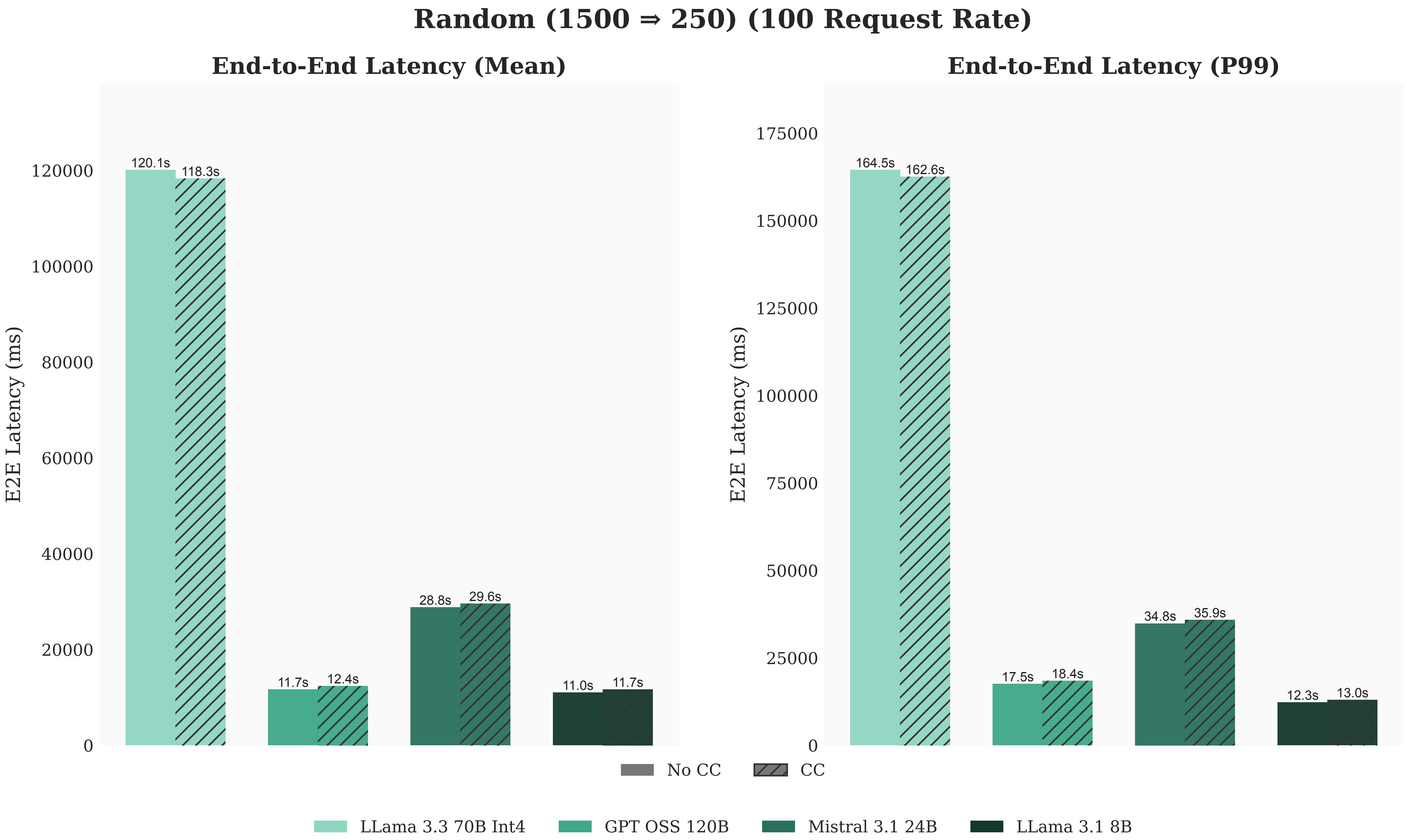
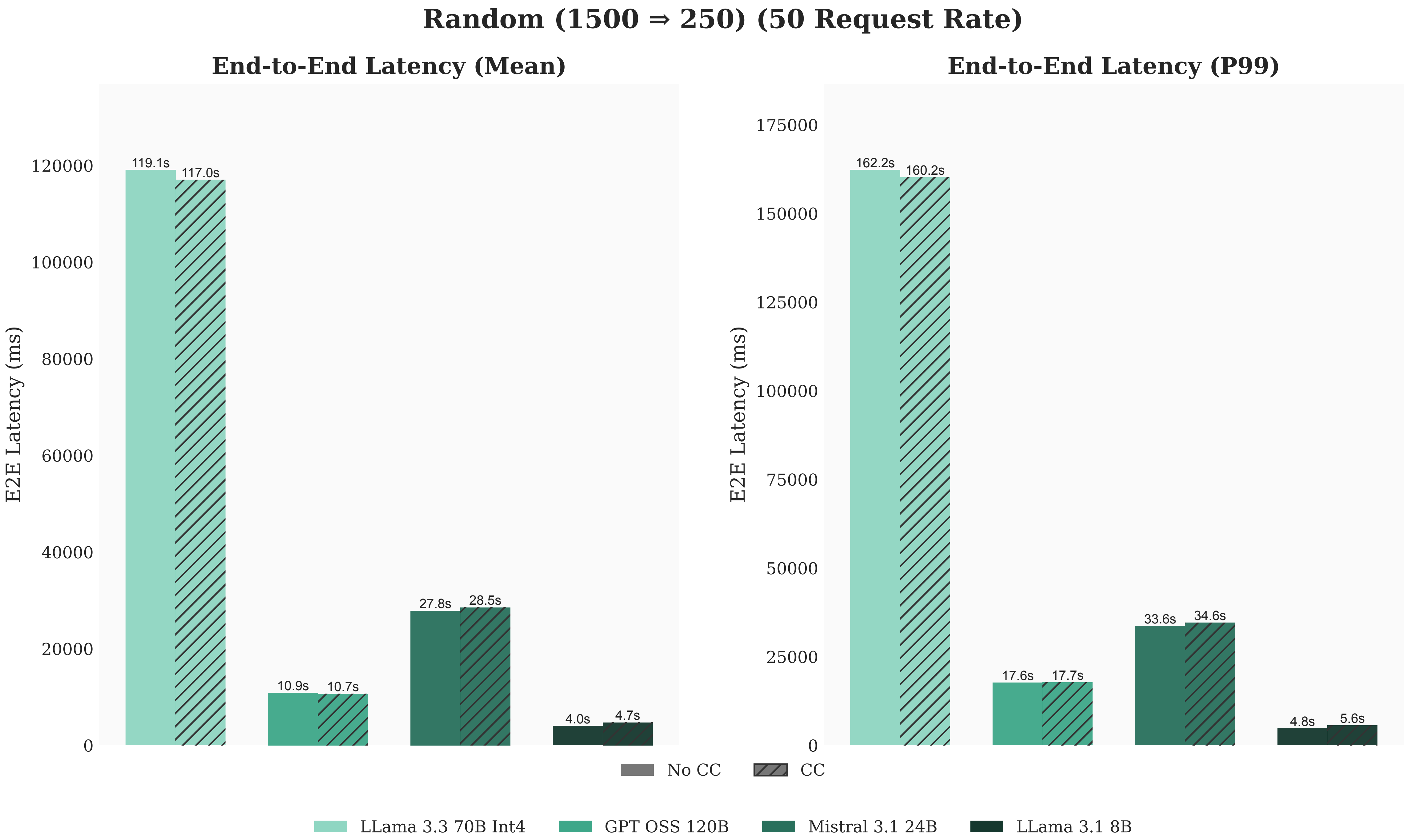
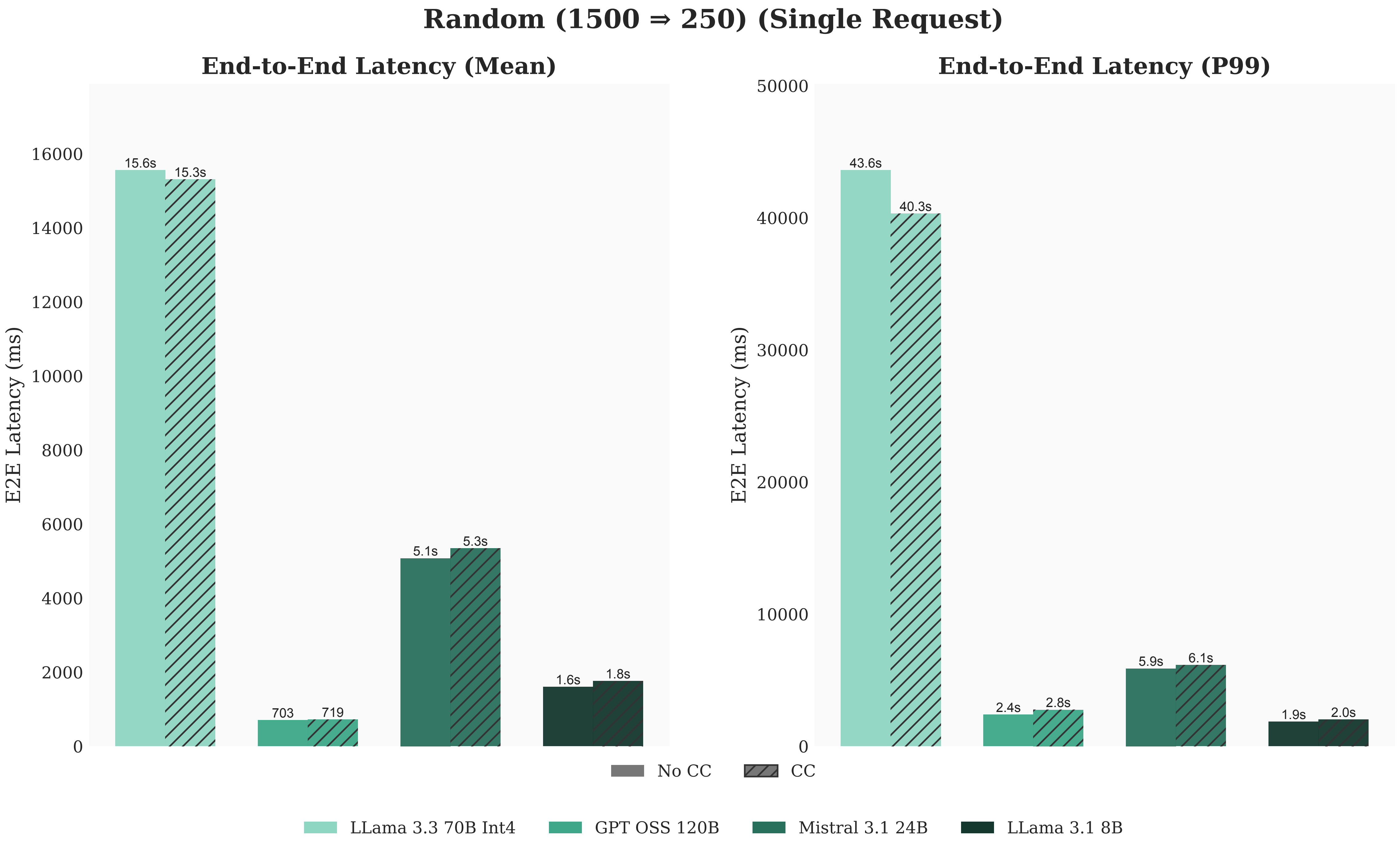
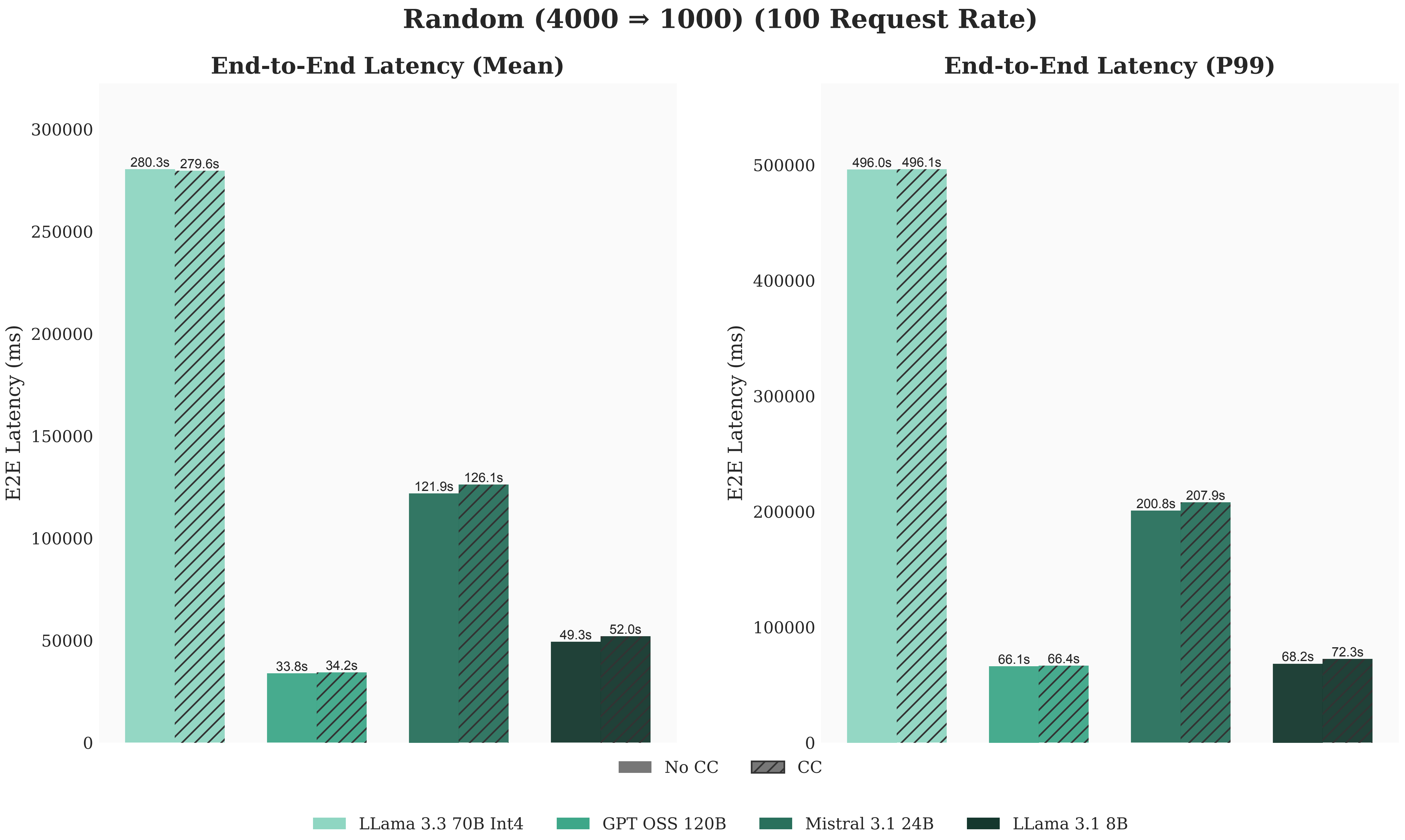
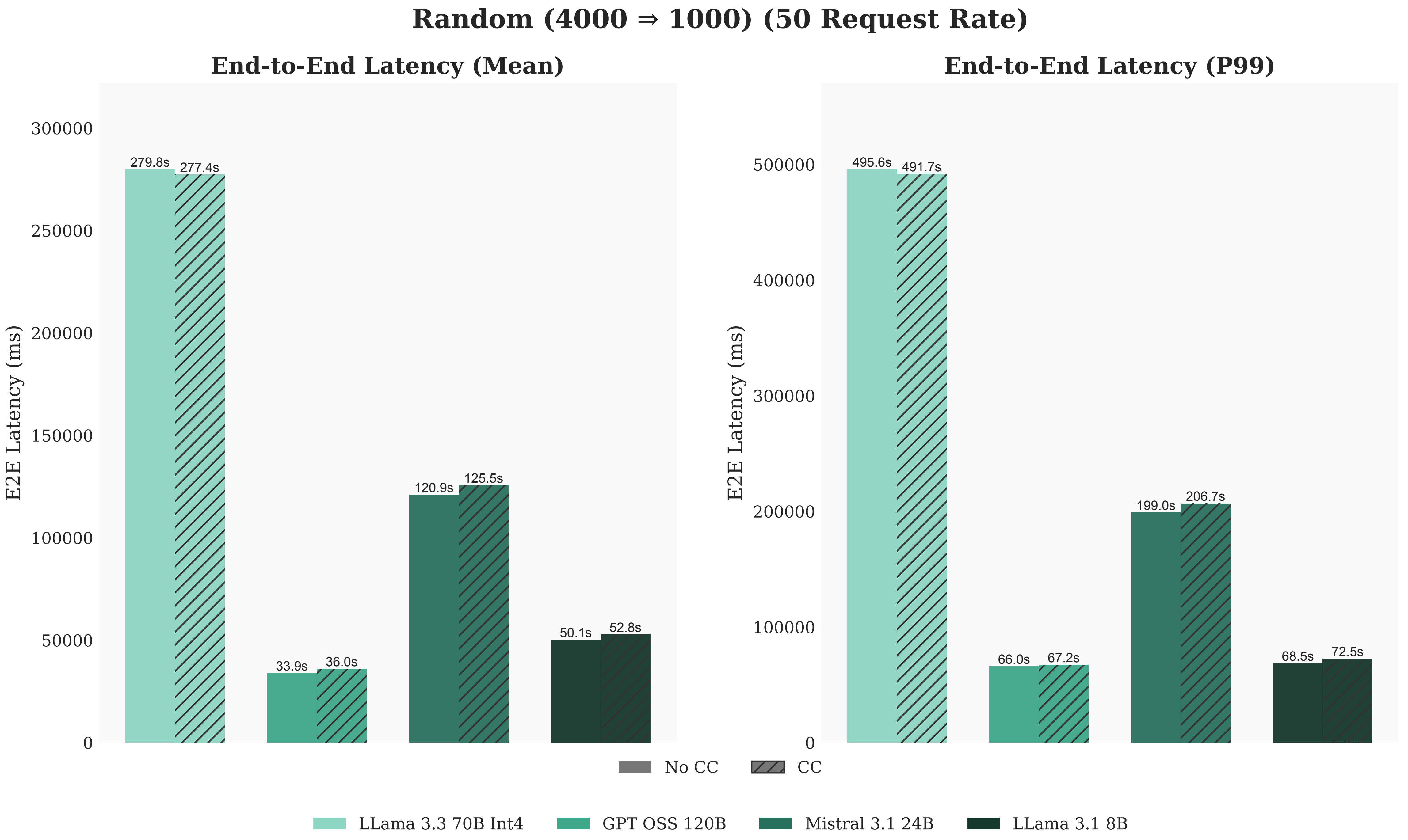
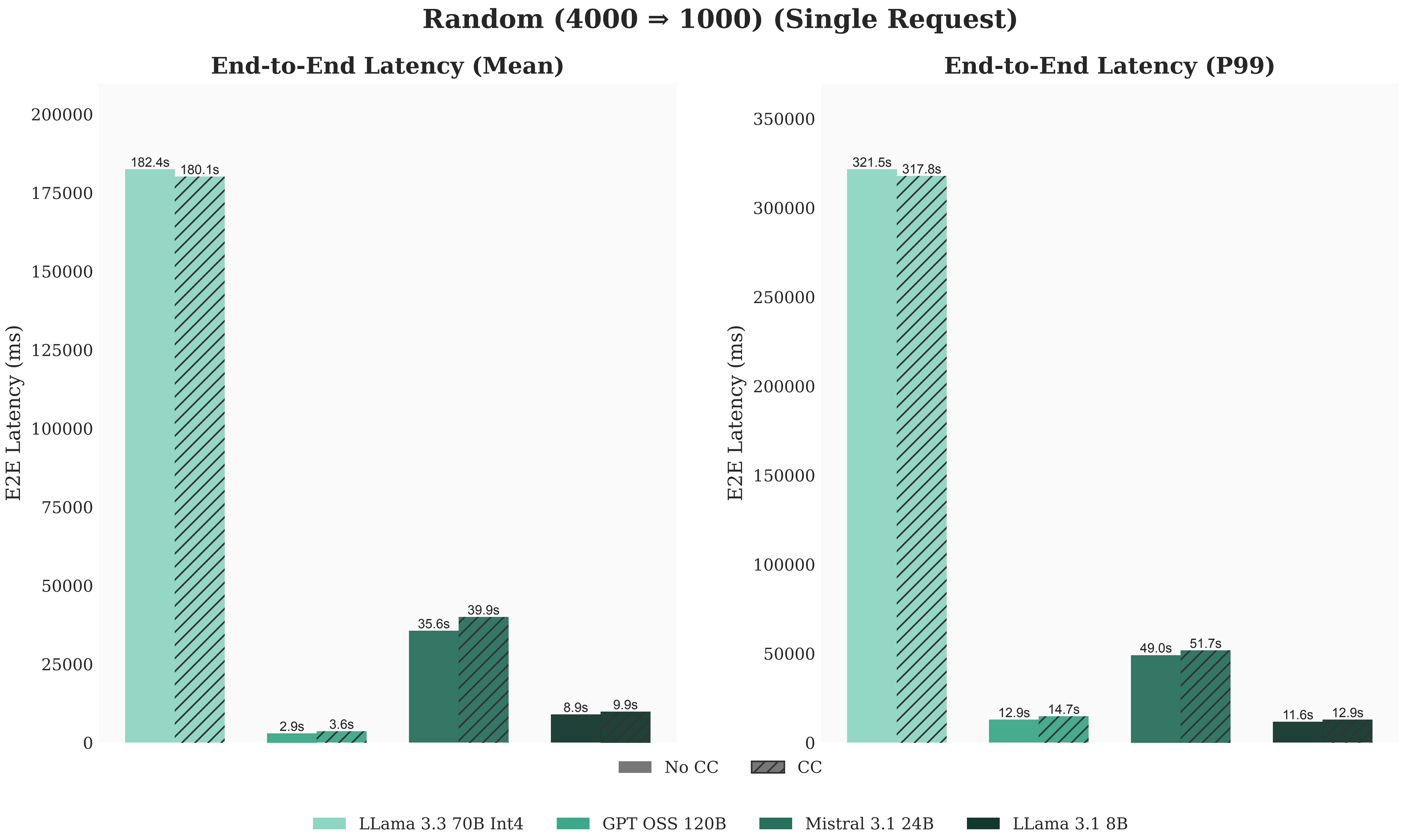

Random (1500 ⇒ 250) (100 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
120.1s
118.3s
11.7s
12.4s
28.8s
29.6s
11.0s
11.7s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
164.5s
162.6s
17.5s
18.4s
34.8s
35.9s
12.3s
13.0s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

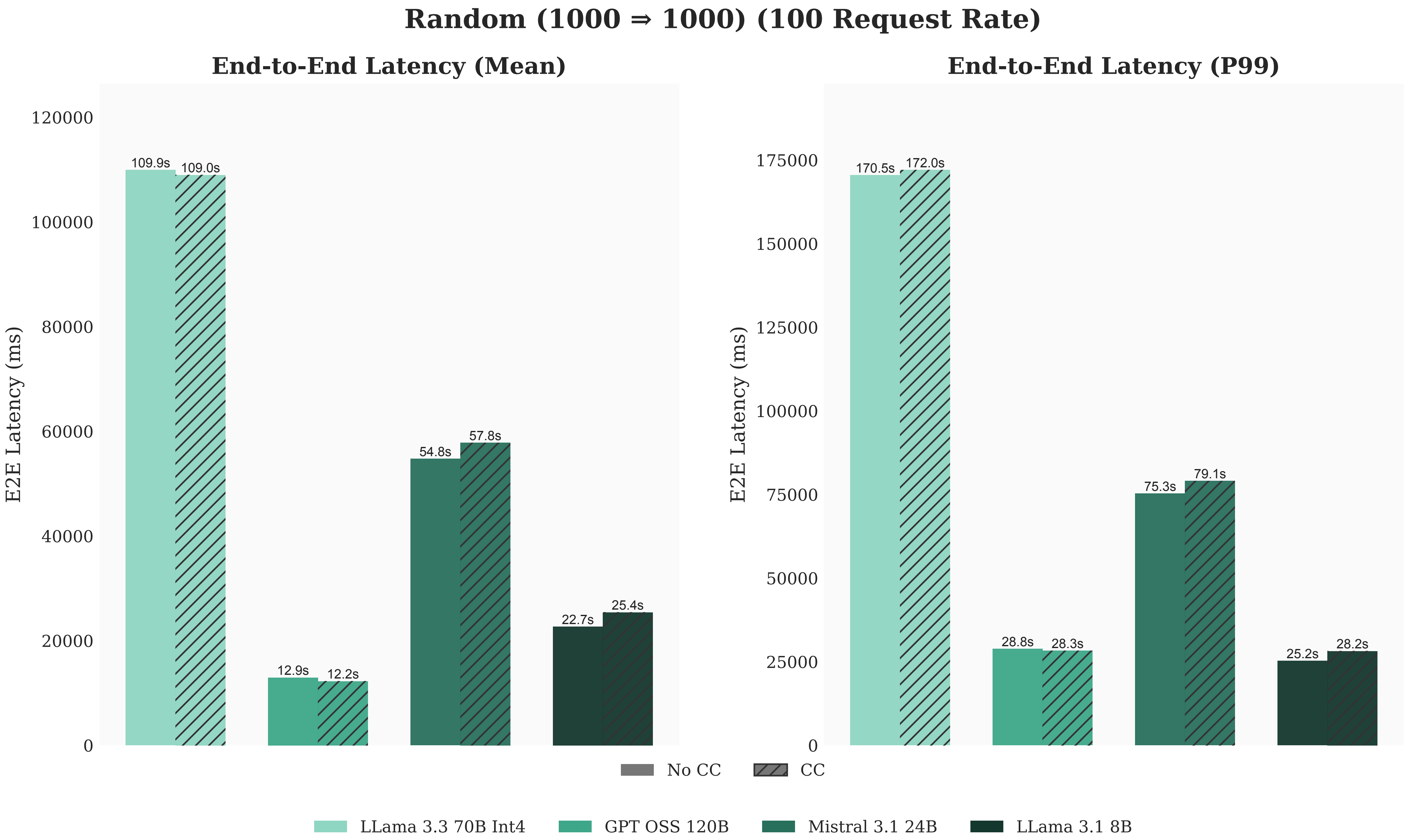

Random (1500 ⇒ 250) (50 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
119.1s
117.0s
10.9s
10.7s
27.8s
28.5s
4.0s
4.7s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
162.2s
160.2s
17.6s
17.7s
33.6s
34.6s
4.8s
5.6s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

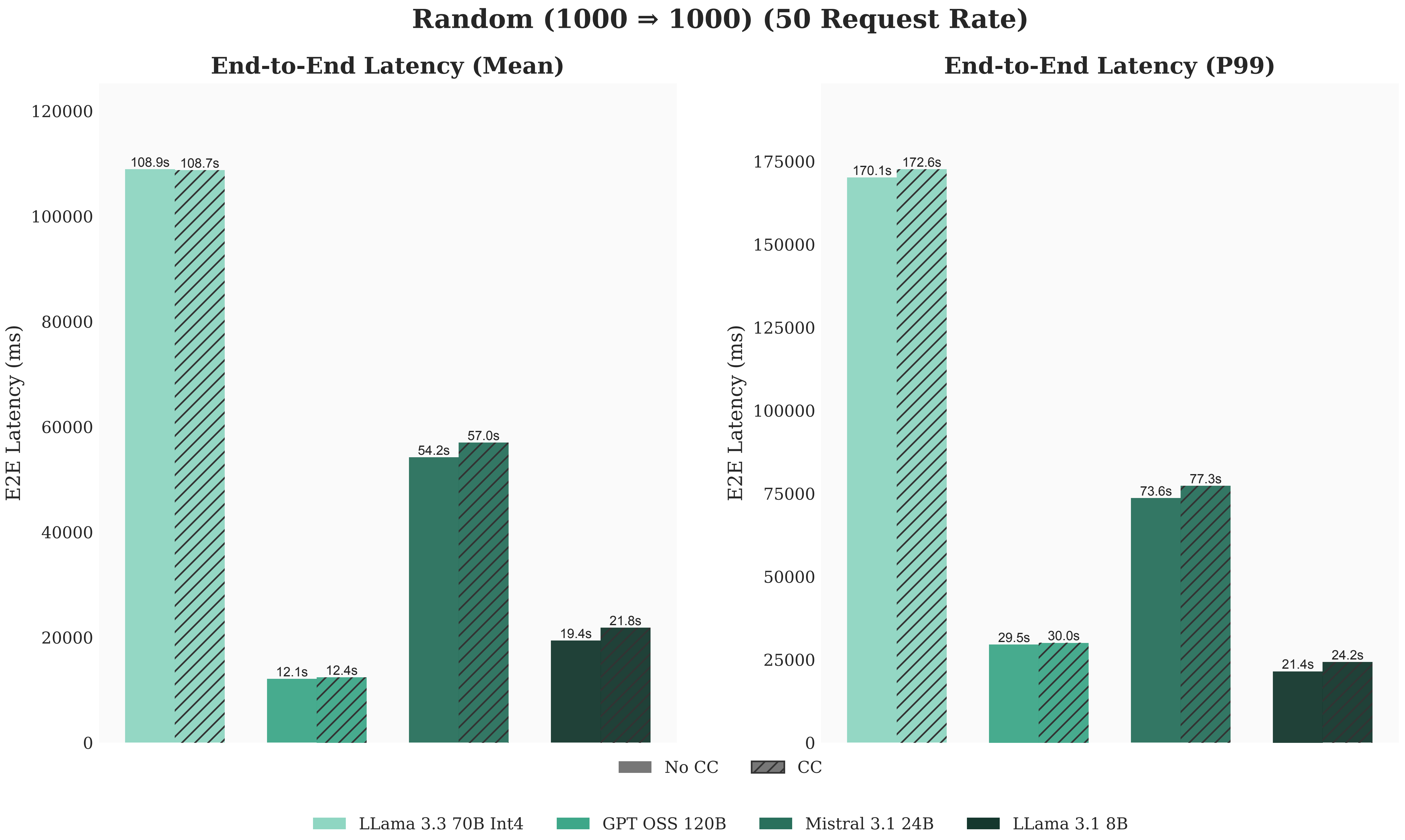

Random (1500 ⇒ 250) (Single Request)
End-to-End Latency (Mean)
End-to-End Latency (P99)
E2E Latency (ms)
0
2000
4000
6000
8000
10000
12000
14000
16000
15.6s
15.3s
703
719
5.1s
5.3s
1.6s
1.8s
E2E Latency (ms)
0
10000
20000
30000
40000
50000
43.6s
40.3s
2.4s
2.8s
5.9s
6.1s
1.9s
2.0s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

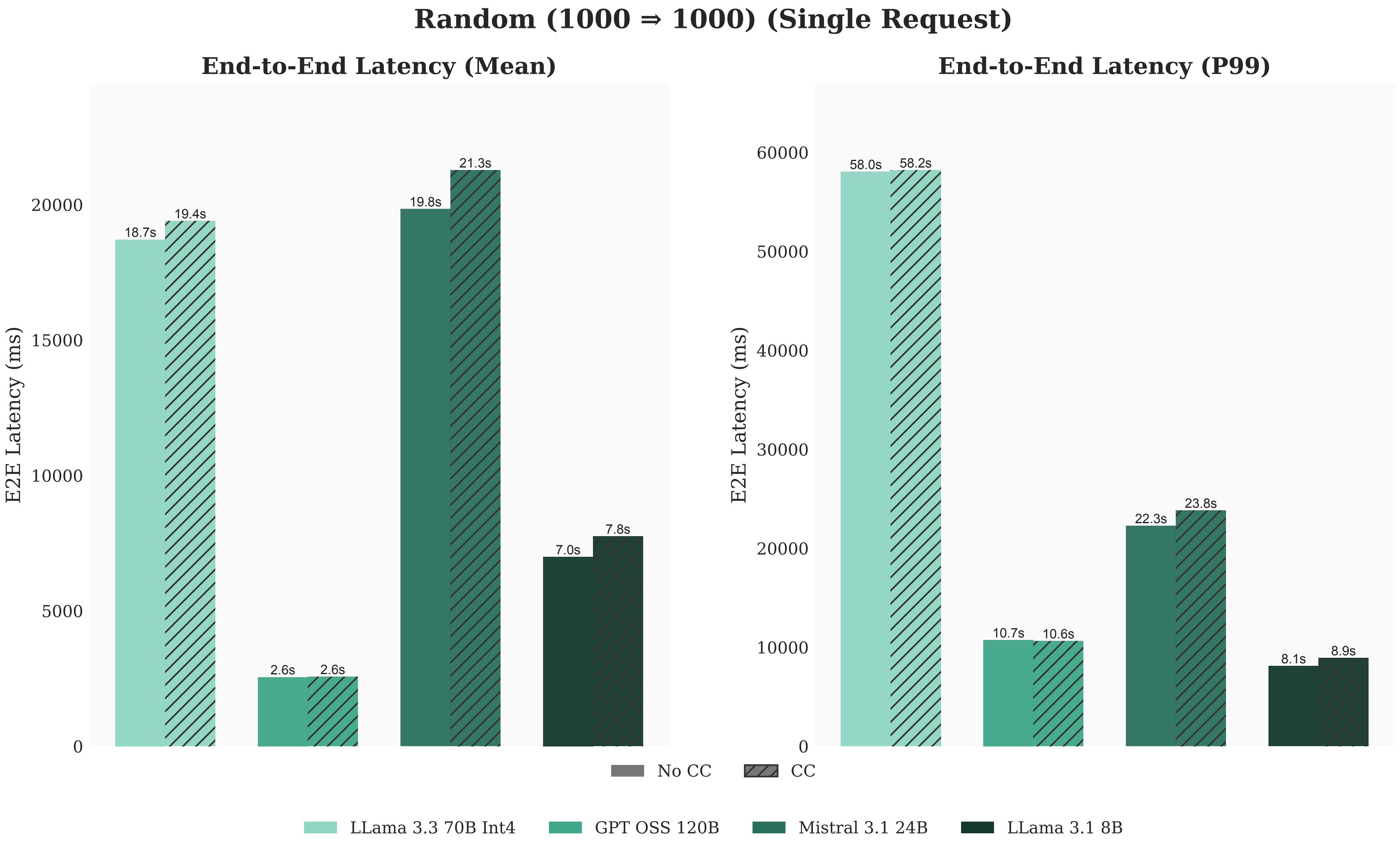

Random (4000 ⇒ 1000) (100 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
300000
250000
200000
150000
100000
50000
0
280.3s
279.6s
33.8s
34.2s
121.9s
126.1s
49.3s
52.0s
End-to-End Latency (P99)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
496.0s
496.1s
66.1s
66.4s
200.8s
207.9s
68.2s
72.3s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

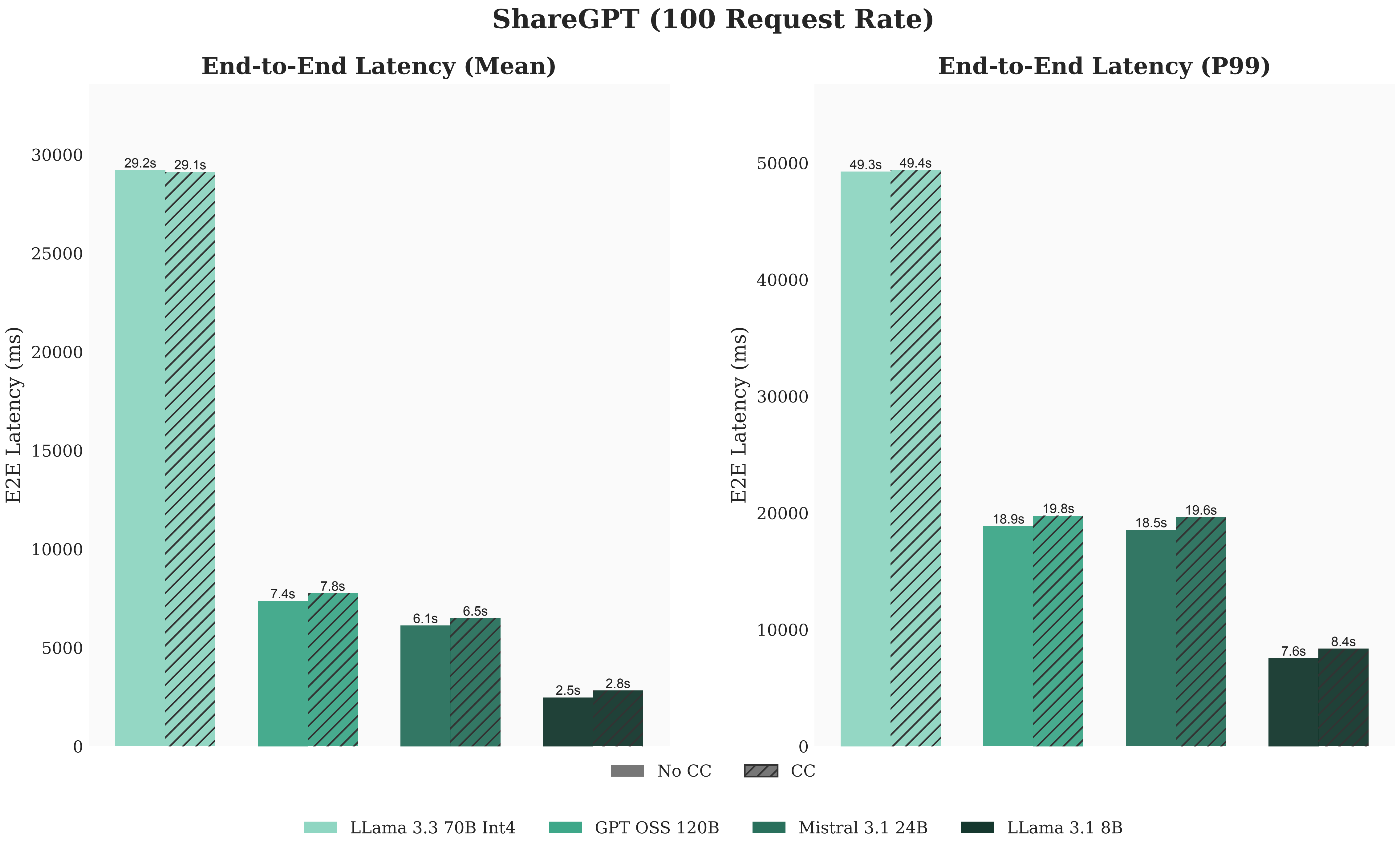

Random (4000 ⇒ 1000) (50 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
300000
250000
200000
150000
100000
50000
0
279.8s
277.4s
33.9s
36.0s
120.9s
125.5s
50.1s
52.8s
End-to-End Latency (P99)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
495.6s
491.7s
66.0s
67.2s
199.0s
206.7s
68.5s
72.5s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

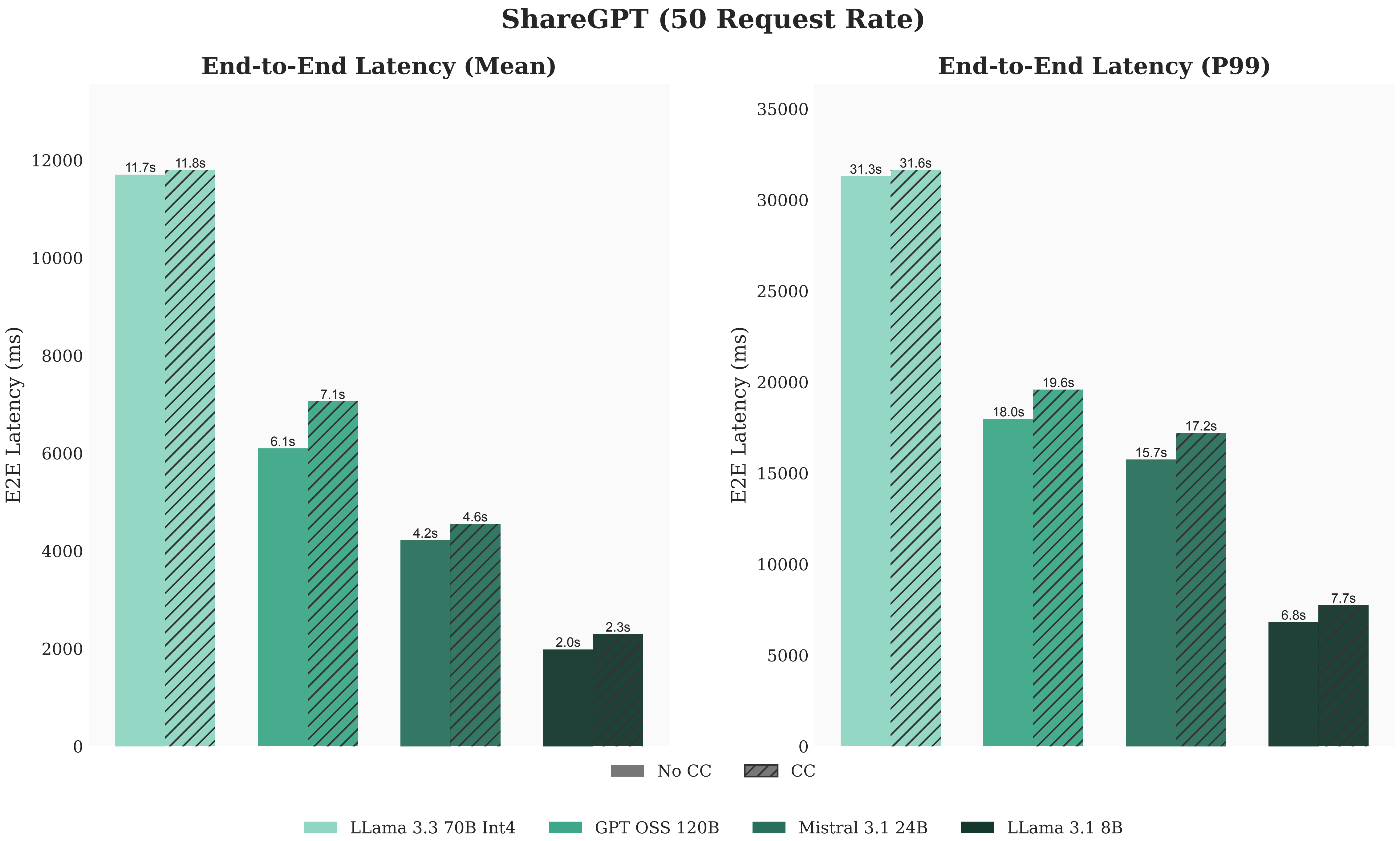

Random (4000 ⇒ 1000) (Single Request)
End-to-End Latency (Mean)
E2E Latency (ms)
200000
175000
150000
125000
100000
75000
50000
25000
0
182.4s
180.1s
2.9s
3.6s
35.6s
39.9s
8.9s
9.9s
End-to-End Latency (P99)
E2E Latency (ms)
350000
300000
250000
200000
150000
100000
50000
0
321.5s
317.8s
12.9s
14.7s
49.0s
51.7s
11.6s
12.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

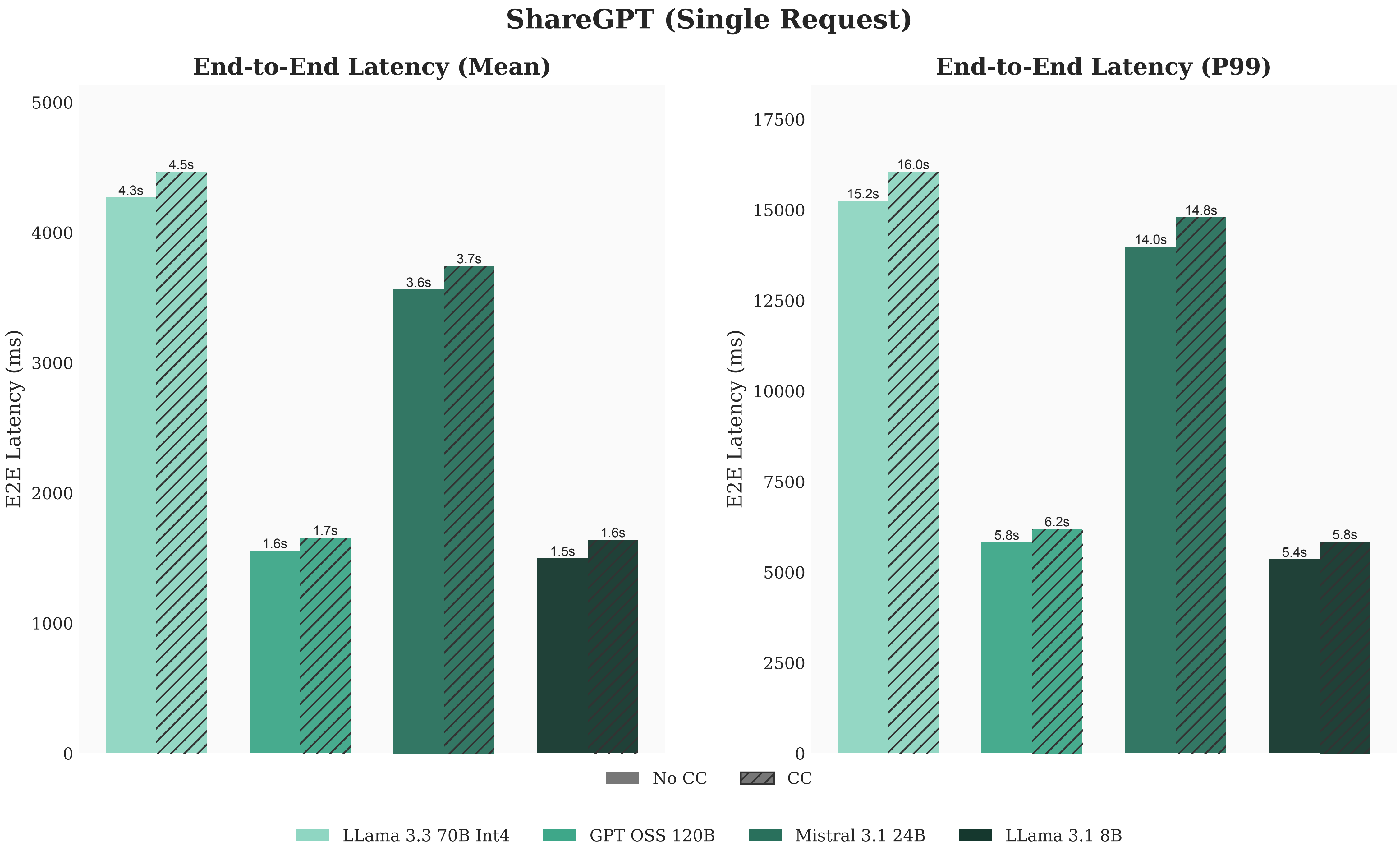

Random (1000 ⇒ 1000) (100 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
109.9s
109.0s
12.9s
12.2s
54.8s
57.8s
22.7s
25.4s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
170.5s
172.0s
28.8s
28.3s
75.3s
79.1s
25.2s
28.2s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

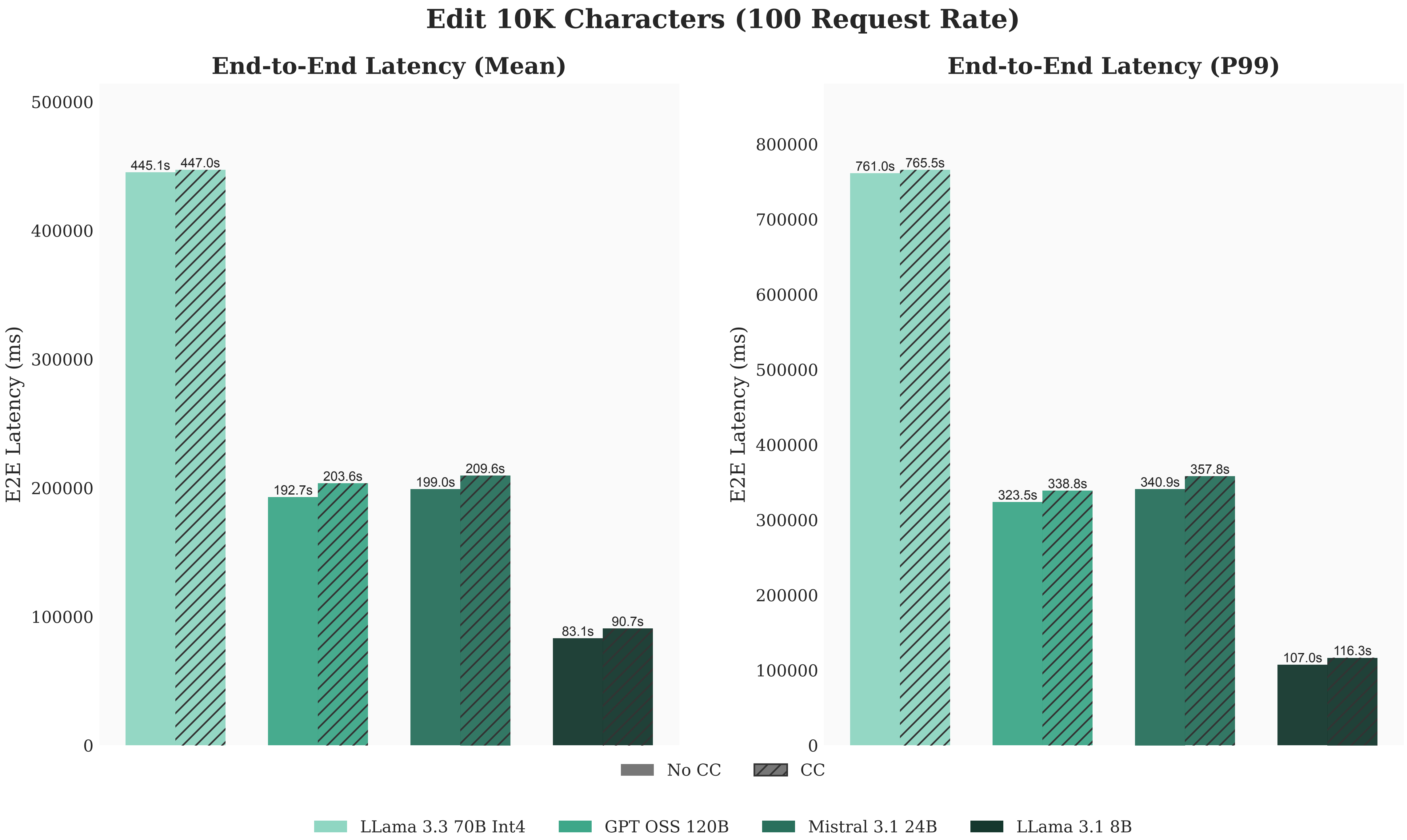

Random (1000 ⇒ 1000) (50 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
0
20000
40000
60000
80000
100000
120000
108.9s
108.7s
12.1s
12.4s
54.2s
57.0s
19.4s
21.8s
End-to-End Latency (P99)
E2E Latency (ms)
0
25000
50000
75000
100000
125000
150000
175000
170.1s
172.6s
29.5s
30.0s
73.6s
77.3s
21.4s
24.2s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

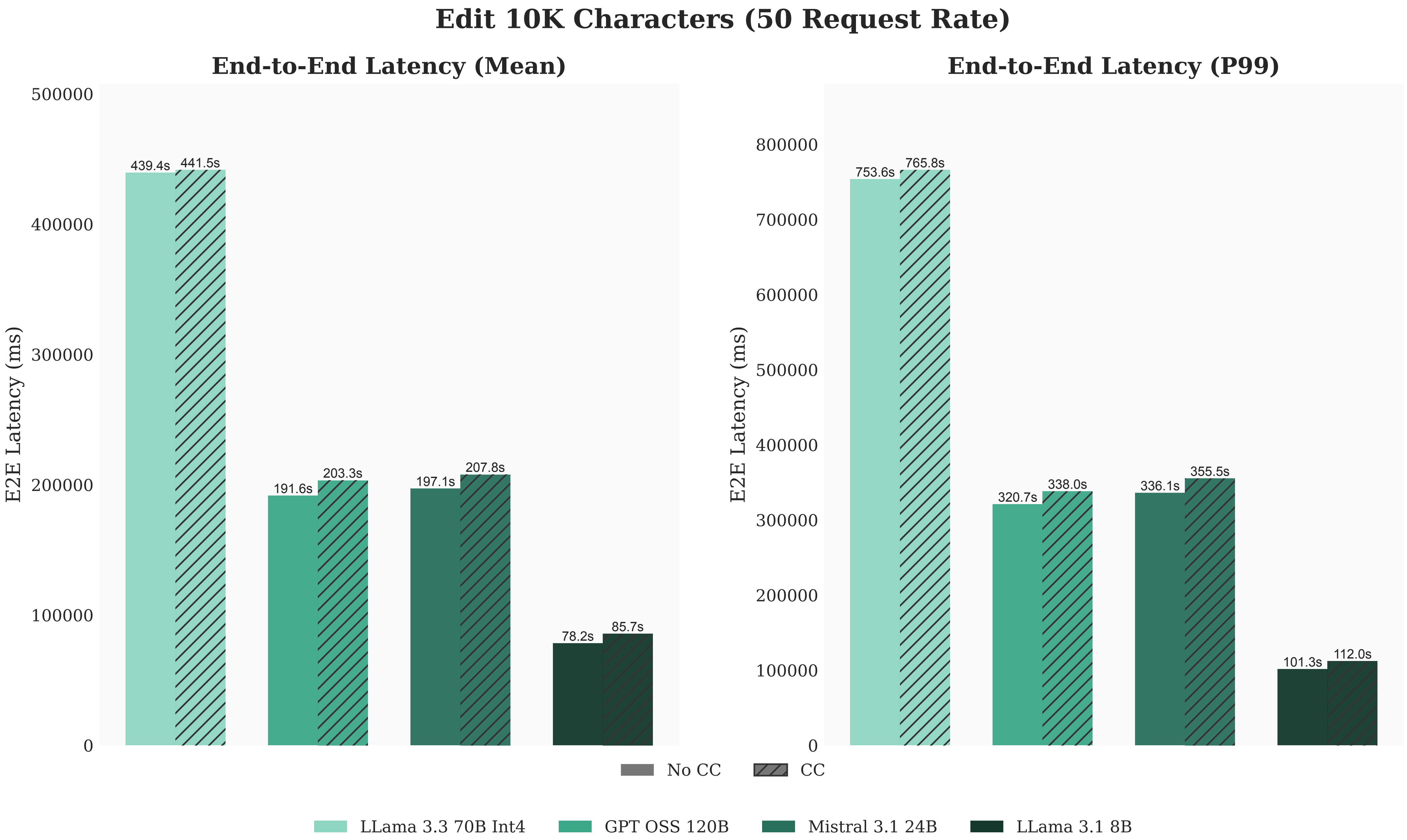

Random (1000 ⇒ 1000) (Single Request)
End-to-End Latency (Mean)
E2E Latency (ms)
0
5000
10000
15000
20000
18.7s
19.4s
2.6s
2.6s
19.8s
21.3s
7.0s
7.8s
End-to-End Latency (P99)
E2E Latency (ms)
0
10000
20000
30000
40000
50000
60000
58.0s
58.2s
10.7s
10.6s
22.3s
23.8s
8.1s
8.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

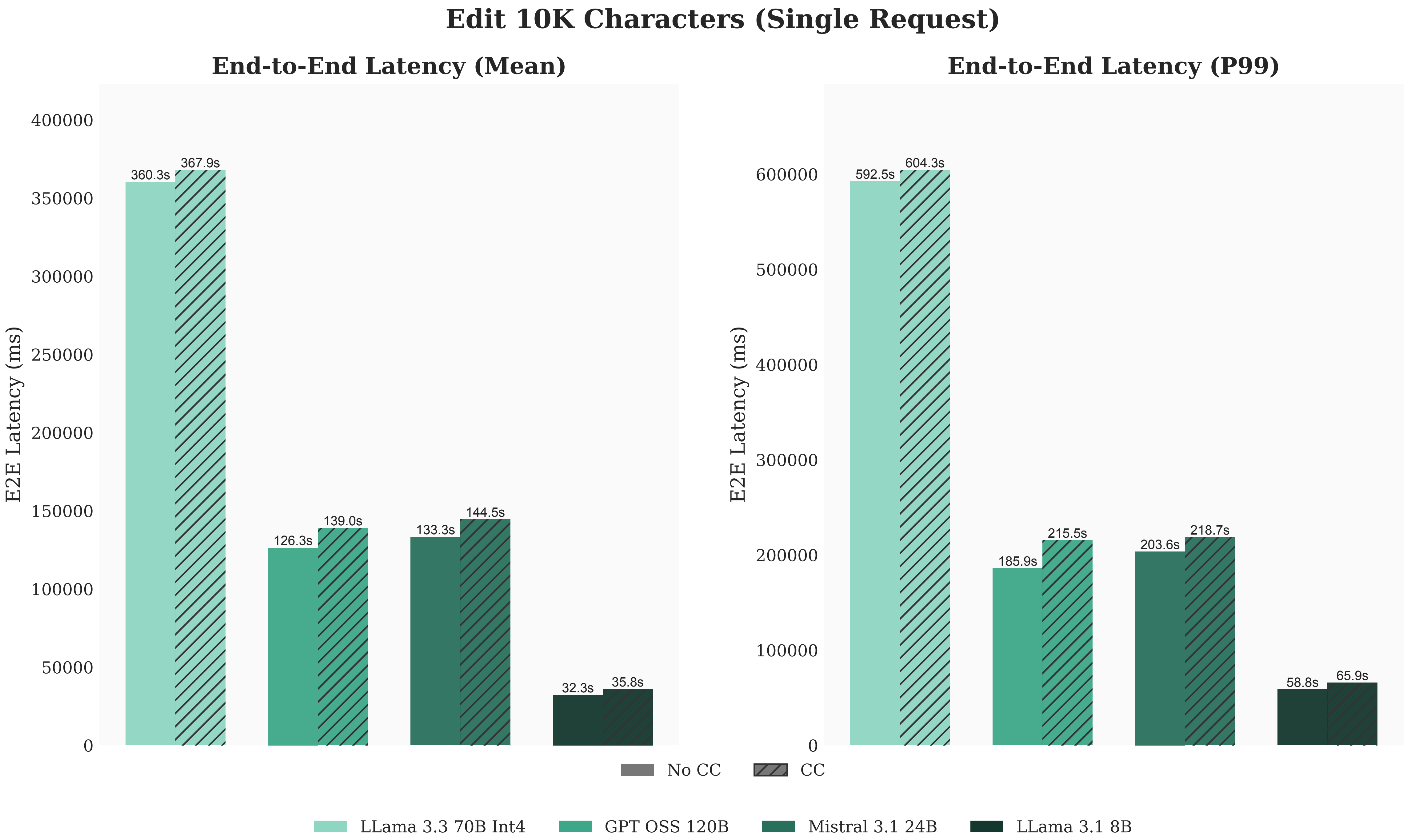

ShareGPT (100 Request Rate)
End-to-End Latency (Mean)
End-to-End Latency (P99)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
29.2s
29.1s
7.4s
7.8s
6.1s
6.5s
2.5s
2.8s
E2E Latency (ms)
0
10000
20000
30000
40000
50000
49.3s
49.4s
18.9s
19.8s
18.5s
19.6s
7.6s
8.4s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

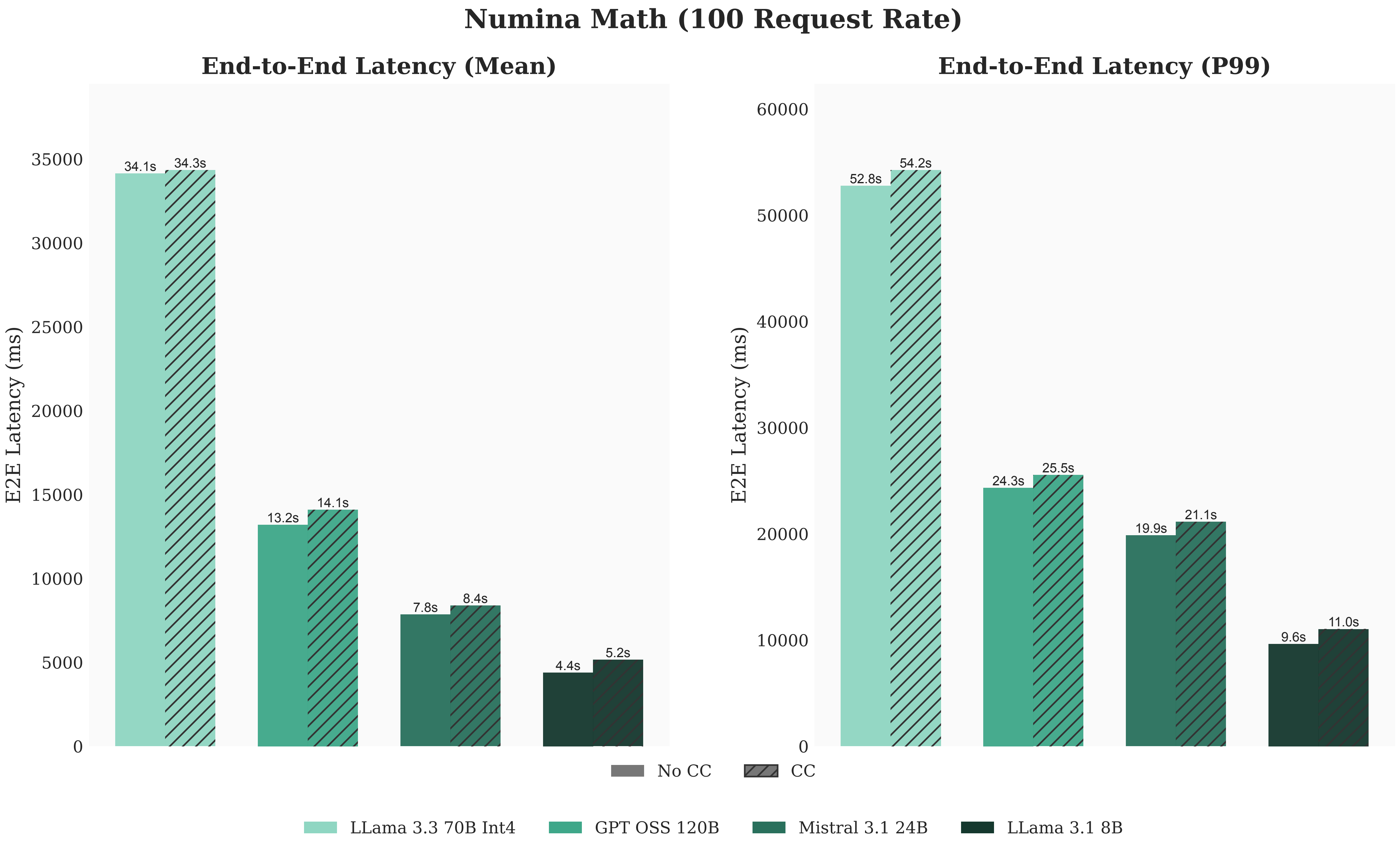

ShareGPT (50 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
0
2000
4000
6000
8000
10000
12000
11.7s
11.8s
6.1s
7.1s
4.2s
4.6s
2.0s
2.3s
End-to-End Latency (P99)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
35000
31.3s
31.6s
18.0s
19.6s
15.7s
17.2s
6.8s
7.7s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

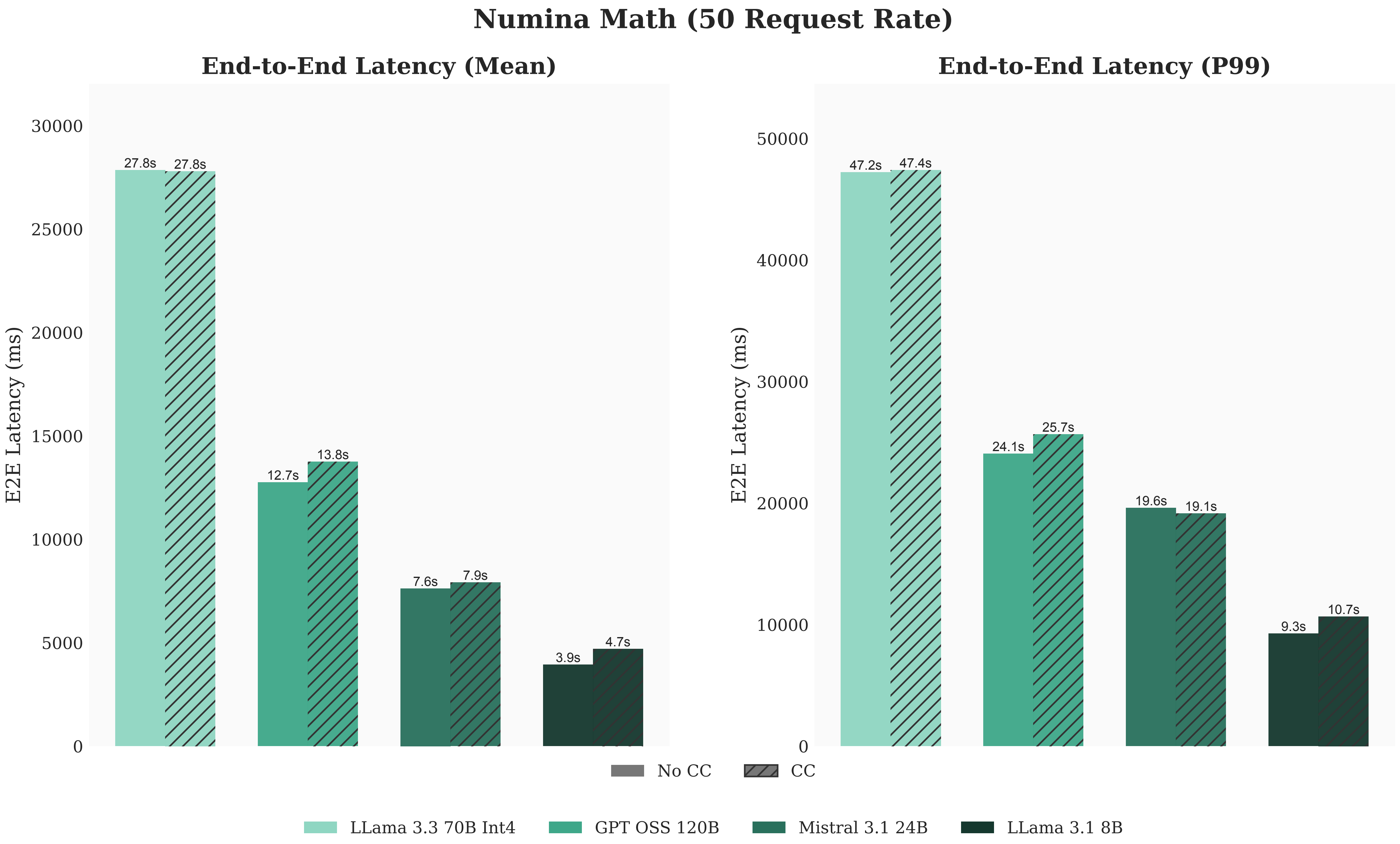

ShareGPT (Single Request)
End-to-End Latency (Mean)
End-to-End Latency (P99)
E2E Latency (ms)
0
1000
2000
3000
4000
5000
4.3s
4.5s
1.6s
1.7s
3.6s
3.7s
1.5s
1.6s
E2E Latency (ms)
0
2500
5000
7500
10000
12500
15000
17500
15.2s
16.0s
5.8s
6.2s
14.0s
14.8s
5.4s
5.8s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

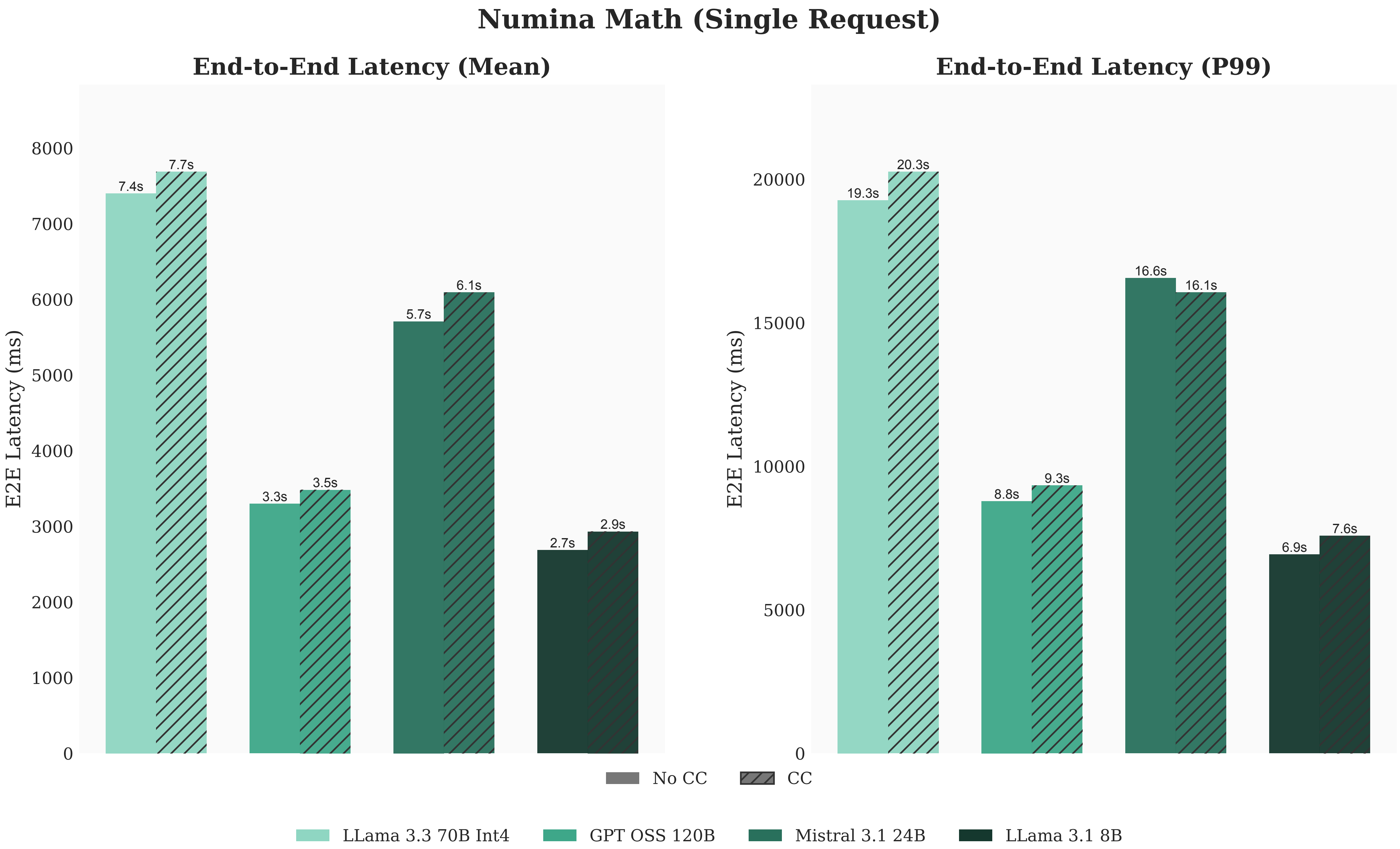

Edit 10K Characters (100 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
445.1s
447.0s
192.7s
203.6s
199.0s
209.6s
83.1s
90.7s
End-to-End Latency (P99)
E2E Latency (ms)
800000
700000
600000
500000
400000
300000
200000
100000
0
761.0s
765.5s
323.5s
338.8s
340.9s
357.8s
107.0s
116.3s
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)
End-to-End Latency (Mean)
End-to-End Latency (P99)
E2E Latency (ms)
500000
400000
300000
200000
100000
0
439.4s
441.5s
191.6s
203.3s
197.1s
207.8s
78.2s
85.7s
E2E Latency (ms)
800000
700000
600000
500000
400000
300000
200000
100000
0
753.6s
765.8s
320.7s
338.0s
336.1s
355.5s
101.3s
112.0s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


Edit 10K Characters (Single Request)
End-to-End Latency (Mean)
End-to-End Latency (P99)
E2E Latency (ms)
400000
350000
300000
250000
200000
150000
100000
50000
0
360.3s
367.9s
126.3s
139.0s
133.3s
144.5s
32.3s
35.8s
E2E Latency (ms)
600000
500000
400000
300000
200000
100000
0
592.5s
604.3s
185.9s
215.5s
203.6s
218.7s
58.8s
65.9s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


Numina Math (100 Request Rate)
End-to-End Latency (Mean)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
35000
34.1s
34.3s
13.2s
14.1s
7.8s
8.4s
4.4s
5.2s
End-to-End Latency (P99)
E2E Latency (ms)
0
10000
20000
30000
40000
50000
60000
52.8s
54.2s
24.3s
25.5s
19.9s
21.1s
9.6s
11.0s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


Numina Math (50 Request Rate)
End-to-End Latency (Mean)
End-to-End Latency (P99)
E2E Latency (ms)
0
5000
10000
15000
20000
25000
30000
27.8s
27.8s
12.7s
13.8s
7.6s
7.9s
3.9s
4.7s
E2E Latency (ms)
0
10000
20000
30000
40000
50000
47.2s
47.4s
24.1s
25.7s
19.6s
19.1s
9.3s
10.7s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


Numina Math (Single Request)
End-to-End Latency (Mean)
E2E Latency (ms)
0
1000
2000
3000
4000
5000
6000
7000
8000
7.4s
7.7s
3.3s
3.5s
5.7s
6.1s
2.7s
2.9s
End-to-End Latency (P99)
E2E Latency (ms)
0
5000
10000
15000
20000
19.3s
20.3s
8.8s
9.3s
16.6s
16.1s
6.9s
7.6s
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B