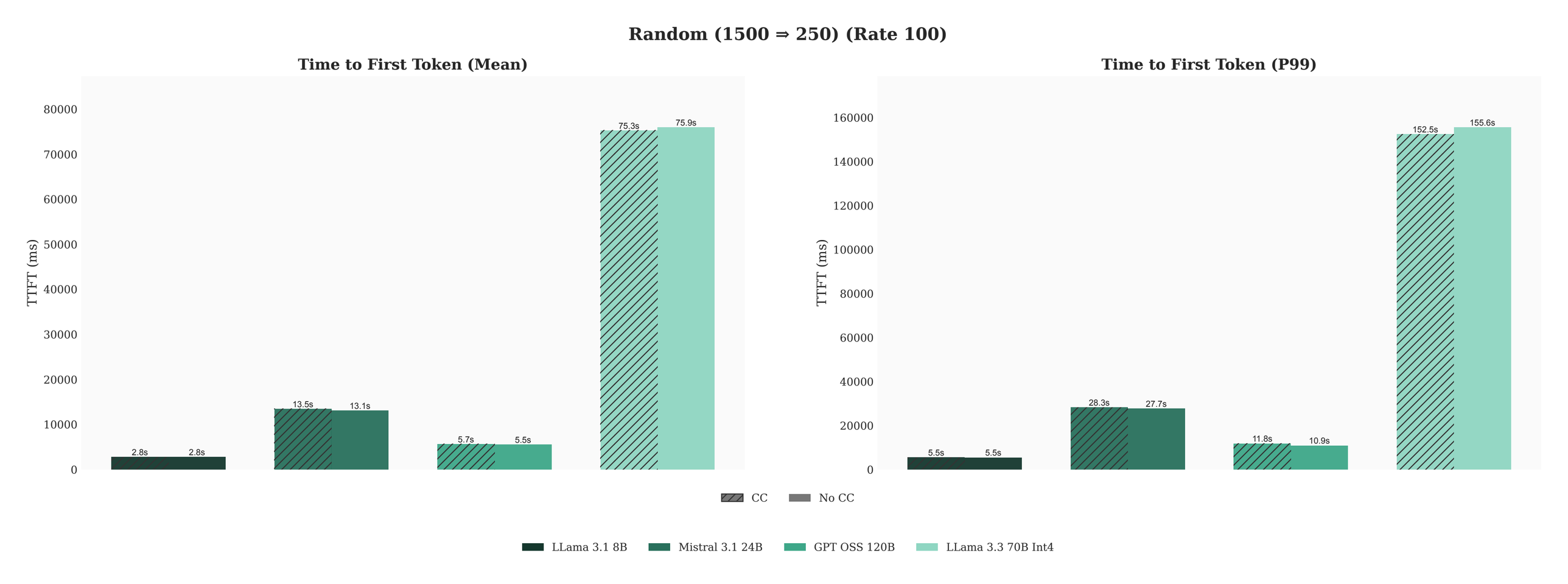
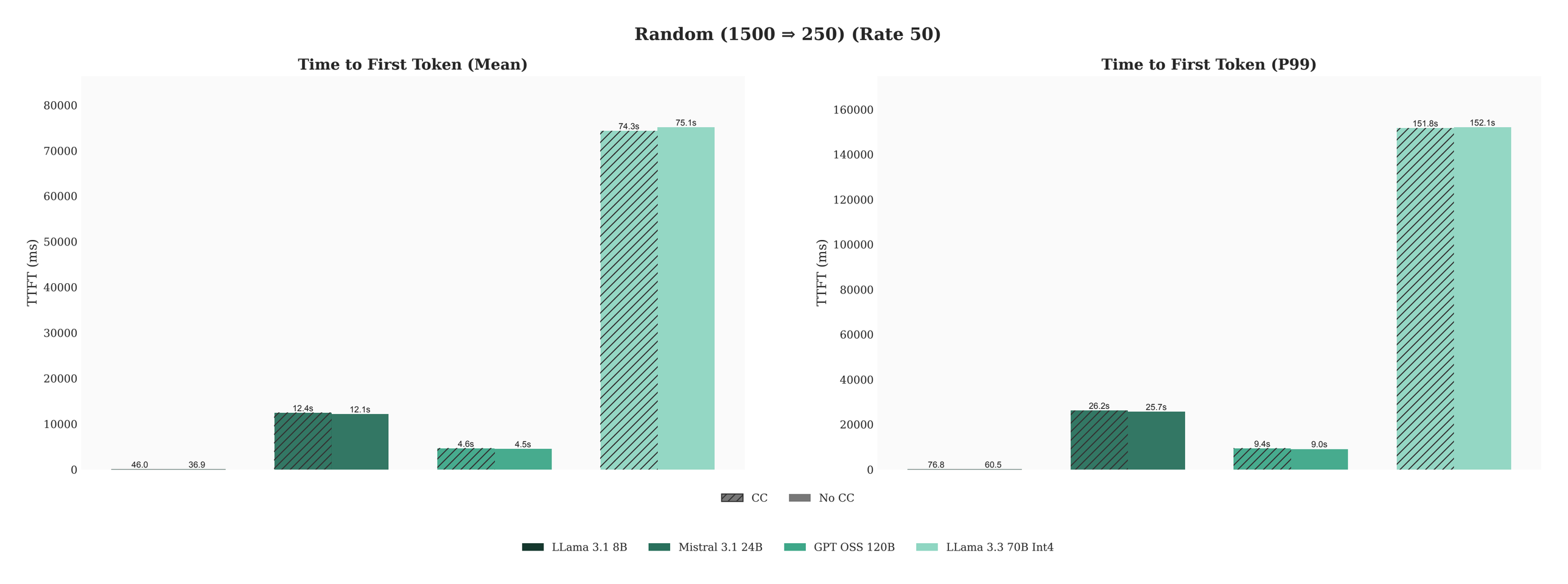
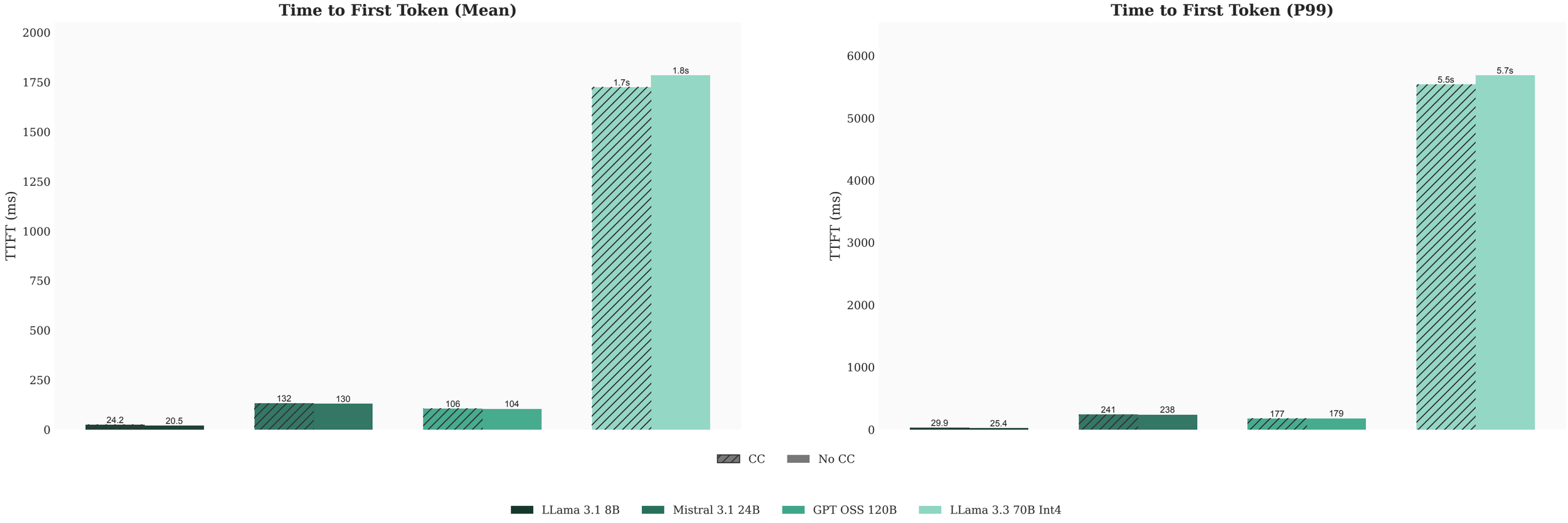
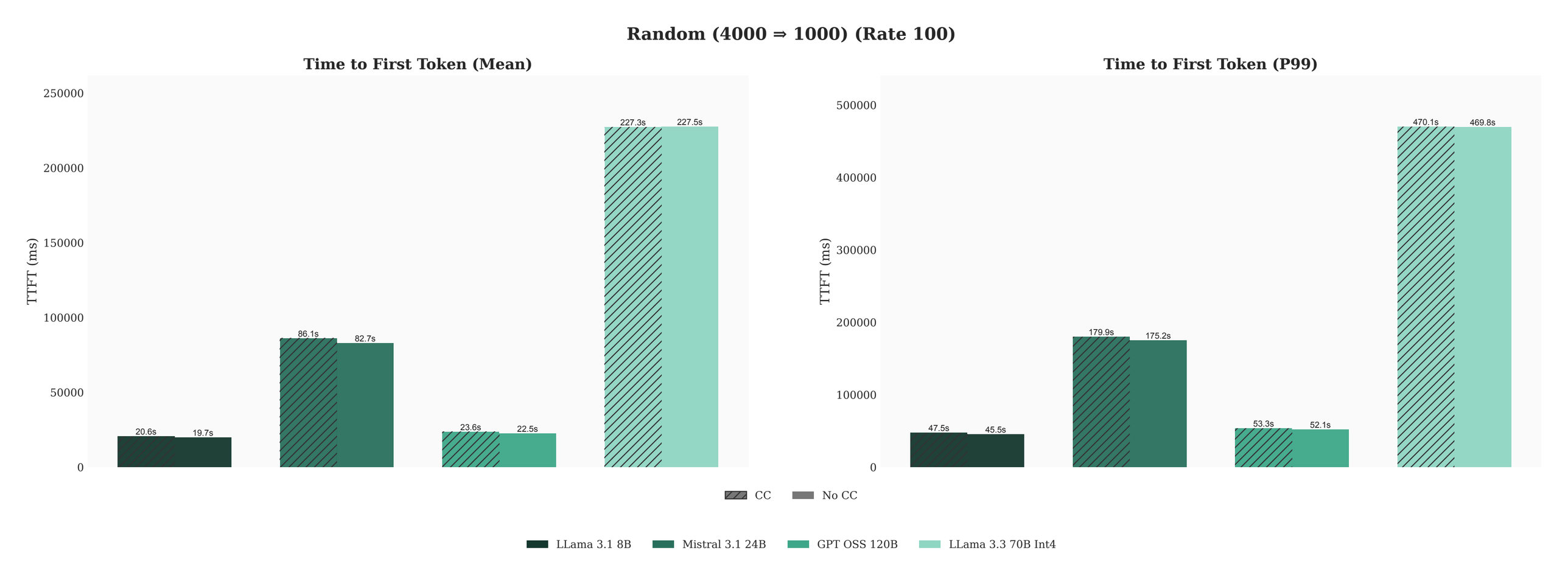

Random (1500 ⇒ 250) (Rate 100)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
60000
70000
80000
2.8s
2.8s
13.5s
13.1s
5.7s
5.5s
75.3s
75.9s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
160000
5.5s
5.5s
28.3s
27.7s
11.8s
10.9s
152.5s
155.6s
CC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

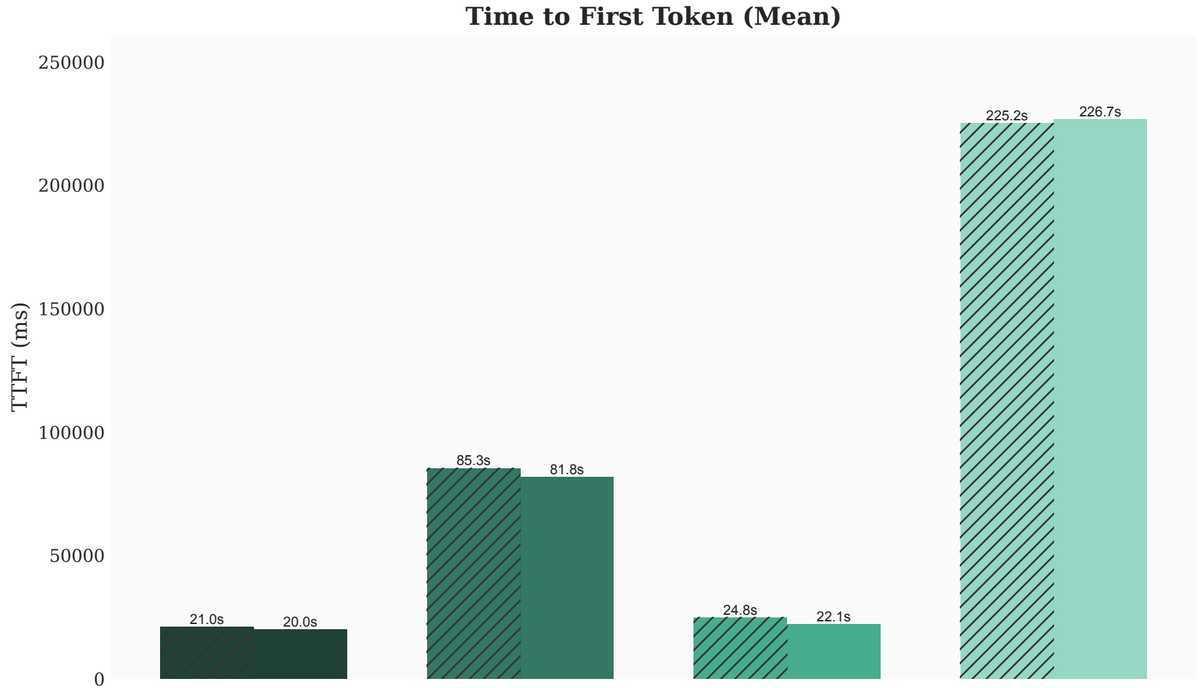

Random (1500 ⇒ 250) (Rate 50)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
60000
70000
80000
46.0
36.9
12.4s
12.1s
4.6s
4.5s
74.3s
75.1s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
160000
76.8
60.5
26.2s
25.7s
9.4s
9.0s
151.8s
152.1s
CC
No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

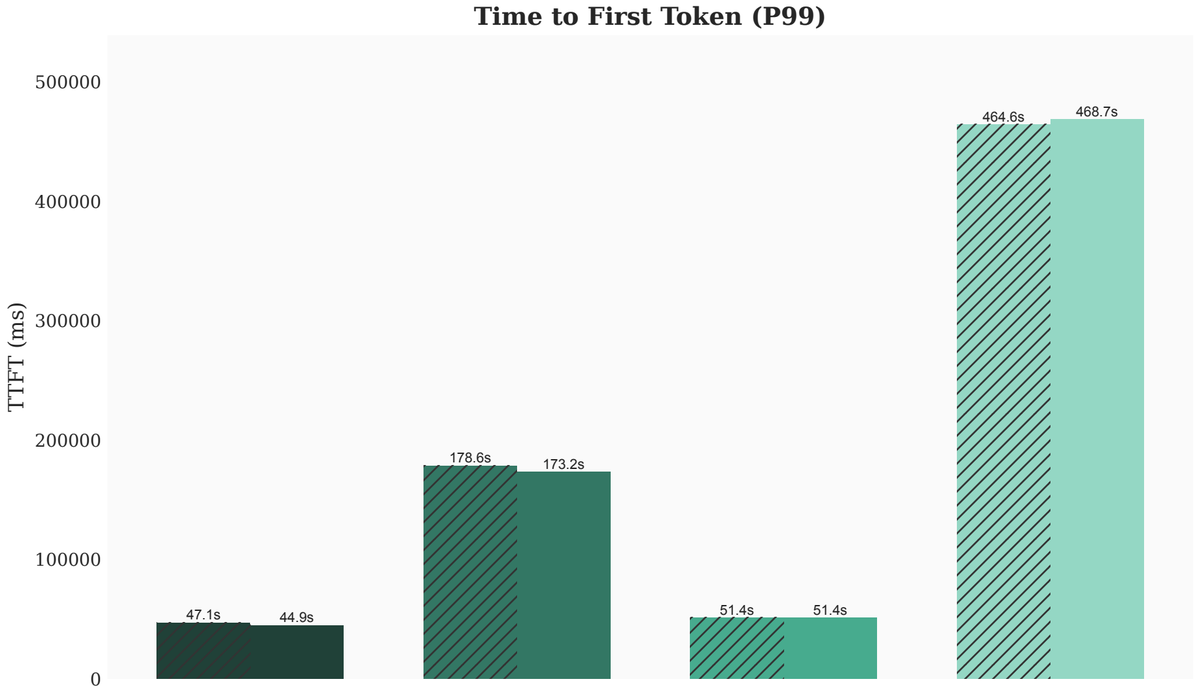
Random (1500 ⇒ 250) (Rate 1)




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
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) (Rate 50)

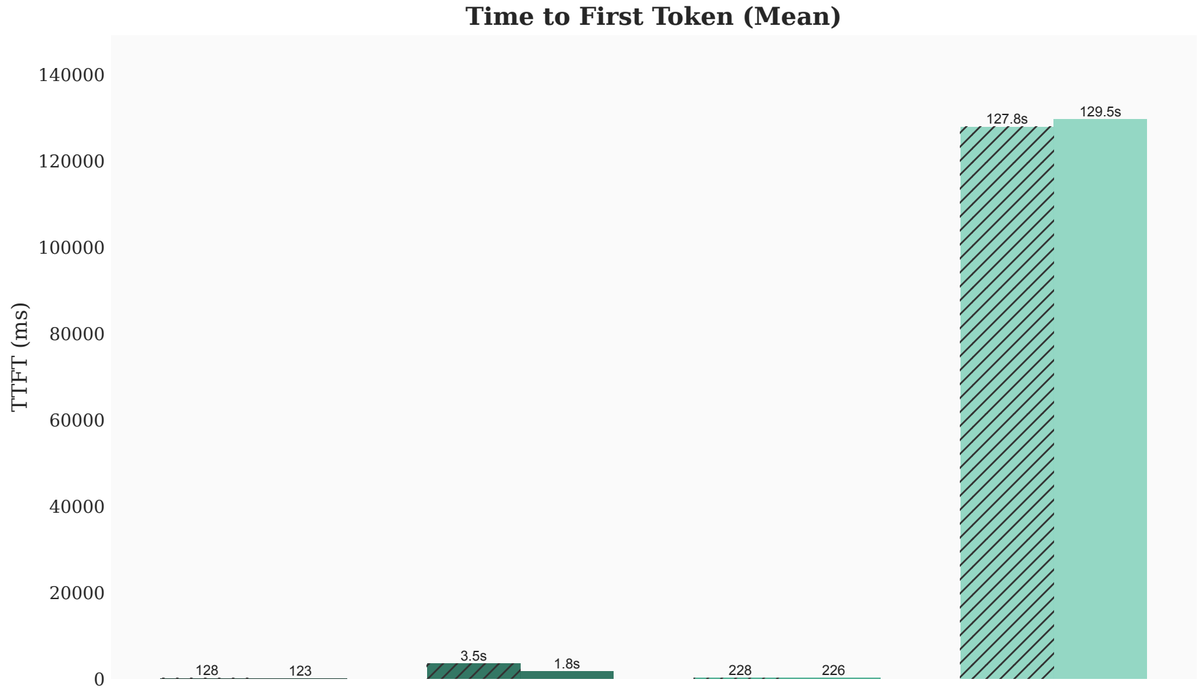




CC
No CC

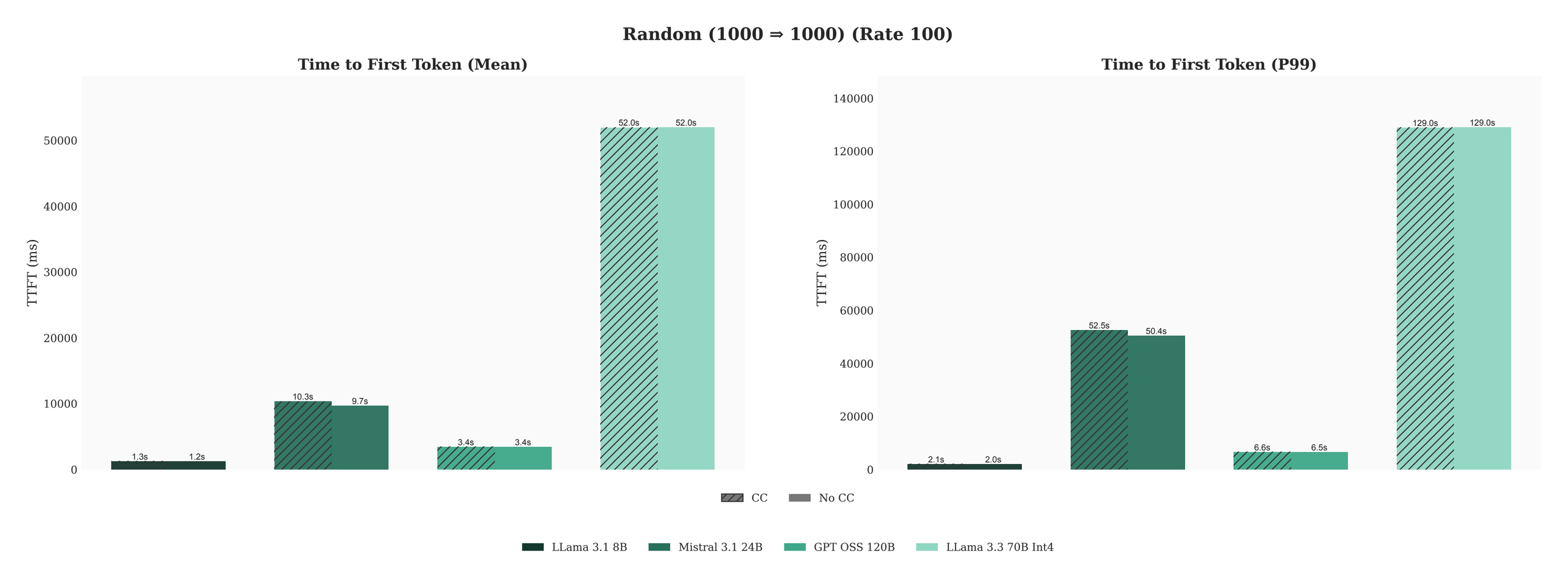
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

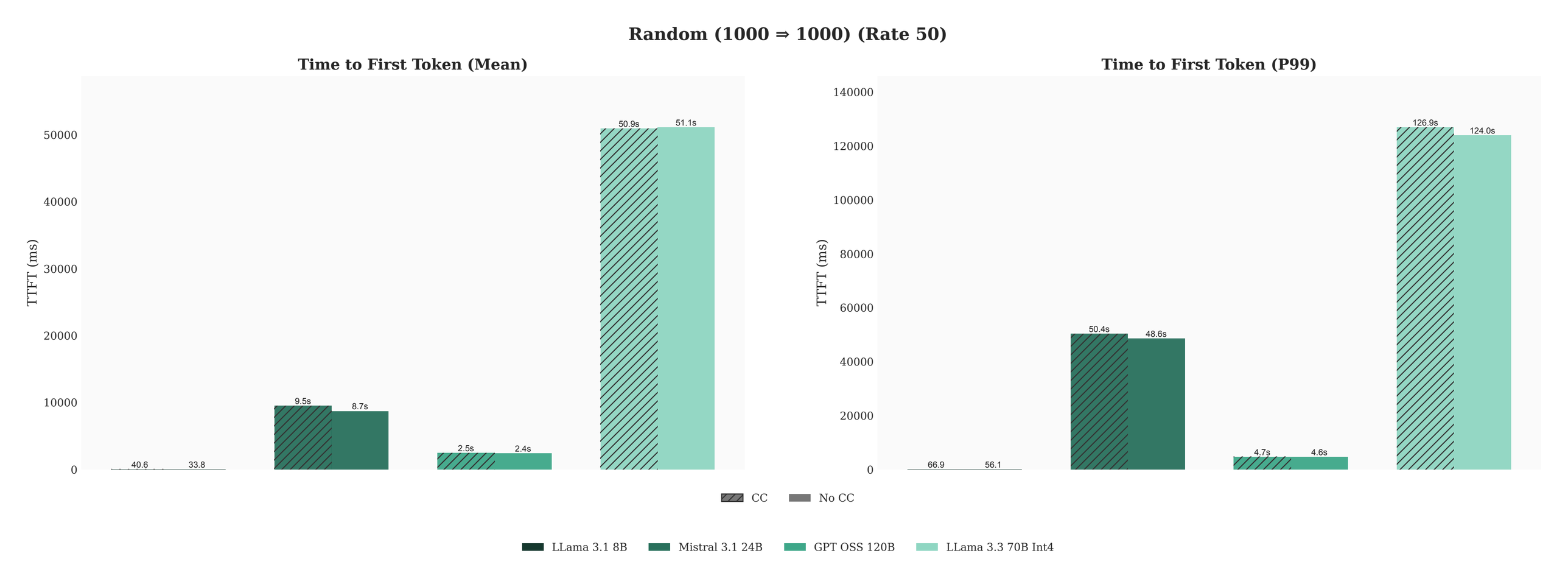
Random (4000 ⇒ 1000) (Rate 1)

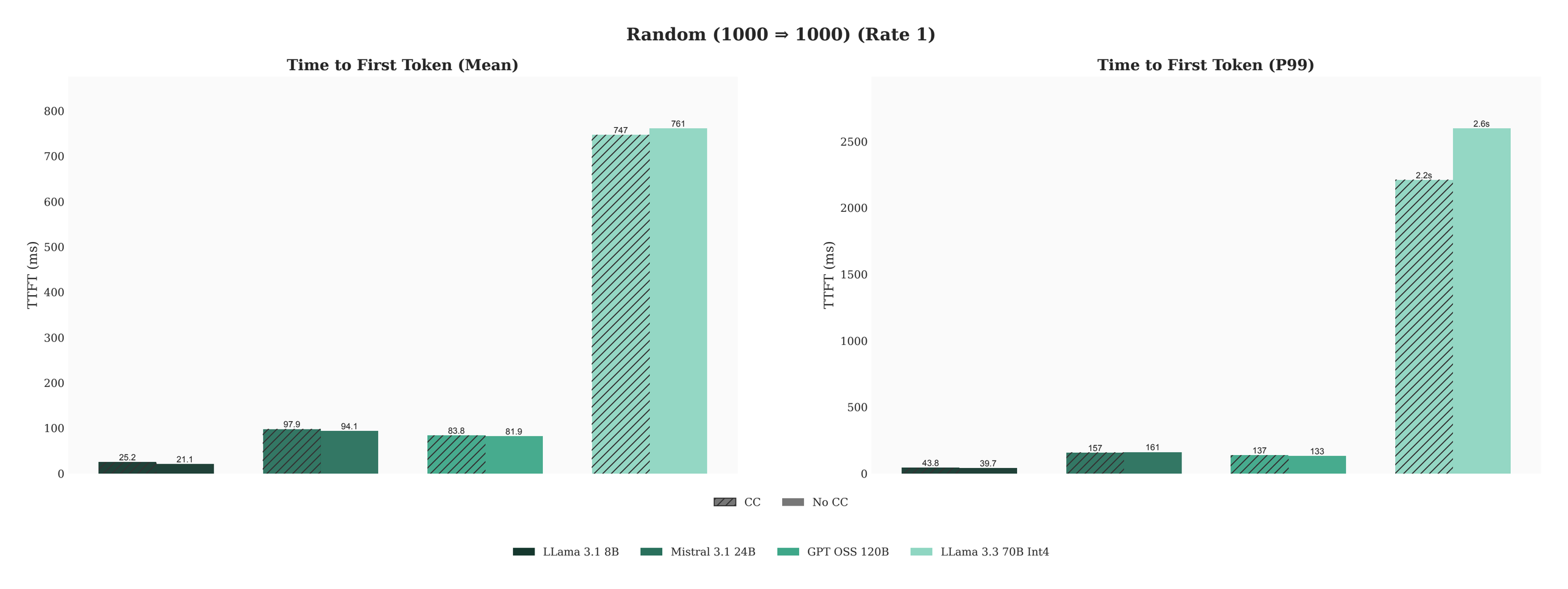


CC No CC

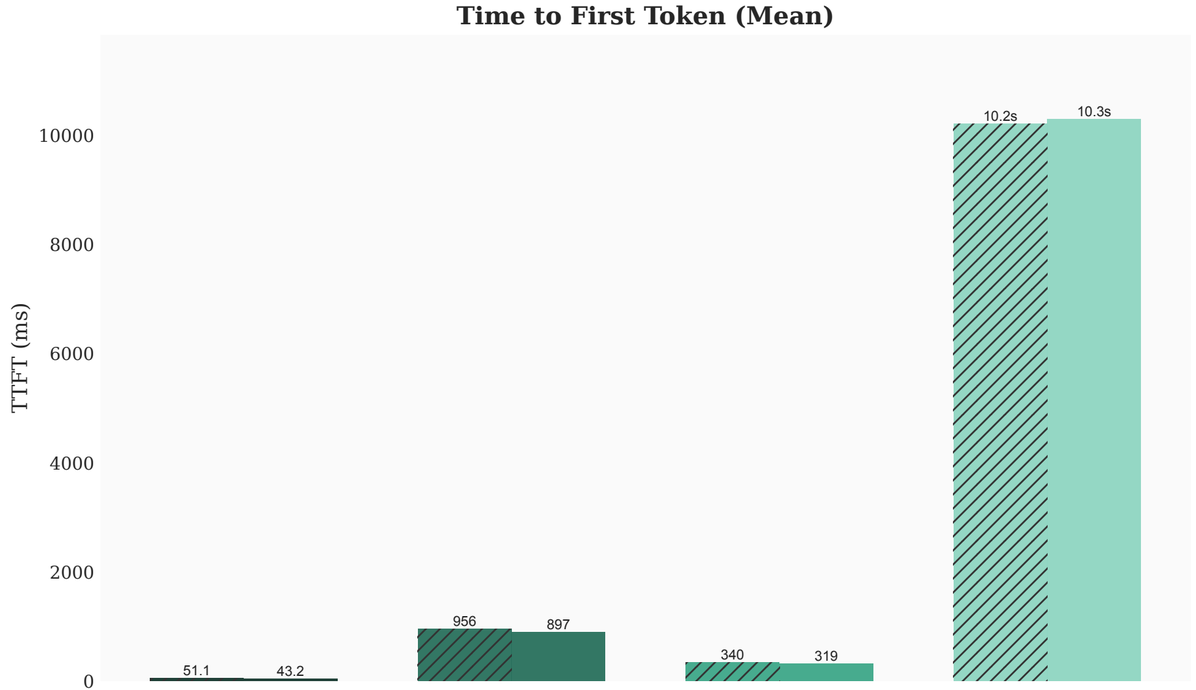
LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

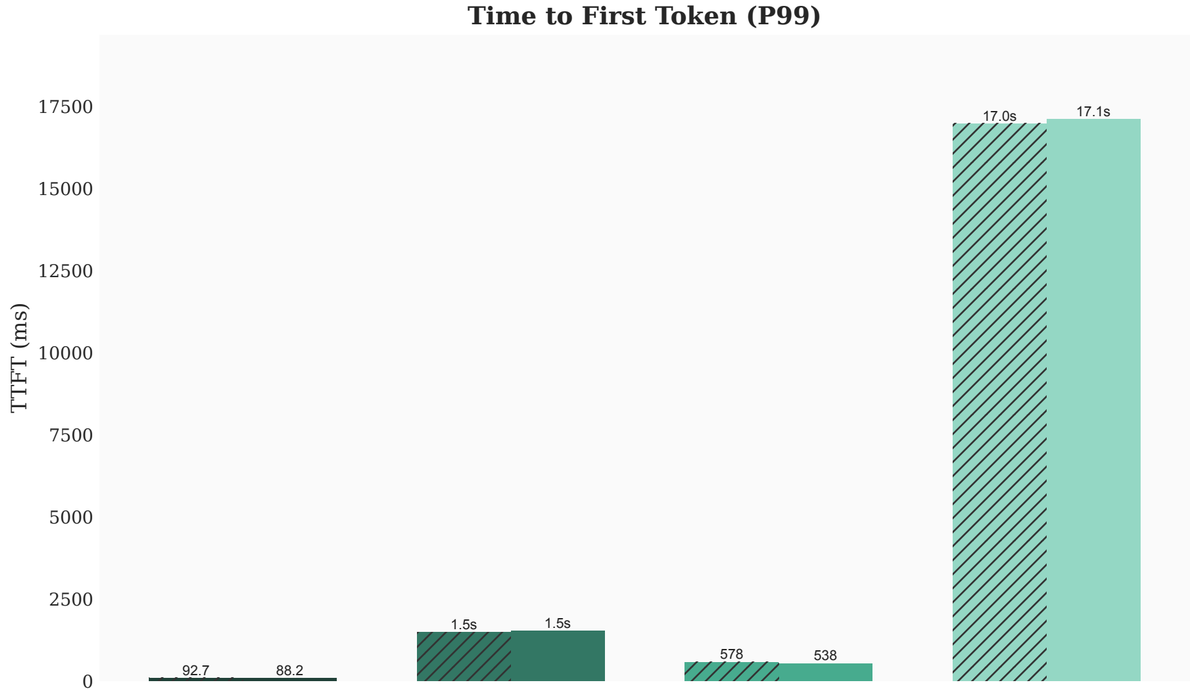

Random (1000 ⇒ 1000) (Rate 100)
Time to First Token (Mean)
TTFT (ms)
0
10000
20000
30000
40000
50000
1.3s
1.2s
10.3s
9.7s
3.4s
3.4s
52.0s
52.0s
Time to First Token (P99)
TTFT (ms)
0
20000
40000
60000
80000
100000
120000
140000
2.1s
2.0s
52.5s
50.4s
6.6s
6.5s
129.0s
129.0s
CC
No CC
LLama 3.1 8B
Mistral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

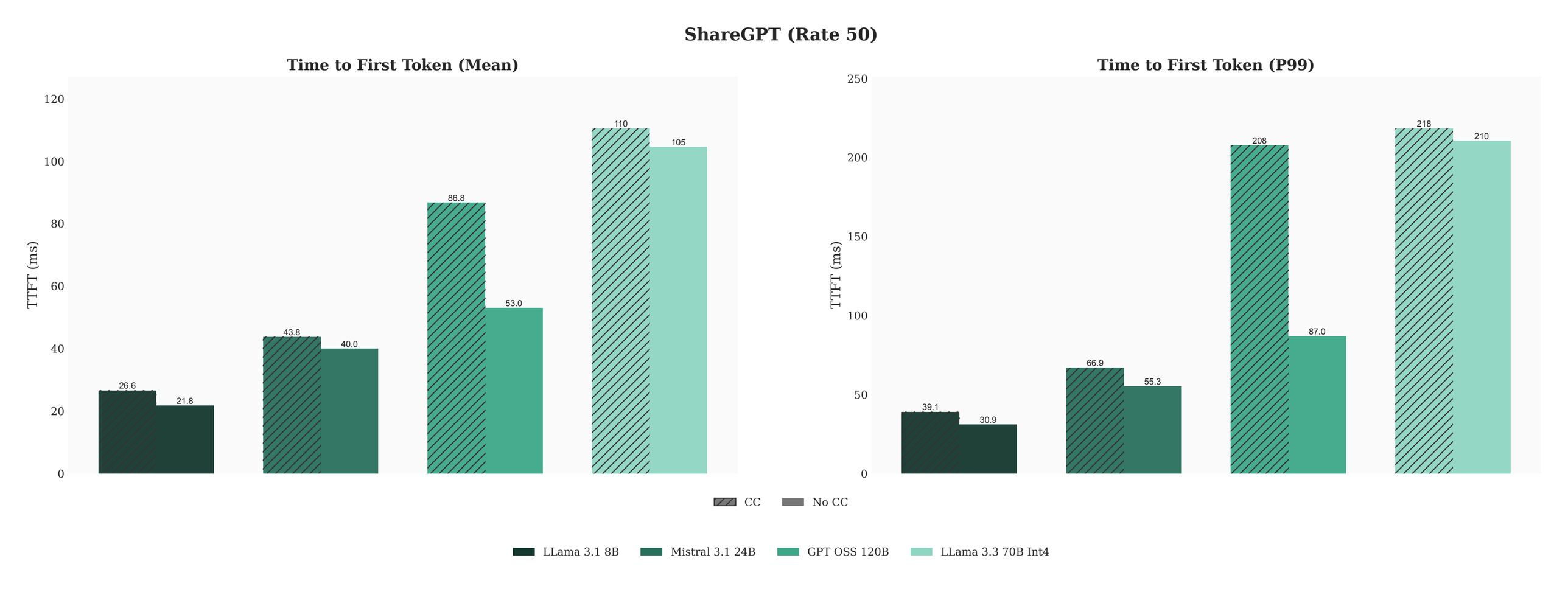
ShareGPT (Rate 100)

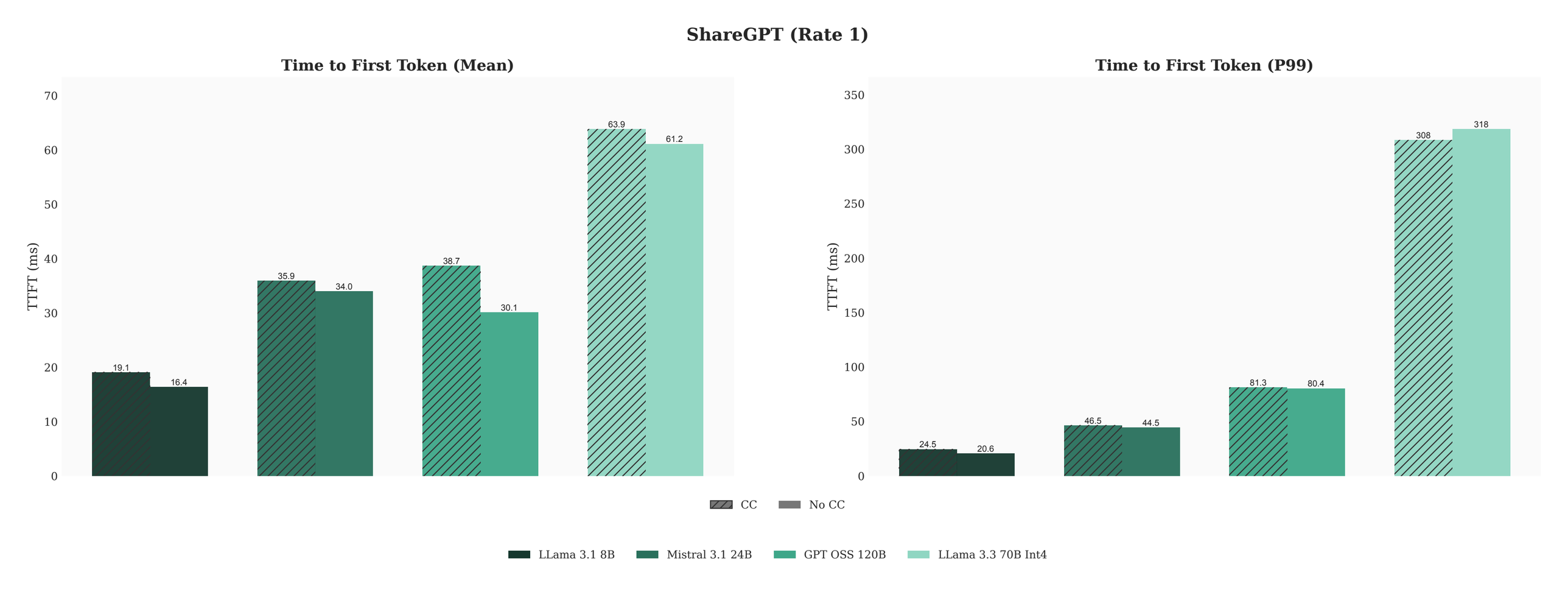


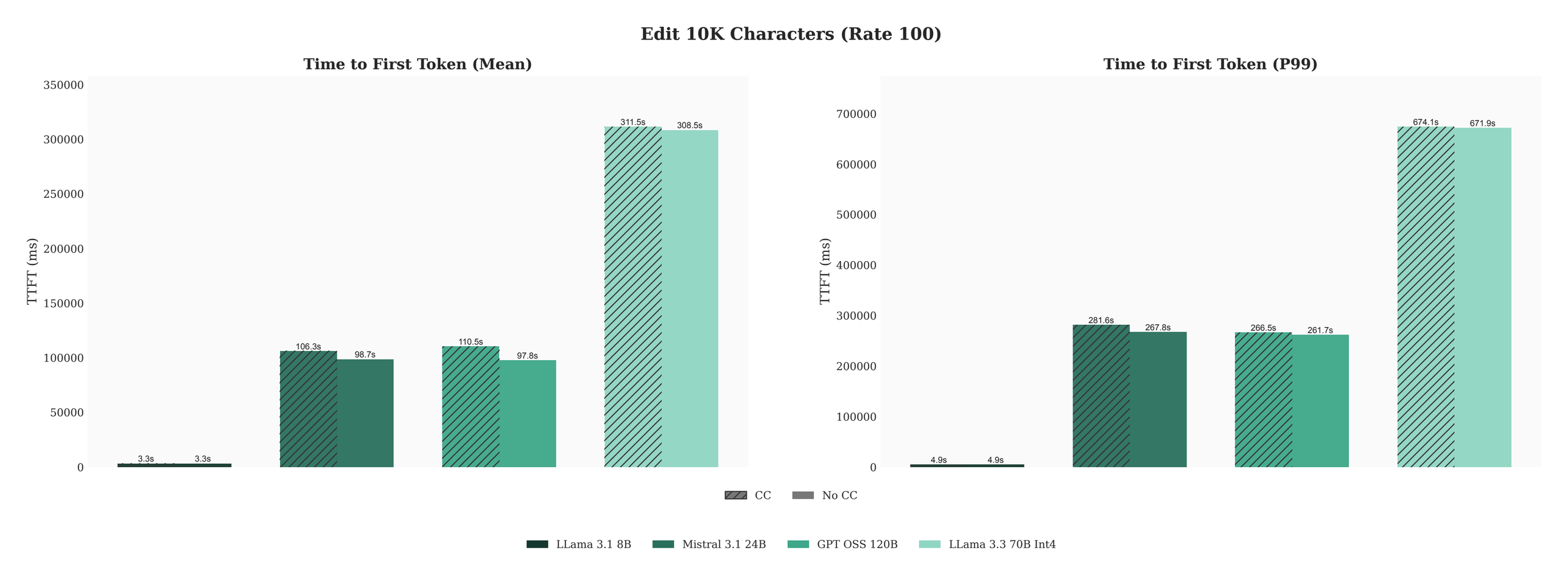


CC No CC

LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

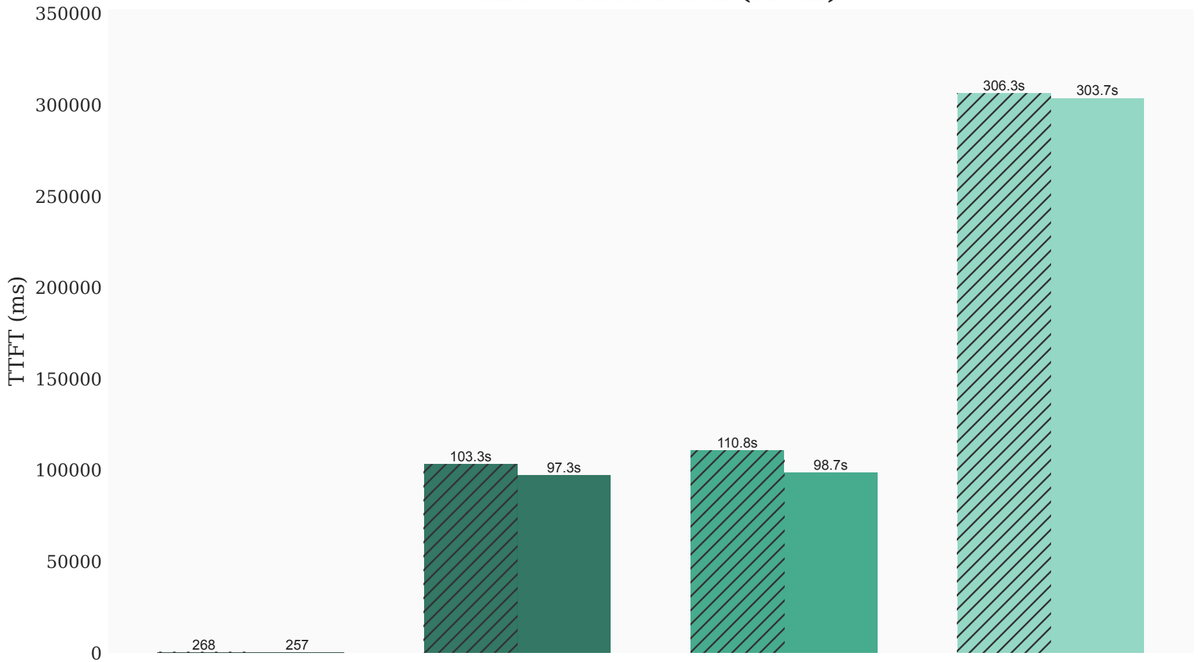

ShareGPT (Rate 50)
Time to First Token (Mean)
TTFT (ms)
0
20
40
60
80
100
120
26.6
21.8
43.8
40.0
86.8
53.0
110
105
Time to First Token (P99)
TTFT (ms)
0
50
100
150
200
250
39.1
30.9
66.9
55.3
208
87.0
218
210
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


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

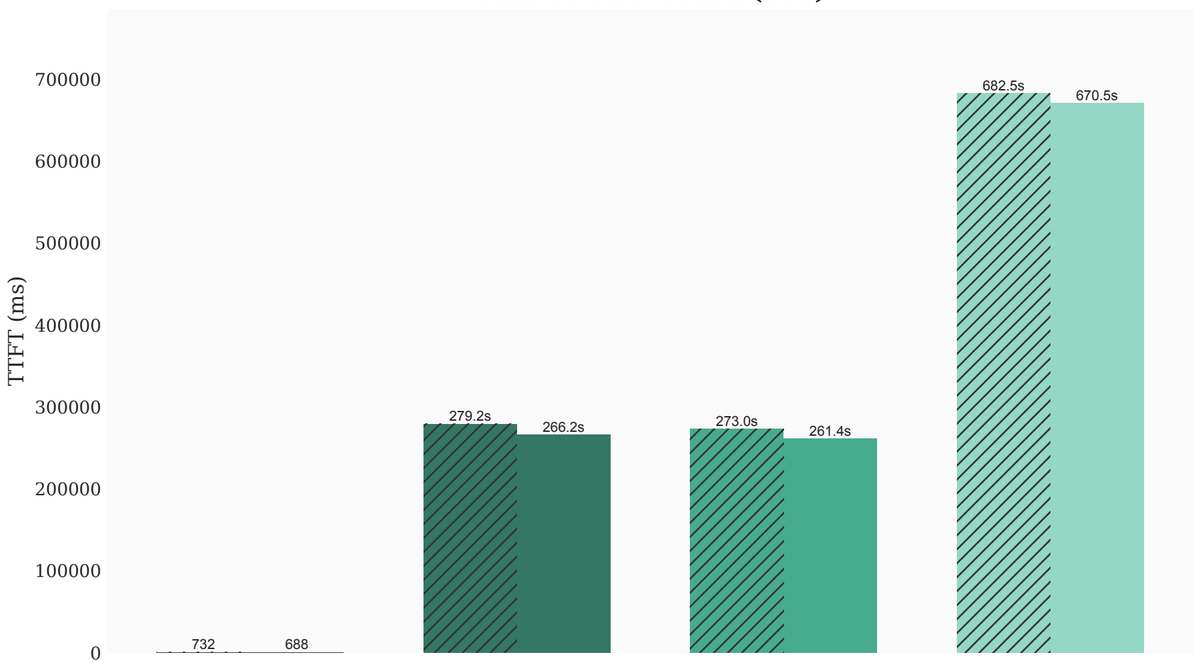

Edit 10K Characters (Rate 100)
Time to First Token (Mean)
Time to First Token (P99)
TTFT (ms)
350000
300000
250000
200000
150000
100000
50000
0
700000
600000
500000
400000
300000
200000
100000
3.3s
3.3s
106.3s
98.7s
110.5s
97.8s
311.5s
308.5s
4.9s
4.9s
281.6s
267.8s
266.5s
261.7s
674.1s
671.9s
CC
No 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)



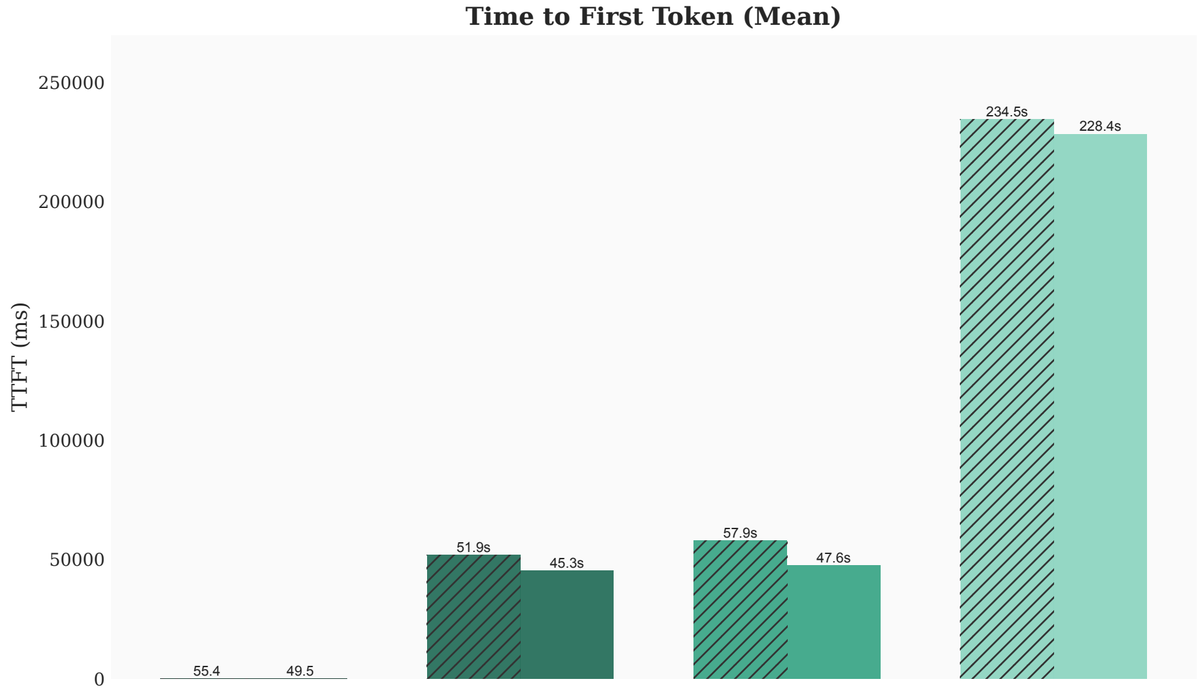
Time to First Token (P99)

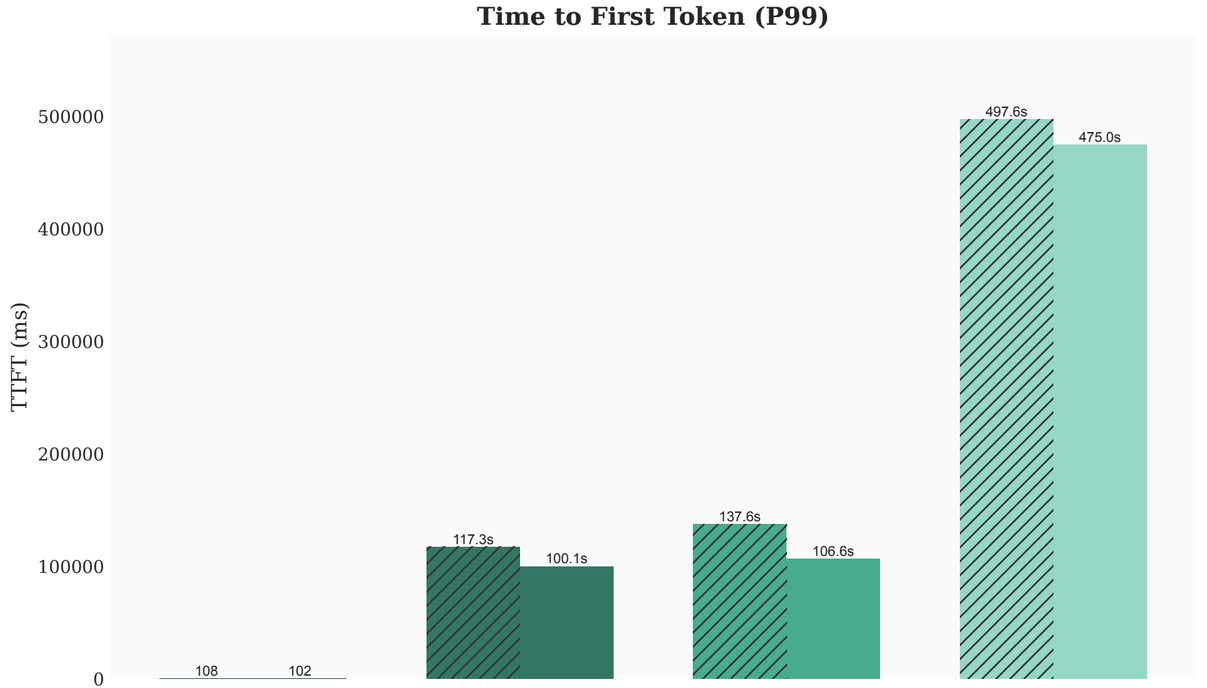


CC No CC

LLama 3.1 8B Mistral 3.1 24B GPT OSS 120B LLama 3.3 70B Int4

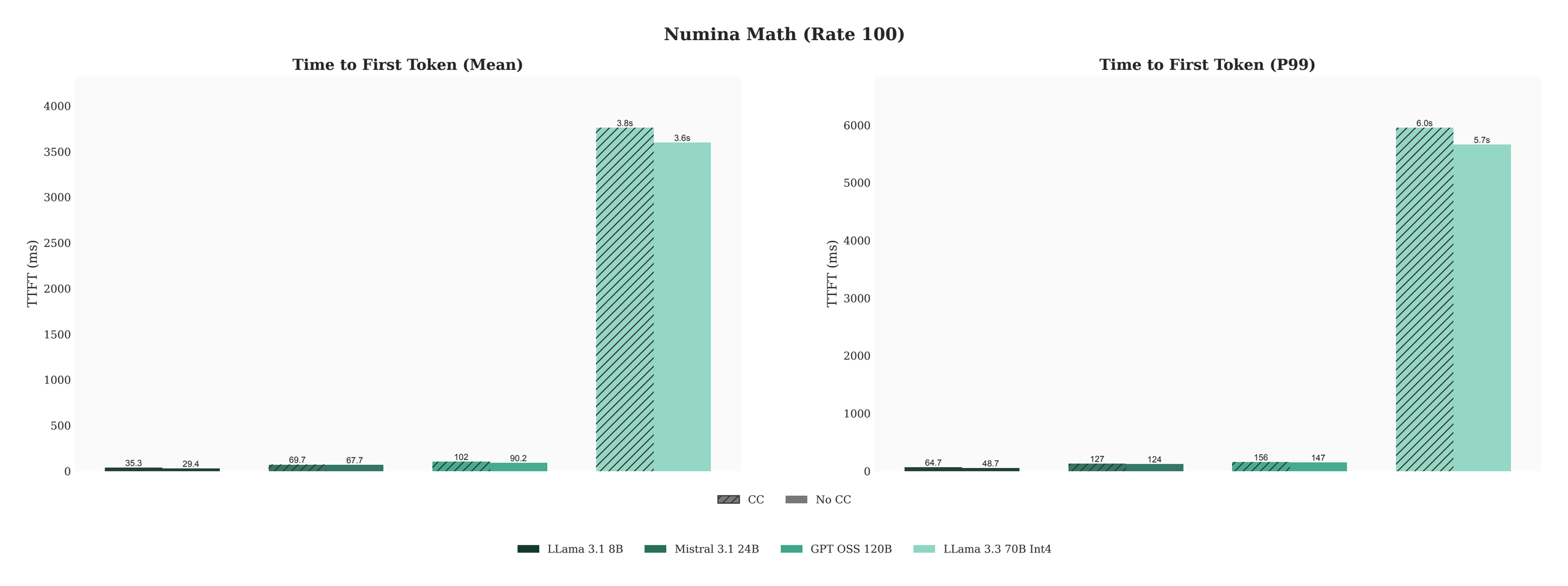
Edit 10K Characters (Rate 1)

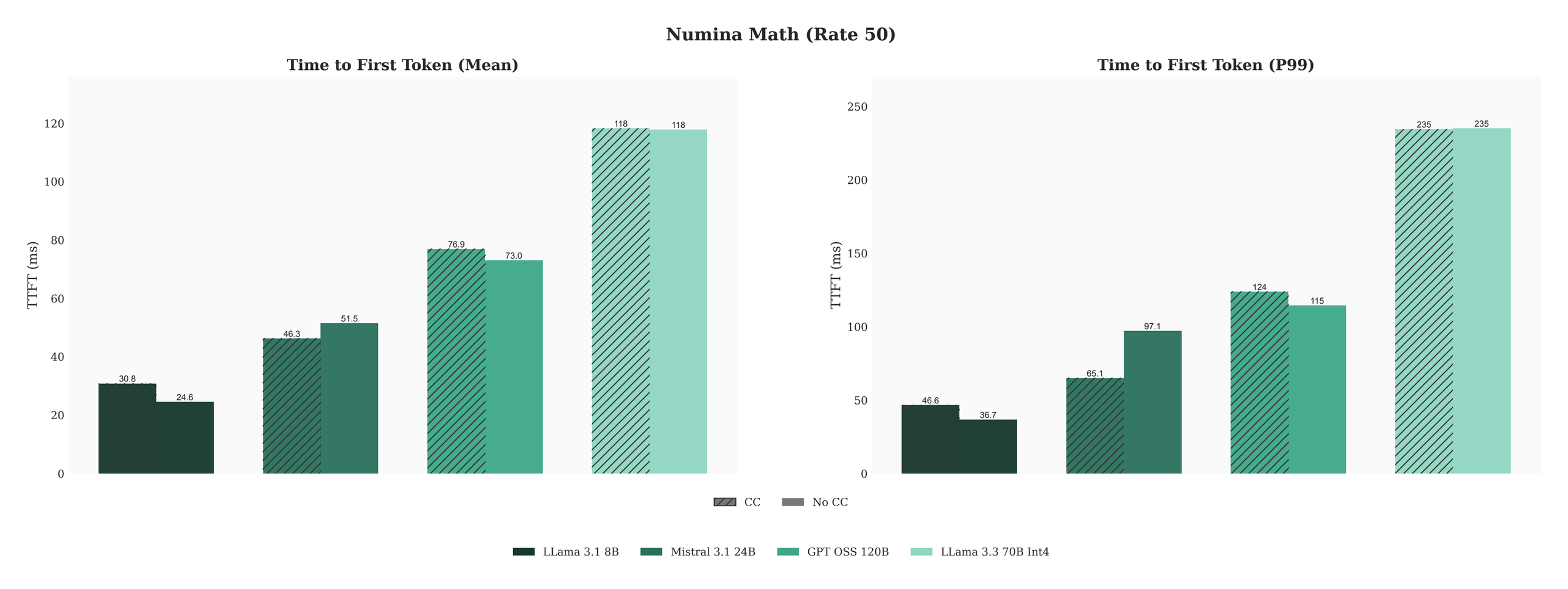




CC No CC

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)
4000
3500
3000
2500
2000
1500
1000
500
0
35.3
29.4
69.7
67.7
102
90.2
3.8s
3.6s
Time to First Token (P99)
TTFT (ms)
6000
5000
4000
3000
2000
1000
0
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 (Rate 50)
Time to First Token (Mean)
Time to First Token (P99)
TTFT (ms)
0
20
40
60
80
100
120
0
50
100
150
200
250
30.8
24.6
46.3
51.5
76.9
73.0
118
118
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 (Rate 1)