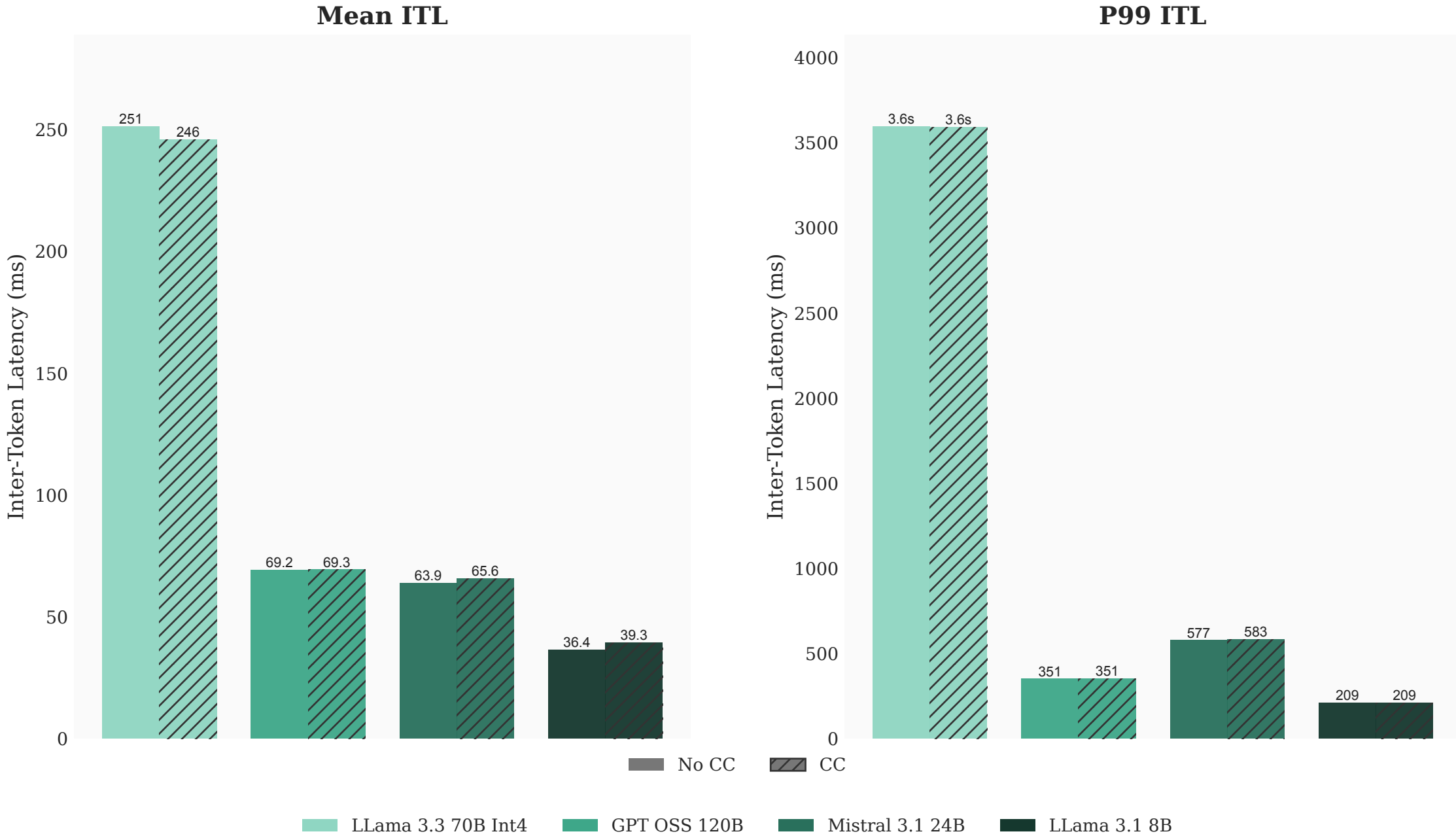
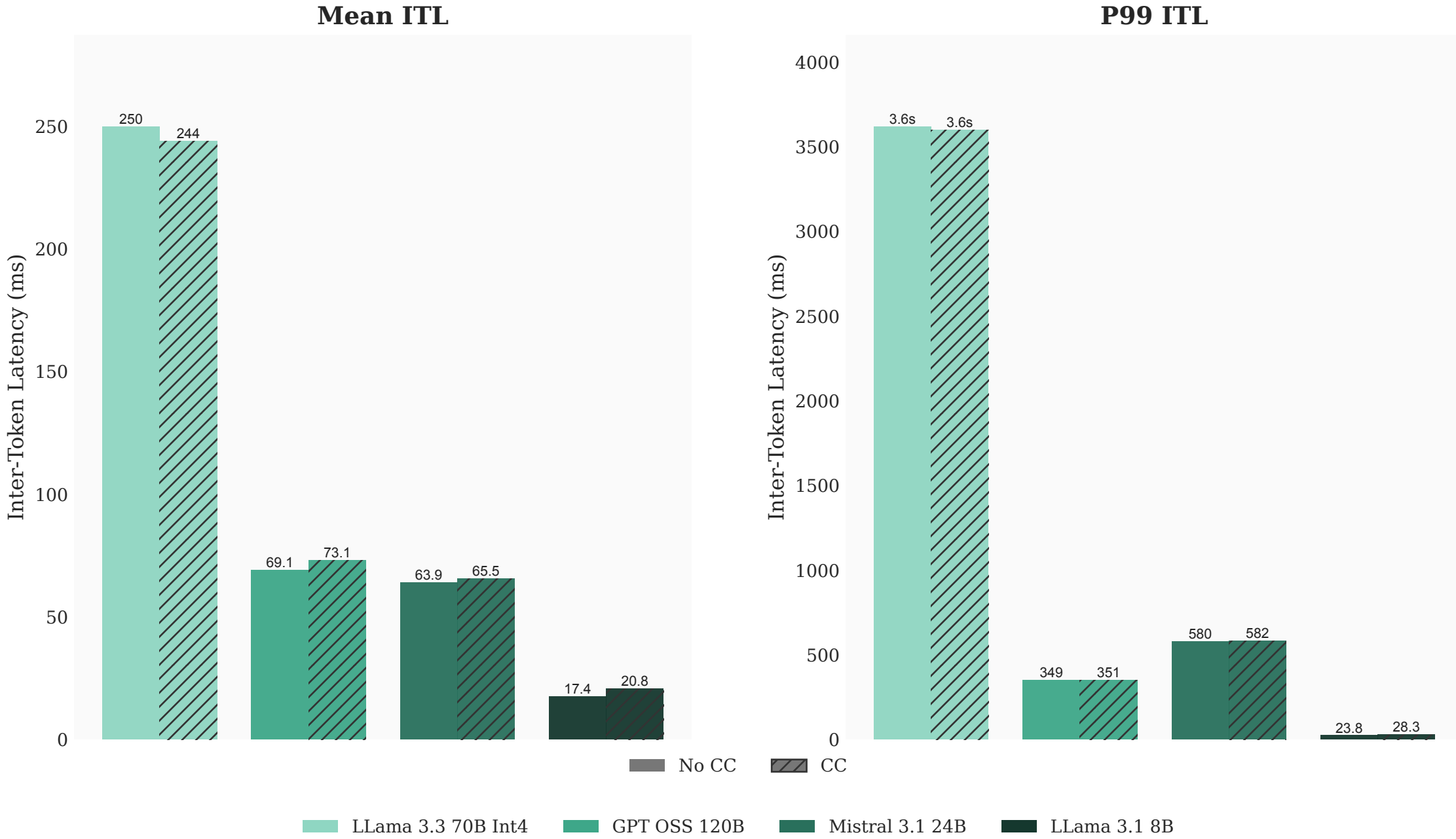
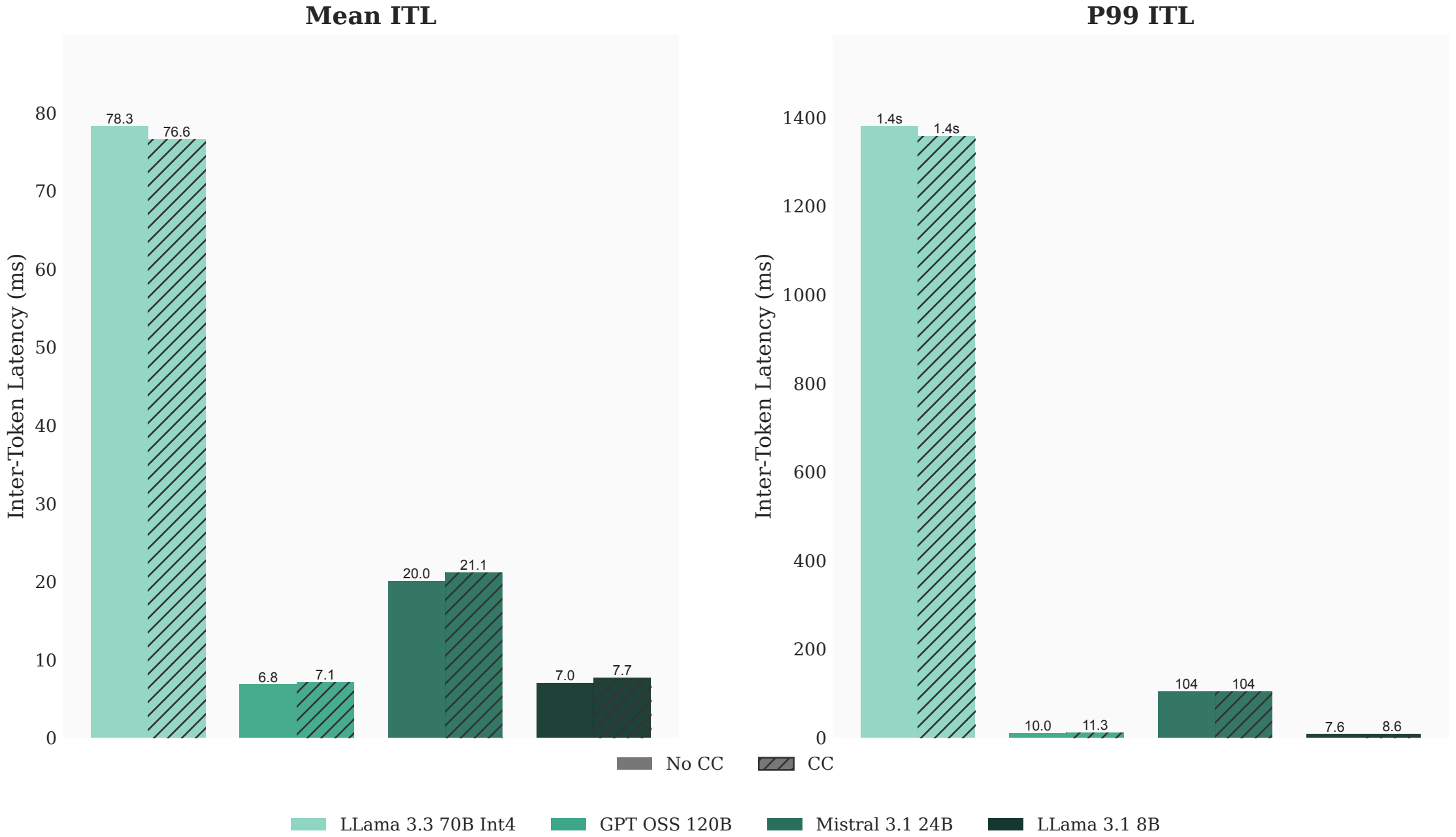
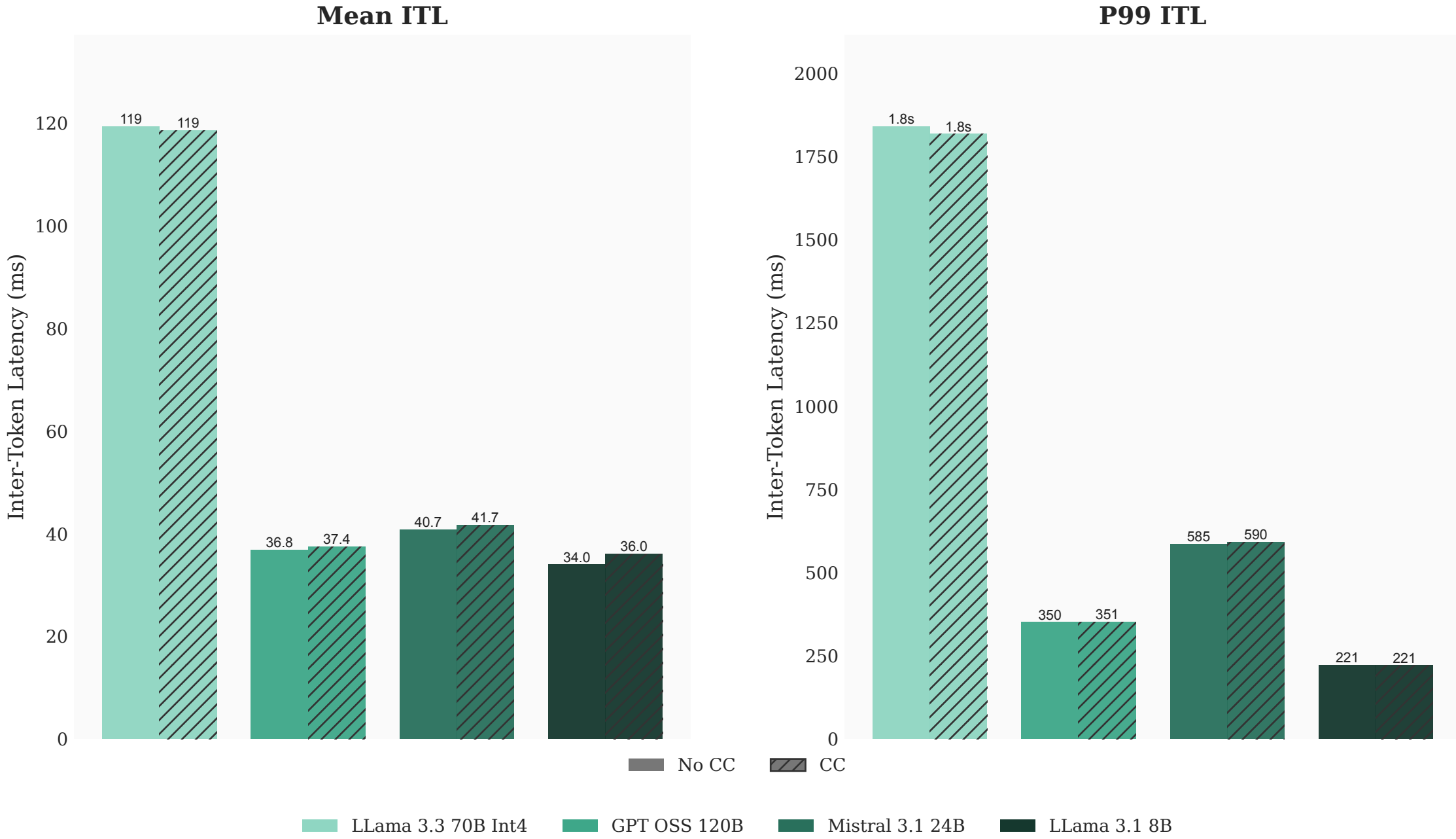
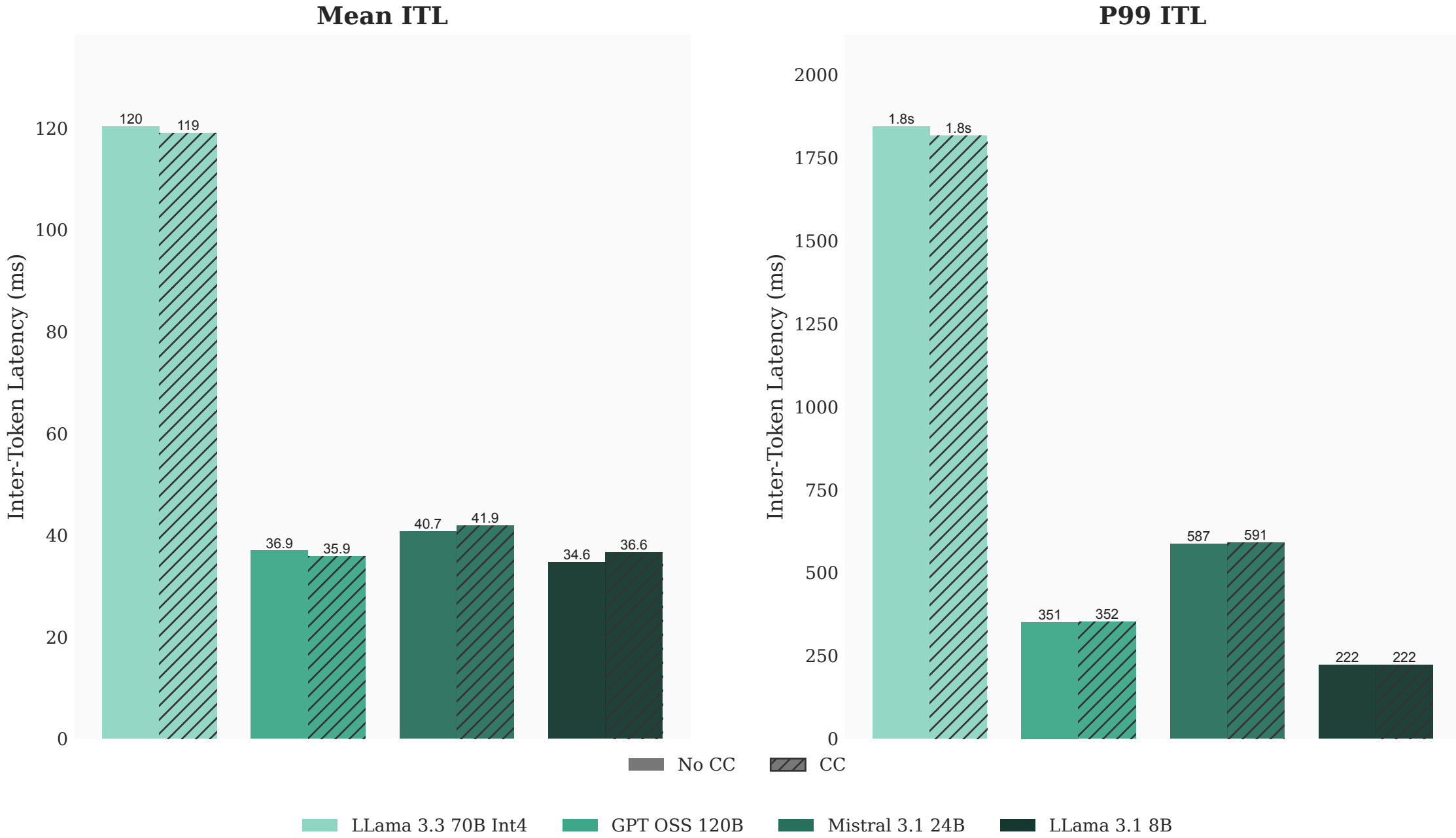
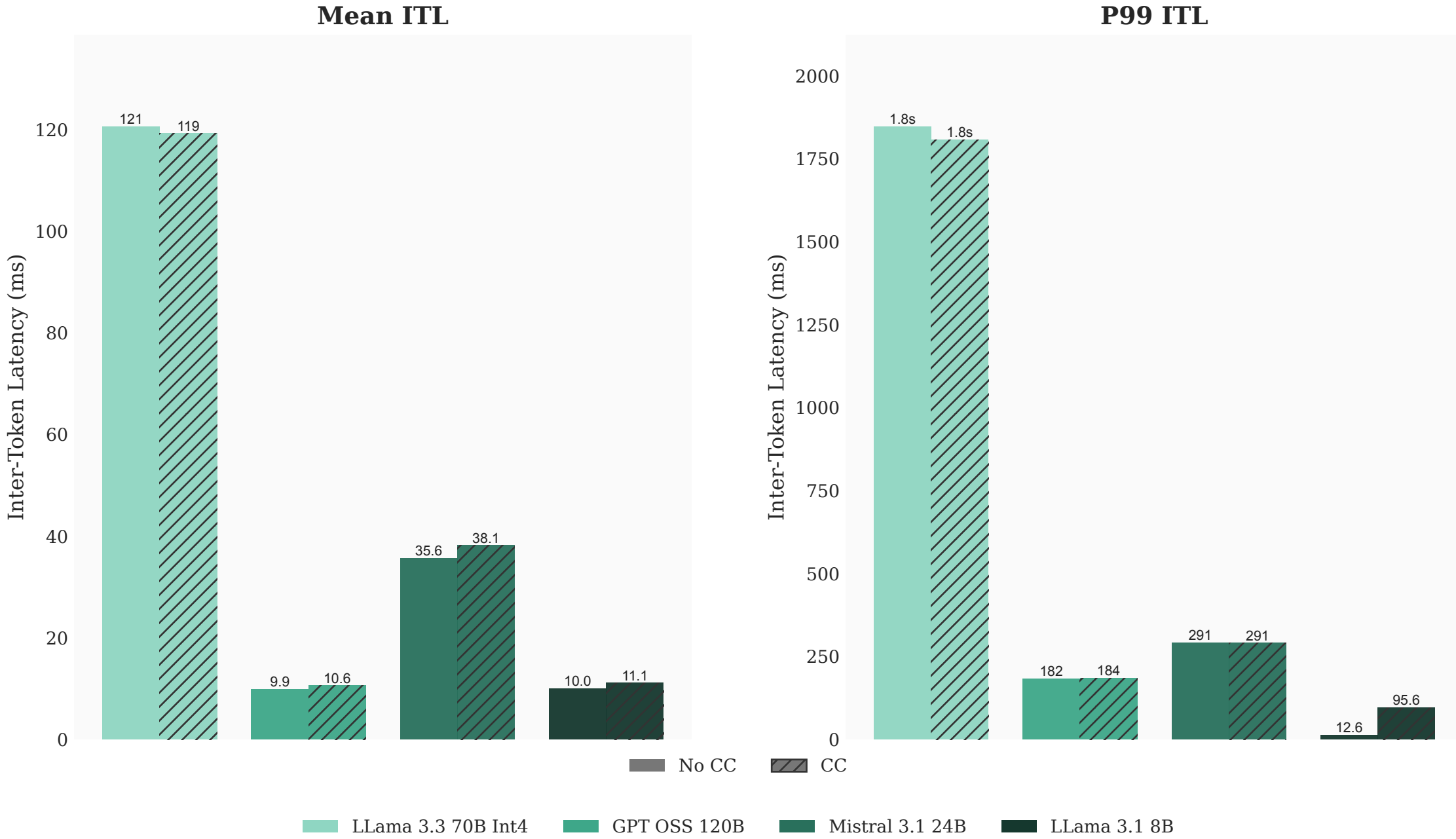

Random (1500 ⇒ 250) (100 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
50
100
150
200
250
500
1000
1500
2000
2500
3000
3500
4000
251
246
69.2
69.3
63.9
65.6
36.4
39.3
3.6s
3.6s
351
351
577
583
209
209
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

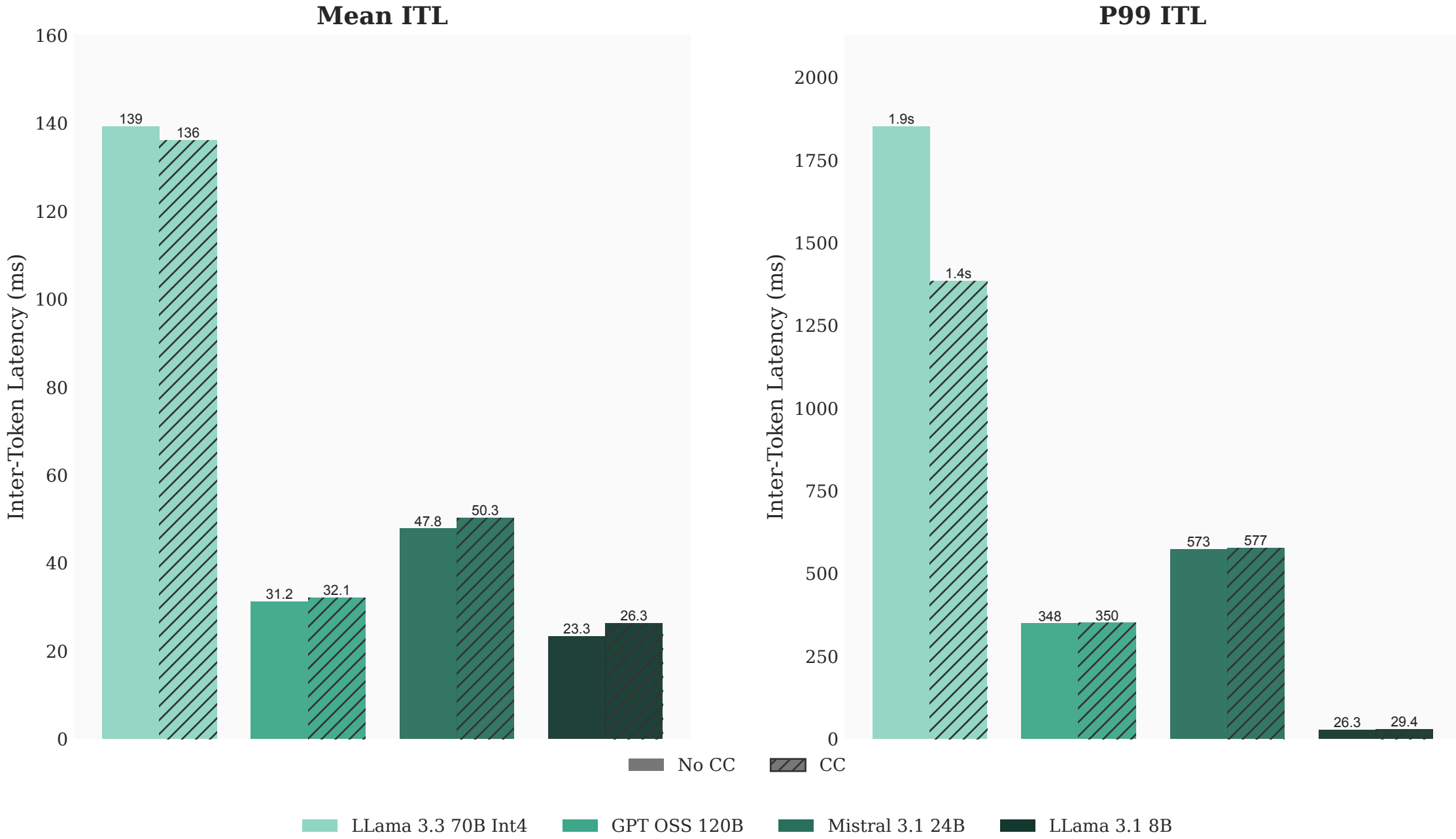

Random (1500 ⇒ 250) (50 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
Inter-Token Latency (ms)
0
50
100
150
200
250
0
500
1000
1500
2000
2500
3000
3500
4000
250
244
69.1
73.1
63.9
65.5
17.4
20.8
3.6s
3.6s
349
351
580
582
23.8
28.3
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama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

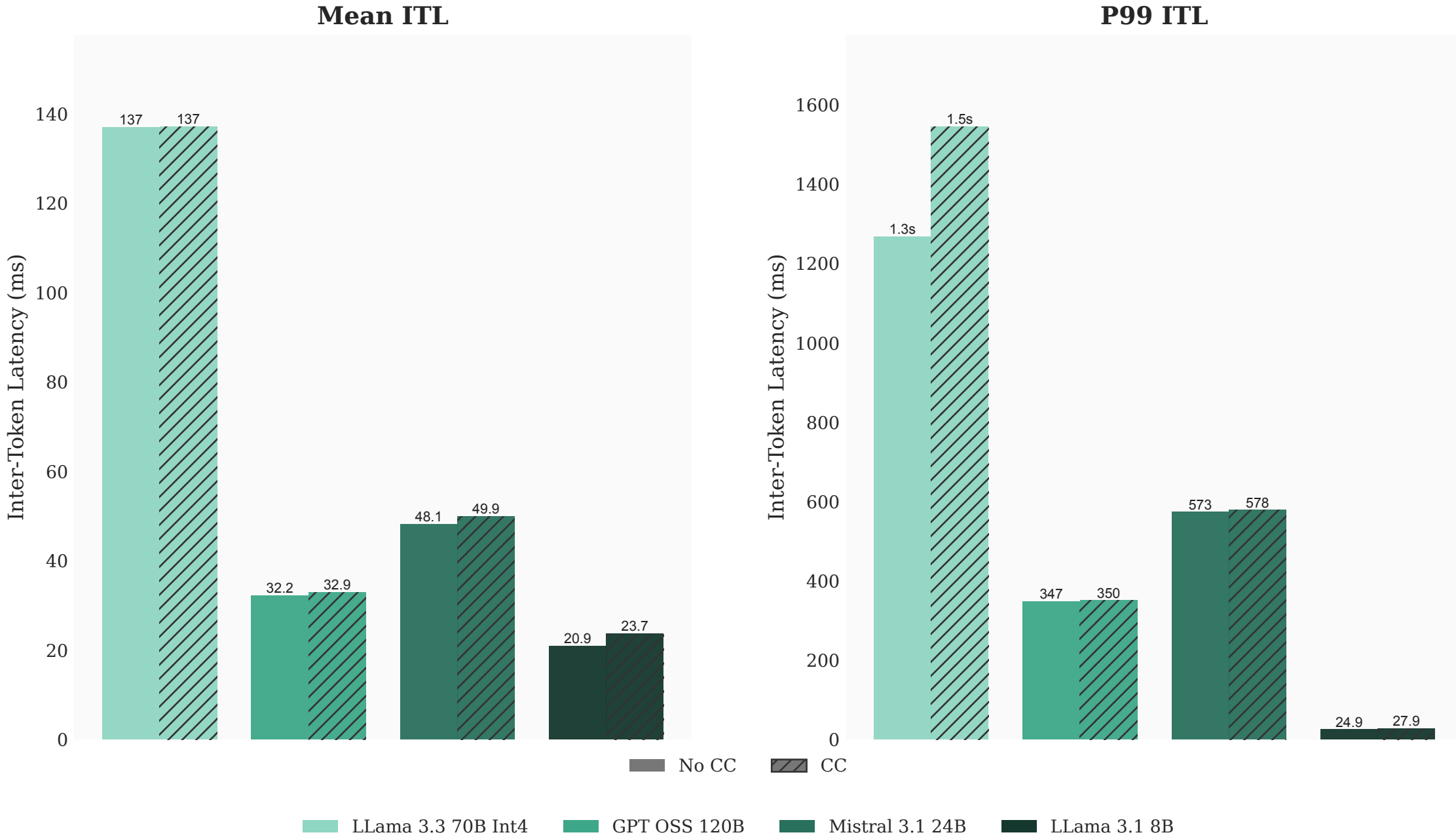

Random (1500 ⇒ 250) (Single Request)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
Inter-Token Latency (ms)
0
10
20
30
40
50
60
70
80
0
200
400
600
800
1000
1200
1400
78.3
76.6
6.8
7.1
20.0
21.1
7.0
7.7
1.4s
1.4s
10.0
11.3
104
104
7.6
8.6
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1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

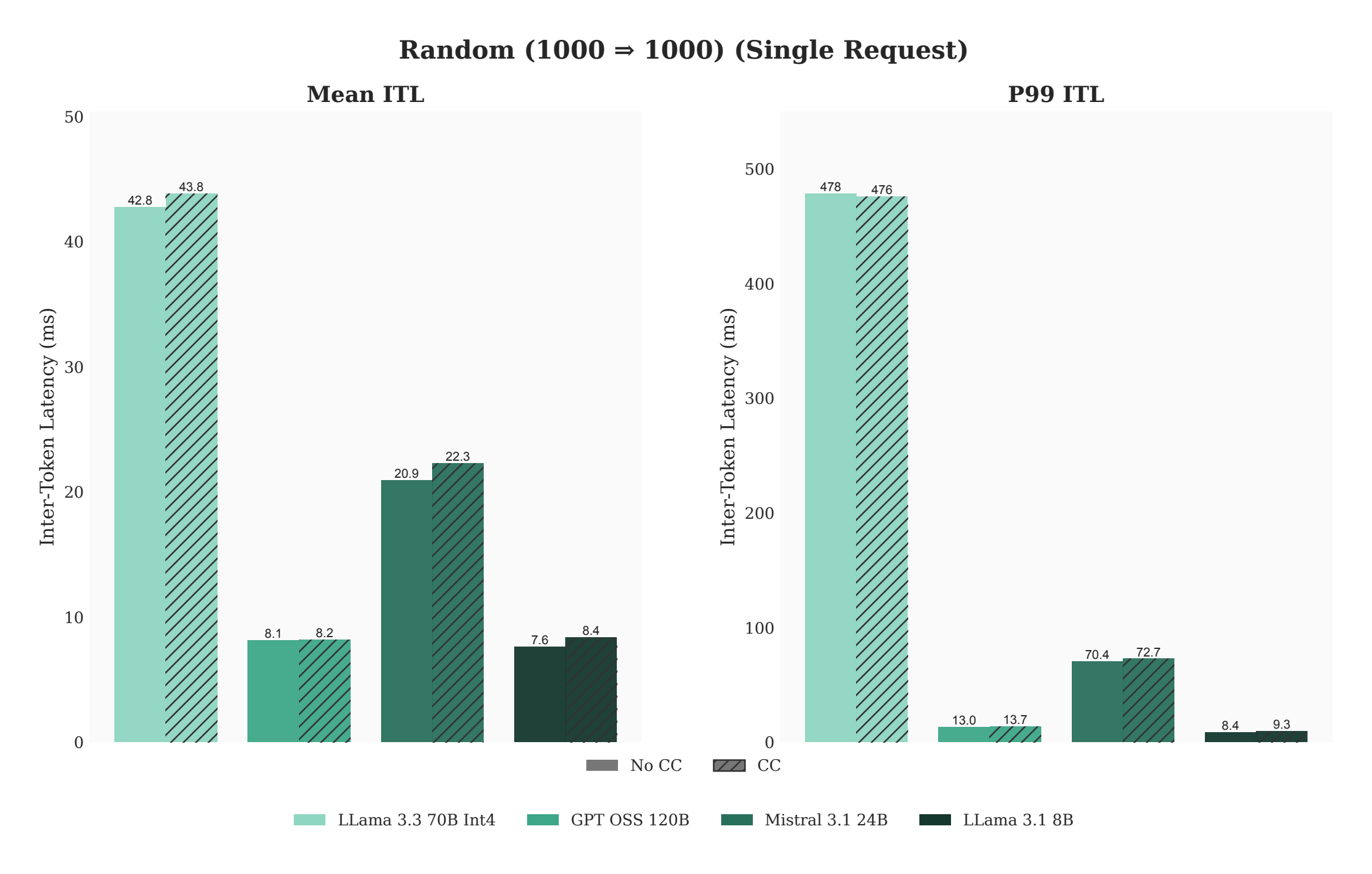

Random (4000 ⇒ 1000) (100 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
250
500
750
1000
1250
1500
1750
2000
119
119
36.8
37.4
40.7
41.7
34.0
36.0
1.8s
1.8s
350
351
585
590
221
221
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

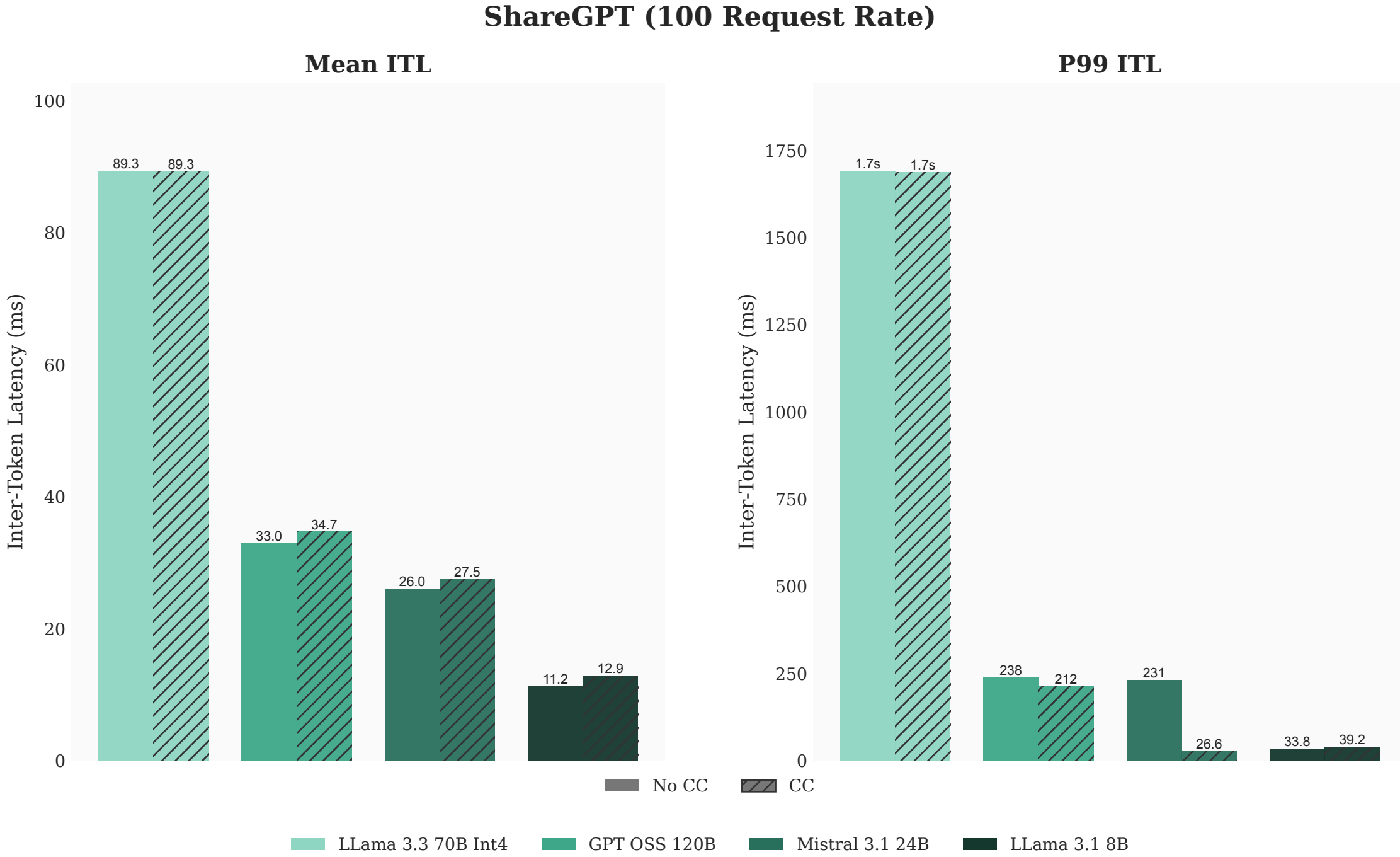

Random (4000 ⇒ 1000) (50 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
250
500
750
1000
1250
1500
1750
2000
120
119
36.9
35.9
40.7
41.9
34.6
36.6
1.8s
1.8s
351
352
587
591
222
222
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

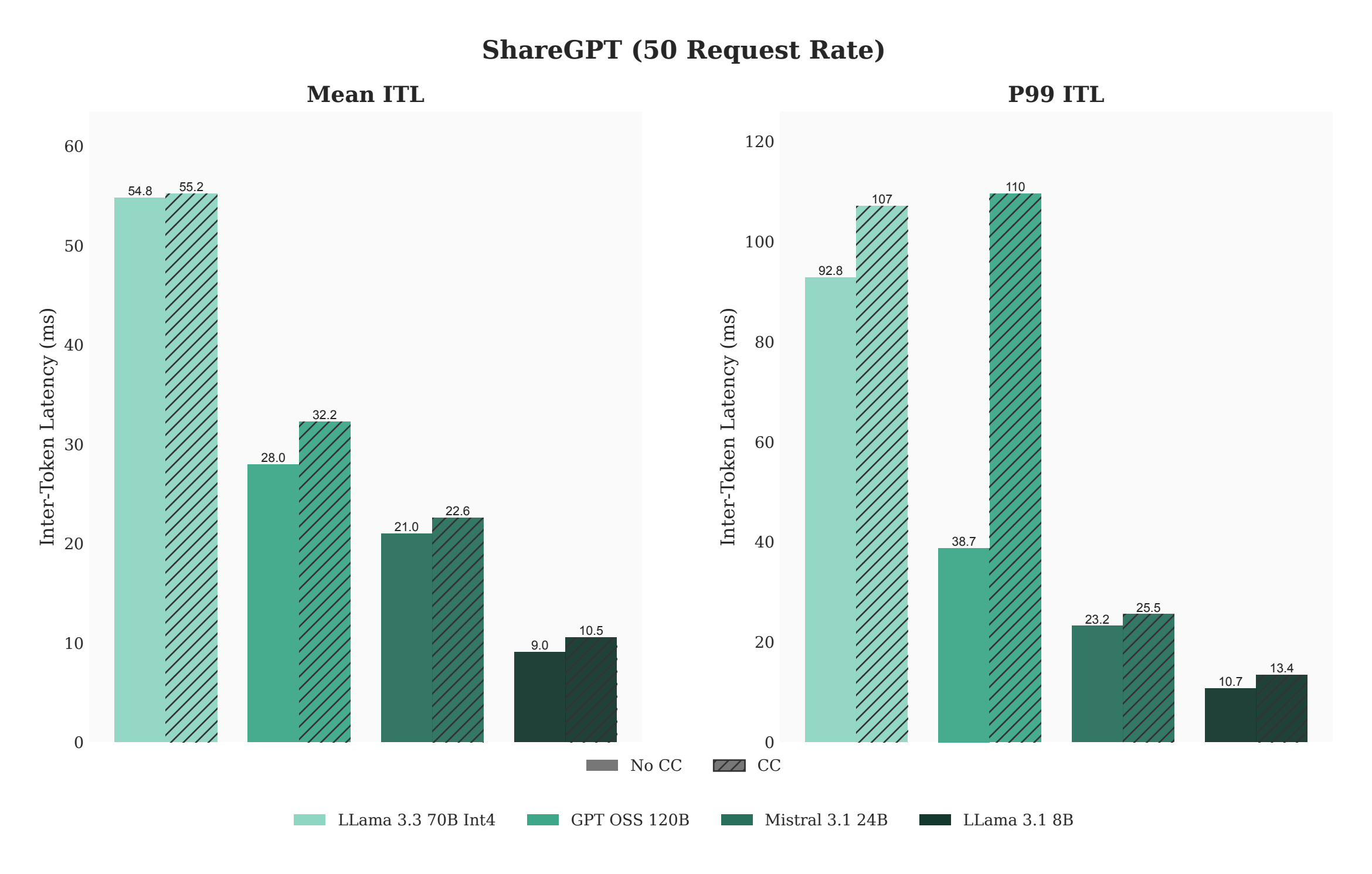

Random (4000 ⇒ 1000) (Single Request)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
20
40
60
80
100
120
0
250
500
750
1000
1250
1500
1750
2000
121
119
9.9
10.6
35.6
38.1
10.0
11.1
1.8s
1.8s
182
184
291
291
12.6
95.6
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

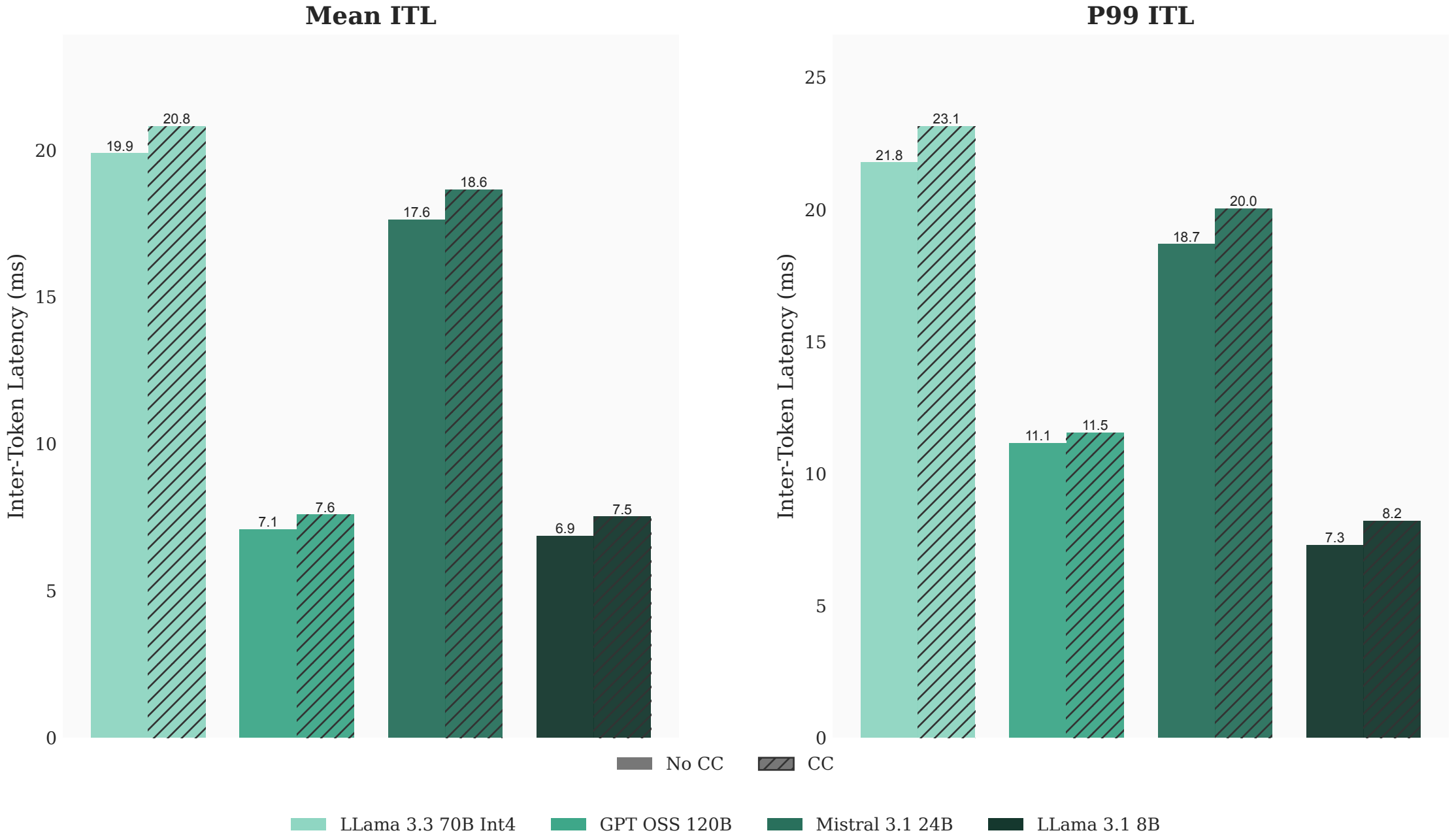

Random (1000 ⇒ 1000) (100 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
Inter-Token Latency (ms)
0
20
40
60
80
100
120
140
160
0
250
500
750
1000
1250
1500
1750
2000
139
136
31.2
32.1
47.8
50.3
23.3
26.3
1.9s
1.4s
348
350
573
577
26.3
29.4
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1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


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


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

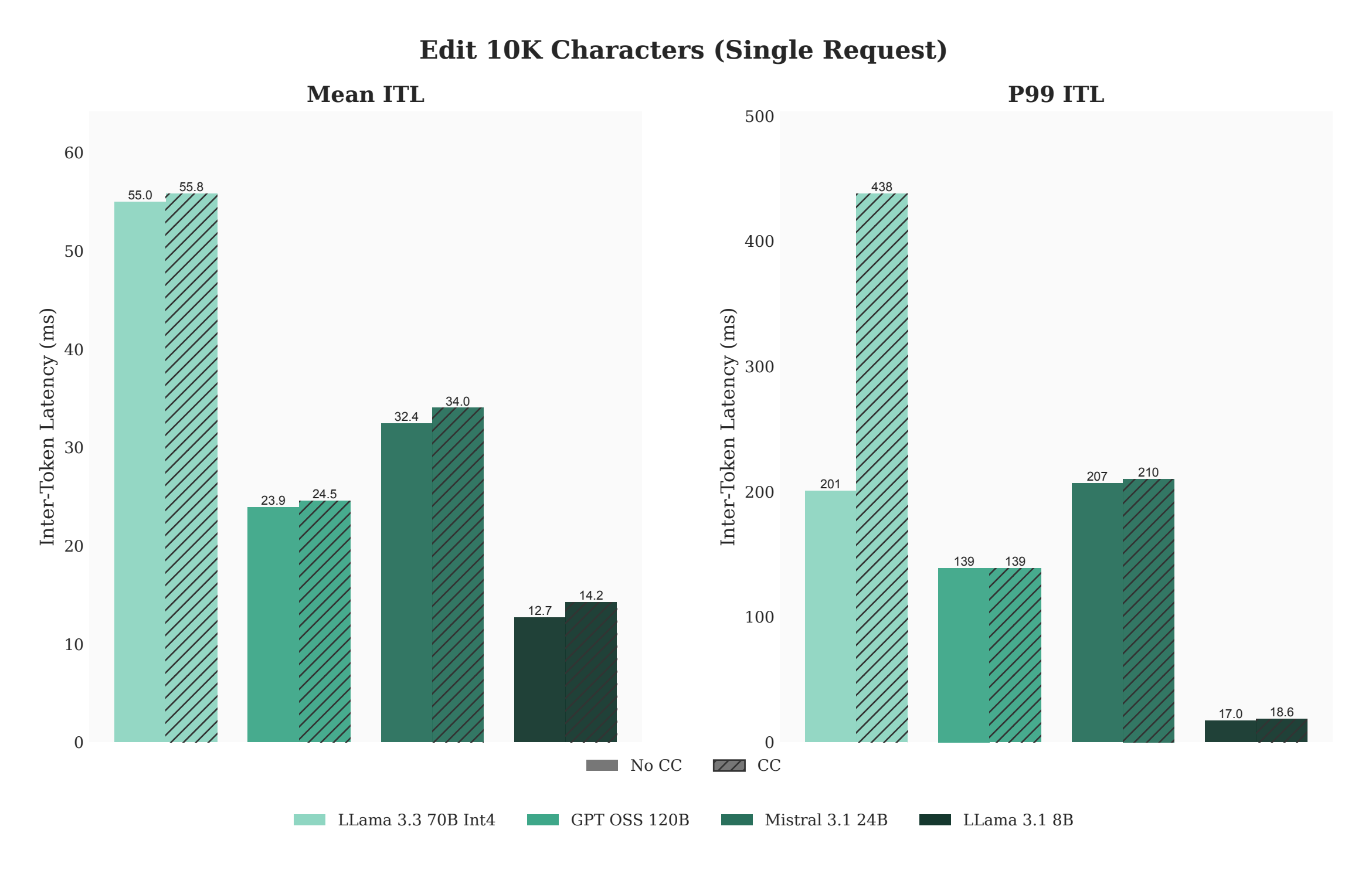

ShareGPT (100 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
Inter-Token Latency (ms)
100
80
60
40
20
0
1750
1500
1250
1000
750
500
250
0
89.3
89.3
33.0
34.7
26.0
27.5
11.2
12.9
1.7s
1.7s
238
212
231
26.6
33.8
39.2
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

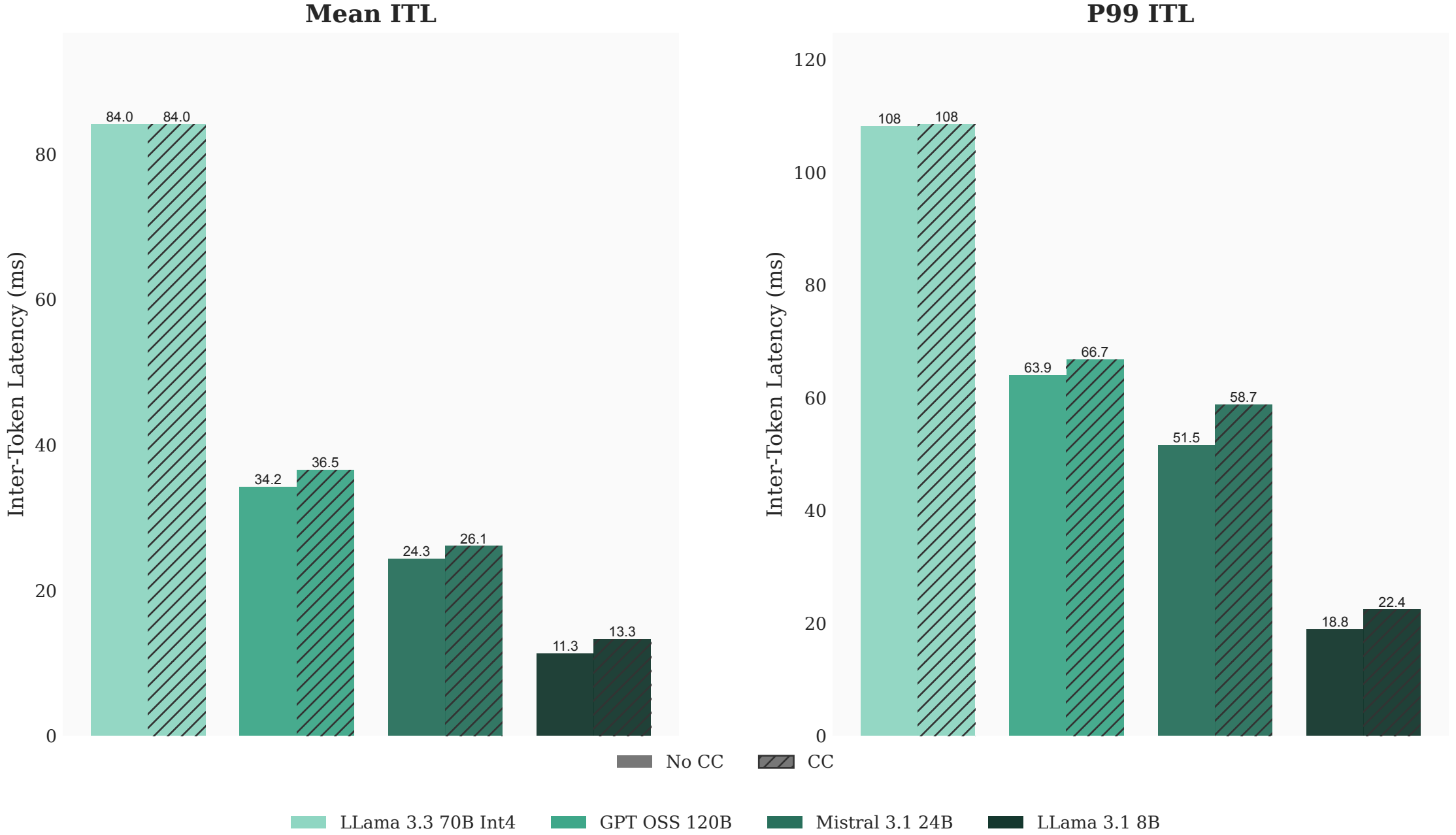

ShareGPT (50 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
Inter-Token Latency (ms)
0
10
20
30
40
50
60
0
20
40
60
80
100
120
54.8
55.2
28.0
32.2
21.0
22.6
9.0
10.5
92.8
107
38.7
110
23.2
25.5
10.7
13.4
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

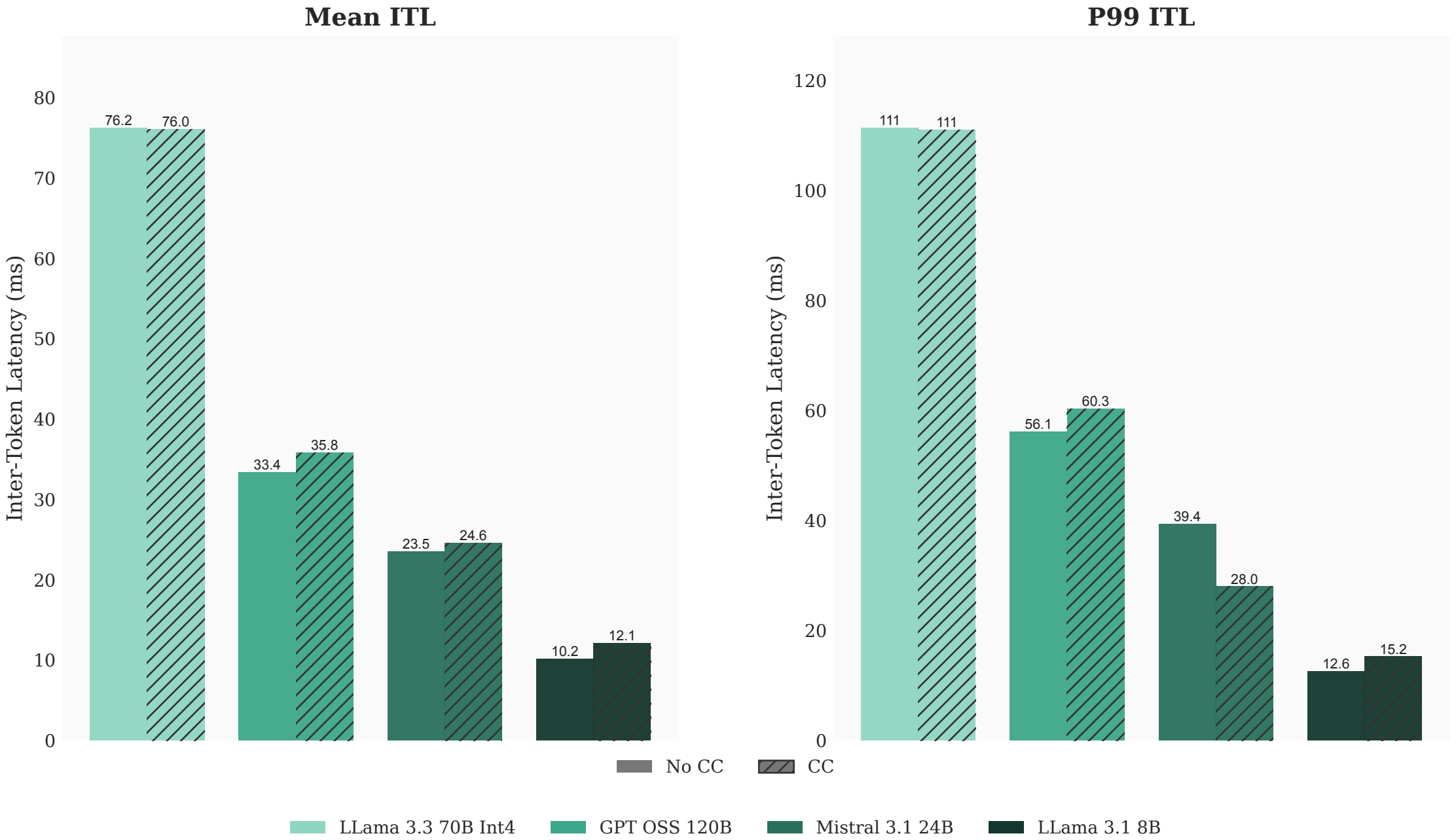

ShareGPT (Single Request)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
0
5
10
15
20
25
19.9
20.8
7.1
7.6
17.6
18.6
6.9
7.5
21.8
23.1
11.1
11.5
18.7
20.0
7.3
8.2
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

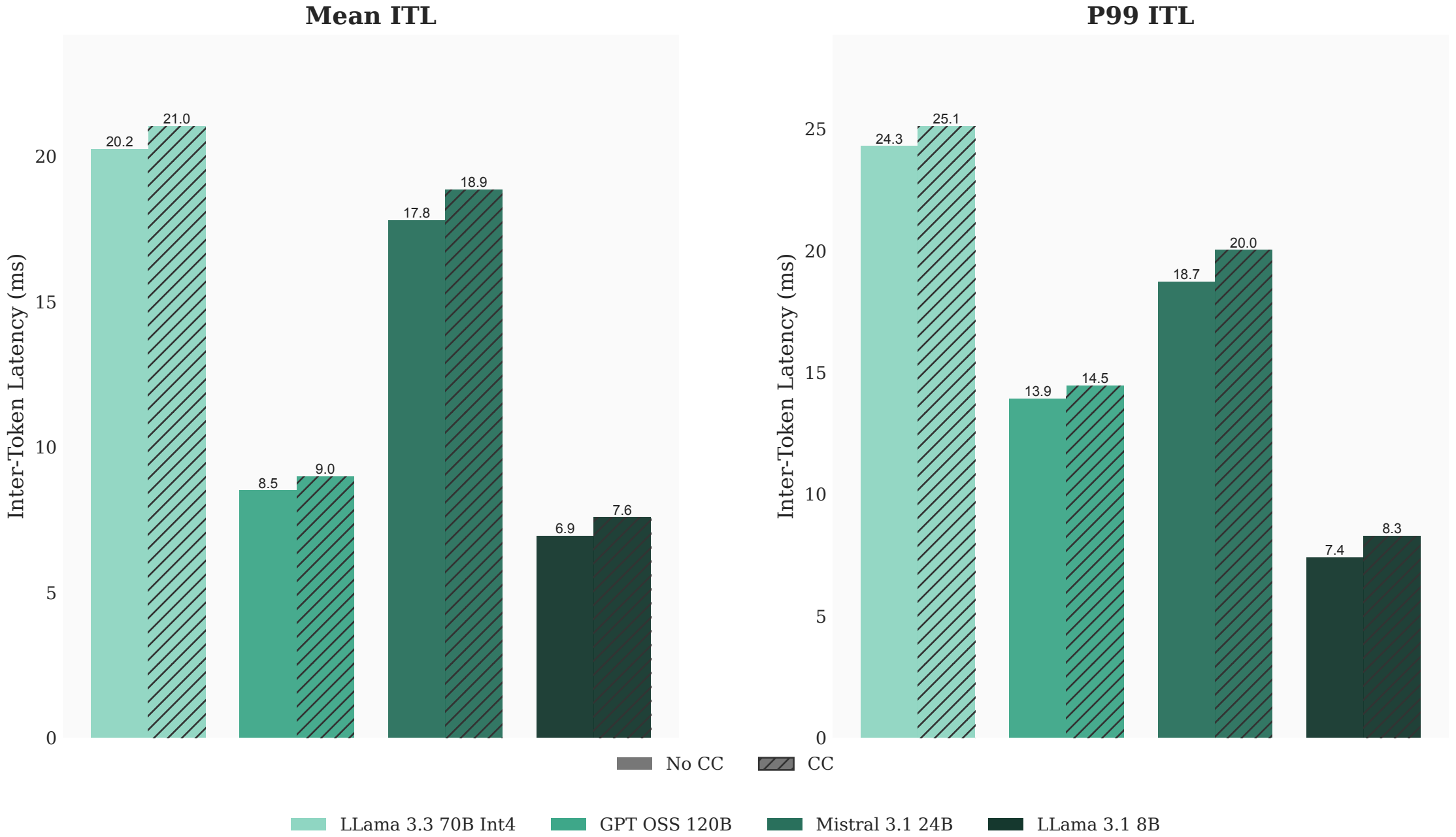

Edit 10K Characters (100 Request Rate)
Mean ITL
P99 ITL
Inter-Token Latency (ms)
Inter-Token Latency (ms)
0
10
20
30
40
50
60
0
50
100
150
200
57.0
57.0
28.6
28.0
36.7
38.0
31.2
34.0
40.8
41.0
139
136
206
202
54.6
57.5
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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


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


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


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


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