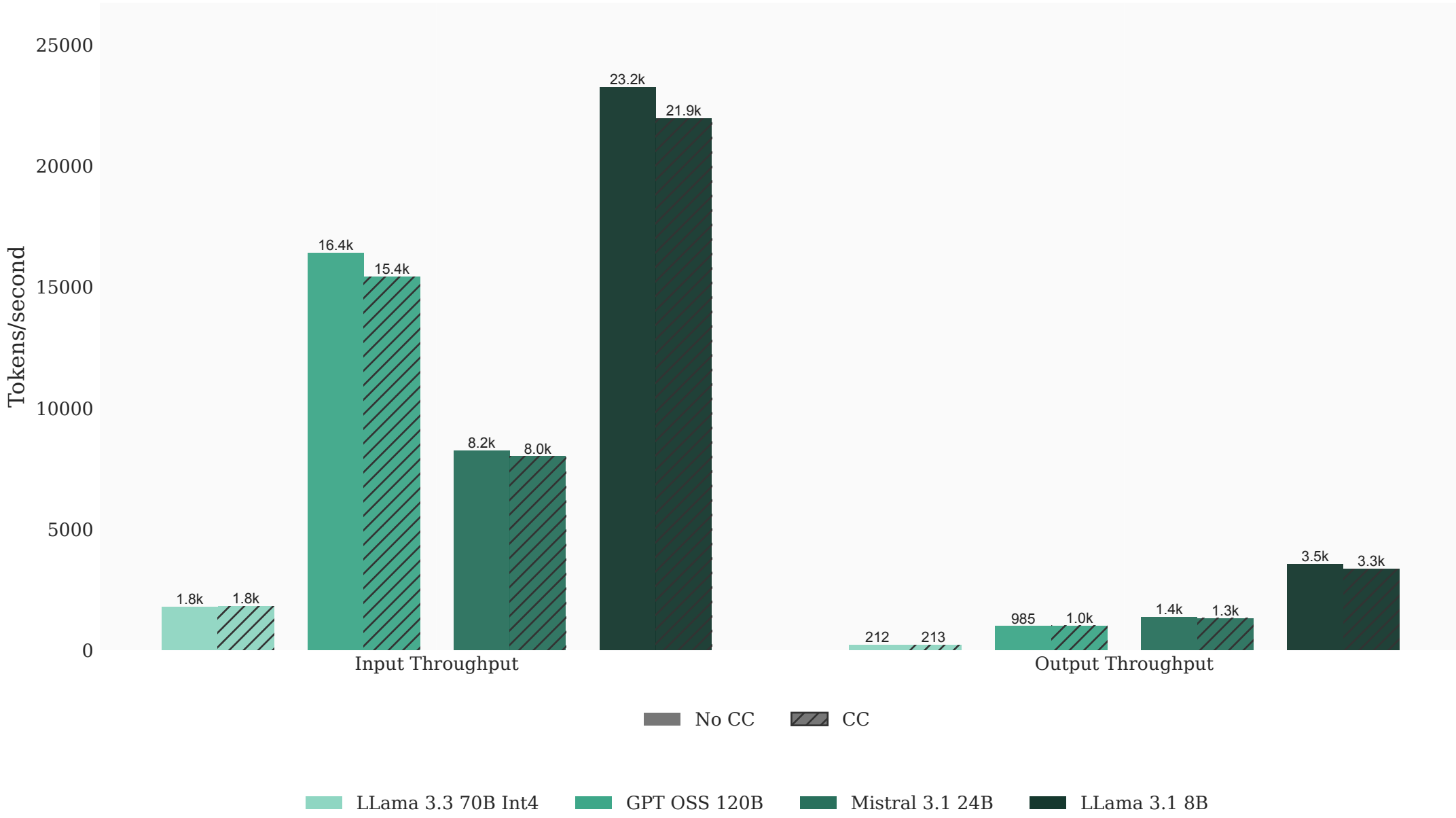
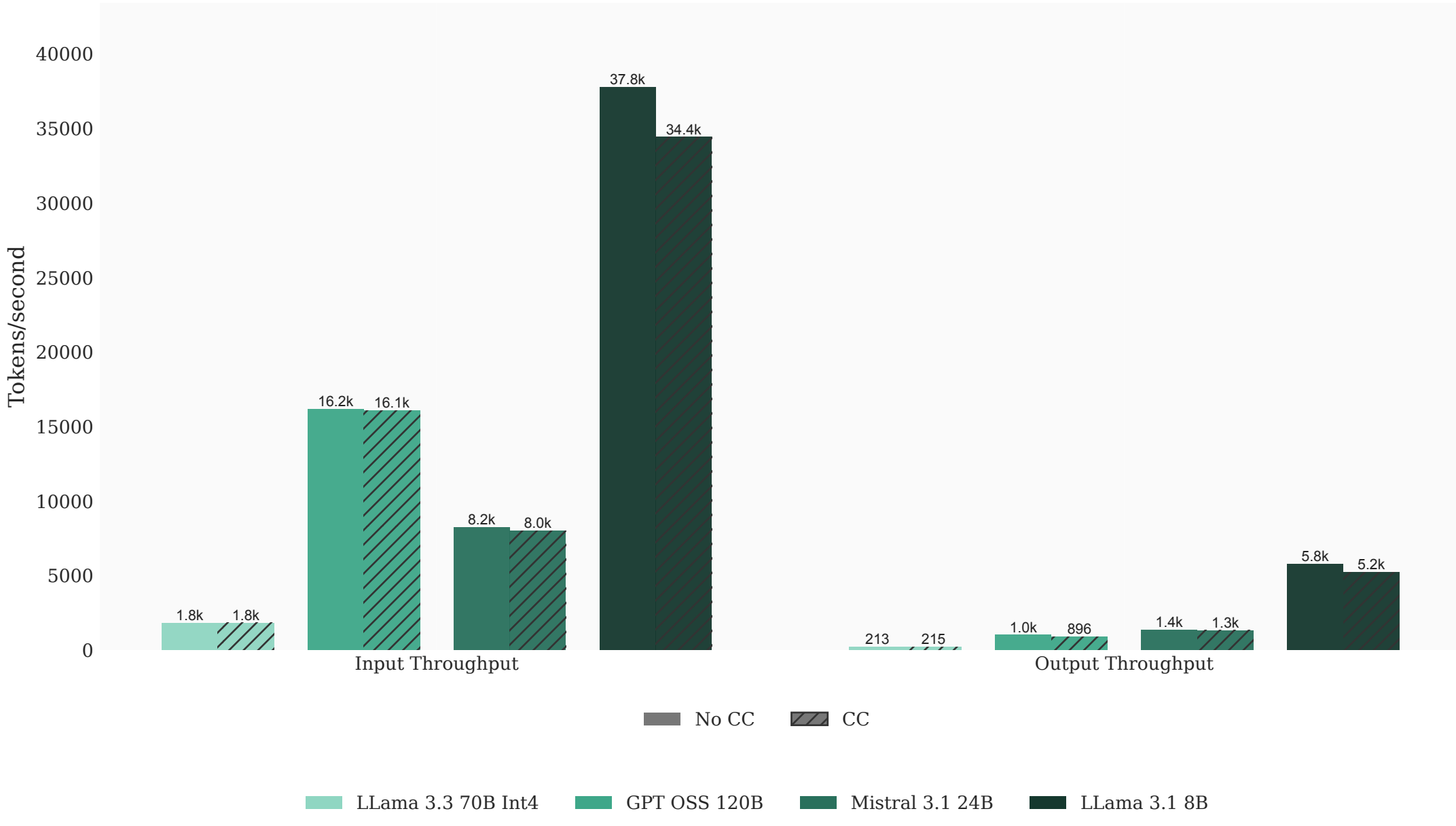
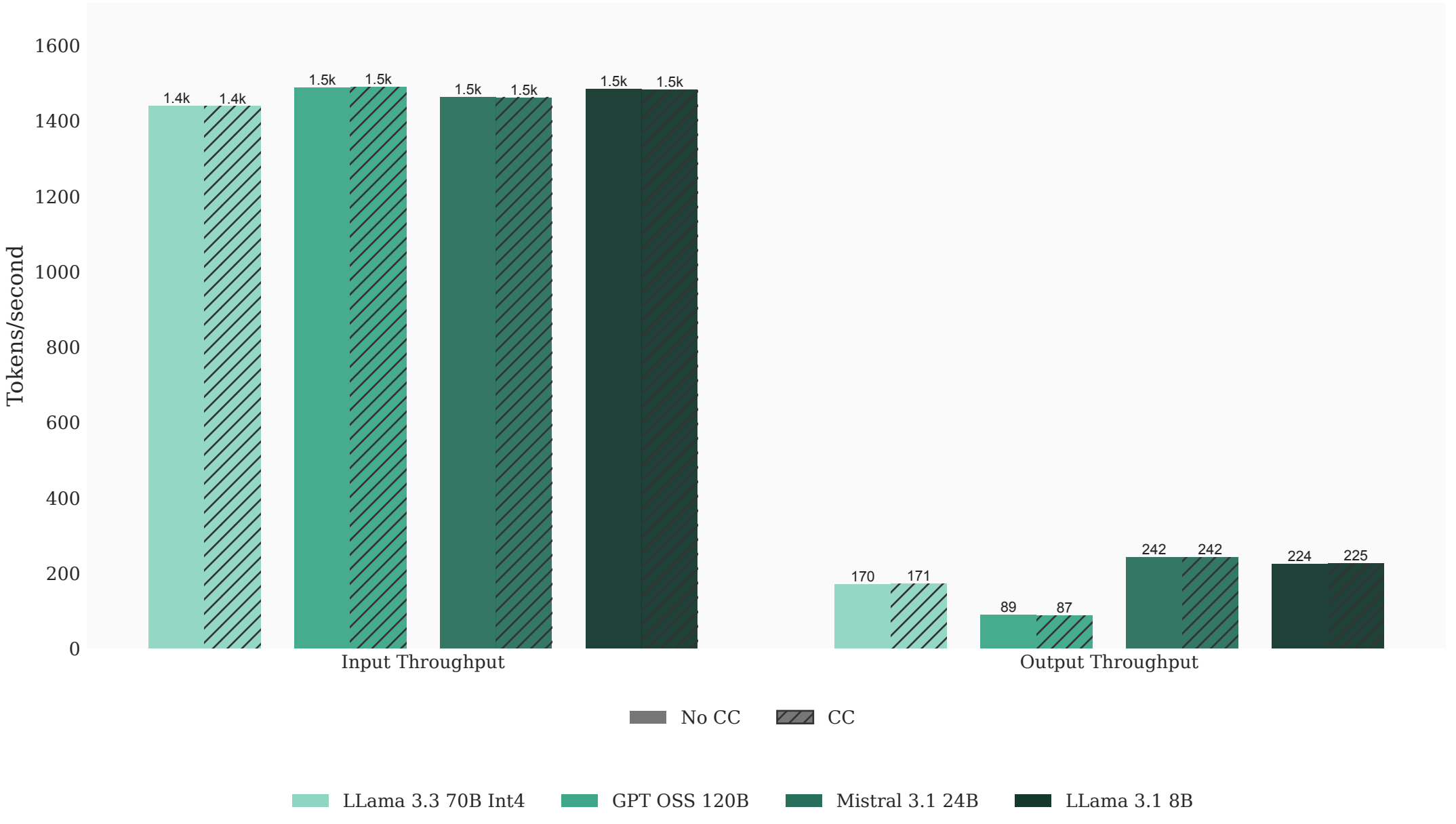
Random (1500 ⇒ 250) (Request Rate 100)

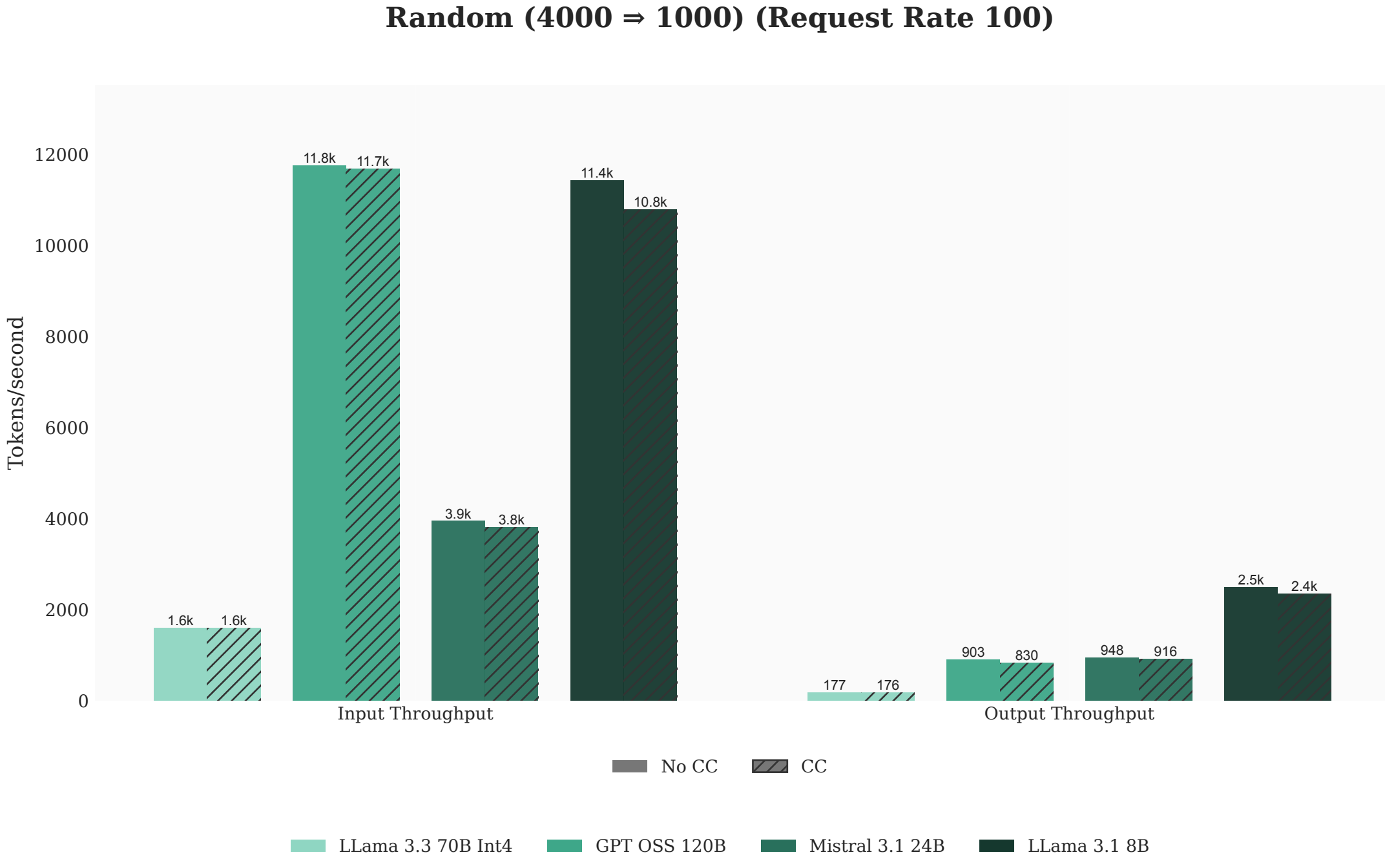


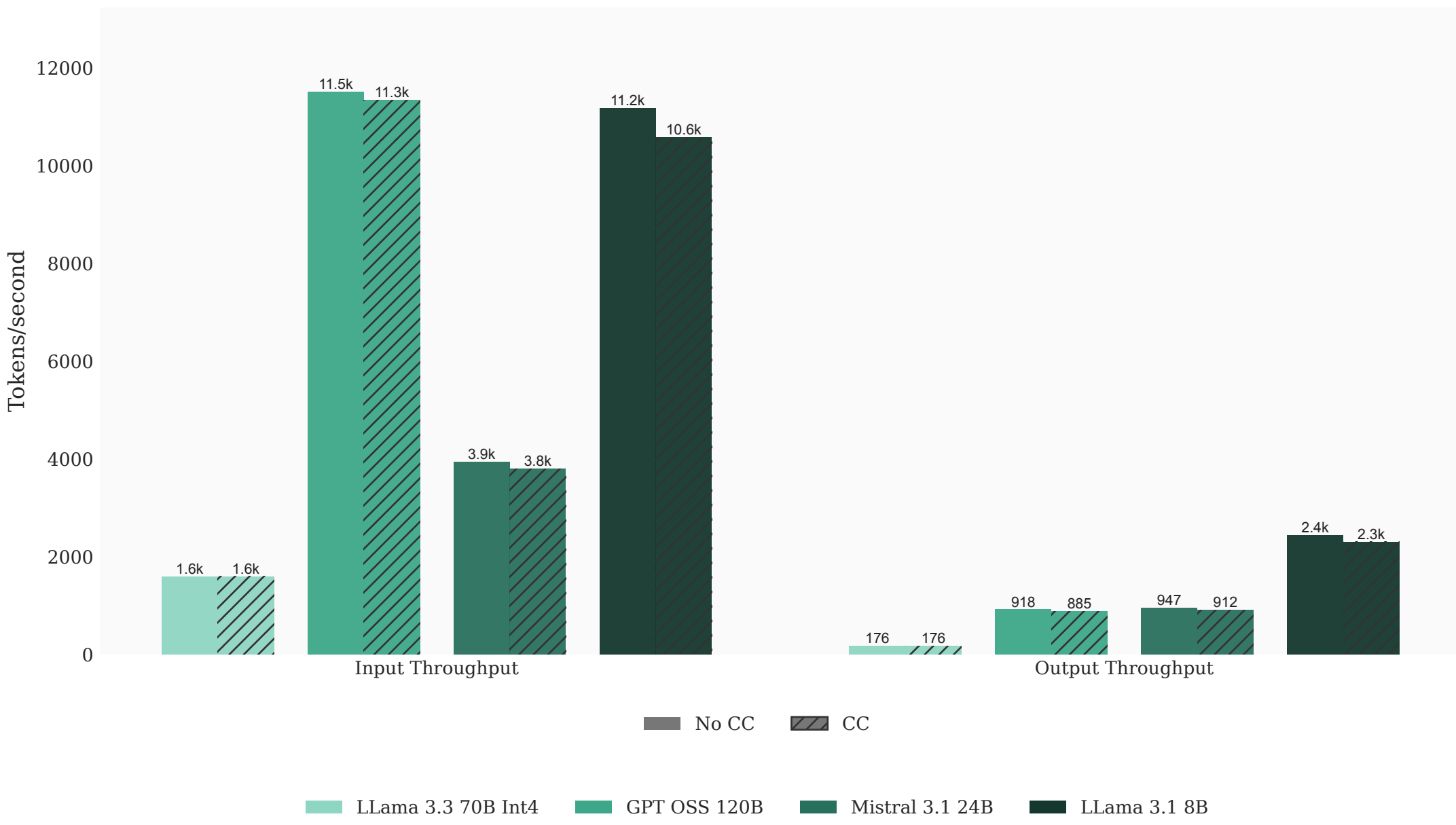
Random (1500 ⇒ 250) (Request Rate 50)

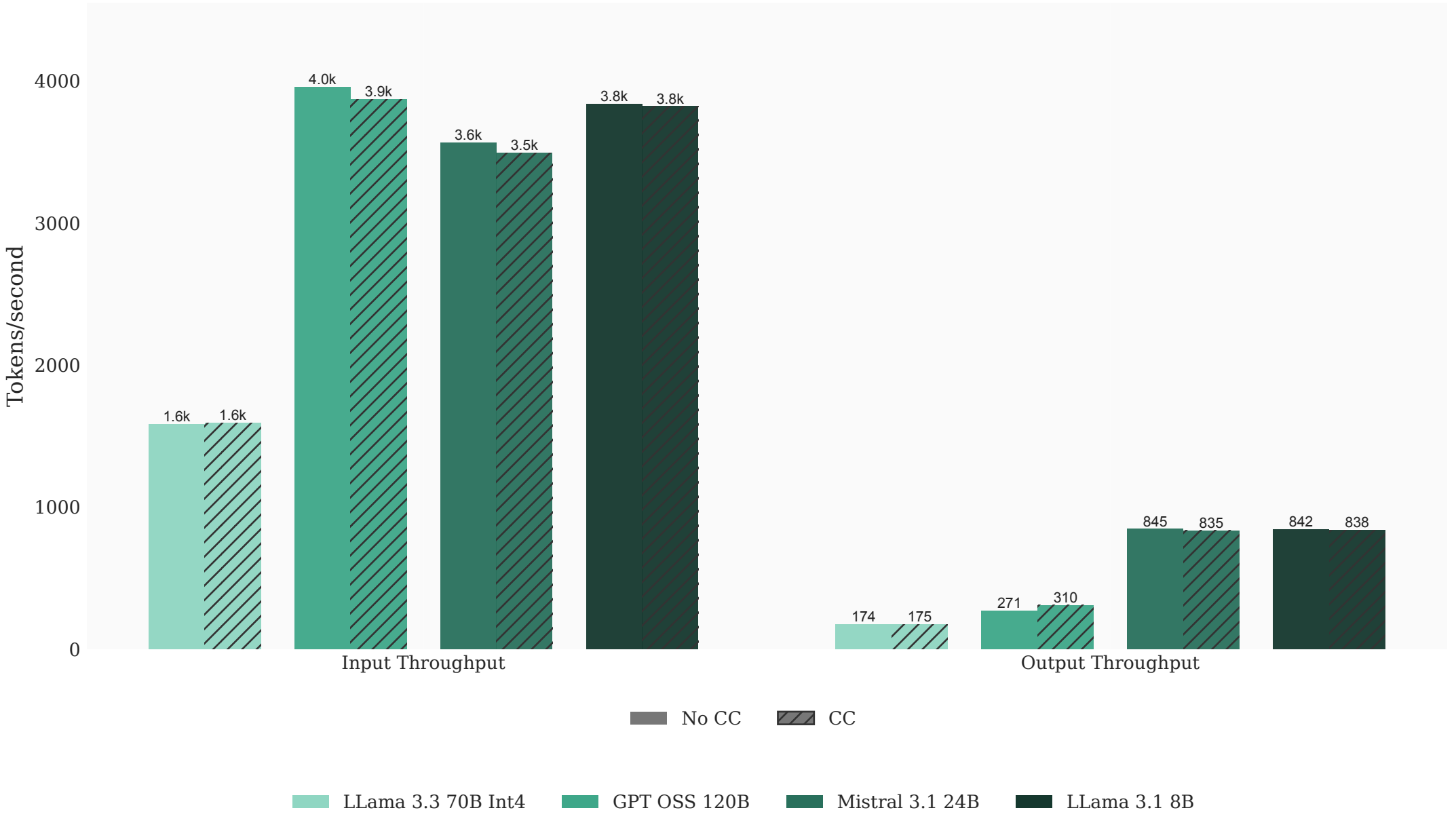


Random (1500 ⇒ 250) (Request Rate 1)

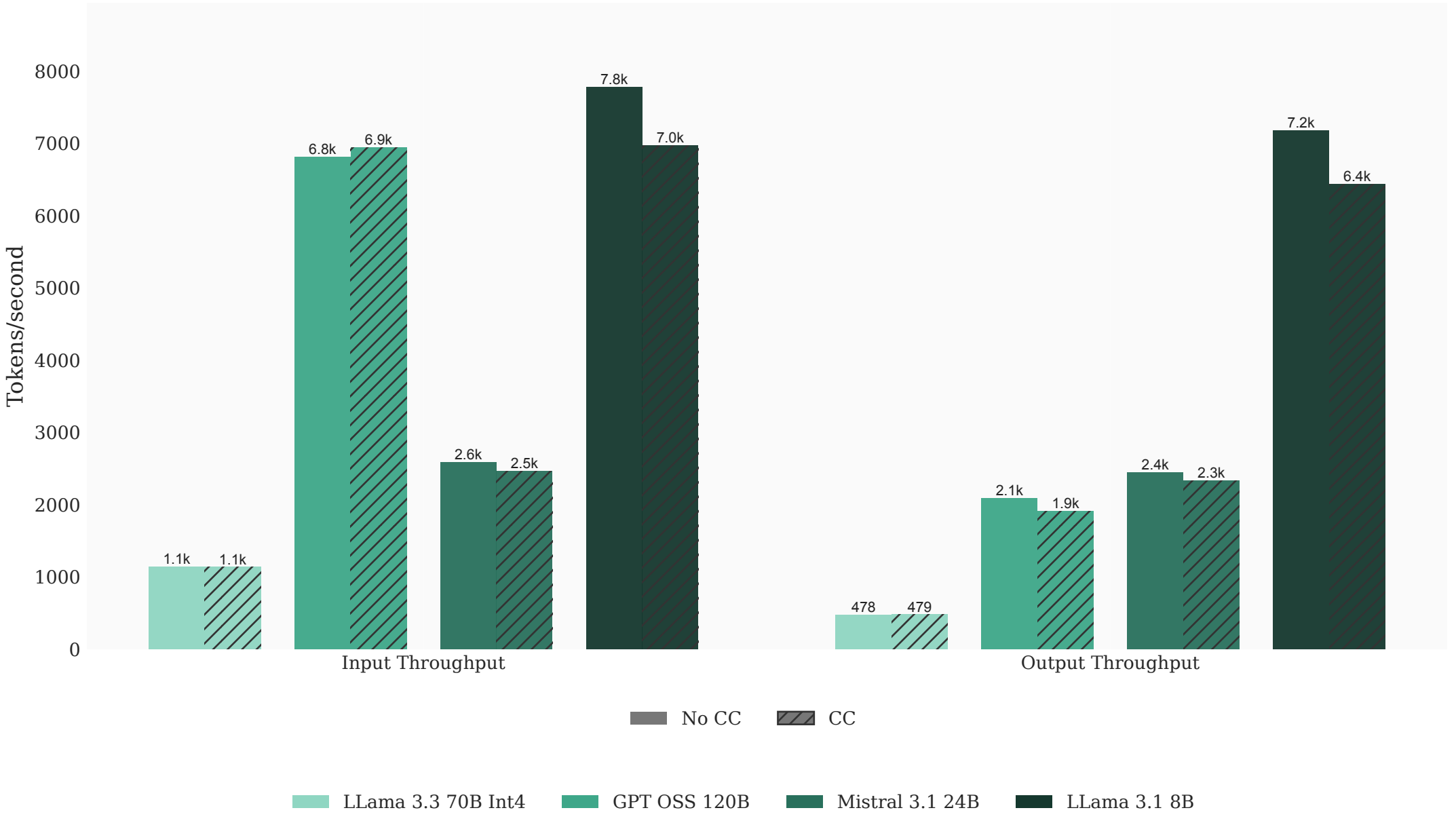


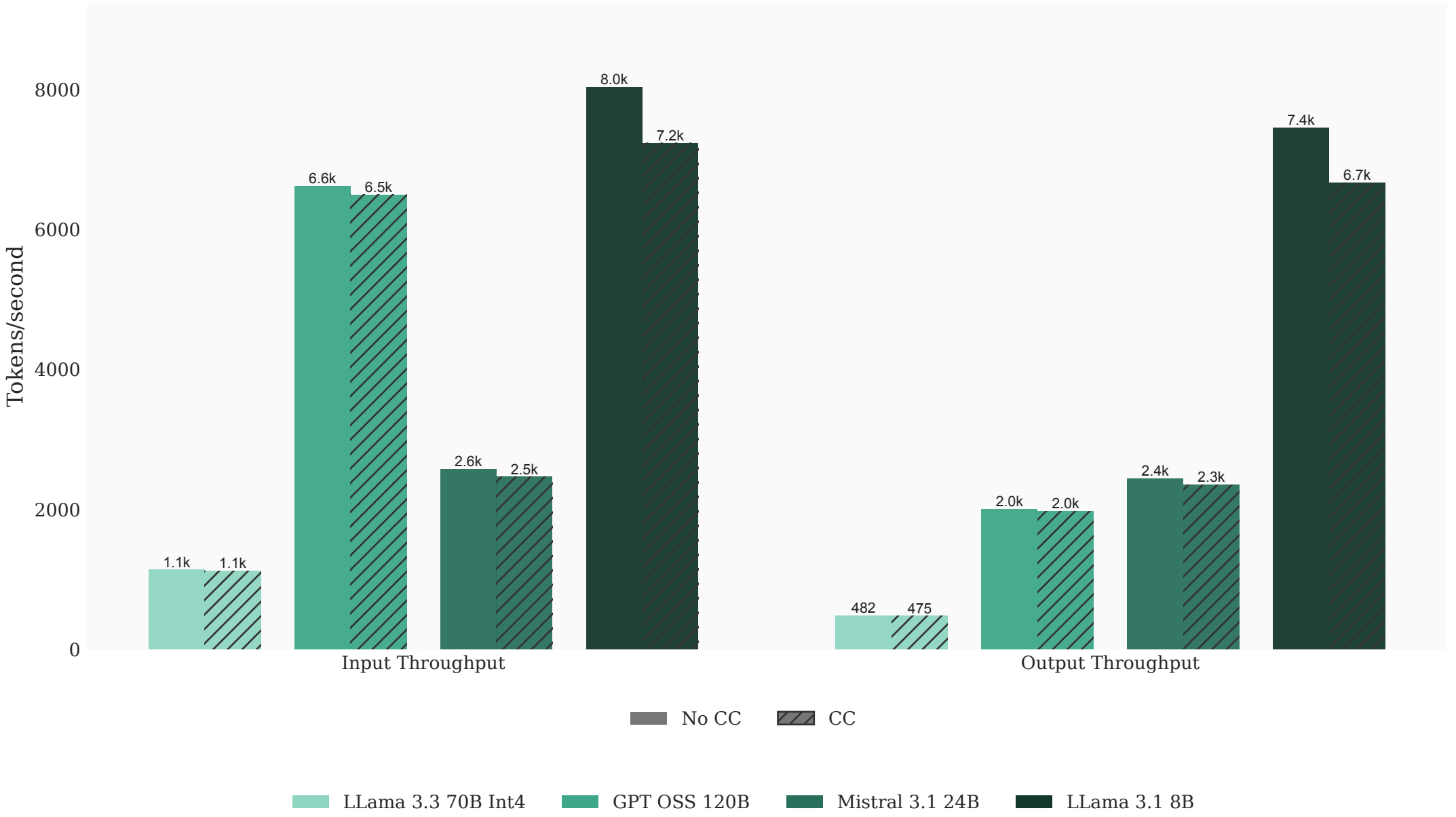

Random (4000 ⇒ 1000) (Request Rate 100)
Tokens/second
0
2000
4000
6000
8000
10000
12000
1.6k
1.6k
11.8k
11.7k
3.9k
3.8k
11.4k
10.8k
177
176
903
830
948
916
2.5k
2.4k
Input Throughput
Output Throughput
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


Random (4000 ⇒ 1000) (Request Rate 50)
Tokens/second
12000
10000
8000
6000
4000
2000
0
11.5k
11.3k
11.2k
10.6k
3.9k
3.8k
1.6k
1.6k
176
176
918
885
947
912
2.4k
2.3k
Input Throughput
Output Throughput
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

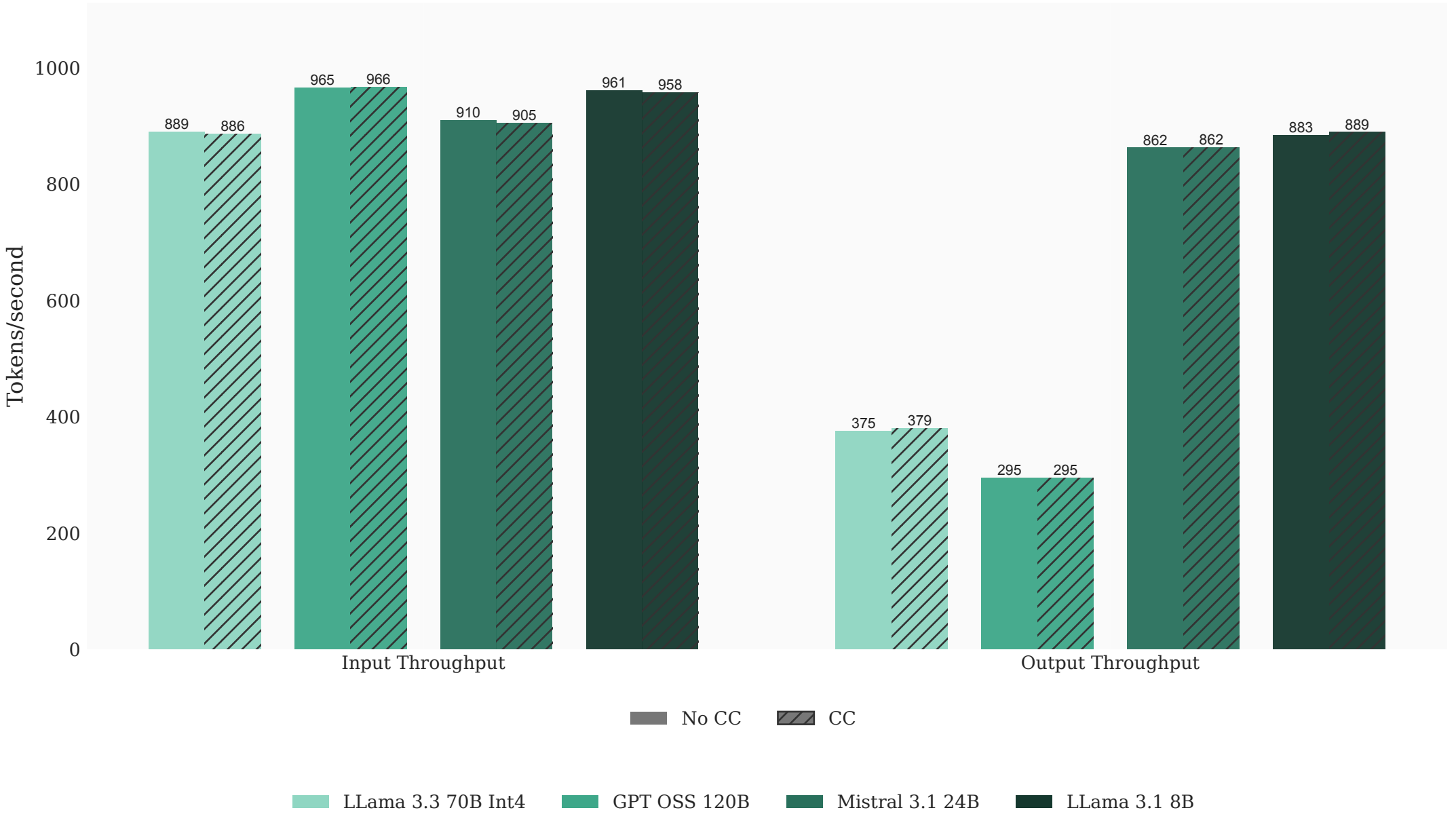

Random (4000 ⇒ 1000) (Request Rate 1)
Tokens/second
4000
3000
2000
1000
0
1.6k
1.6k
4.0k
3.9k
3.6k
3.5k
3.8k
3.8k
174
175
271
310
845
835
842
838
Input Throughput
Output Throughput
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

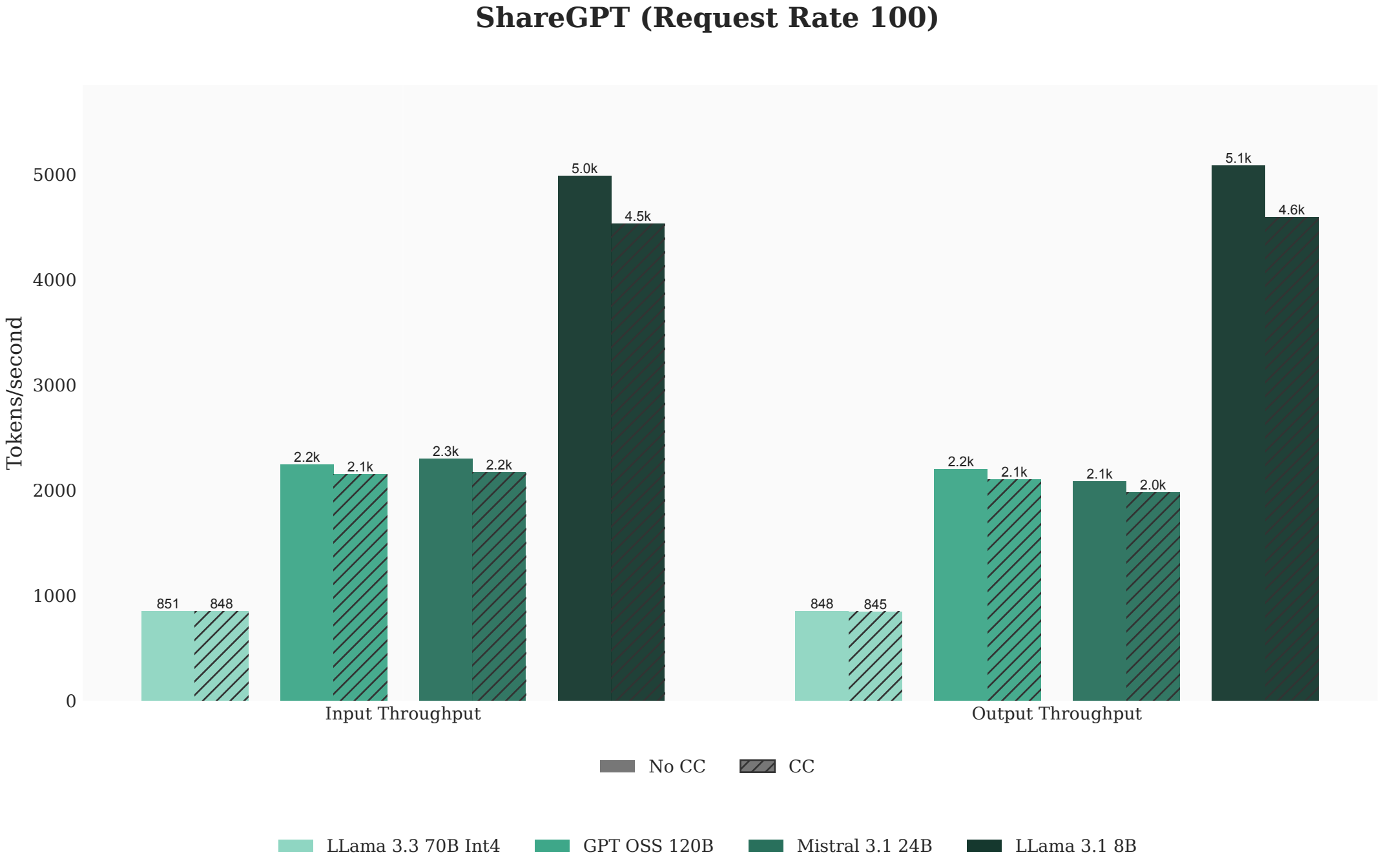
Random (1000 ⇒ 1000) (Request Rate 100)

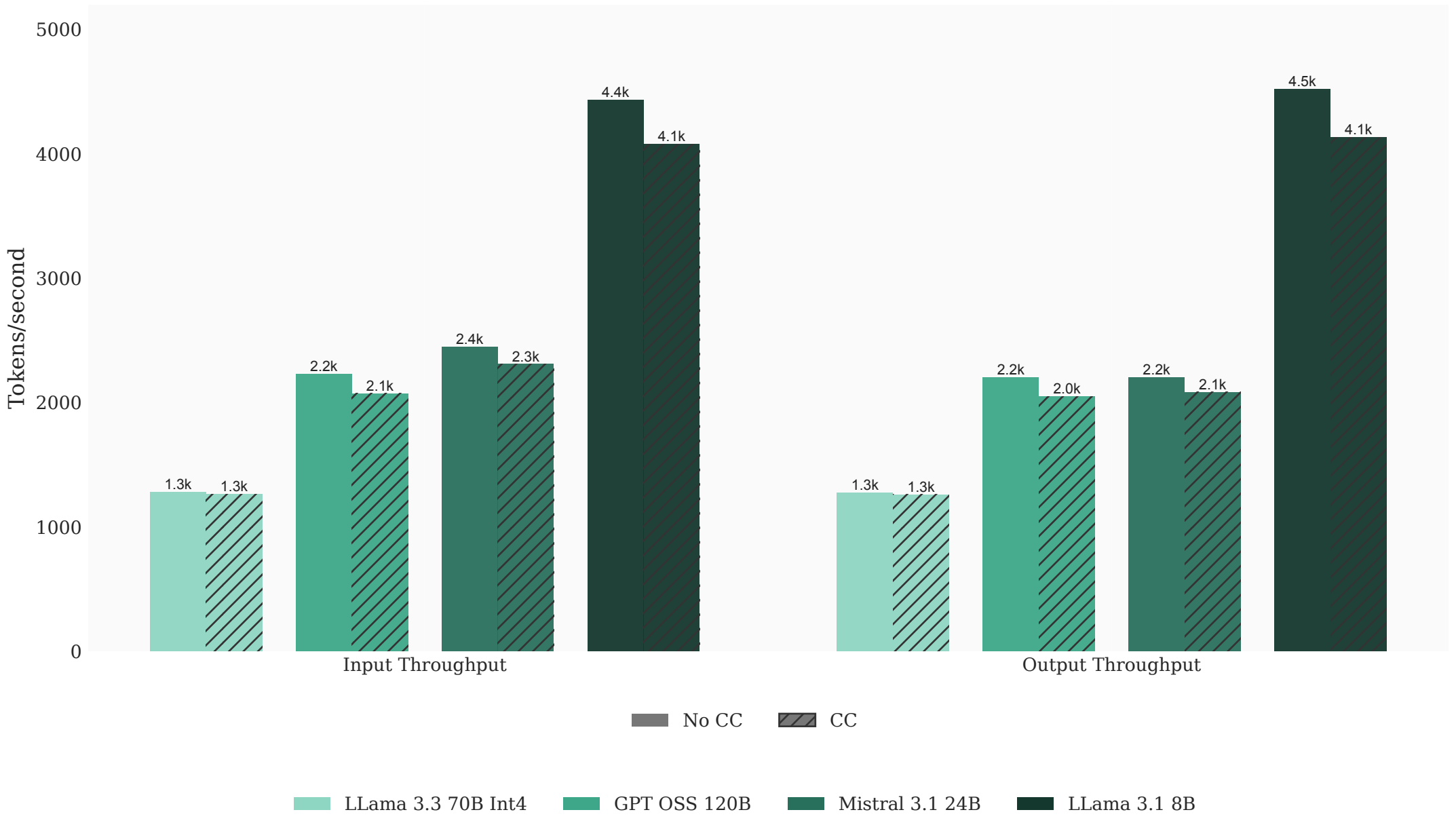


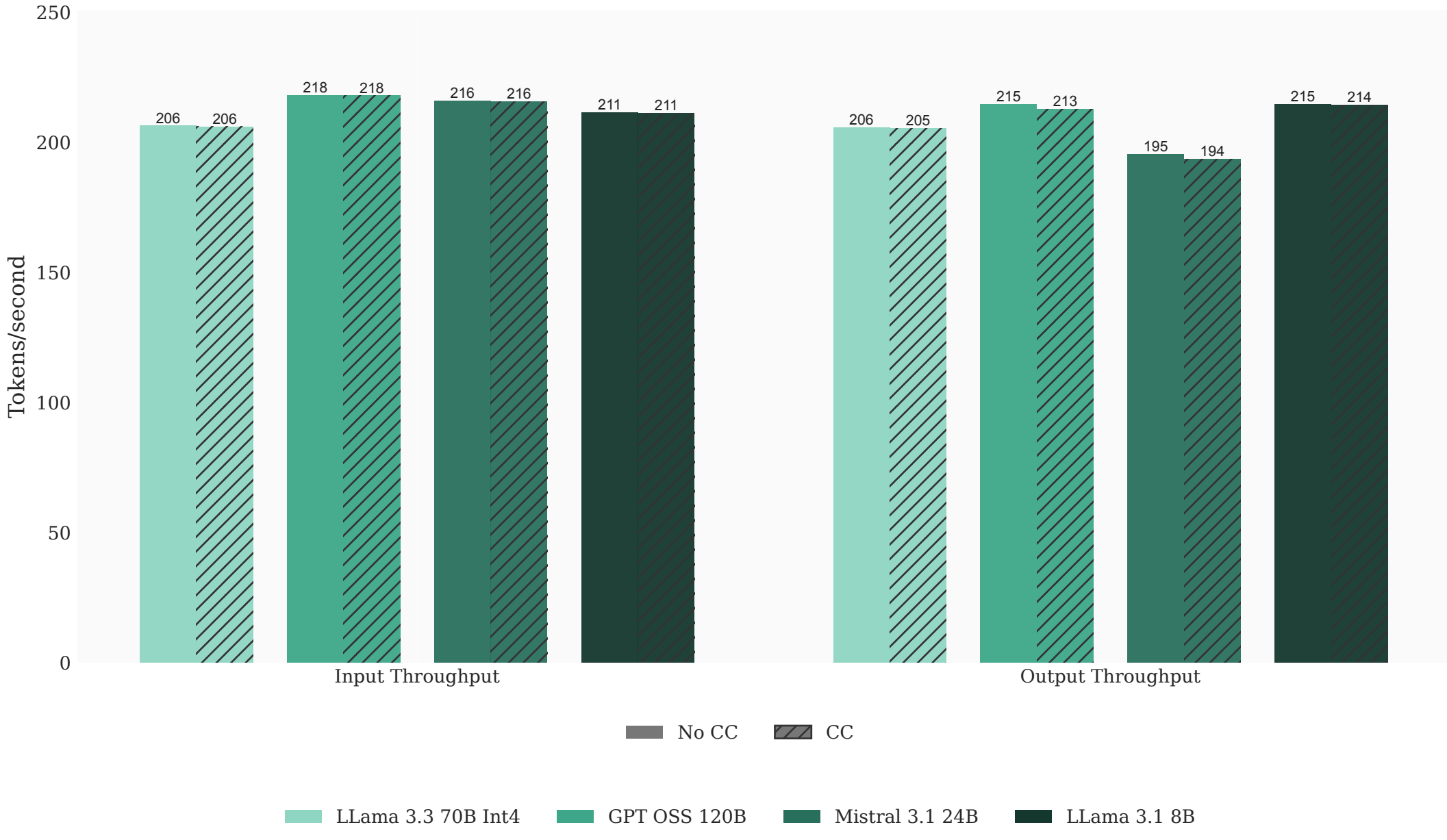

Random (1000 ⇒ 1000) (Request Rate 50)
Tokens/second
0
2000
4000
6000
8000
1.1k
1.1k
6.6k
6.5k
2.6k
2.5k
8.0k
7.2k
482
475
2.0k
2.0k
2.4k
2.3k
7.4k
6.7k
Input Throughput
Output Throughput
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

Random (1000 ⇒ 1000) (Request Rate 1)

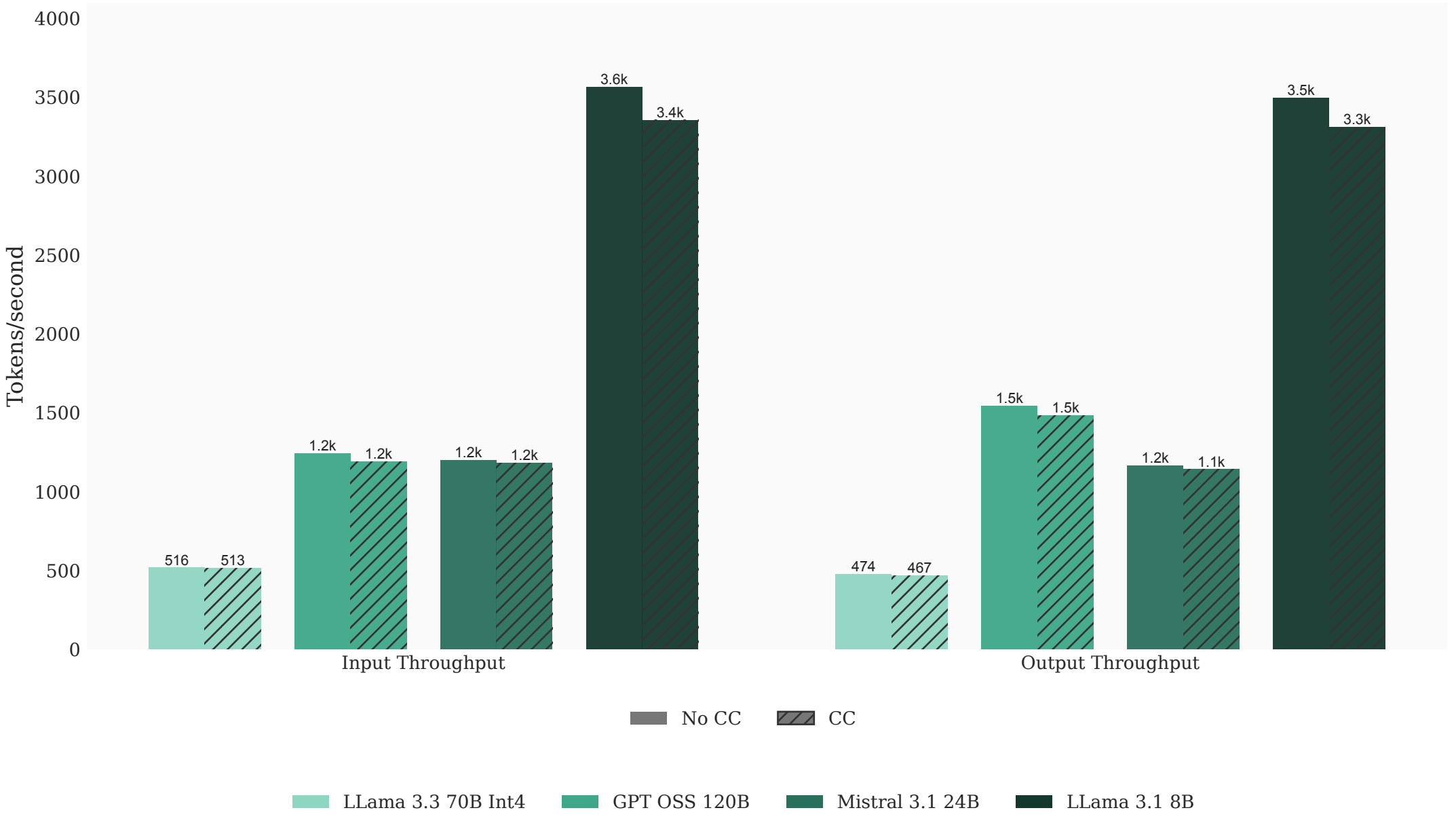



ShareGPT (Request Rate 100)
Tokens/second
0
1000
2000
3000
4000
5000
851
848
2.2k
2.1k
2.3k
2.2k
5.0k
4.5k
848
845
2.2k
2.1k
2.1k
2.0k
5.1k
4.6k
Input Throughput
Output Throughput
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B

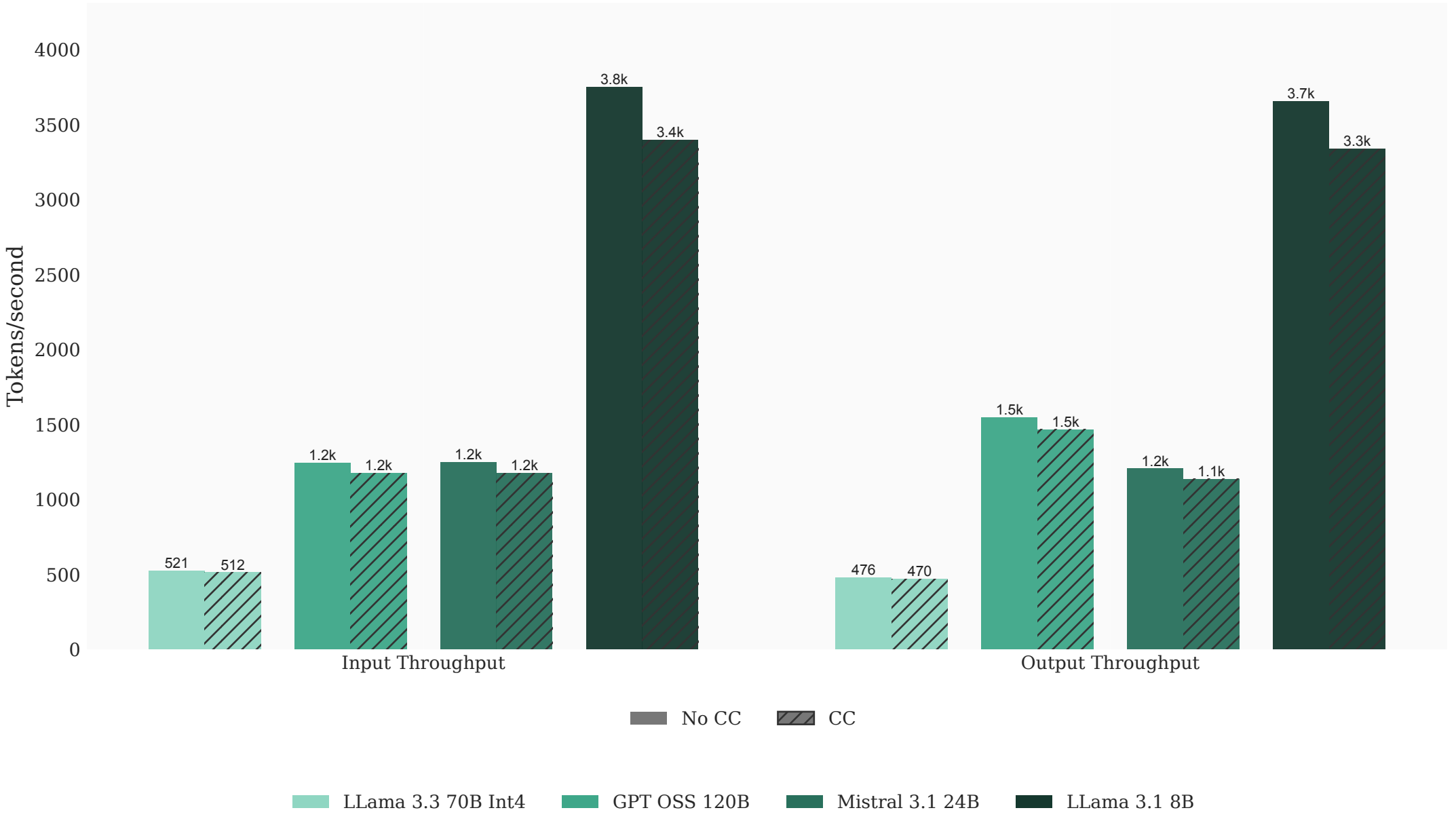

ShareGPT (Request Rate 50)
Tokens/second
5000
4000
3000
2000
1000
0
1.3k
1.3k
2.2k
2.1k
2.4k
2.3k
4.4k
4.1k
1.3k
1.3k
2.2k
2.0k
2.2k
2.1k
4.5k
4.1k
Input Throughput
Output Throughput
No CC
CC
LLama 3.3 70B Int4
GPT OSS 120B
Mistral 3.1 24B
LLama 3.1 8B


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

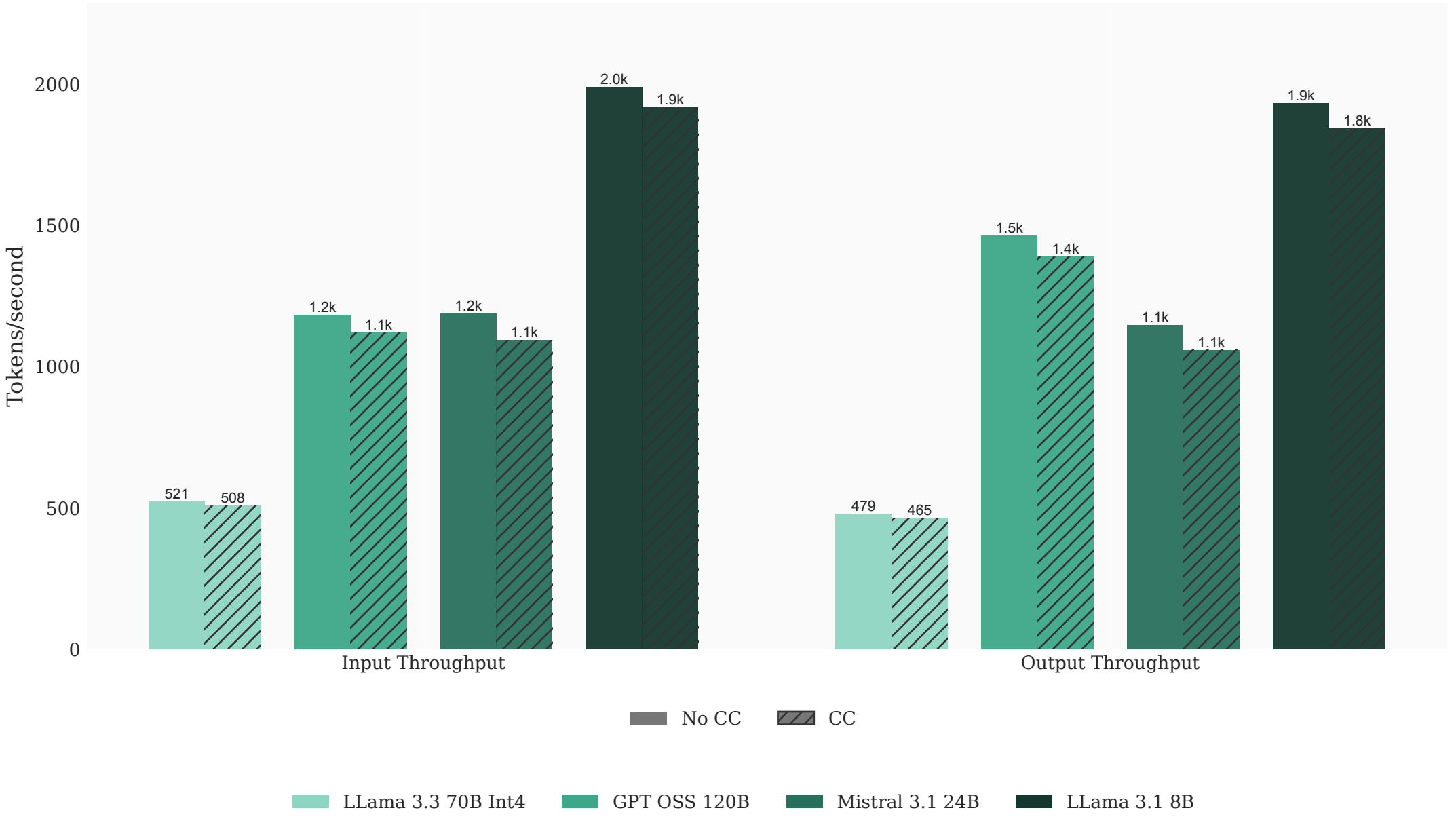
Edit 10K Characters (Request Rate 100)



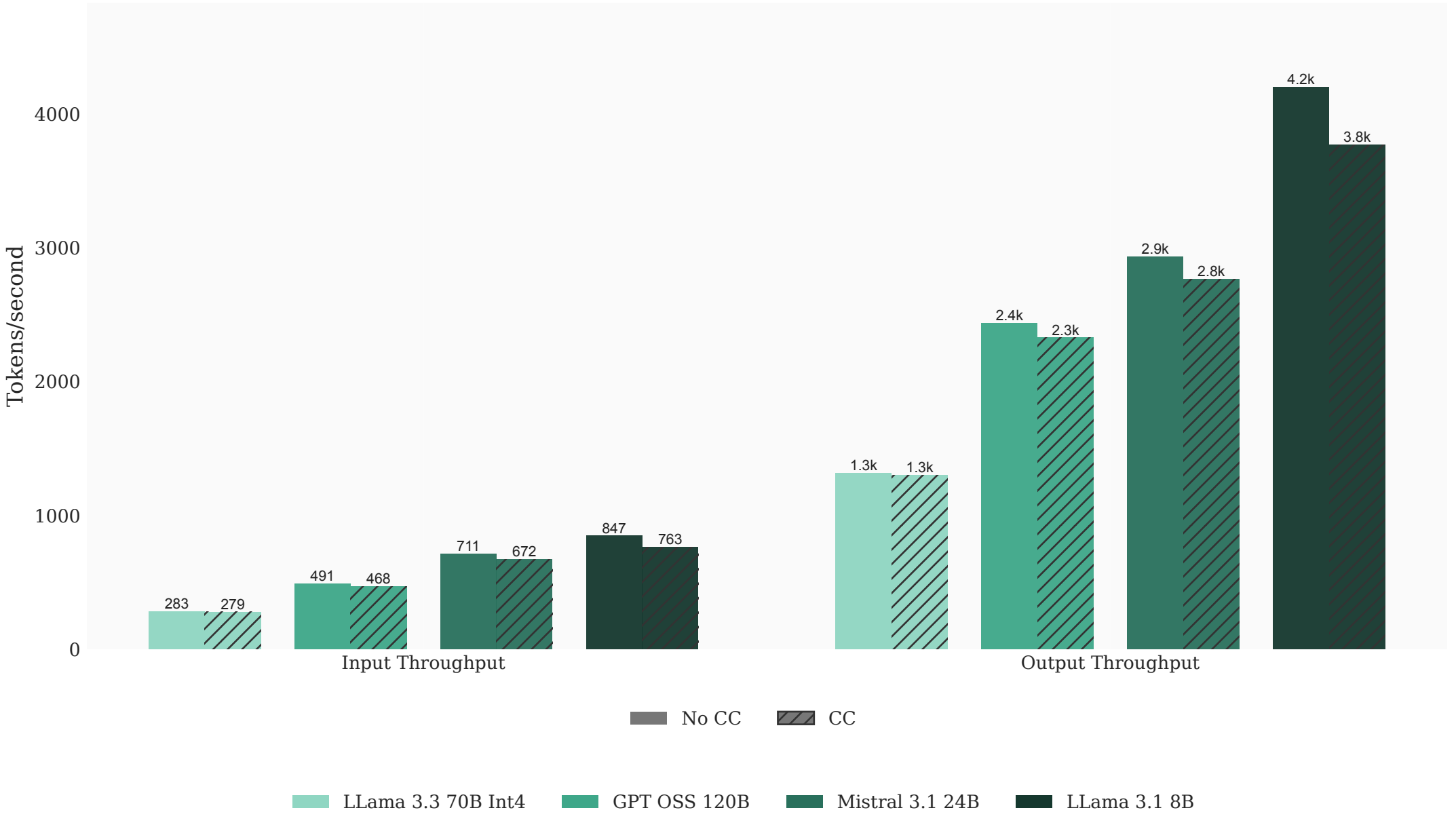
Edit 10K Characters (Request Rate 50)



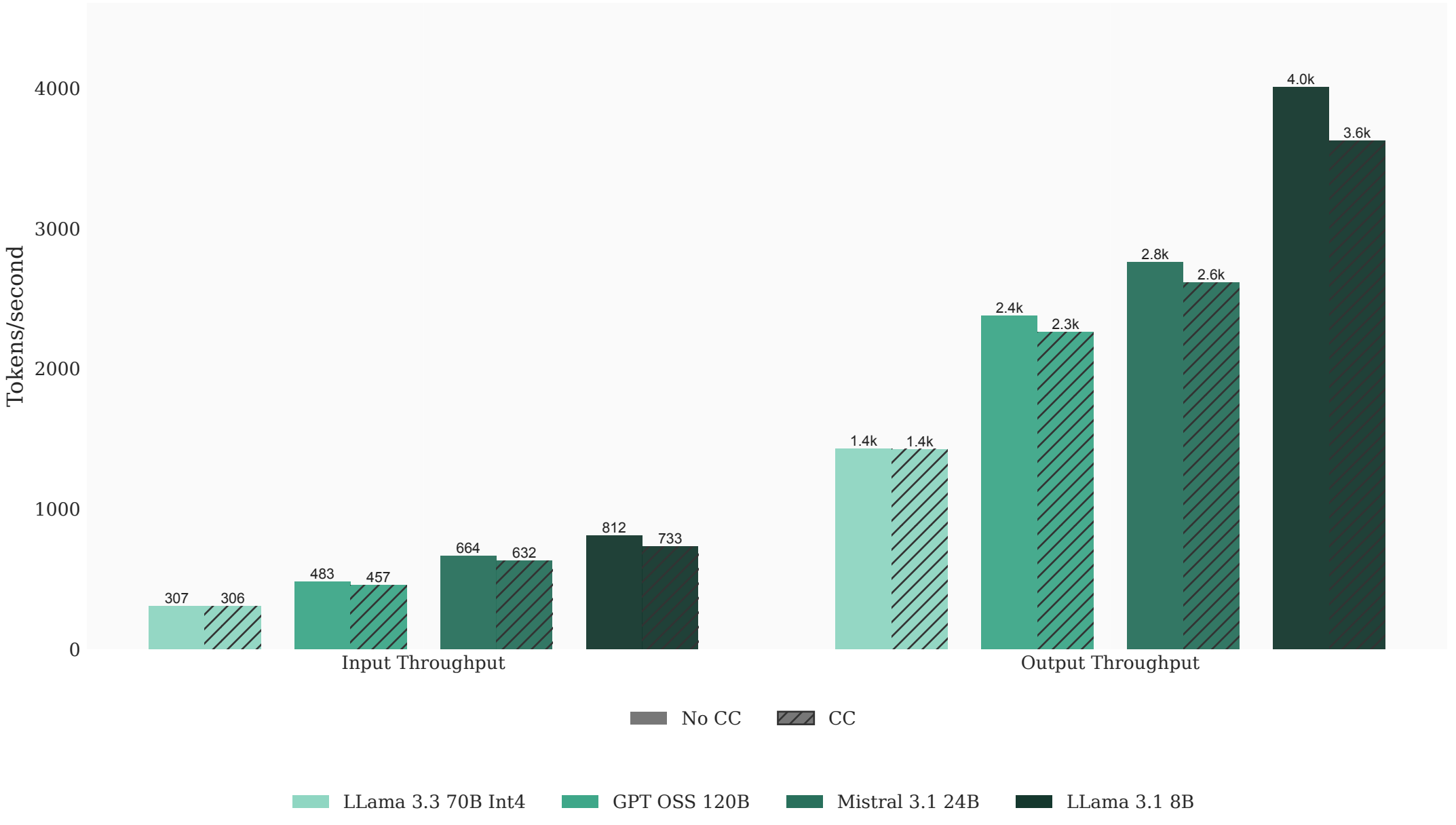
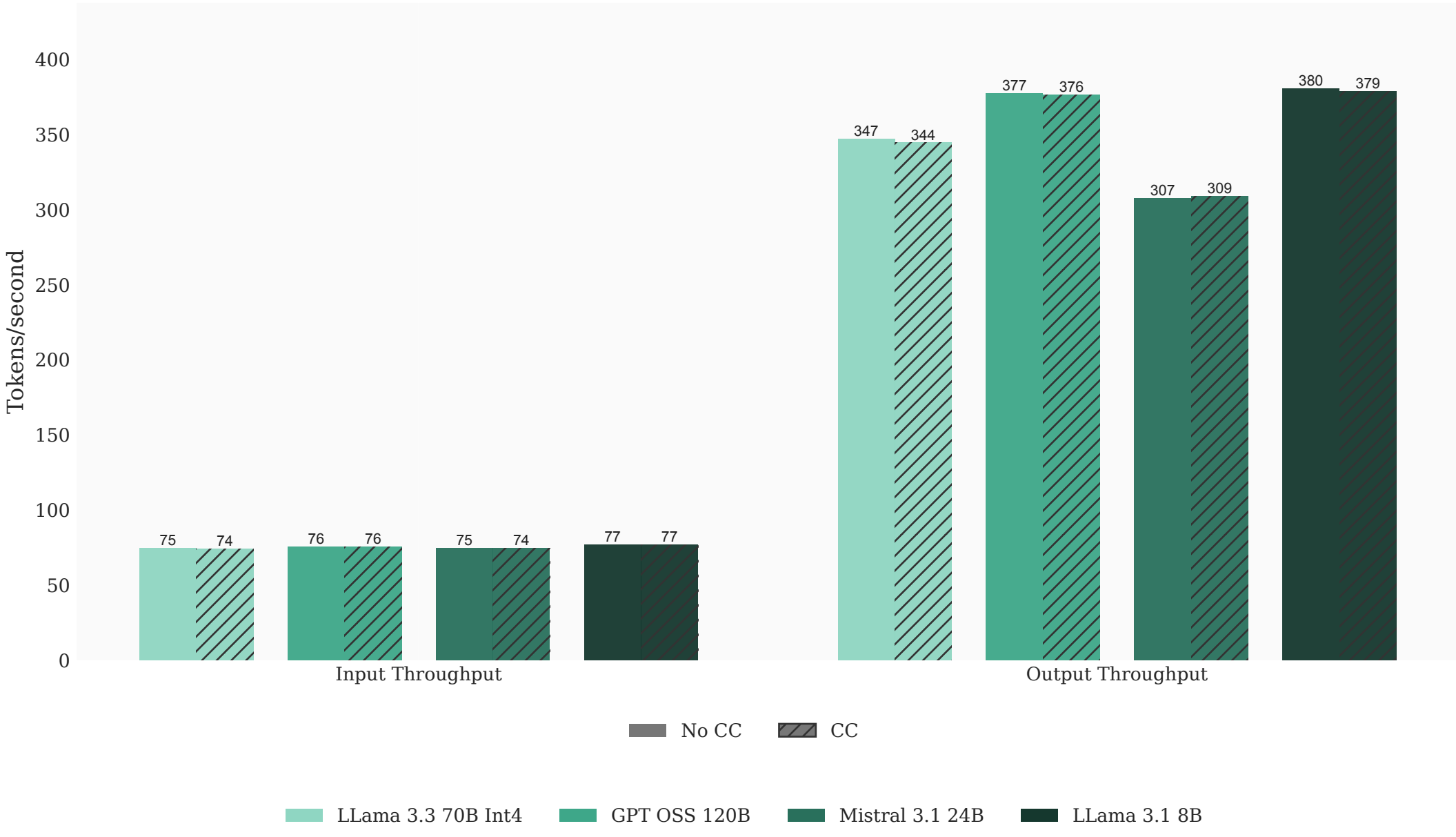
Edit 10K Characters (Request Rate 1)



Numina Math (Request Rate 100)



Numina Math (Request Rate 50)



Numina Math (Request Rate 1)