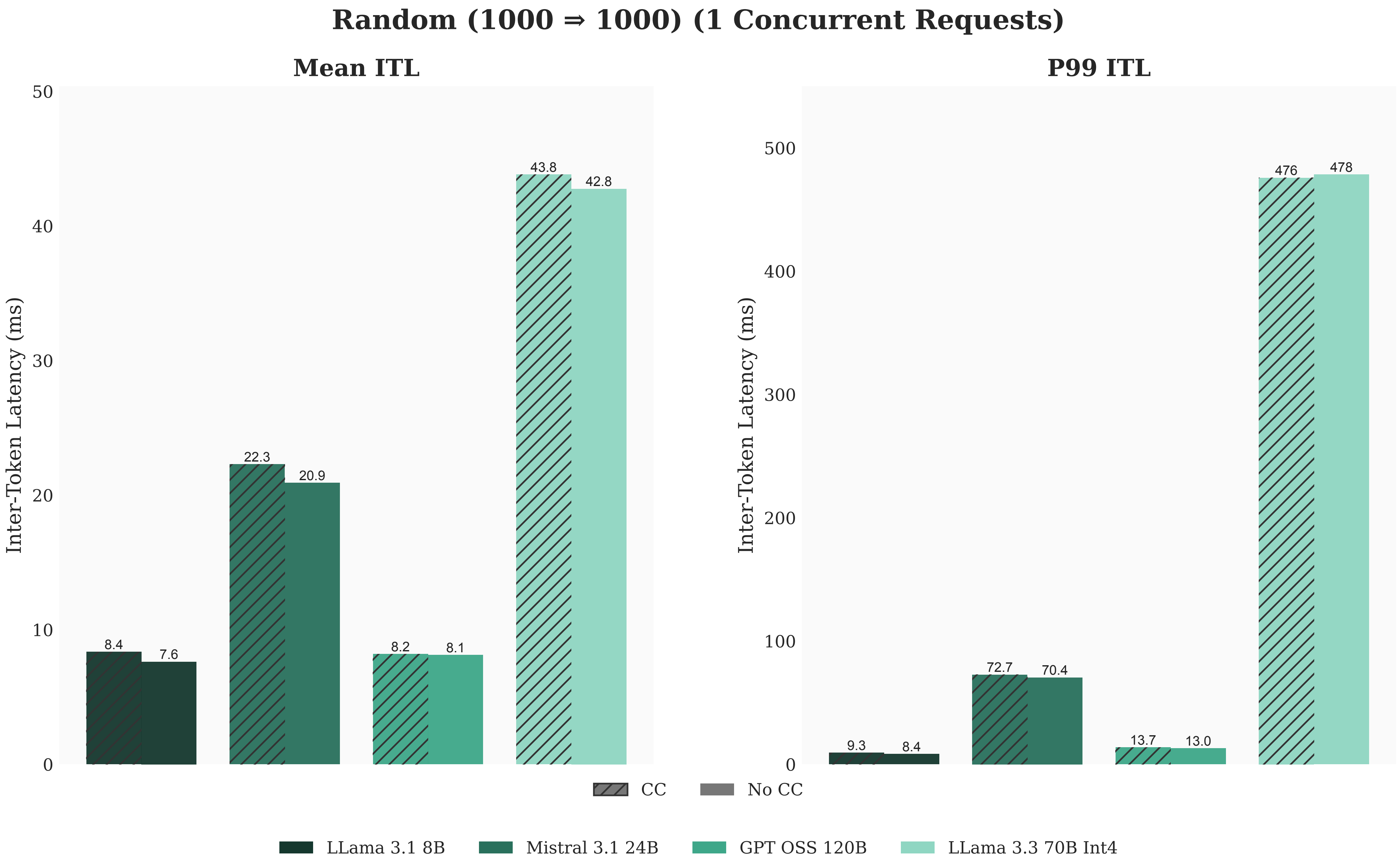

Random (4000 ⇒ 1000) (50 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
36.6
34.6
41.9
40.7
35.9
36.9
119
120
Inter-Token Latency (ms)
0
250
500
750
1000
1250
1500
1750
2000
222
222
591
587
352
351
1.8s
1.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

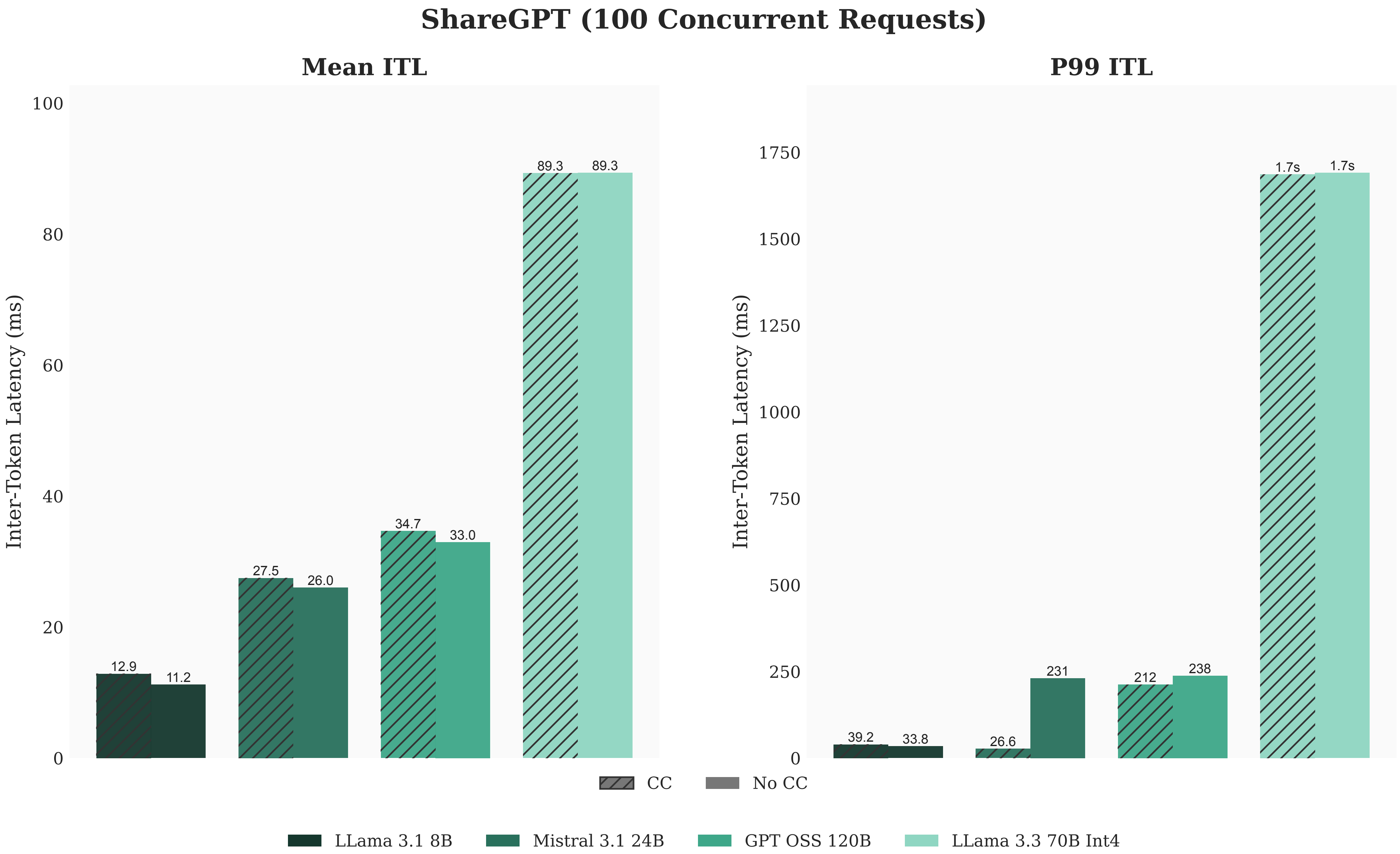


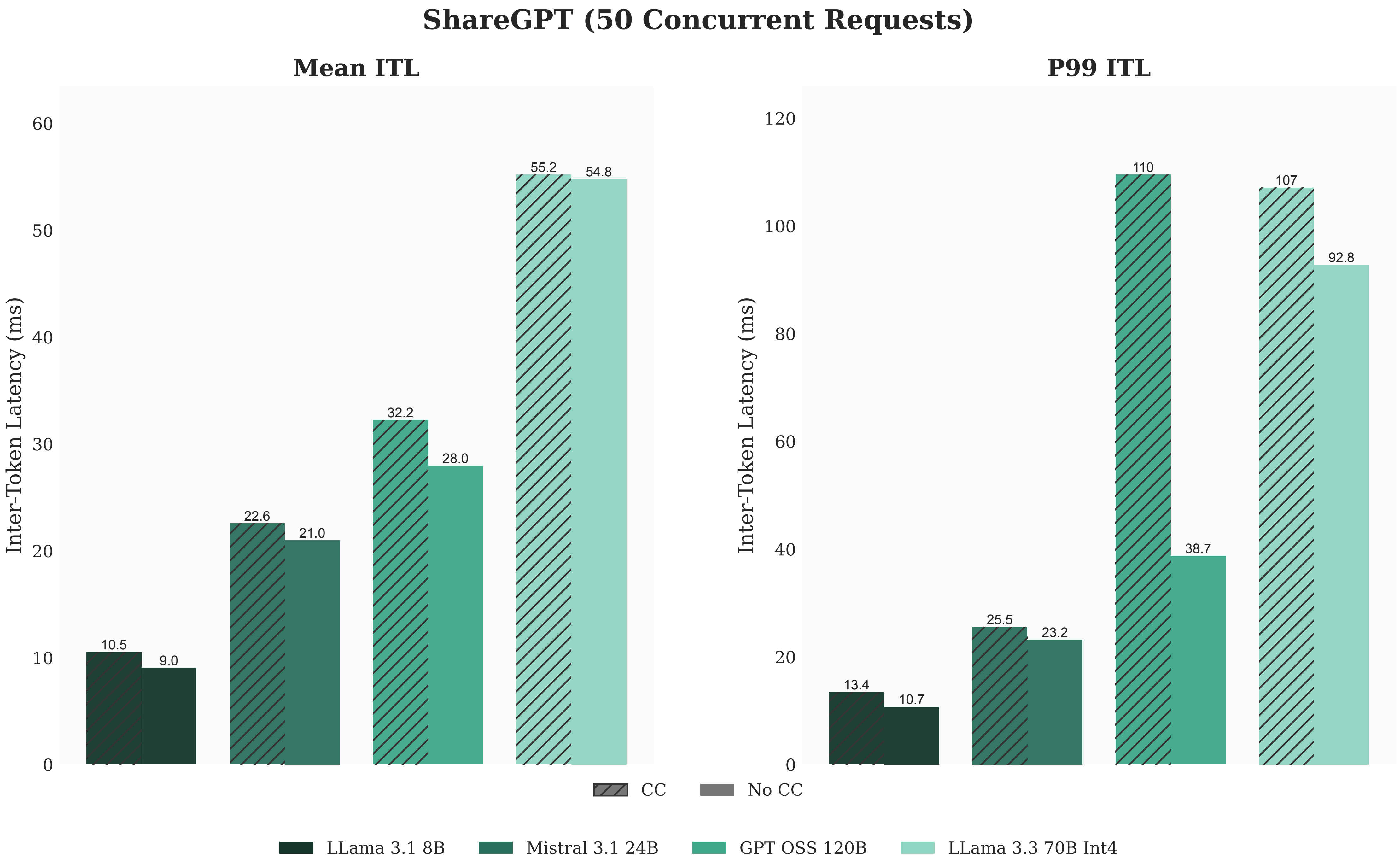
Random (4000 ⇒ 1000) (1 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
11.1
10.0
38.1
35.6
10.6
9.9
119
121
Inter-Token Latency (ms)
0
250
500
750
1000
1250
1500
1750
2000
95.6
12.6
291
291
184
182
1.8s
1.8s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Random (1000 ⇒ 1000) (100 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
140
160
26.3
23.3
50.3
47.8
32.1
31.2
136
139
Inter-Token Latency (ms)
0
250
500
750
1000
1250
1500
1750
2000
29.4
26.3
577
573
350
348
1.4s
1.9s
▨ CC
■ No CC
■ LLama 3.1 8B
■ Mistral 3.1 24B
■ GPT OSS 120B
■ LLama 3.3 70B Int4

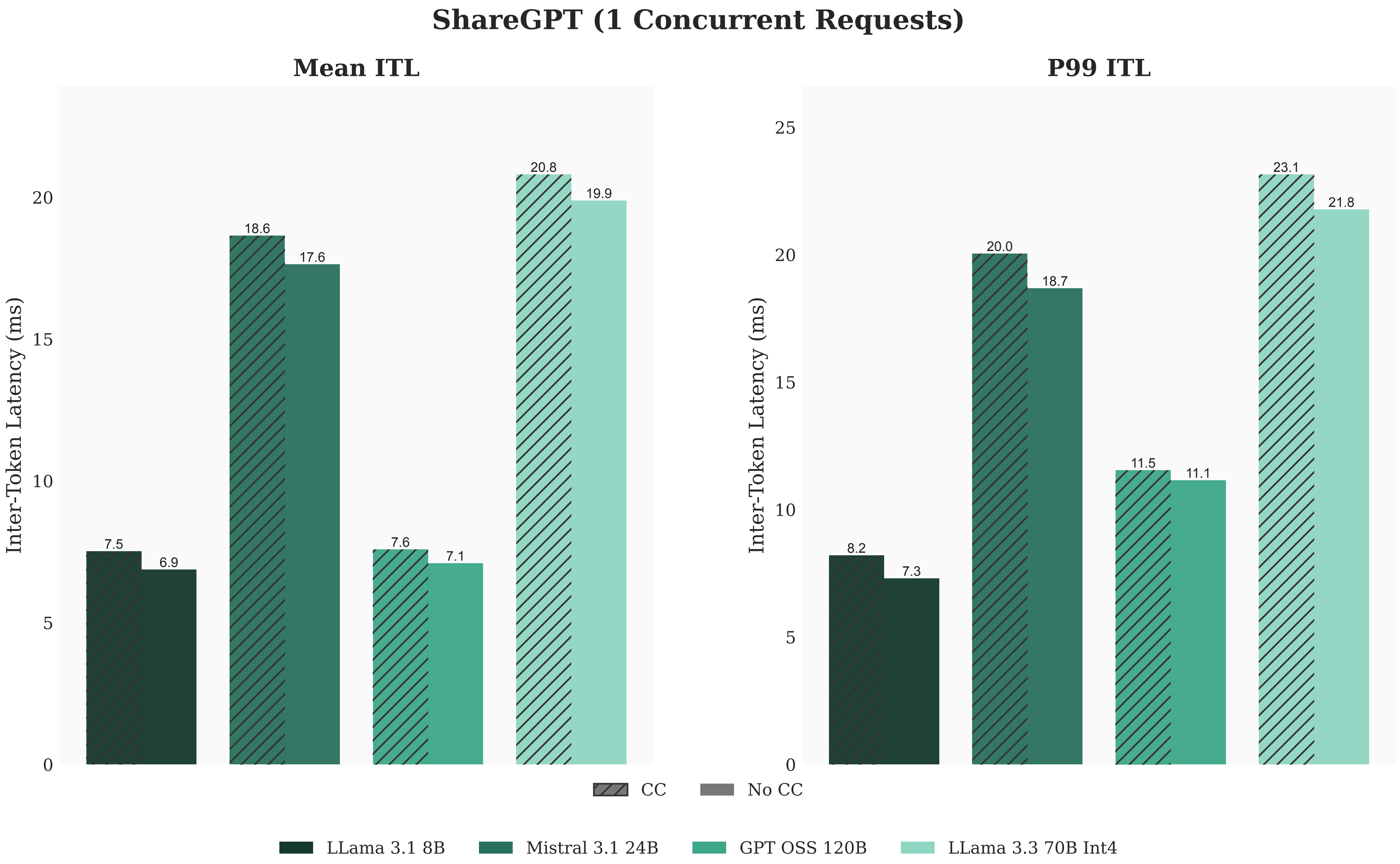

Random (1000 ⇒ 1000) (50 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
140
23.7
20.9
49.9
48.1
32.9
32.2
137
137
Inter-Token Latency (ms)
0
200
400
600
800
1000
1200
1400
1600
27.9
24.9
578
573
350
347
1.5s
1.3s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

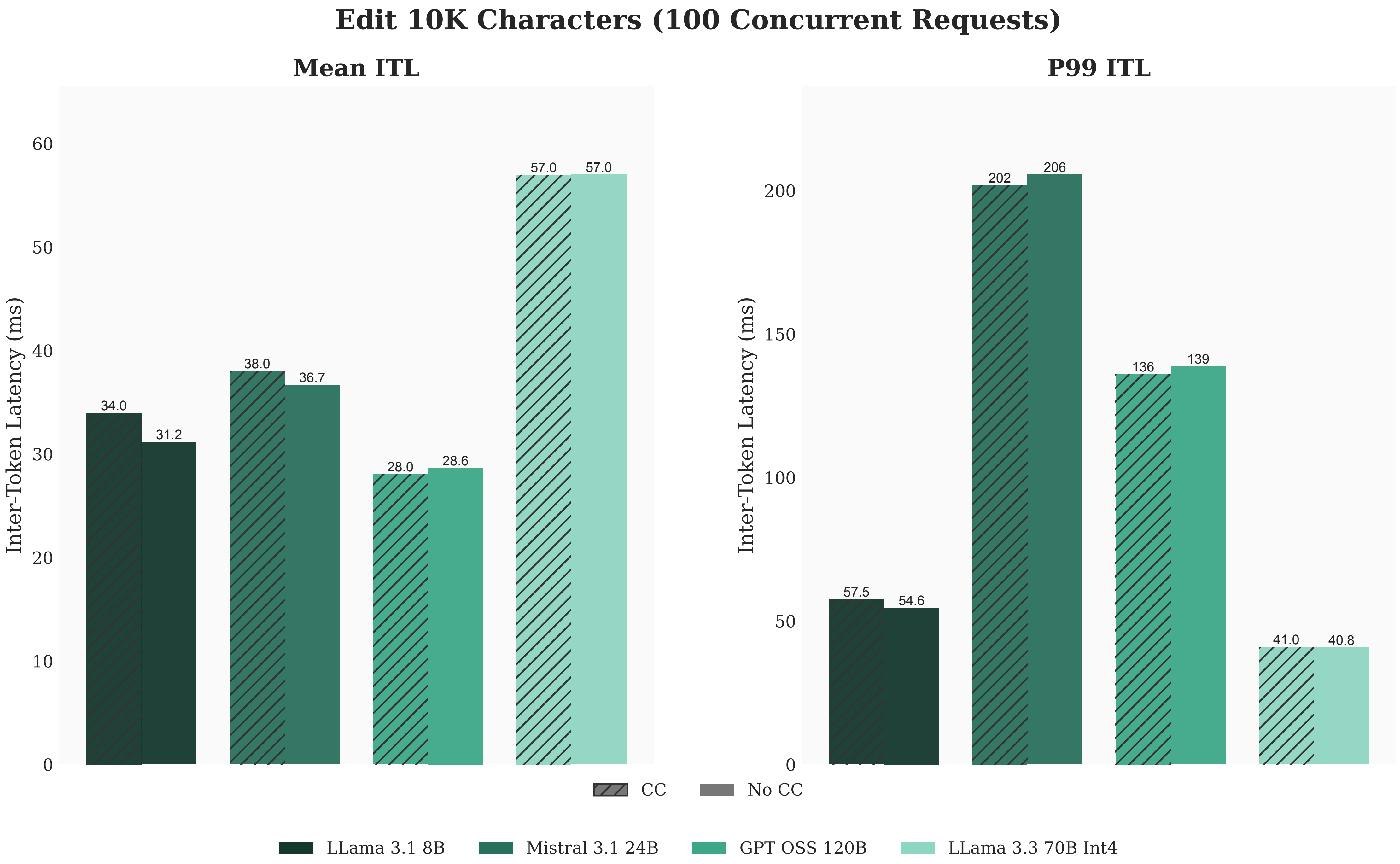

Random (1000 ⇒ 1000) (1 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
8.4
7.6
22.3
20.9
8.2
8.1
43.8
42.8
Inter-Token Latency (ms)
0
100
200
300
400
500
9.3
8.4
72.7
70.4
13.7
13.0
476
478
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

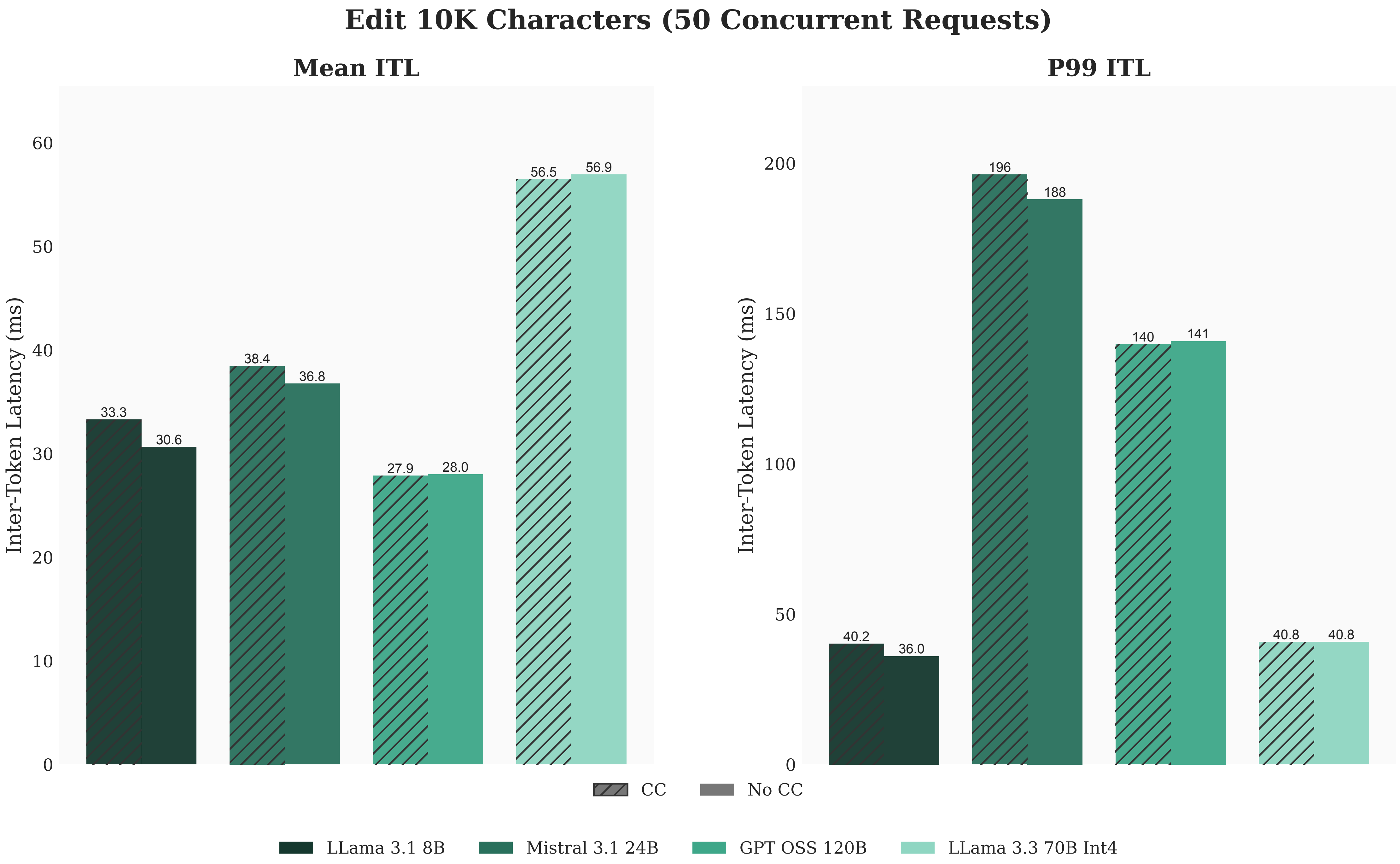

ShareGPT (100 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
12.9
11.2
27.5
26.0
34.7
33.0
89.3
89.3
Inter-Token Latency (ms)
0
250
500
750
1000
1250
1500
1750
39.2
33.8
26.6
231
212
238
1.7s
1.7s
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

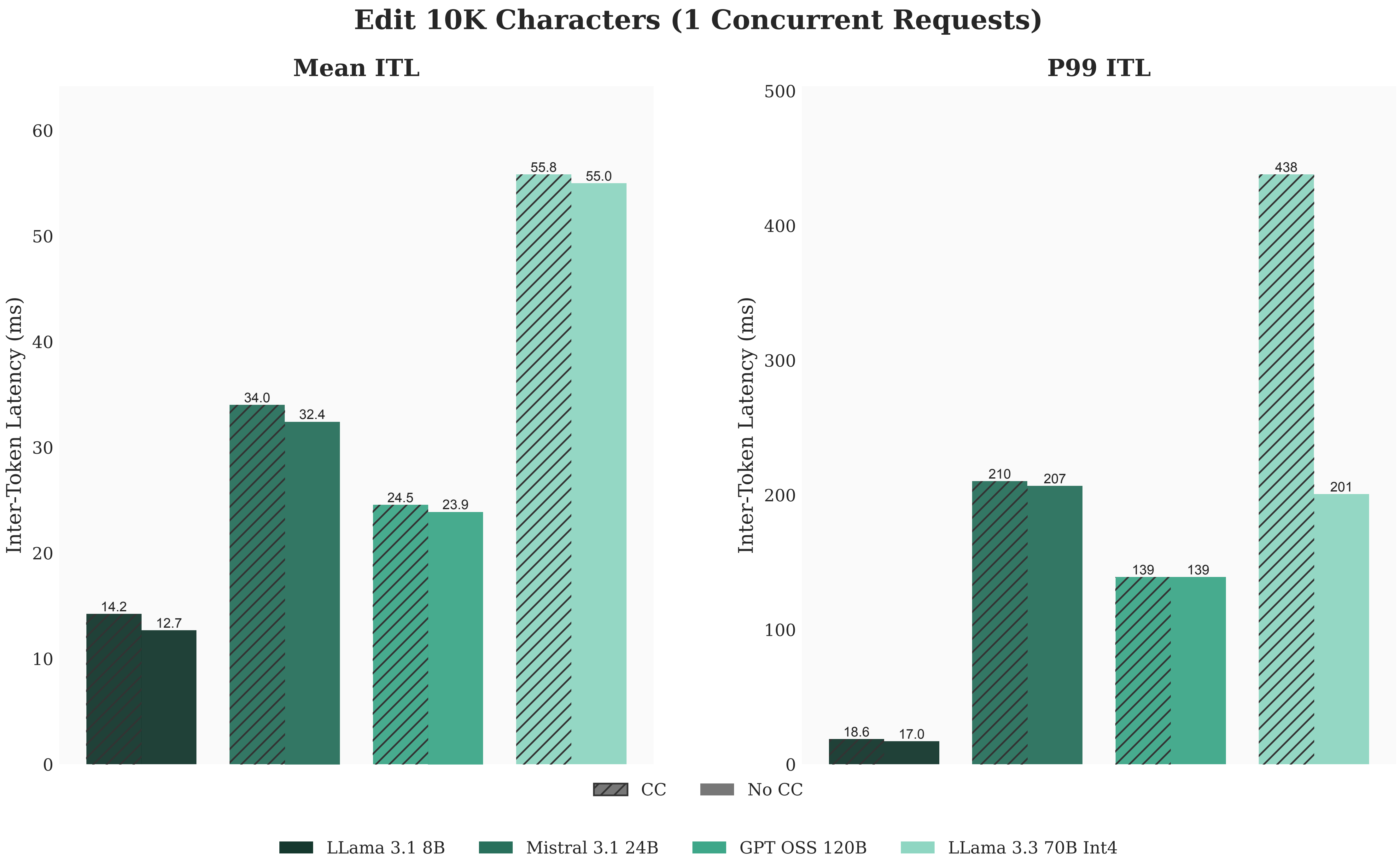


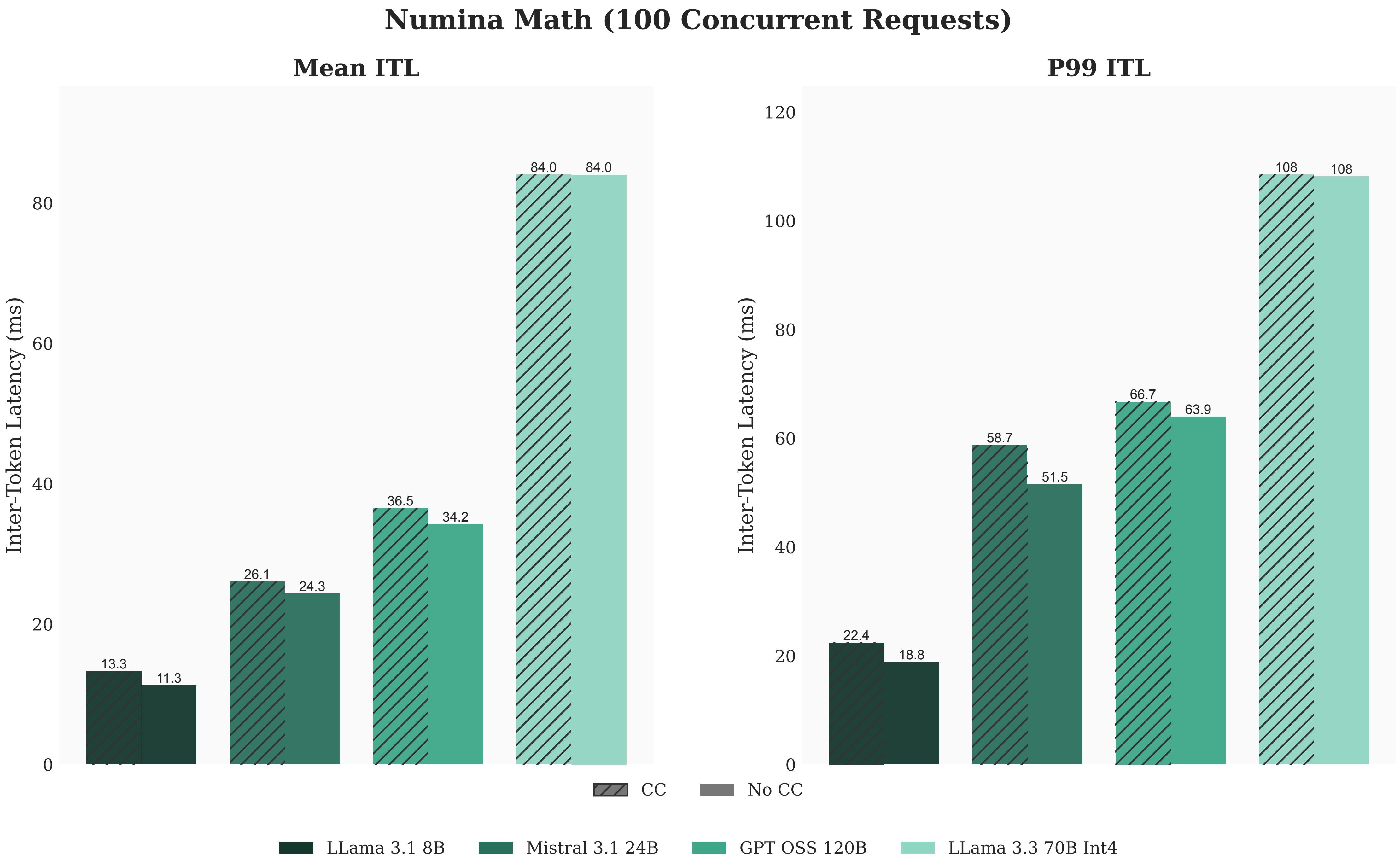
ShareGPT (50 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
60
10.5
9.0
22.6
21.0
32.2
28.0
55.2
54.8
Inter-Token Latency (ms)
0
20
40
60
80
100
120
13.4
10.7
25.5
23.2
110
38.7
107
92.8
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

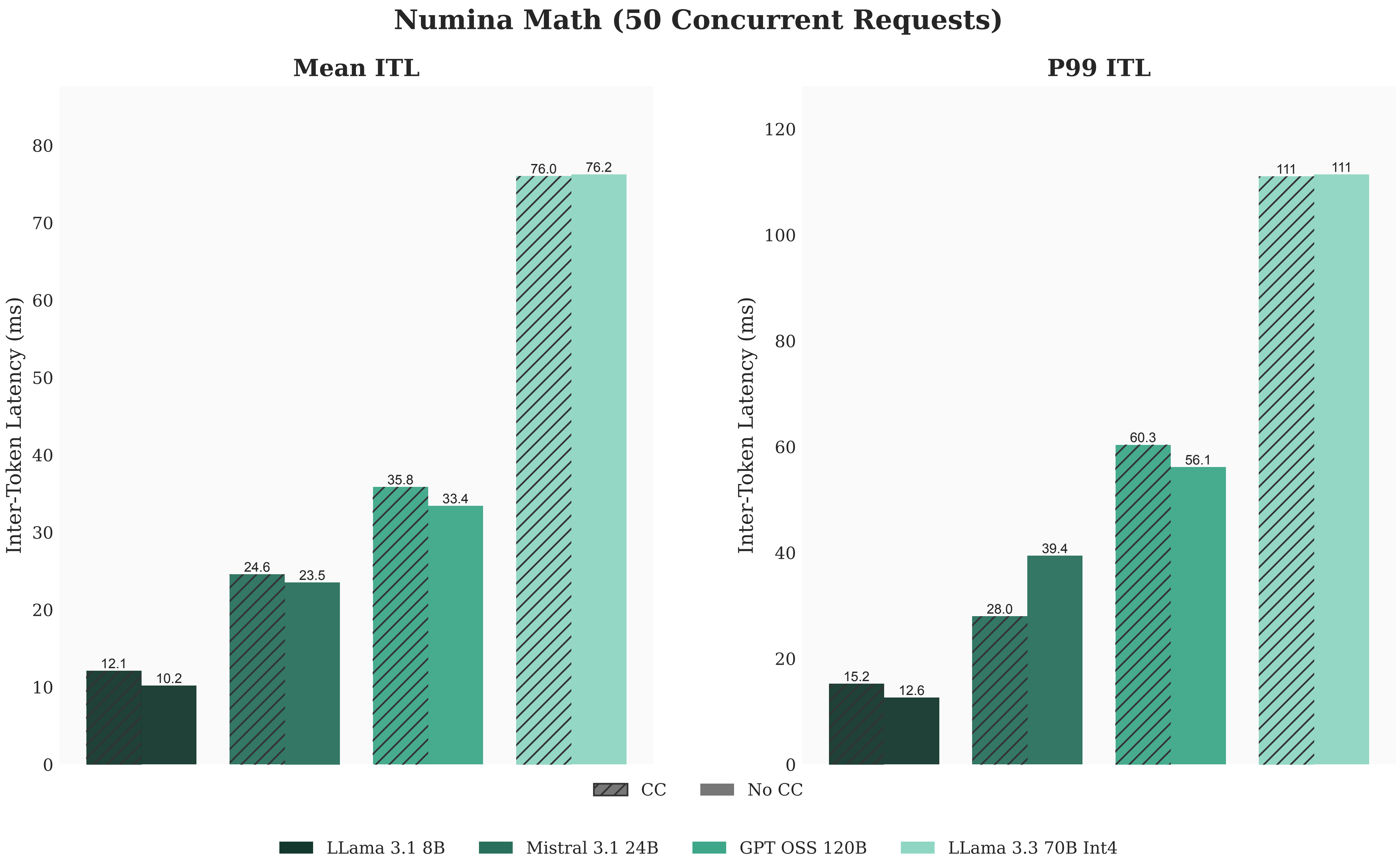

ShareGPT (1 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
5
10
15
20
25
7.5
6.9
18.6
17.6
7.6
7.1
20.8
19.9
8.2
7.3
20.0
18.7
11.5
11.1
23.1
21.8
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

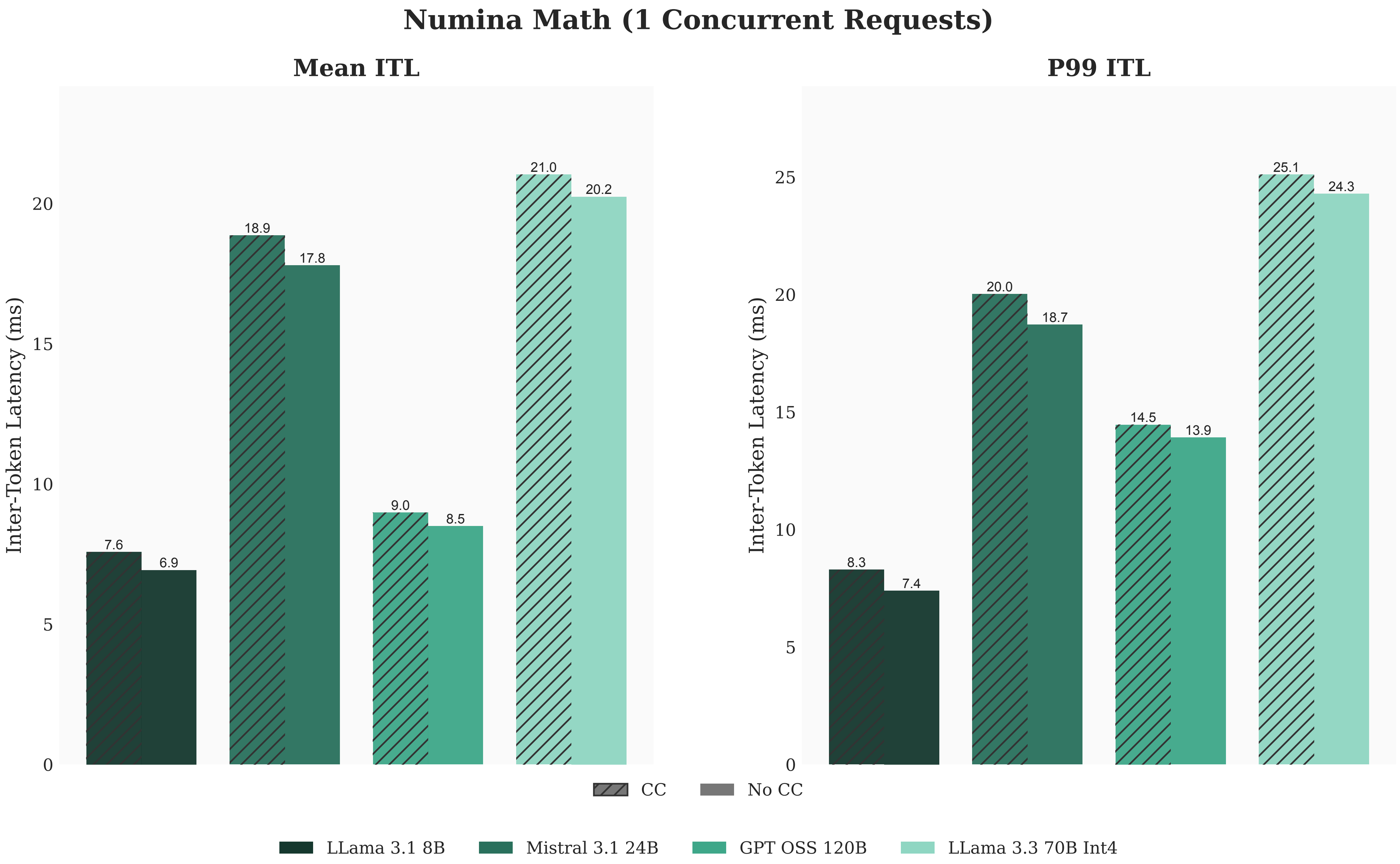

Edit 10K Characters (100 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
60
34.0
31.2
38.0
36.7
28.0
28.6
57.0
57.0
Inter-Token Latency (ms)
0
50
100
150
200
57.5
54.6
202
206
136
139
41.0
40.8
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Edit 10K Characters (50 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
60
33.3
30.6
38.4
36.8
27.9
28.0
56.5
56.9
Inter-Token Latency (ms)
0
50
100
150
200
40.2
36.0
196
188
140
141
40.8
40.8
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Edit 10K Characters (1 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
60
14.2
12.7
34.0
32.4
24.5
23.9
55.8
55.0
Inter-Token Latency (ms)
0
100
200
300
400
500
18.6
17.0
210
207
139
139
438
201
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (100 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
13.3
11.3
26.1
24.3
36.5
34.2
84.0
84.0
22.4
18.8
58.7
51.5
66.7
63.9
108
108
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (50 Concurrent Requests)
Mean ITL
Inter-Token Latency (ms)
0
10
20
30
40
50
60
70
80
12.1
10.2
24.6
23.5
35.8
33.4
76.0
76.2
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
15.2
12.6
28.0
39.4
60.3
56.1
111
111
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


Numina Math (1 Concurrent Requests)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
5
10
15
20
25
7.6
6.9
18.9
17.8
9.0
8.5
21.0
20.2
8.3
7.4
20.0
18.7
14.5
13.9
25.1
24.3
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4