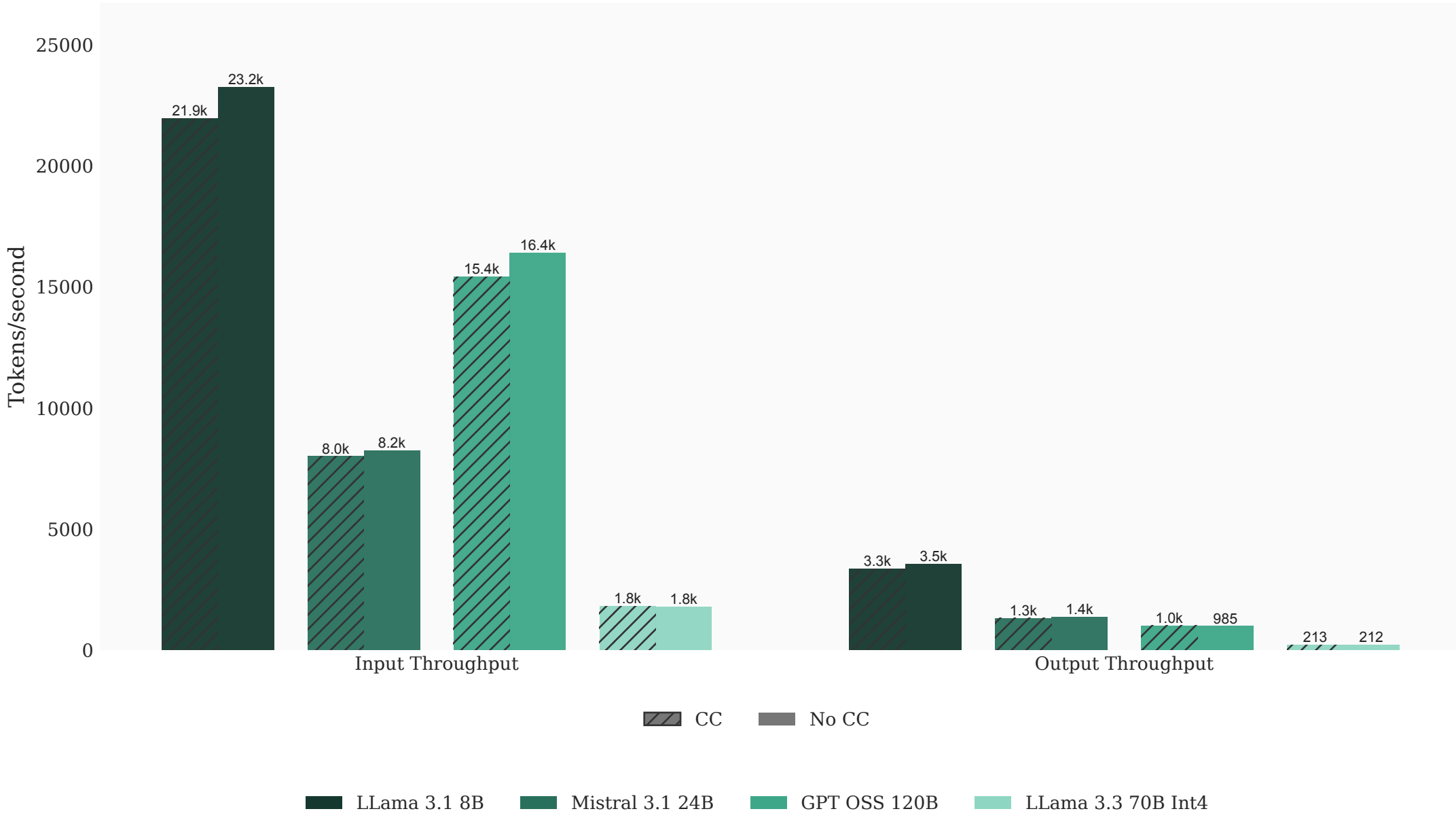
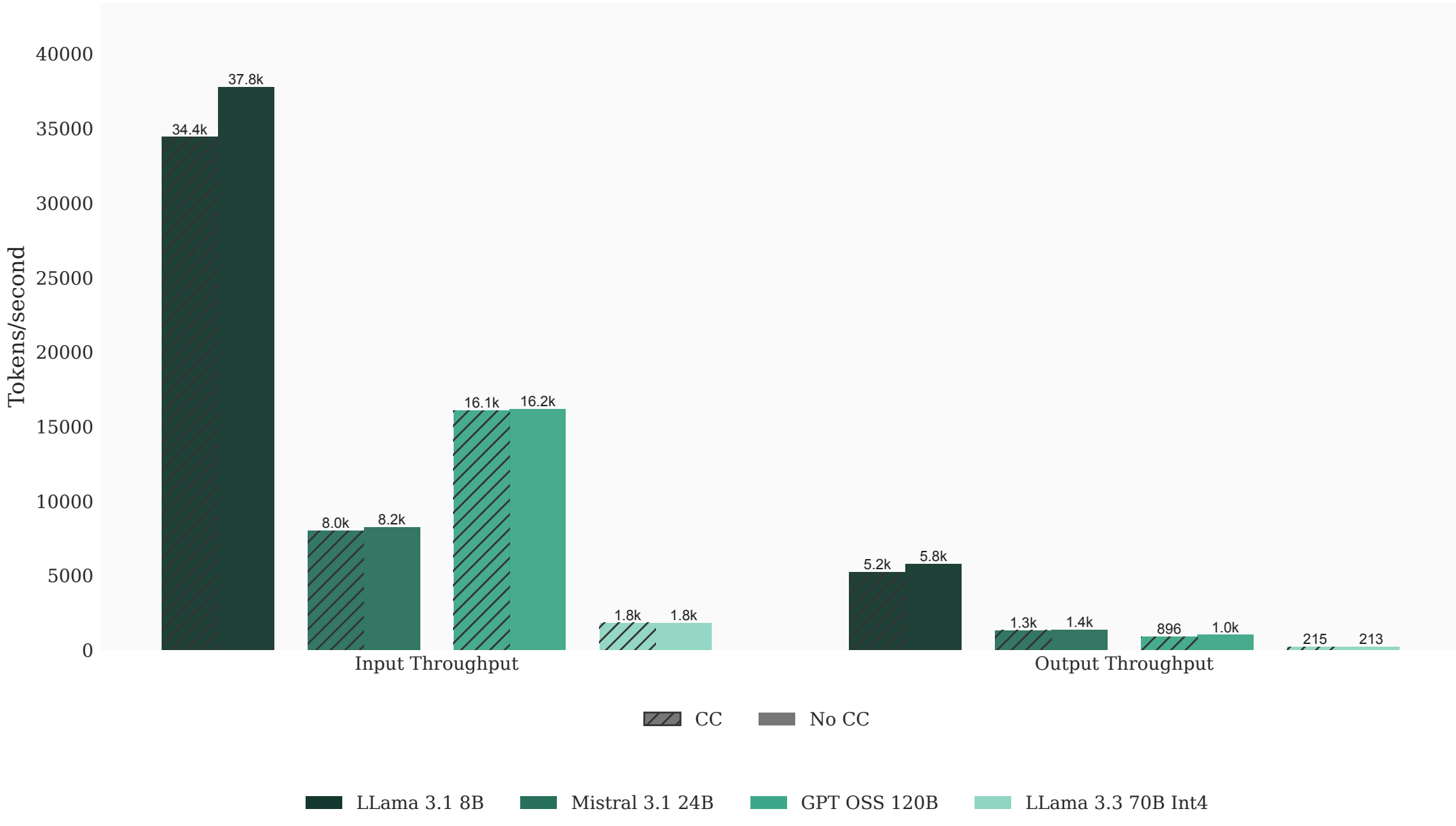
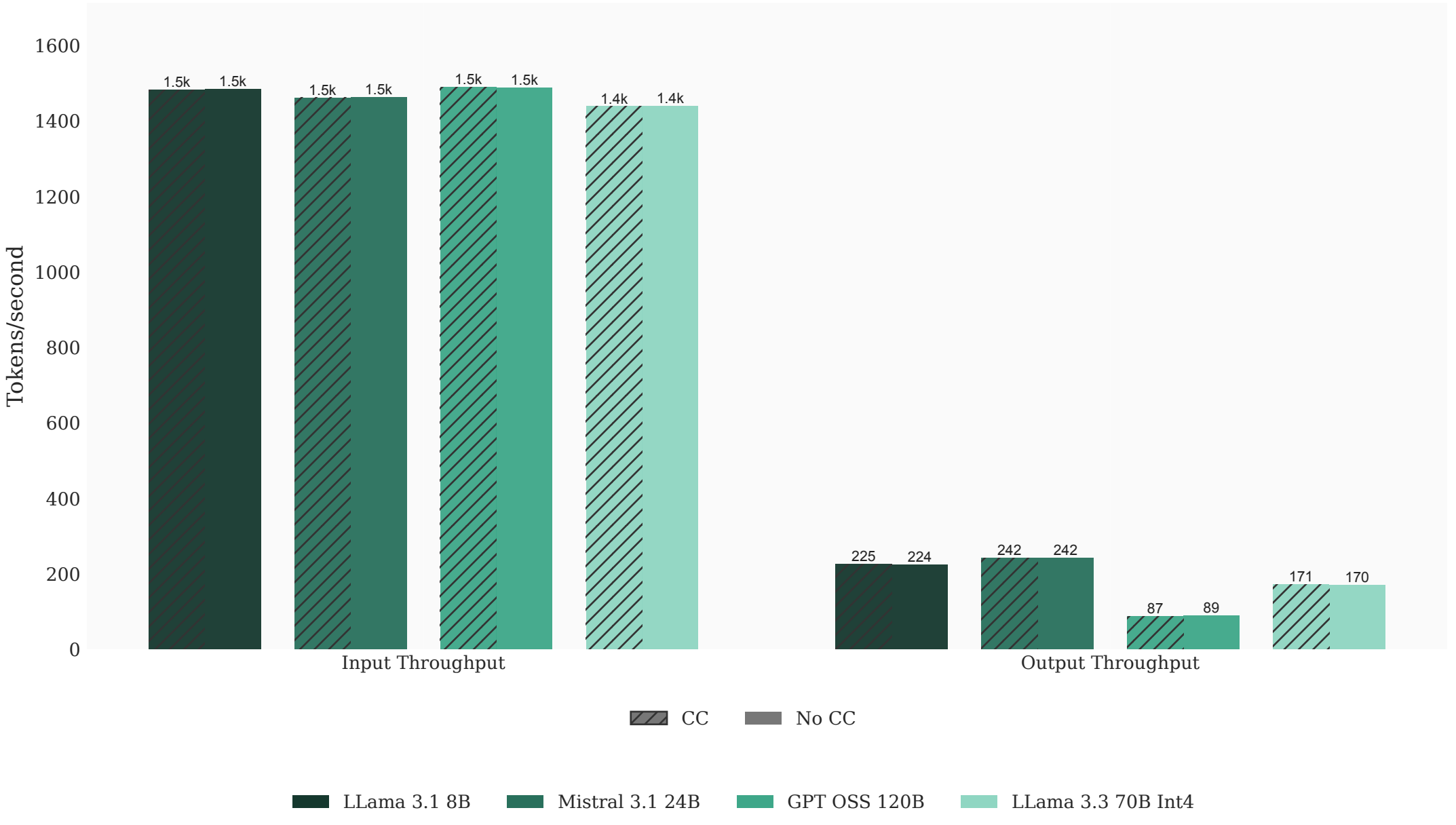
Random (1500 ⇒ 250) (100 Concurrent Requests)



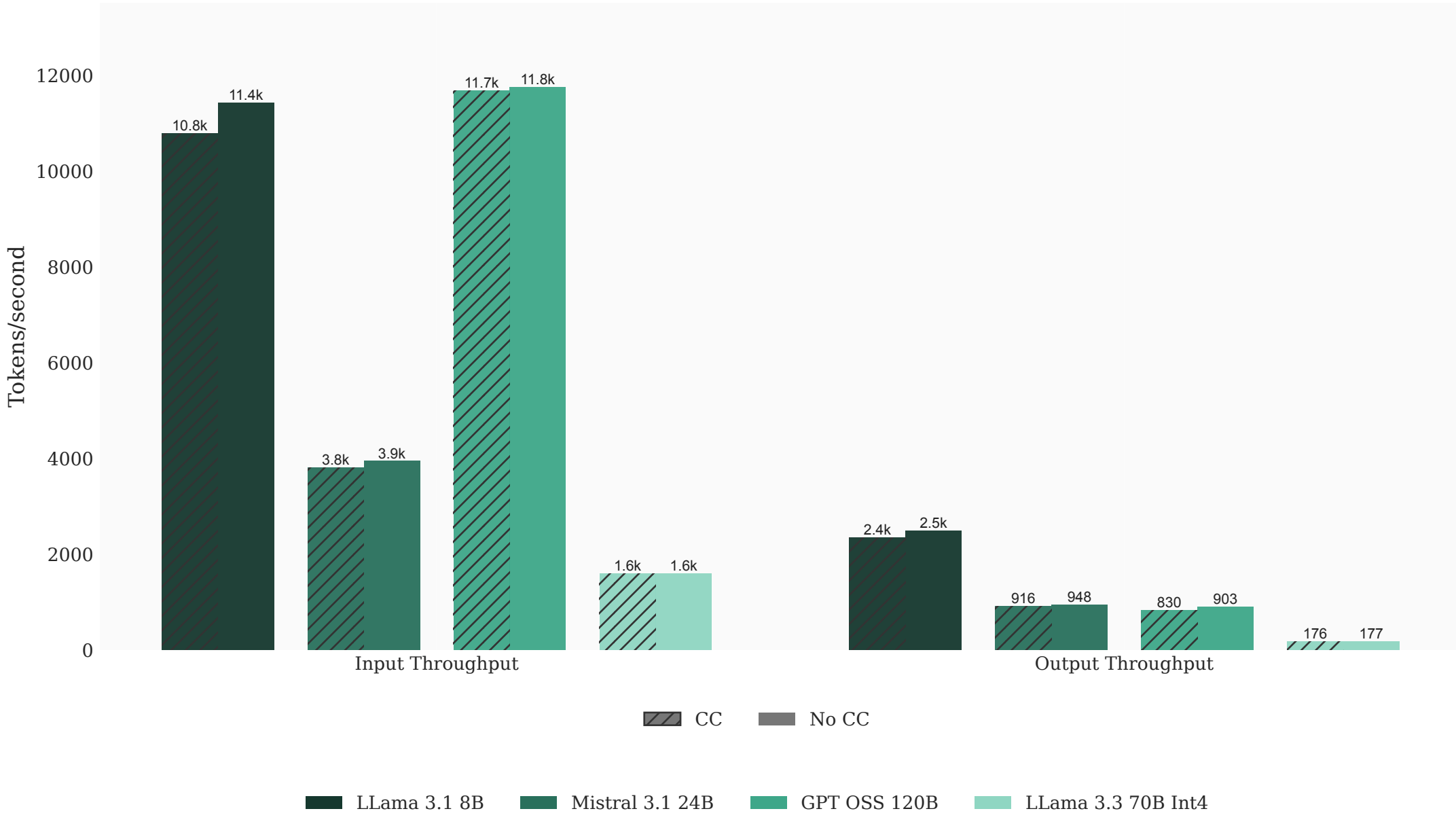
Random (1500 ⇒ 250) (50 Concurrent Requests)



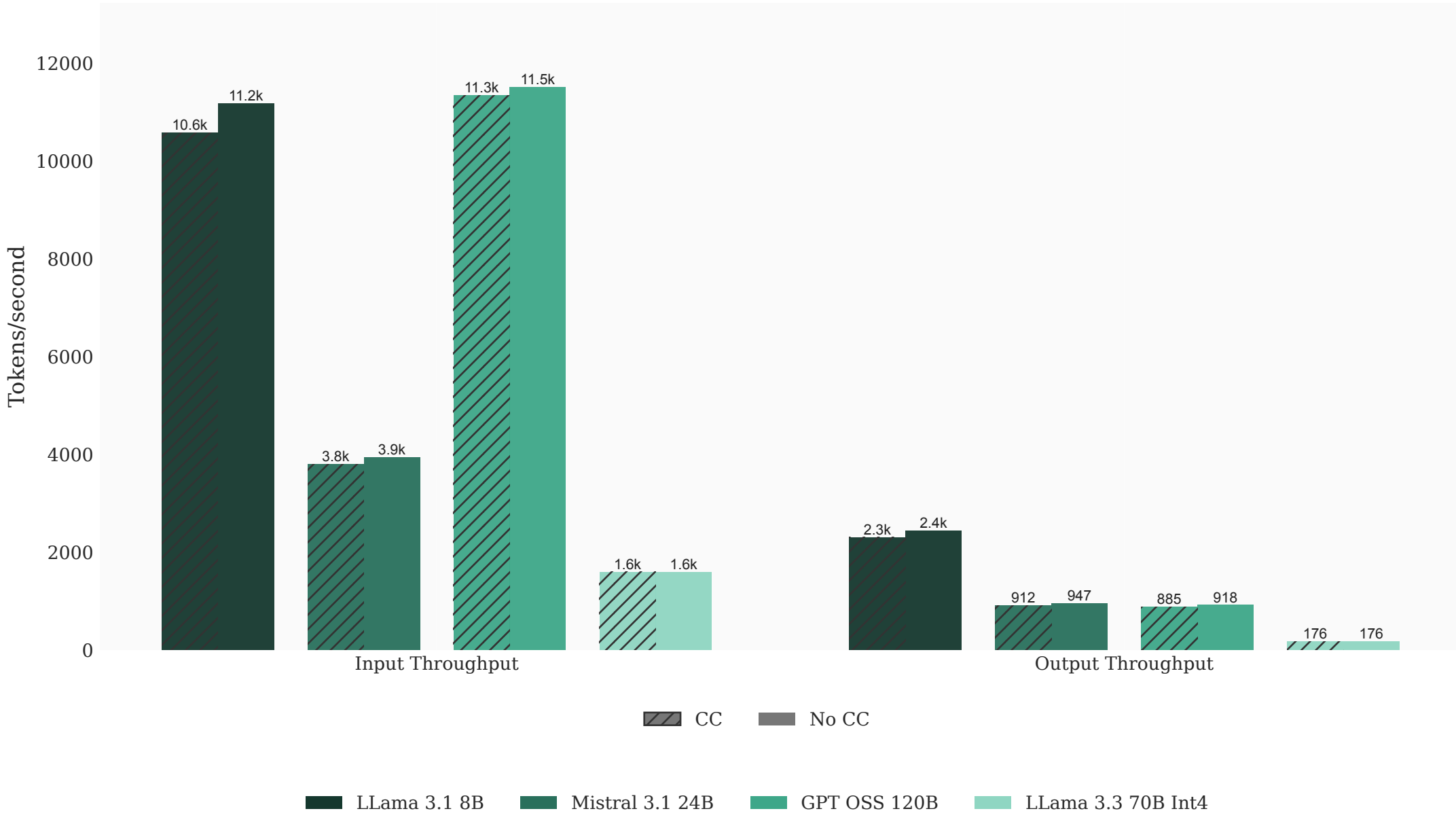
Random (1500 ⇒ 250) (1 Concurrent Requests)

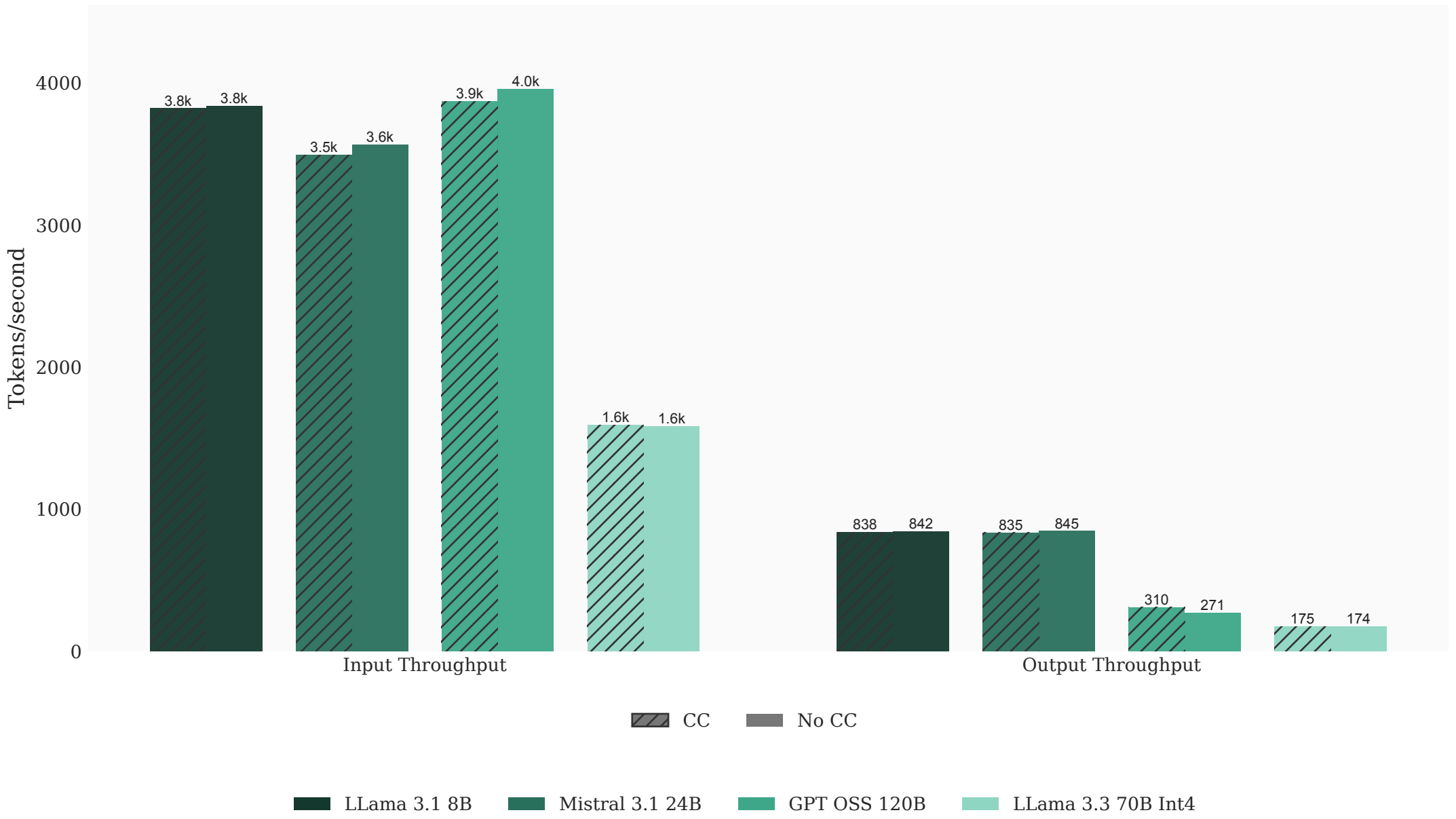


Random (4000 ⇒ 1000) (100 Concurrent Requests)

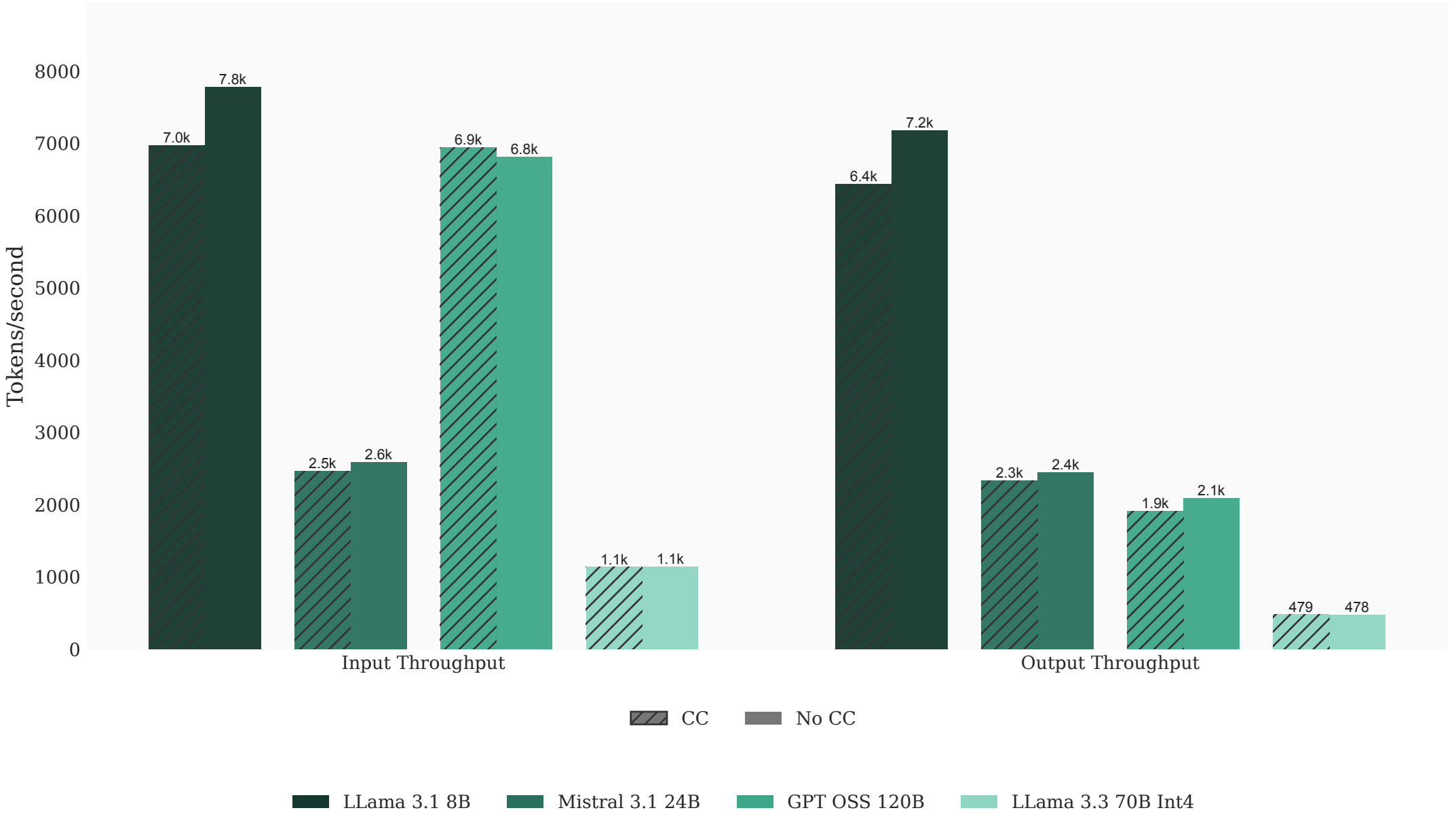


Random (4000 ⇒ 1000) (50 Concurrent Requests)

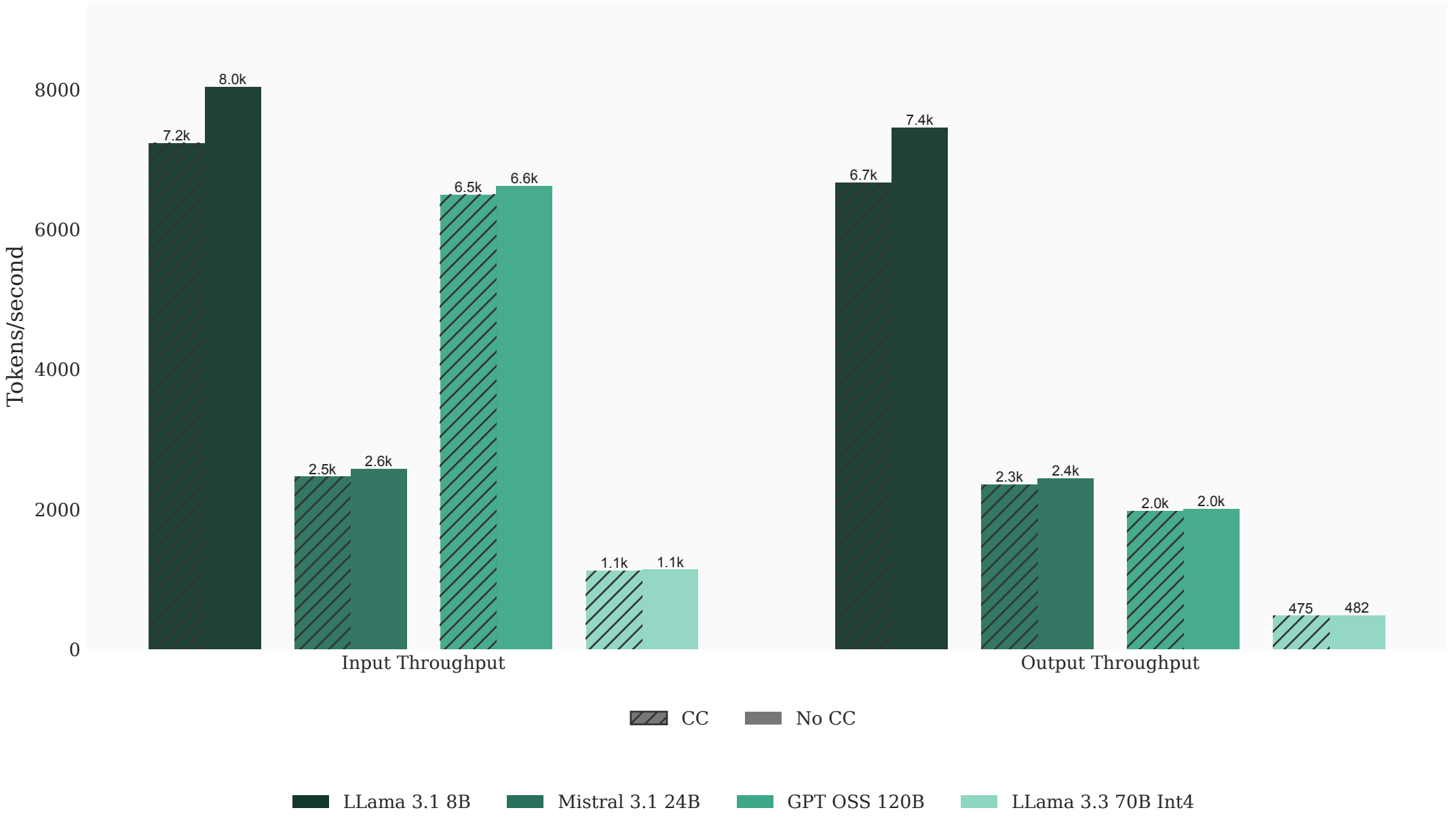



Random (4000 ⇒ 1000) (1 Concurrent Requests)
Tokens/second
4000
3000
2000
1000
0
3.8k
3.8k
3.5k
3.6k
3.9k
4.0k
1.6k
1.6k
838
842
835
845
310
271
175
174
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

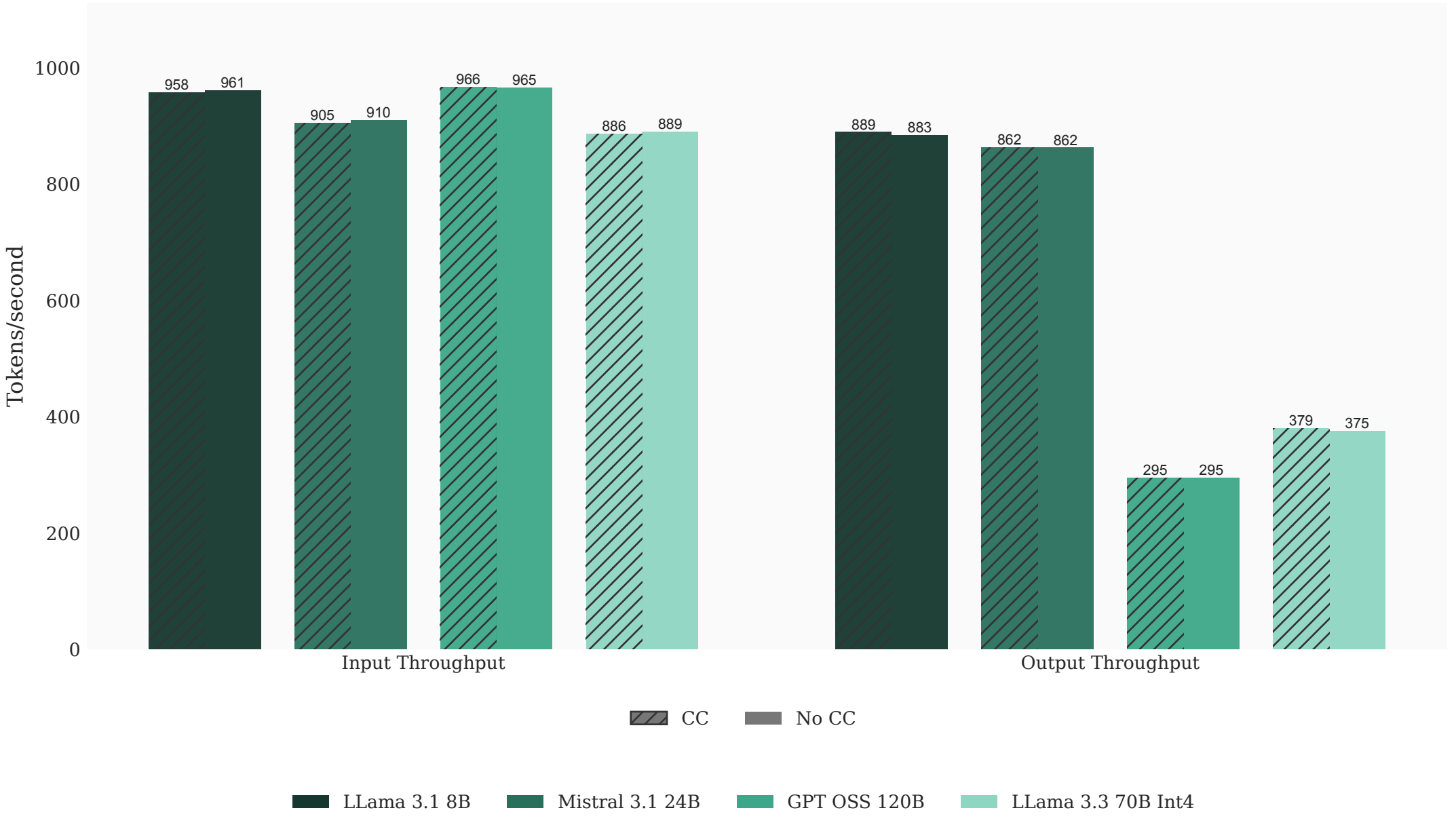
Random (1000 ⇒ 1000) (100 Concurrent Requests)

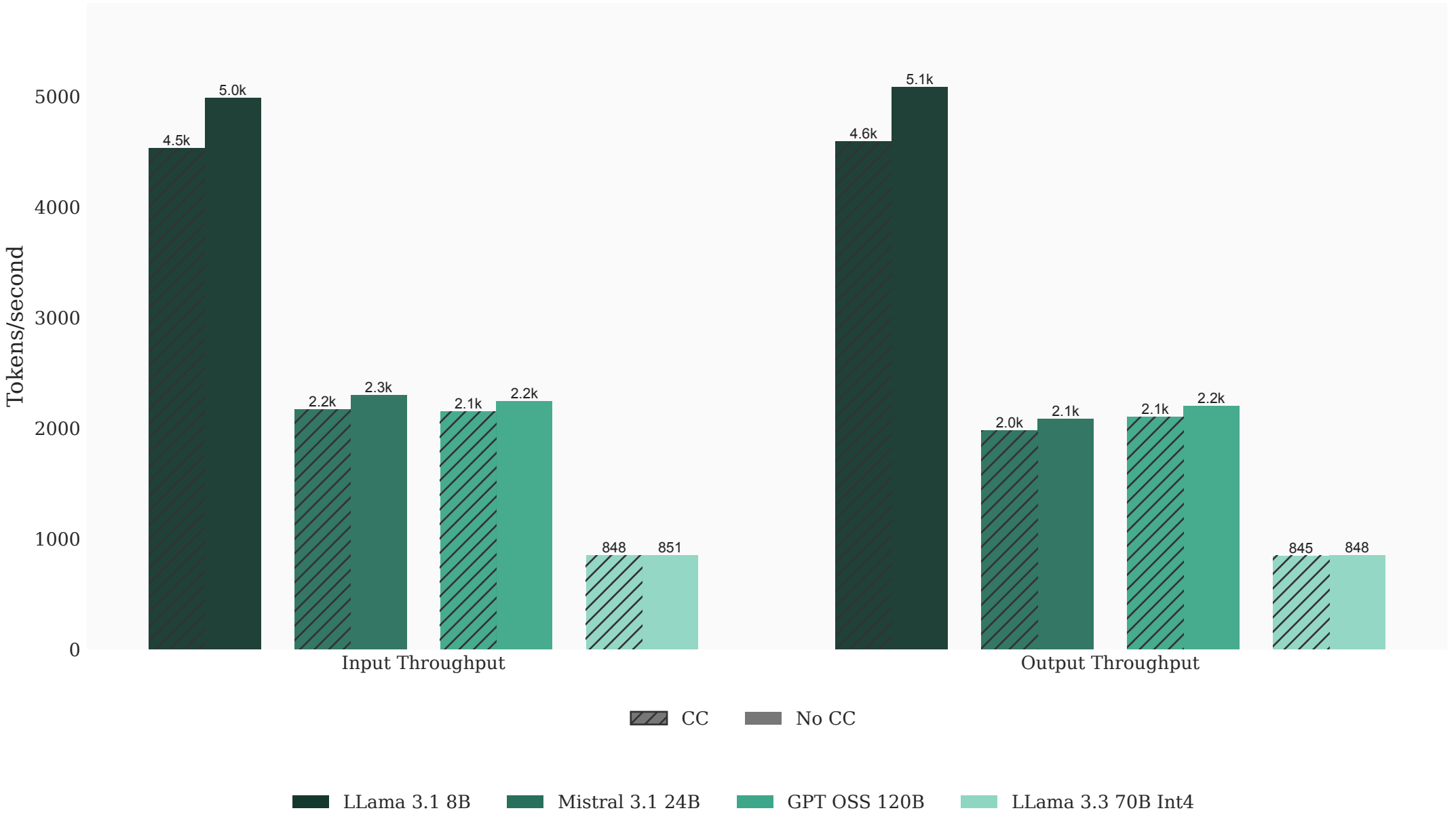


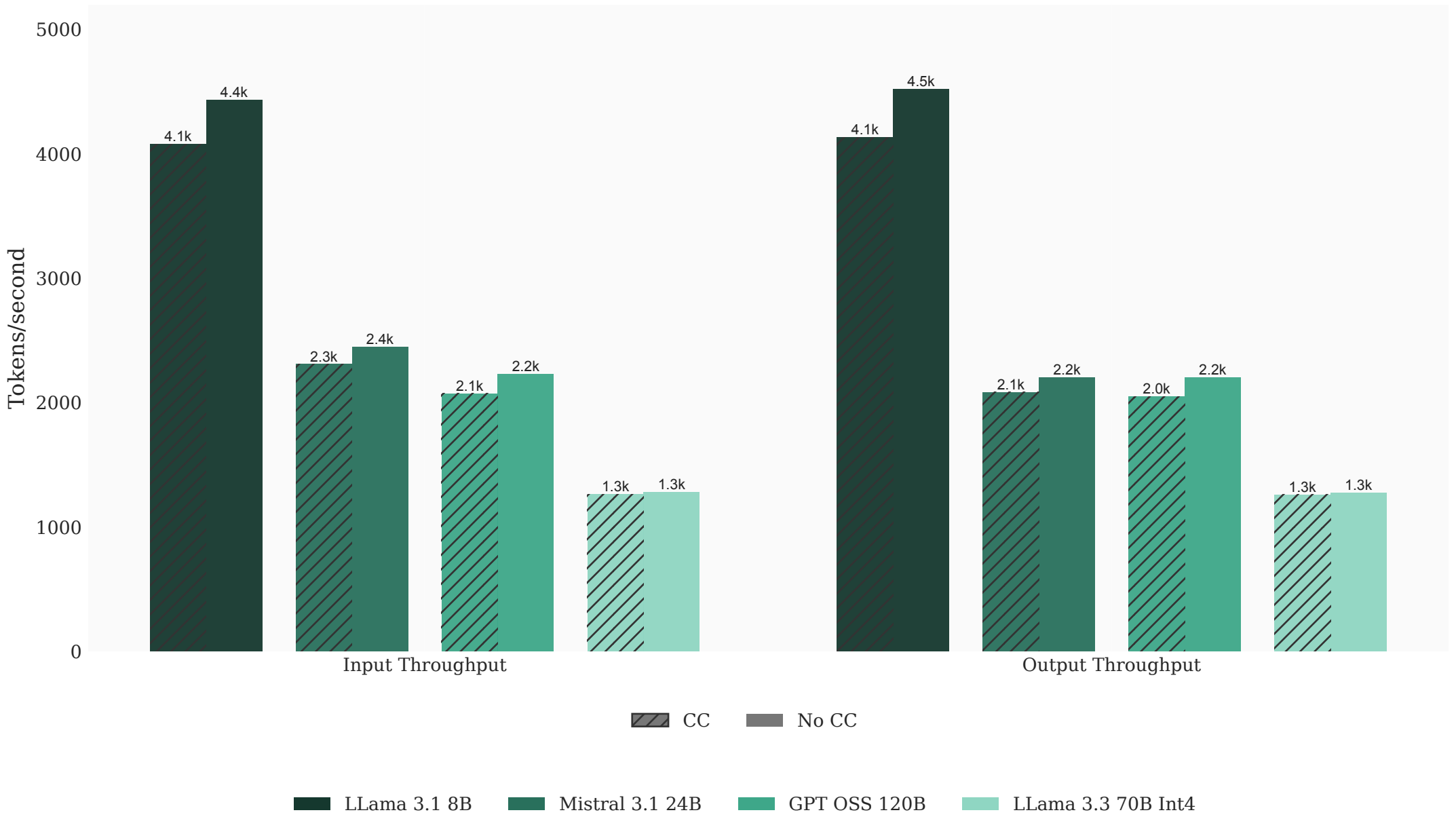
Random (1000 ⇒ 1000) (50 Concurrent Requests)

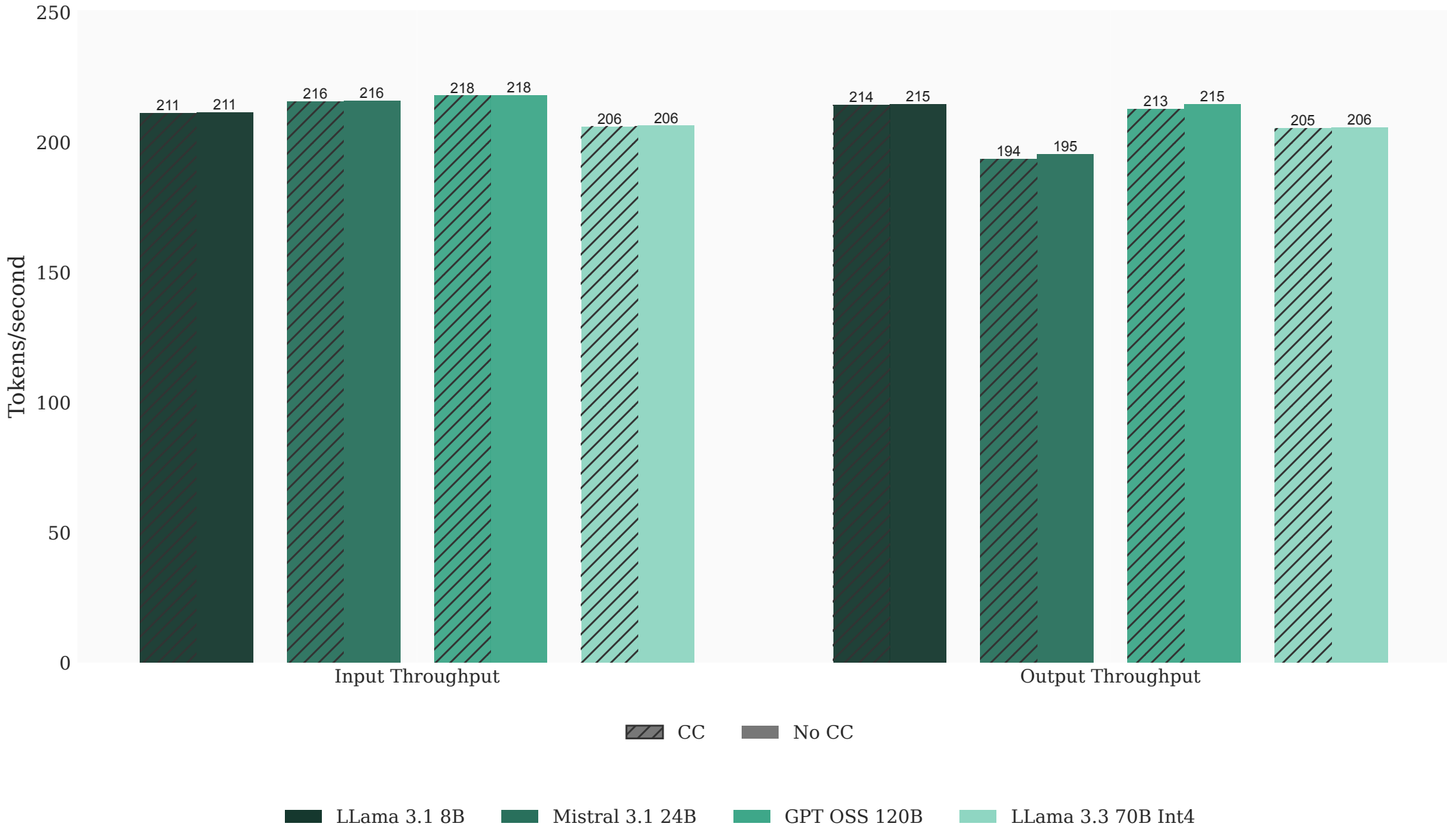


Random (1000 ⇒ 1000) (1 Concurrent Requests)

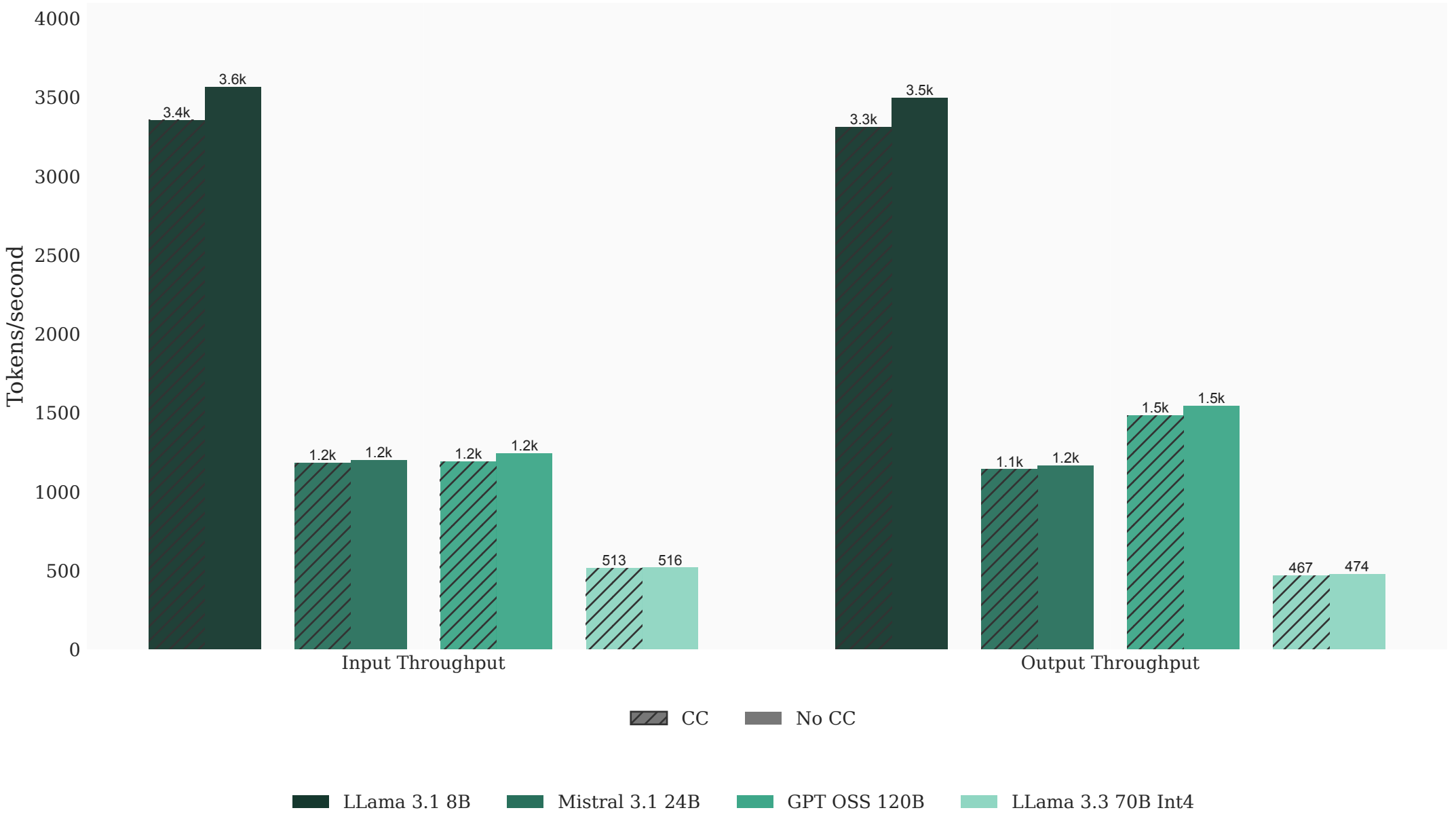



ShareGPT (100 Concurrent Requests)
Tokens/second
0
1000
2000
3000
4000
5000
4.5k
5.0k
2.2k
2.3k
2.1k
2.2k
848
851
4.6k
5.1k
2.0k
2.1k
2.1k
2.2k
845
848
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

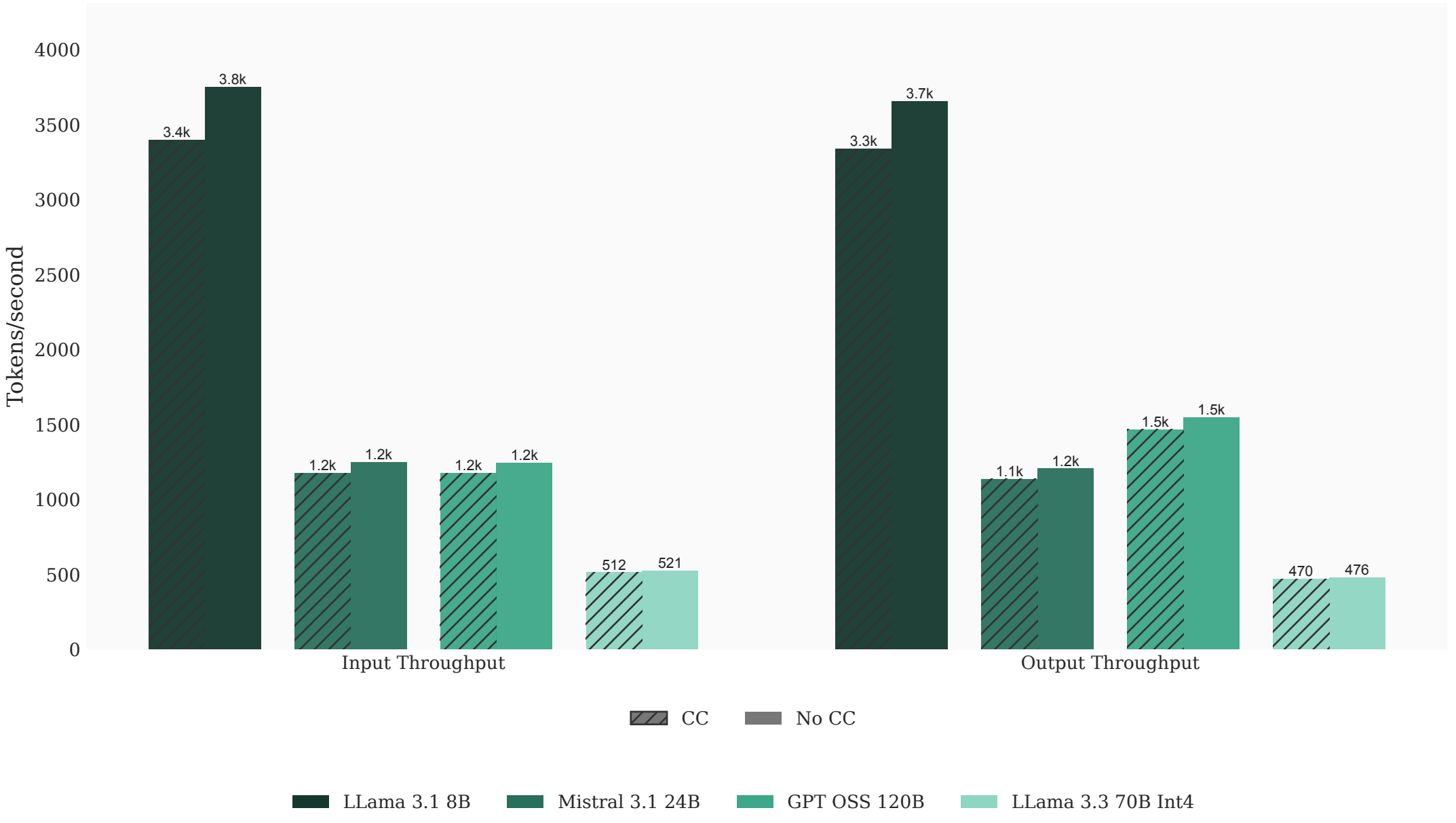

ShareGPT (50 Concurrent Requests)
Tokens/second
5000
4000
3000
2000
1000
0
4.1k
4.4k
2.3k
2.4k
2.1k
2.2k
1.3k
1.3k
4.1k
4.5k
2.1k
2.2k
2.0k
2.2k
1.3k
1.3k
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


ShareGPT (1 Concurrent Requests)
Tokens/second
250
200
150
100
50
0
211
211
216
216
218
218
206
206
214
215
194
195
213
215
205
206
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

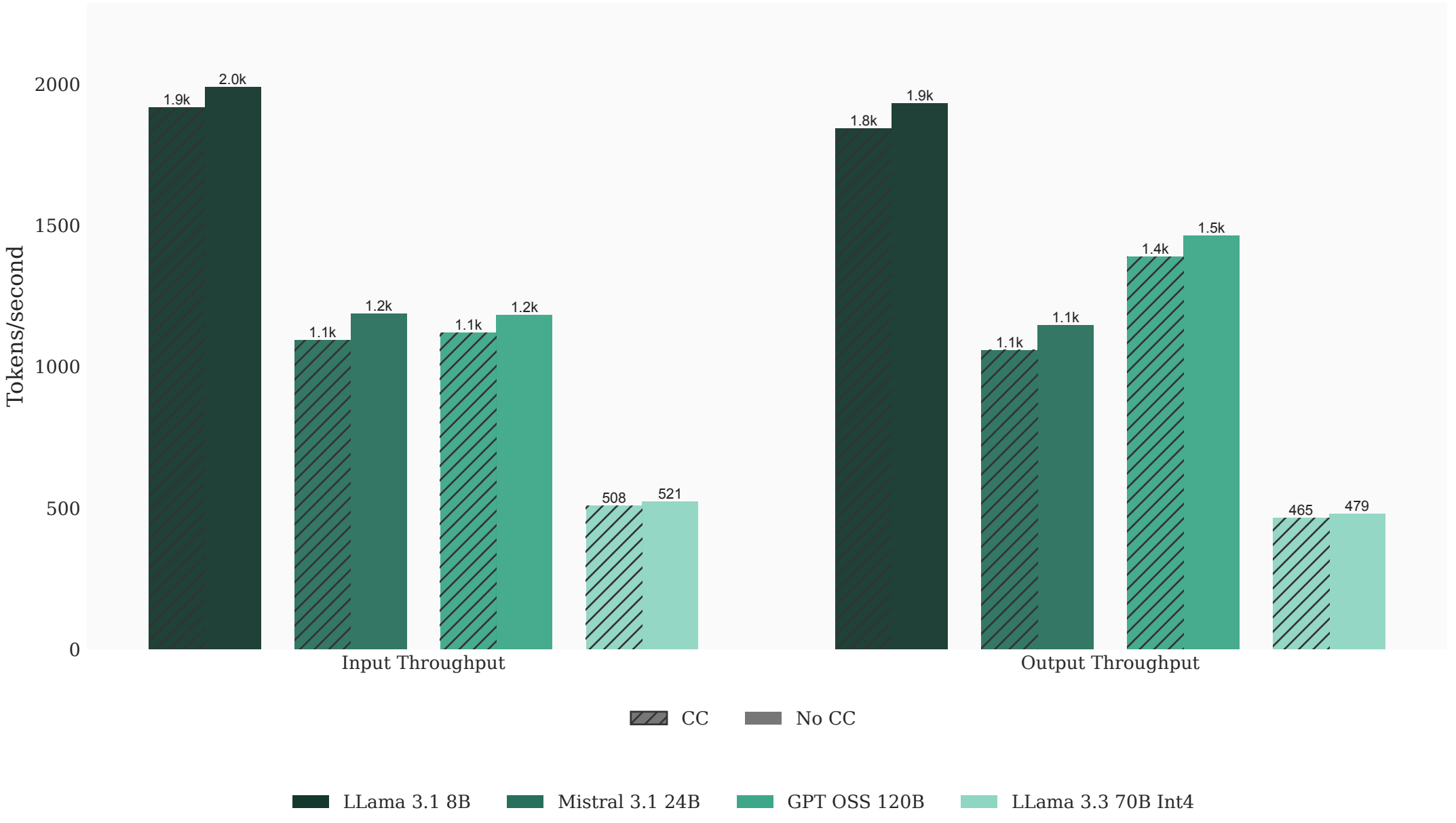
Edit 10K Characters (100 Concurrent Requests)



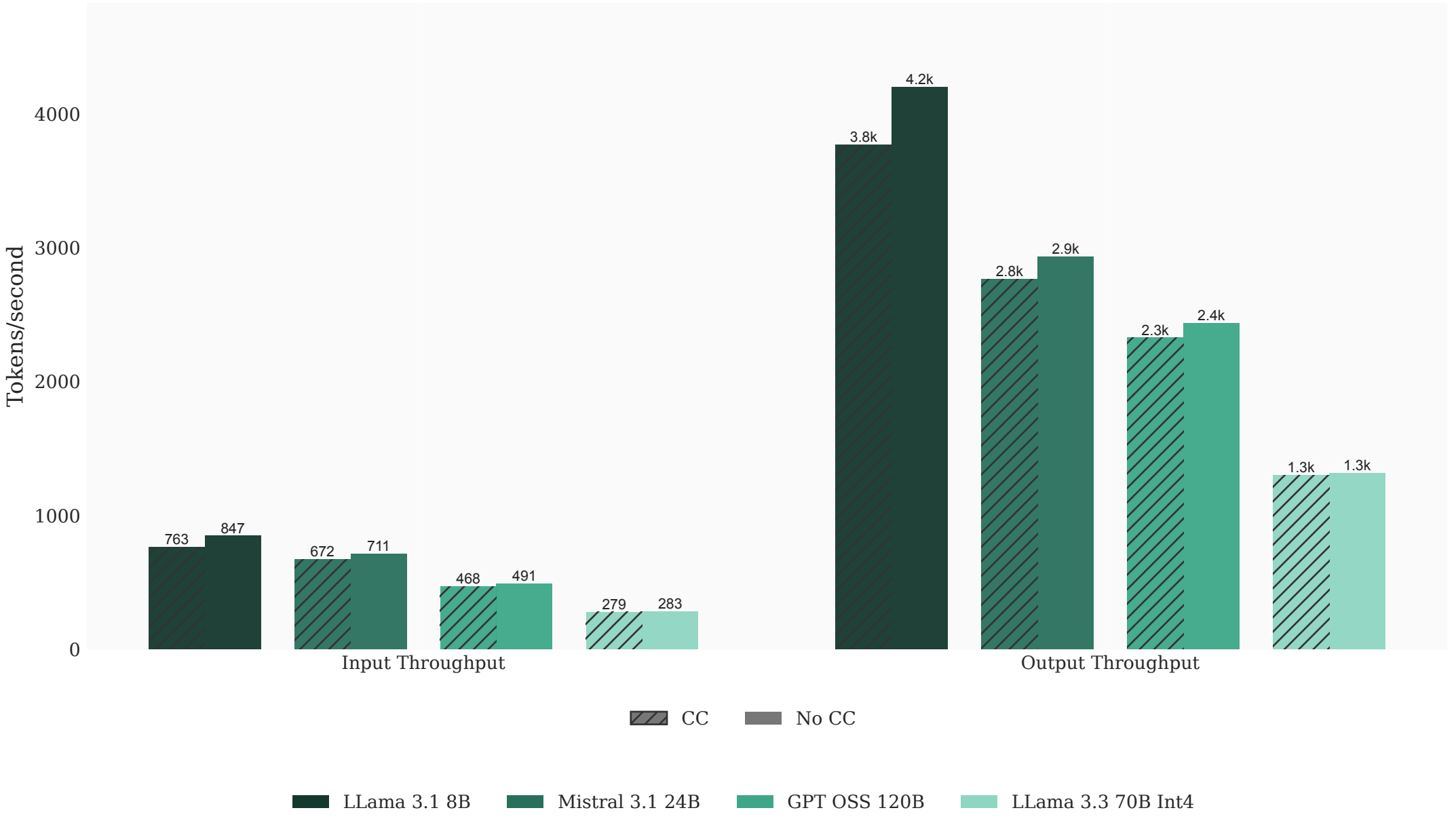
Edit 10K Characters (50 Concurrent Requests)



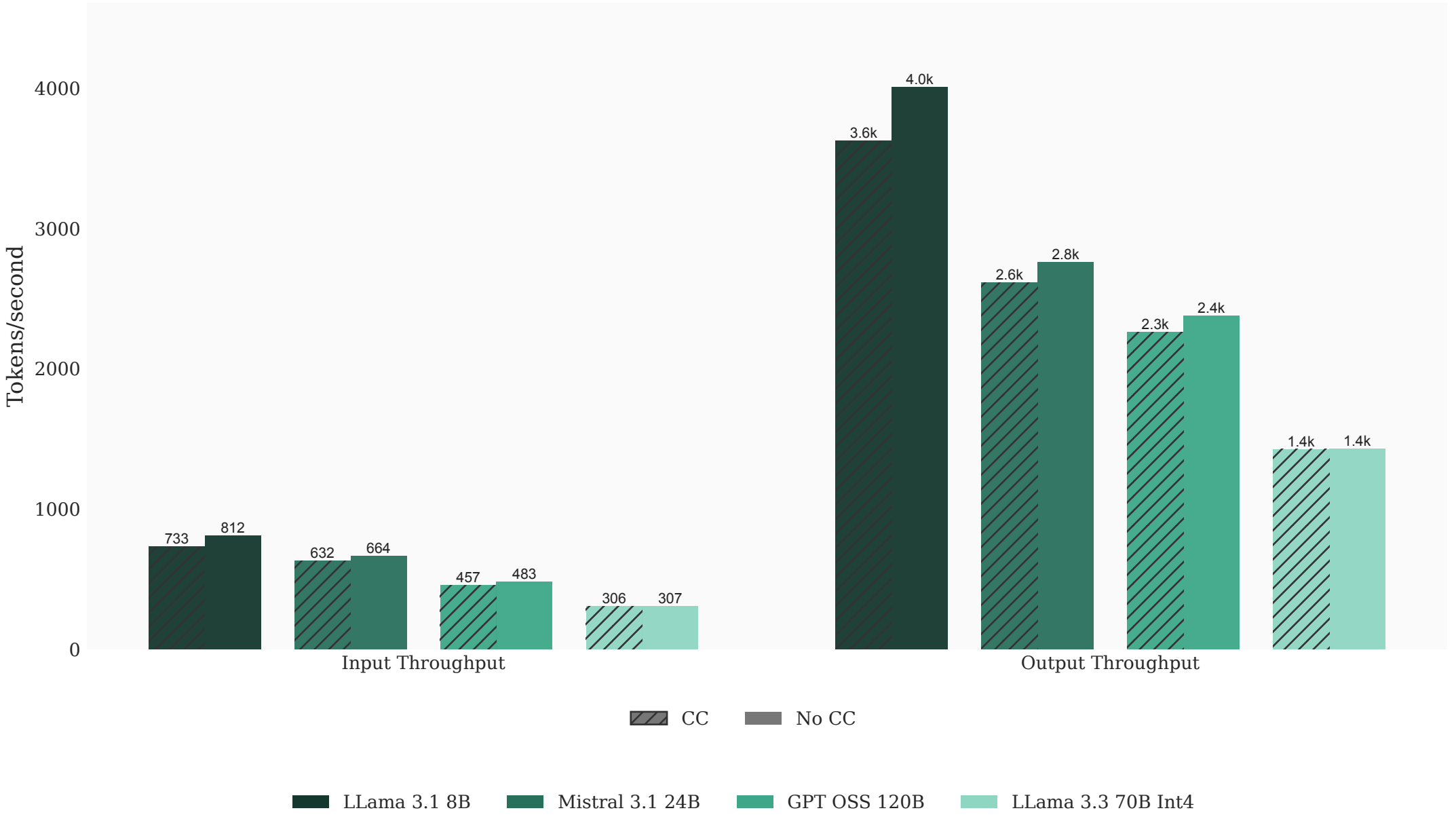
Edit 10K Characters (1 Concurrent Requests)

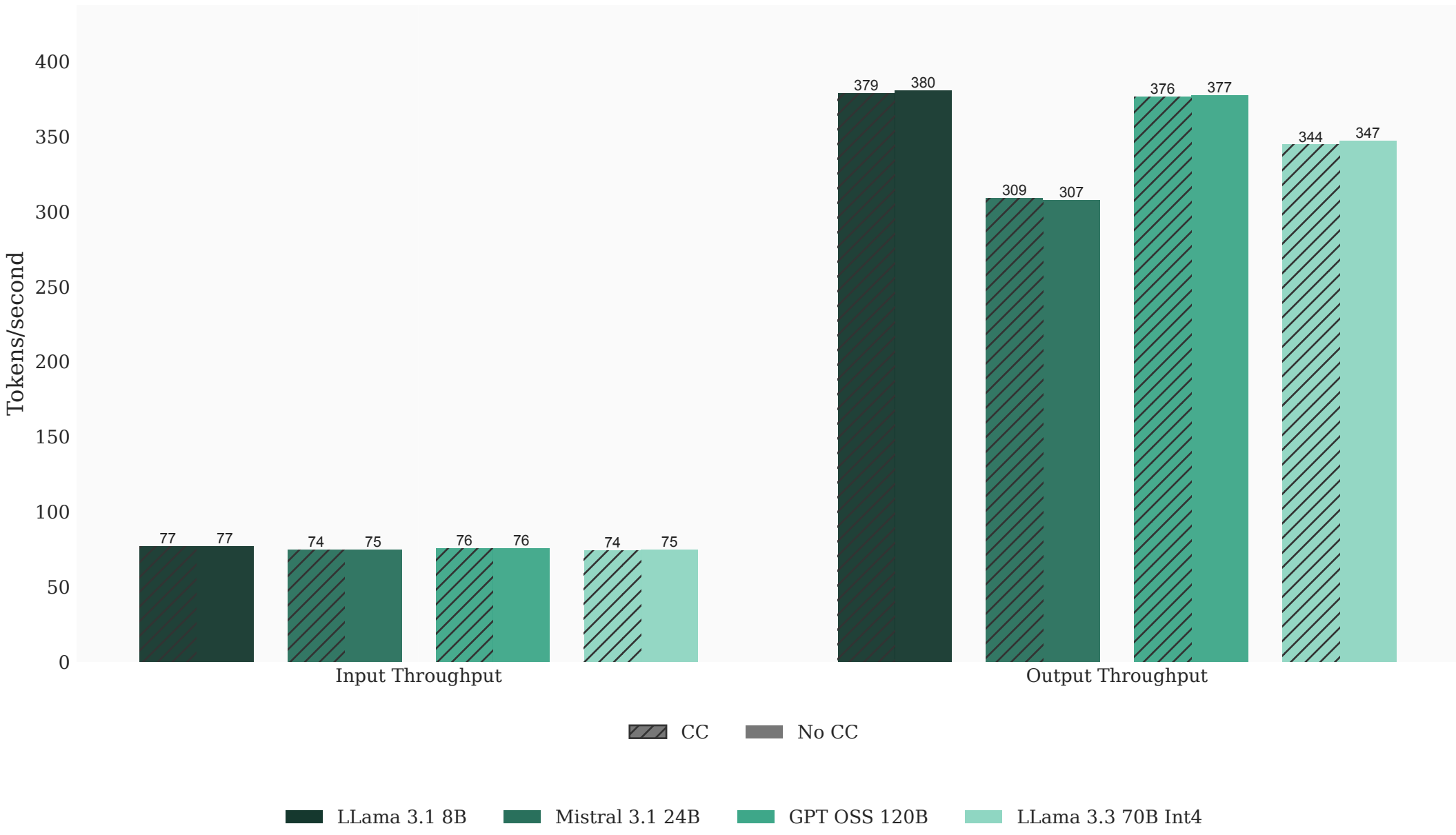


Numina Math (100 Concurrent Requests)



Numina Math (50 Concurrent Requests)




Numina Math (1 Concurrent Requests)
Tokens/second
400
350
300
250
200
150
100
50
0
77
77
74
75
76
76
74
75
379
380
309
307
376
377
344
347
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4