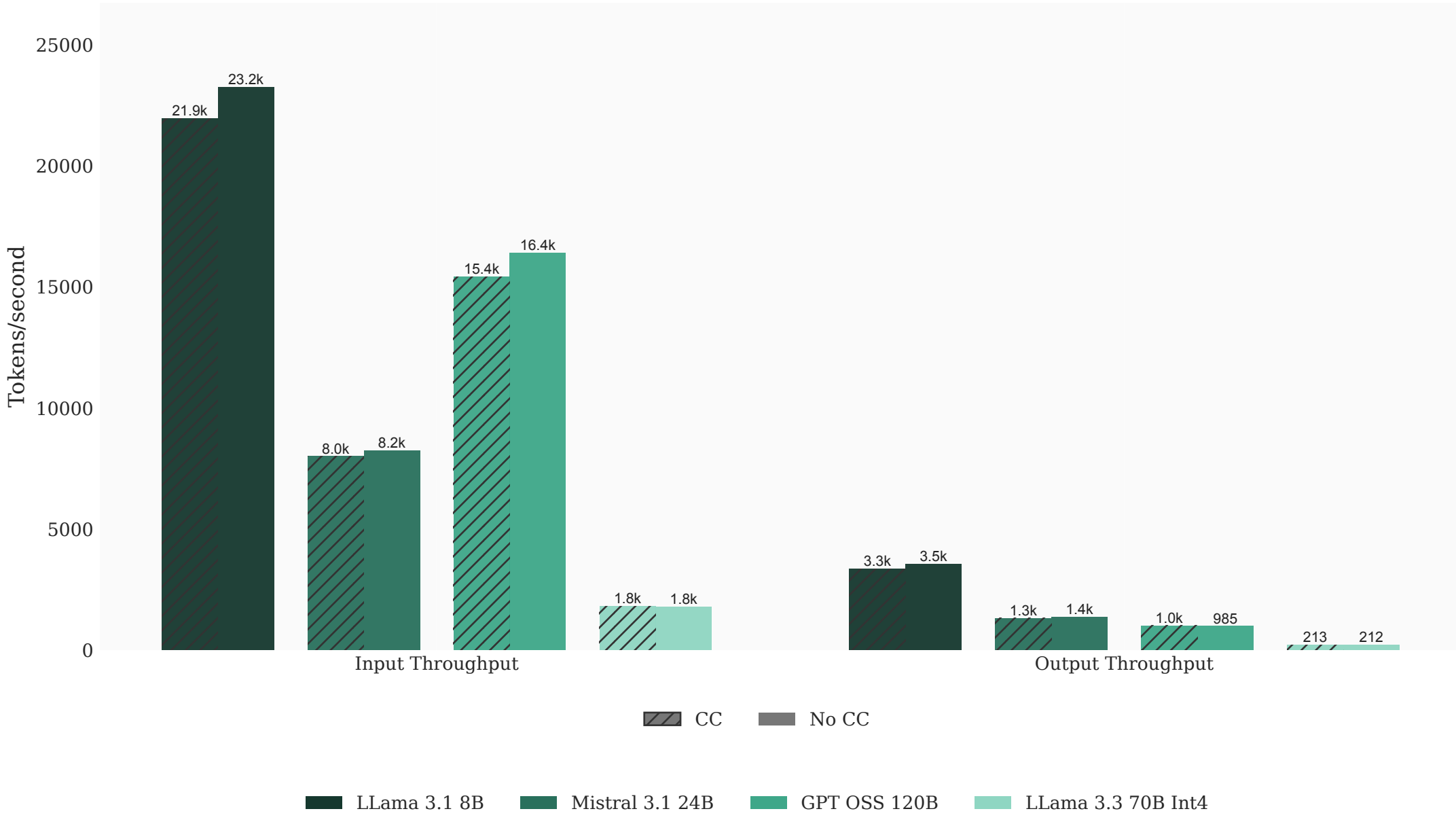
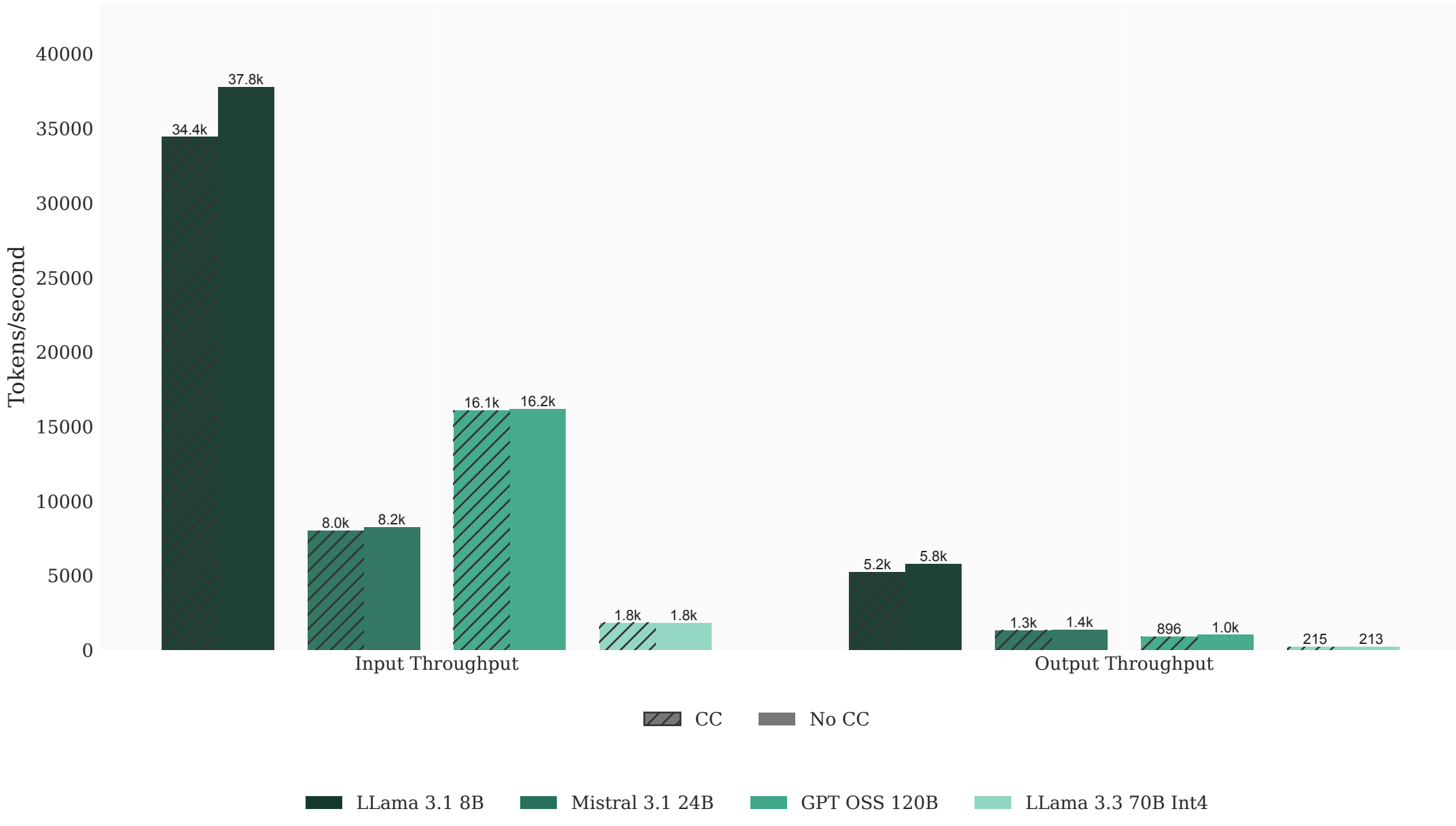
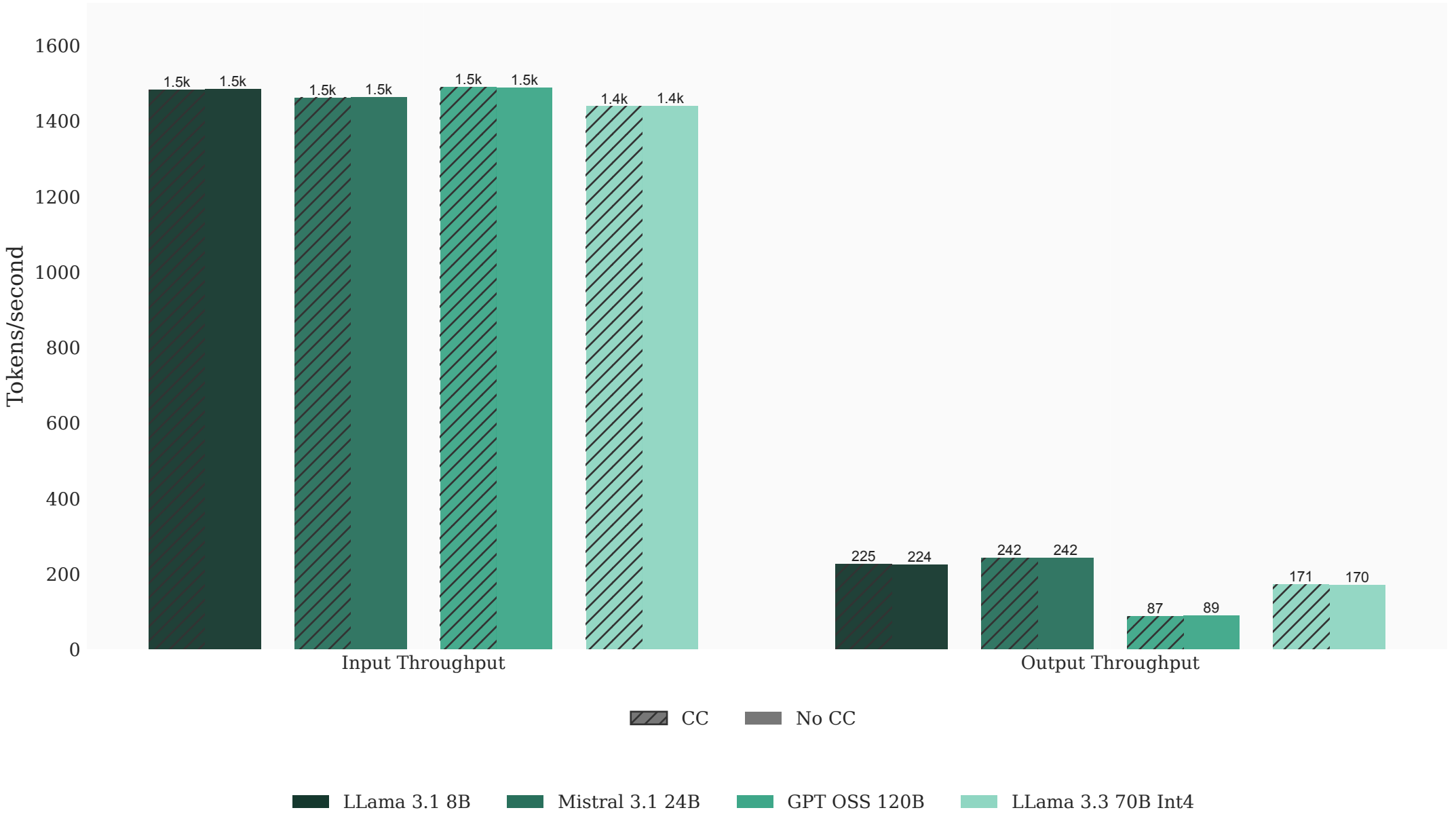
Random (1500 ⇒ 250) (Rate 100)



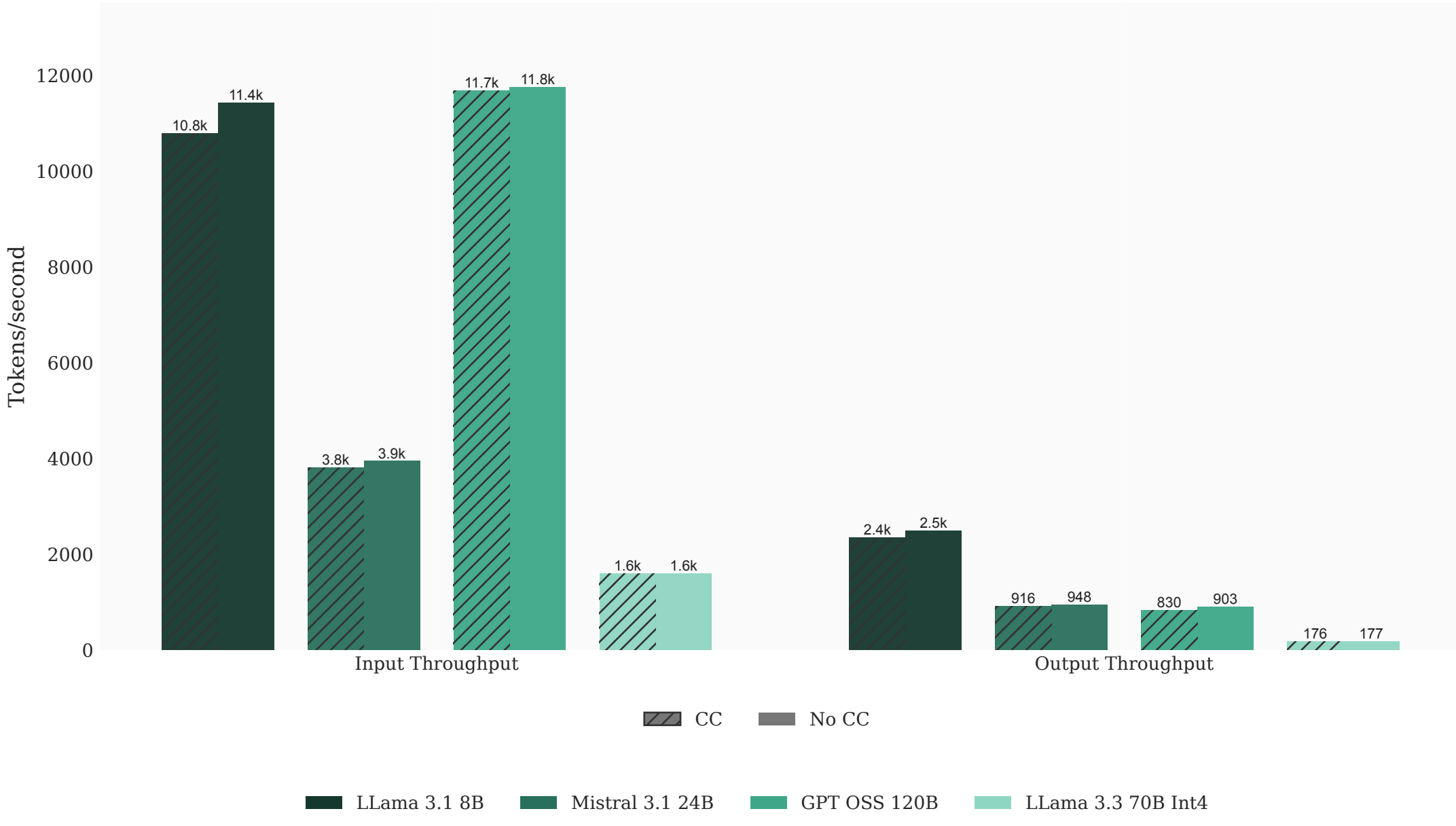
Random (1500 ⇒ 250) (Rate 50)



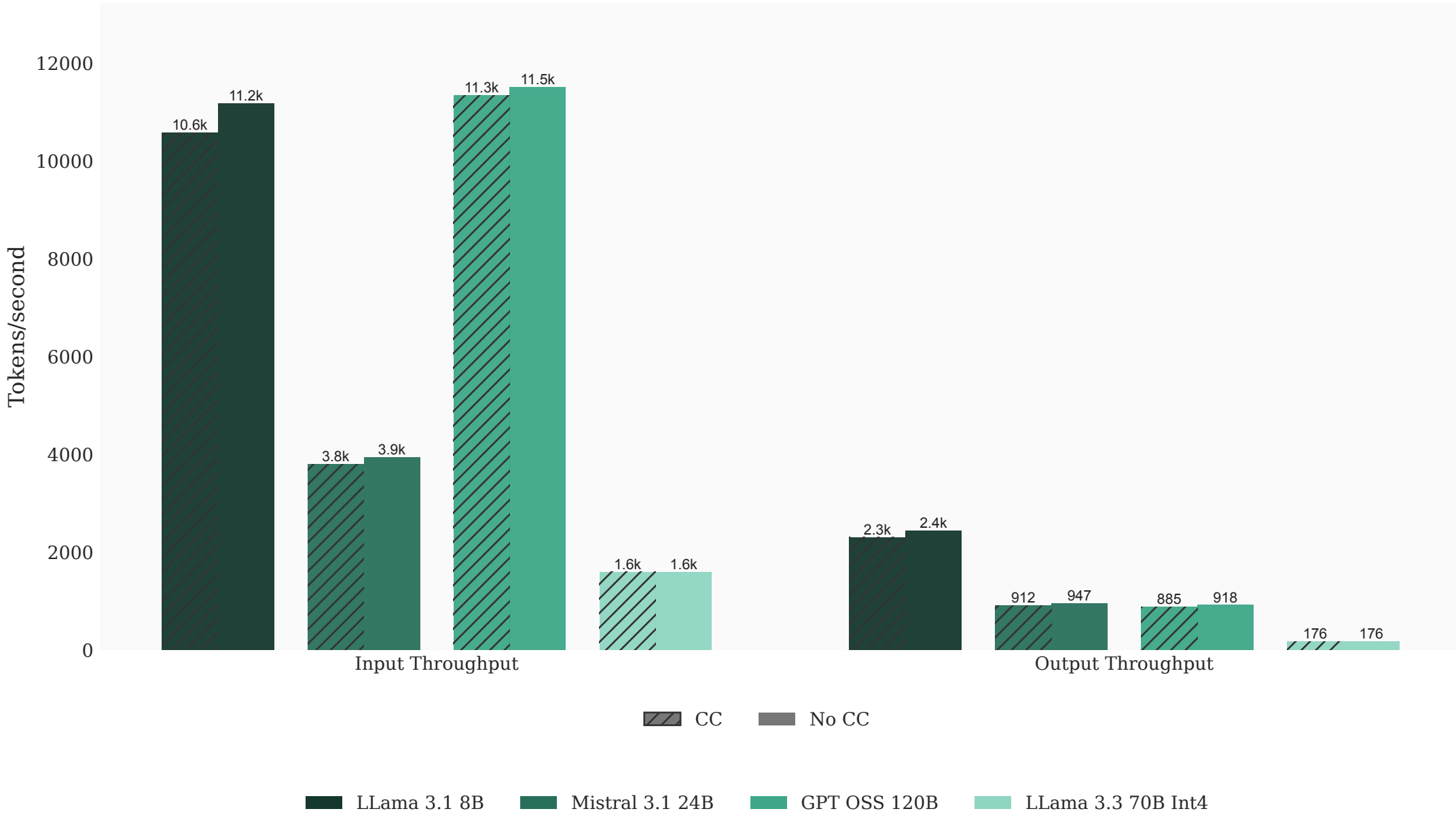
Random (1500 ⇒ 250) (Rate 1)



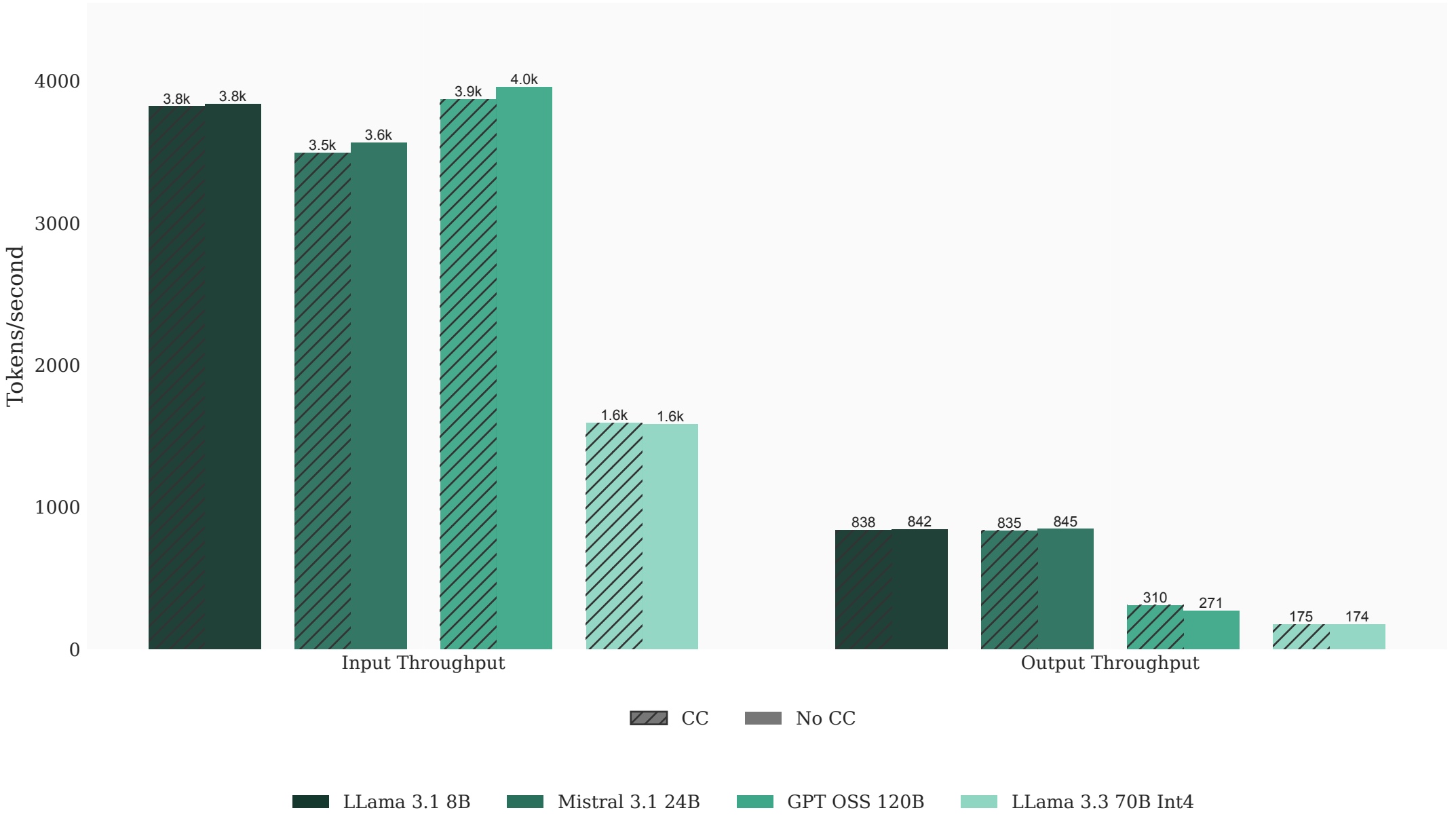
Random (4000 ⇒ 1000) (Rate 100)



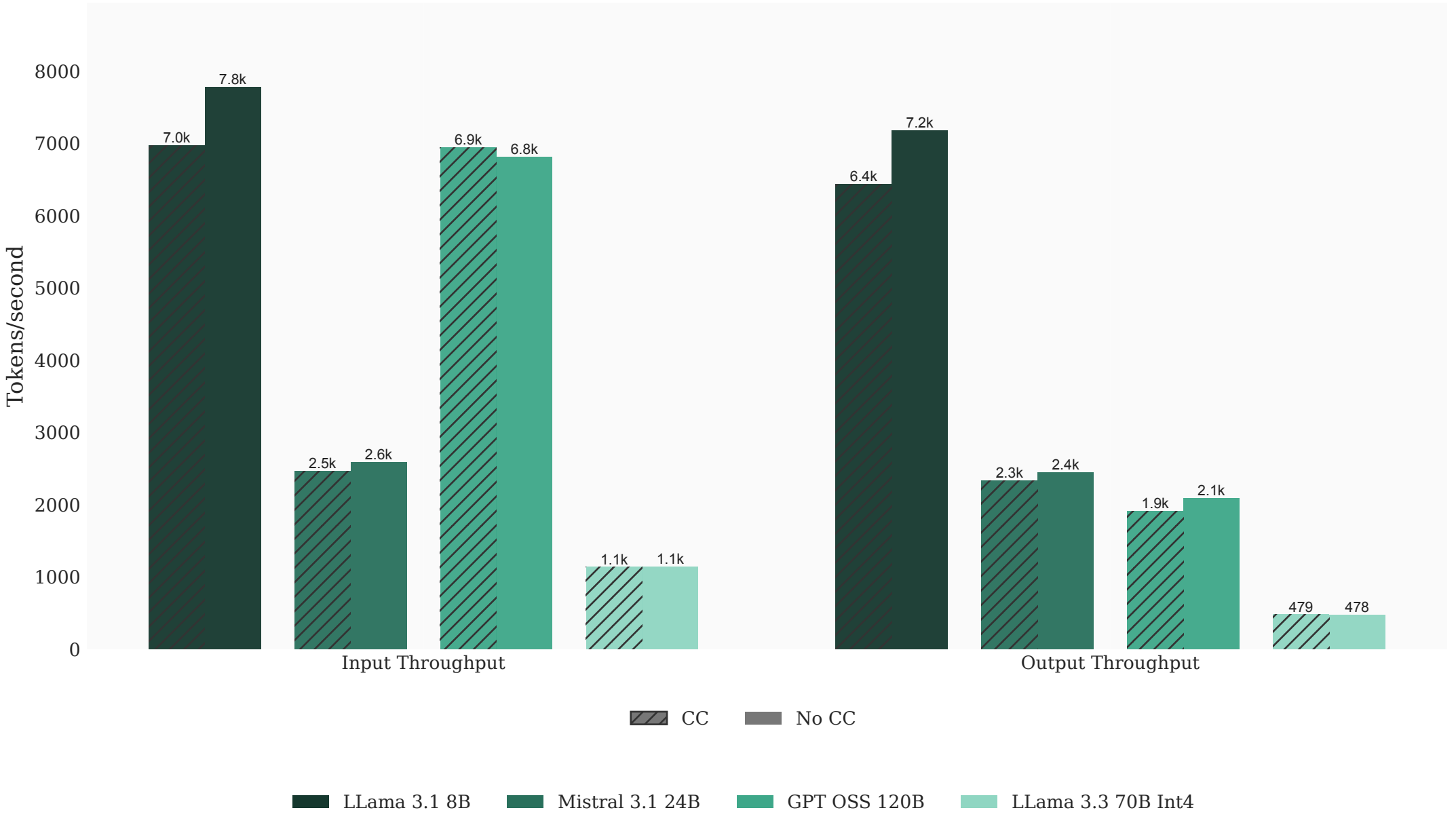
Random (4000 ⇒ 1000) (Rate 50)

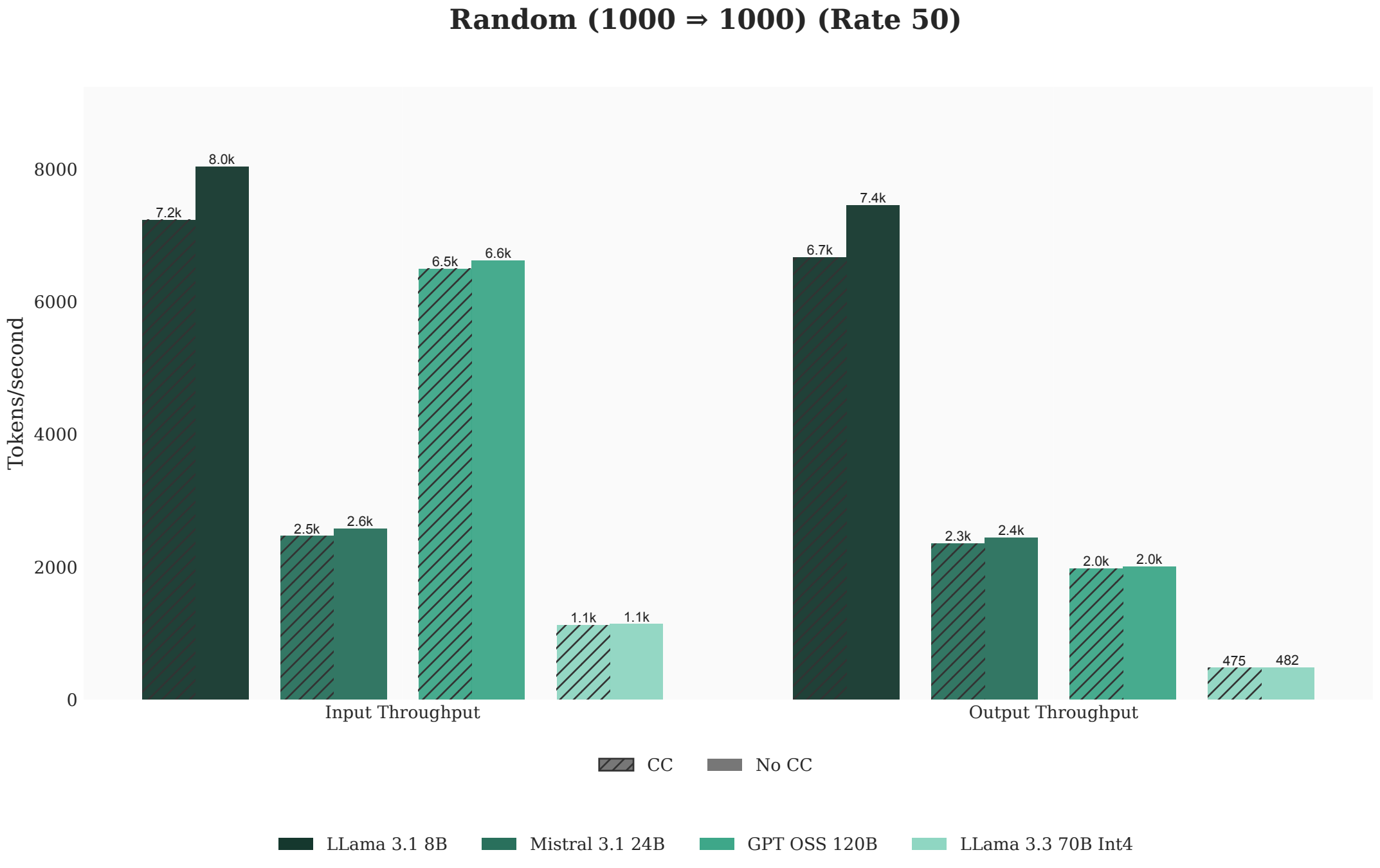


Random (4000 ⇒ 1000) (Rate 1)

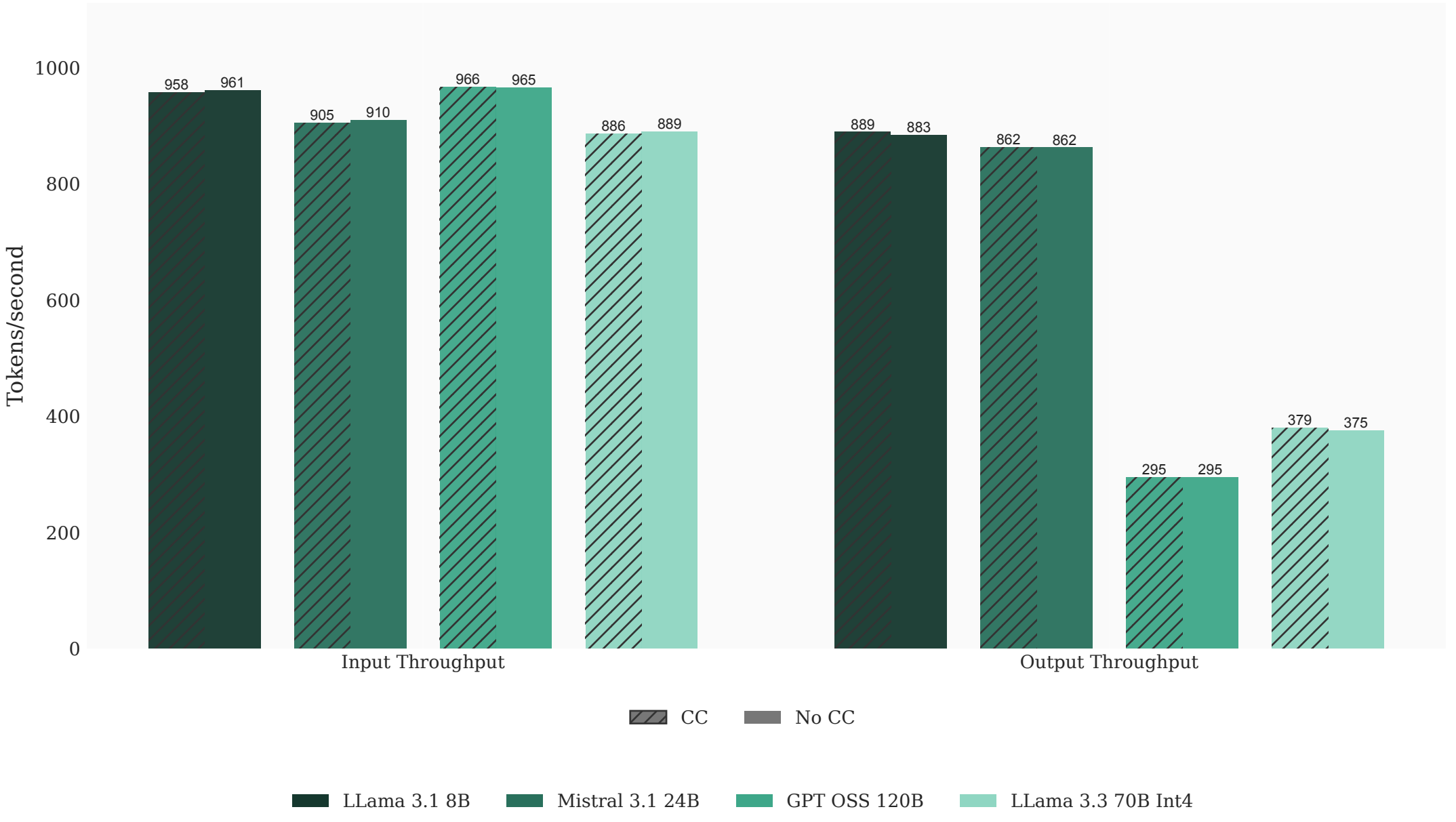


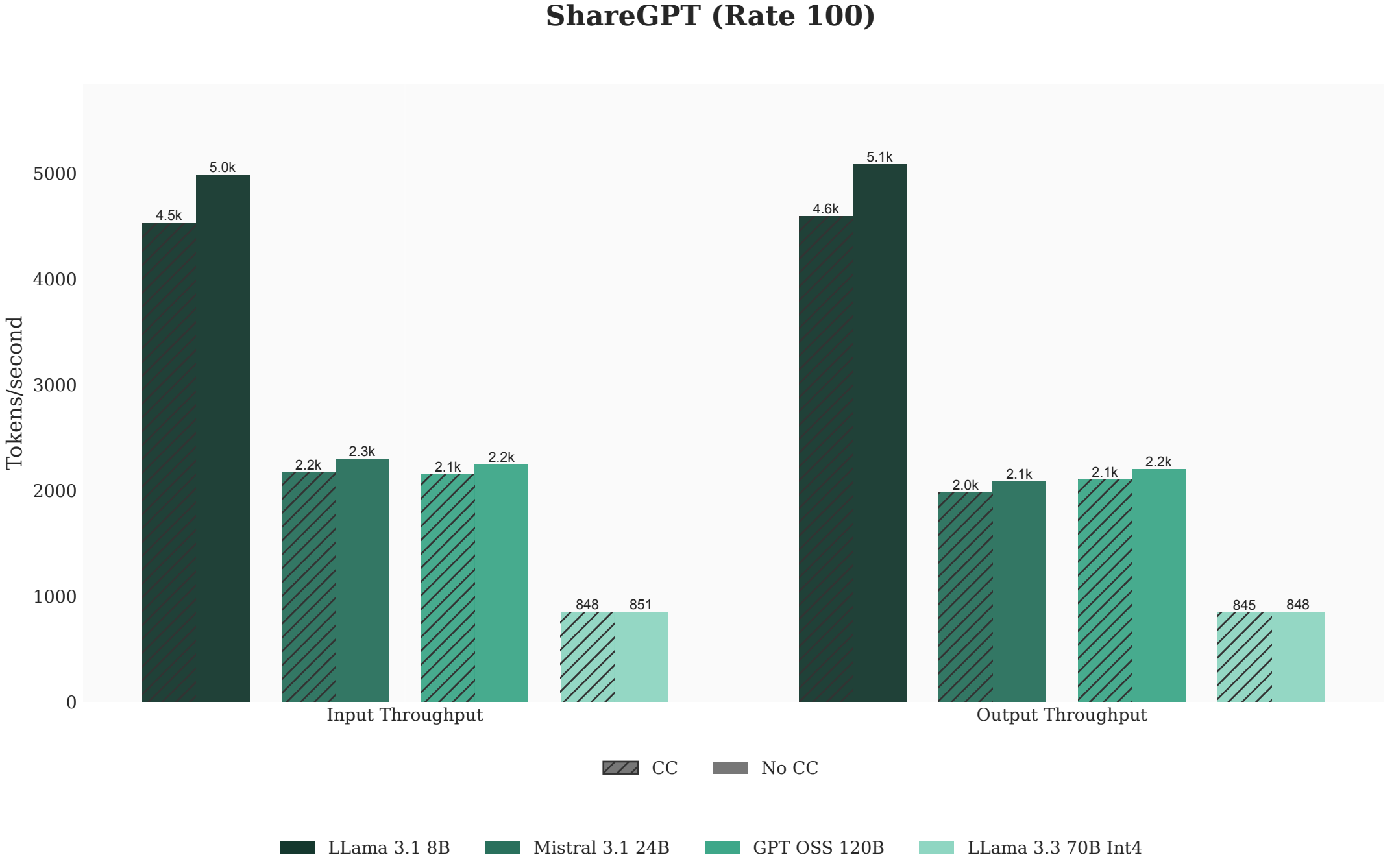
Random (1000 ⇒ 1000) (Rate 100)

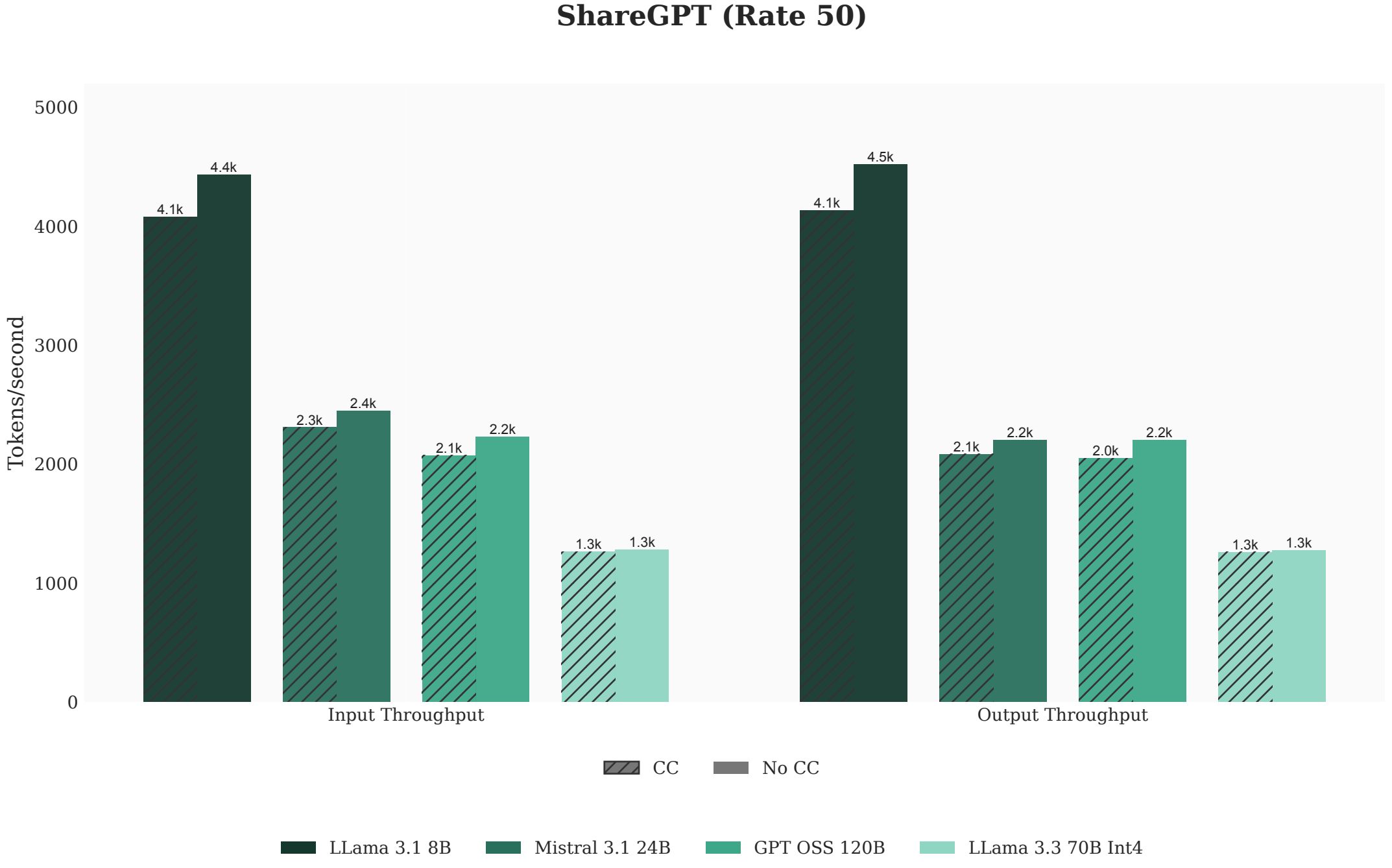


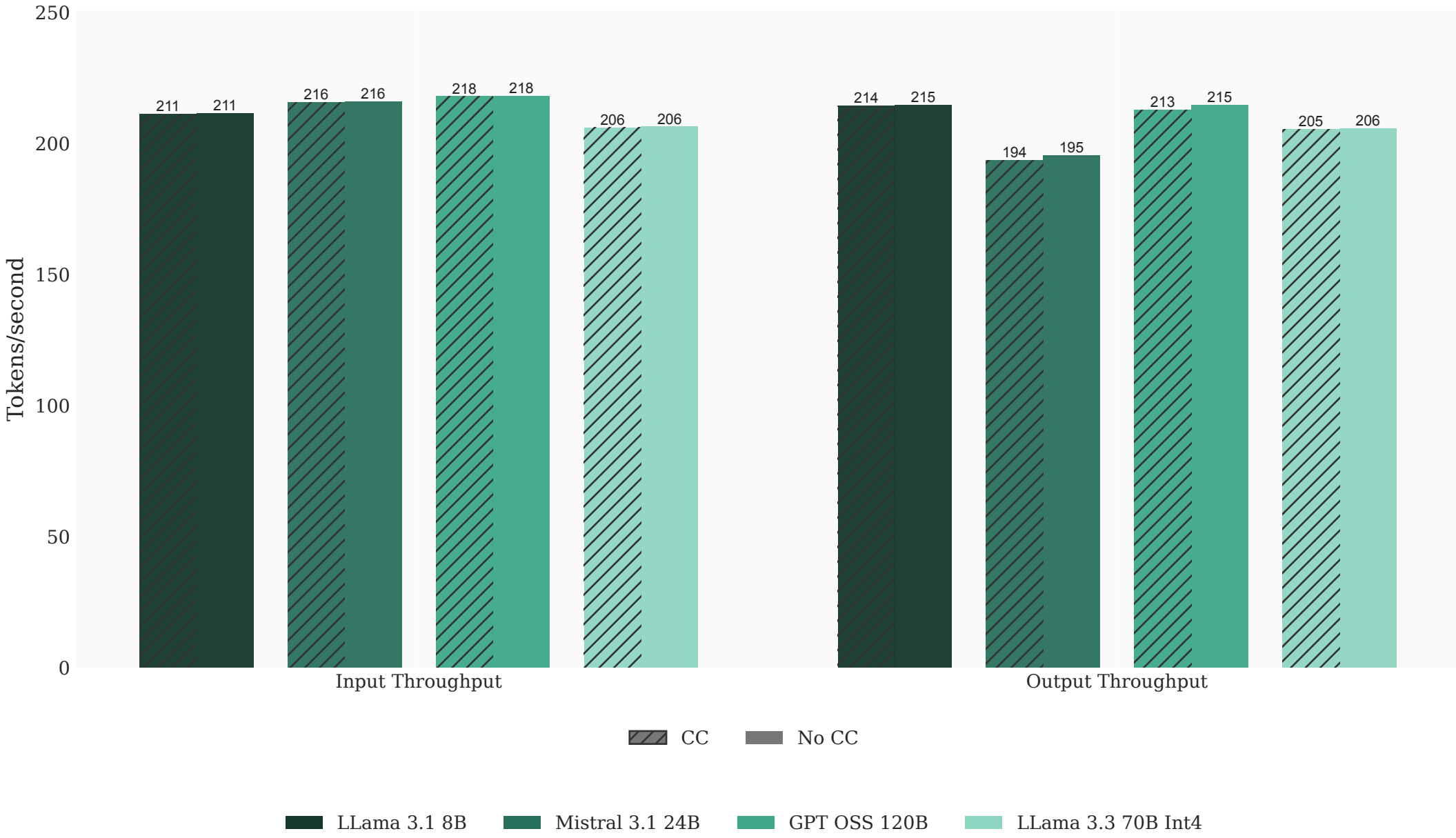

Random (1000 ⇒ 1000) (Rate 50)
Tokens/second
0
2000
4000
6000
8000
7.2k
8.0k
2.5k
2.6k
6.5k
6.6k
1.1k
1.1k
6.7k
7.4k
2.3k
2.4k
2.0k
2.0k
475
482
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

Random (1000 ⇒ 1000) (Rate 1)

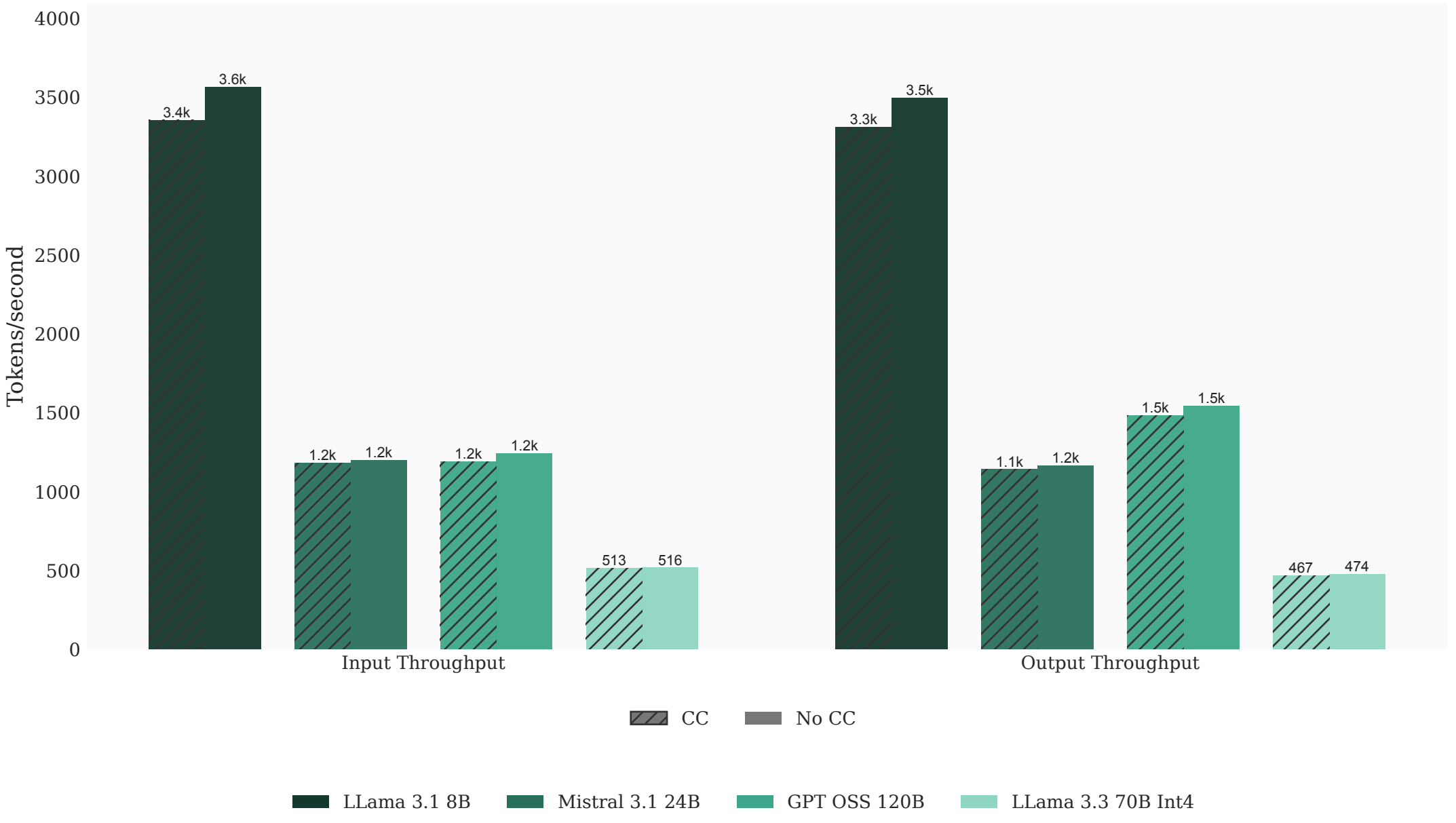



ShareGPT (Rate 100)
Tokens/second
0
1000
2000
3000
4000
5000
4.5k
5.0k
2.2k
2.3k
2.1k
2.2k
848
851
4.6k
5.1k
2.0k
2.1k
2.1k
2.2k
845
848
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4

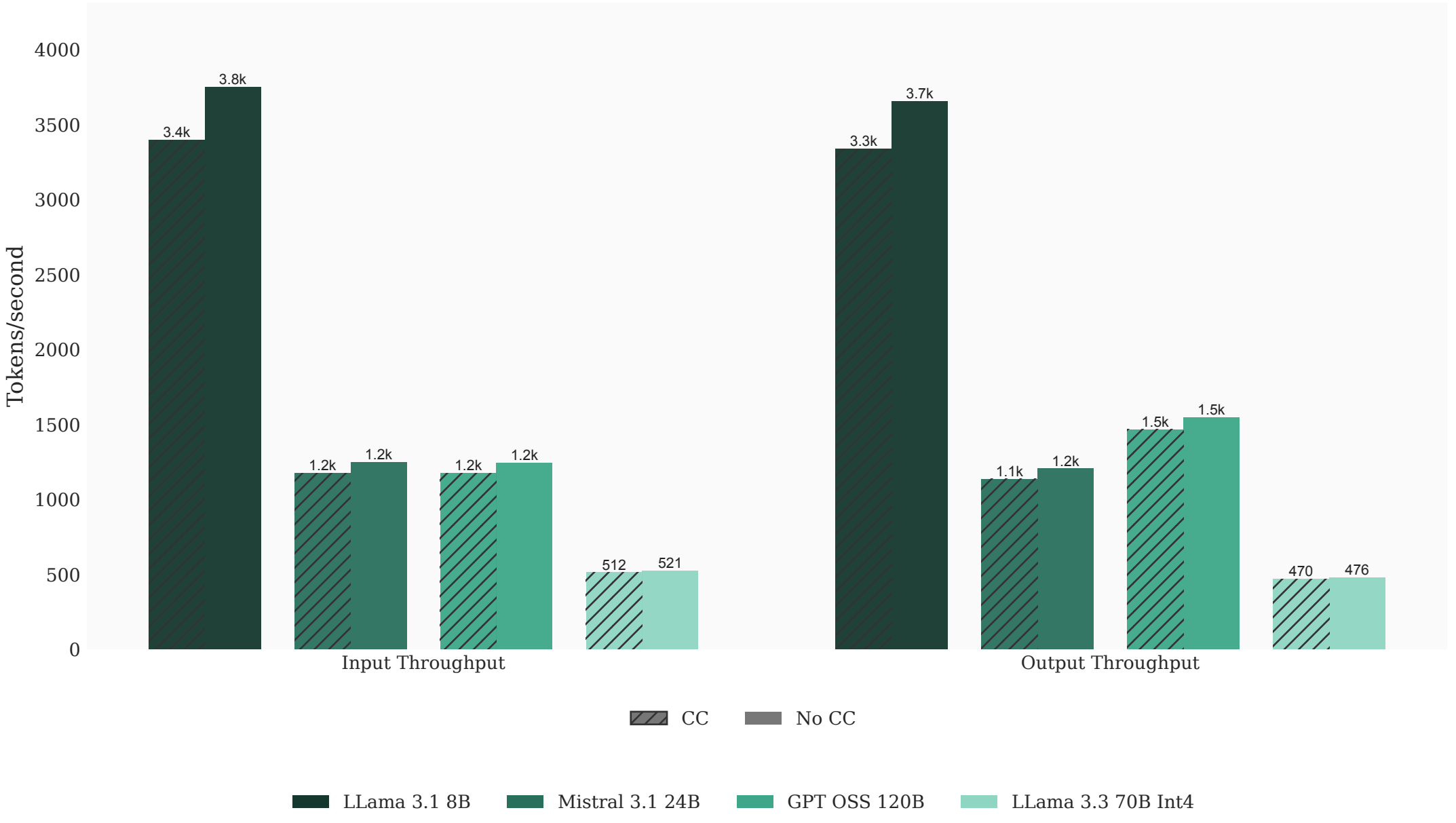

ShareGPT (Rate 50)
Tokens/second
5000
4000
3000
2000
1000
0
4.1k
4.4k
2.3k
2.4k
2.1k
2.2k
1.3k
1.3k
4.1k
4.5k
2.1k
2.2k
2.0k
2.2k
1.3k
1.3k
Input Throughput
Output Throughput
CC
No CC
LLama 3.1 8B
Mistral 3.1 24B
GPT OSS 120B
LLama 3.3 70B Int4


ShareGPT (Rate 1)
250
200
150
100
50
0
Tokens/second
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

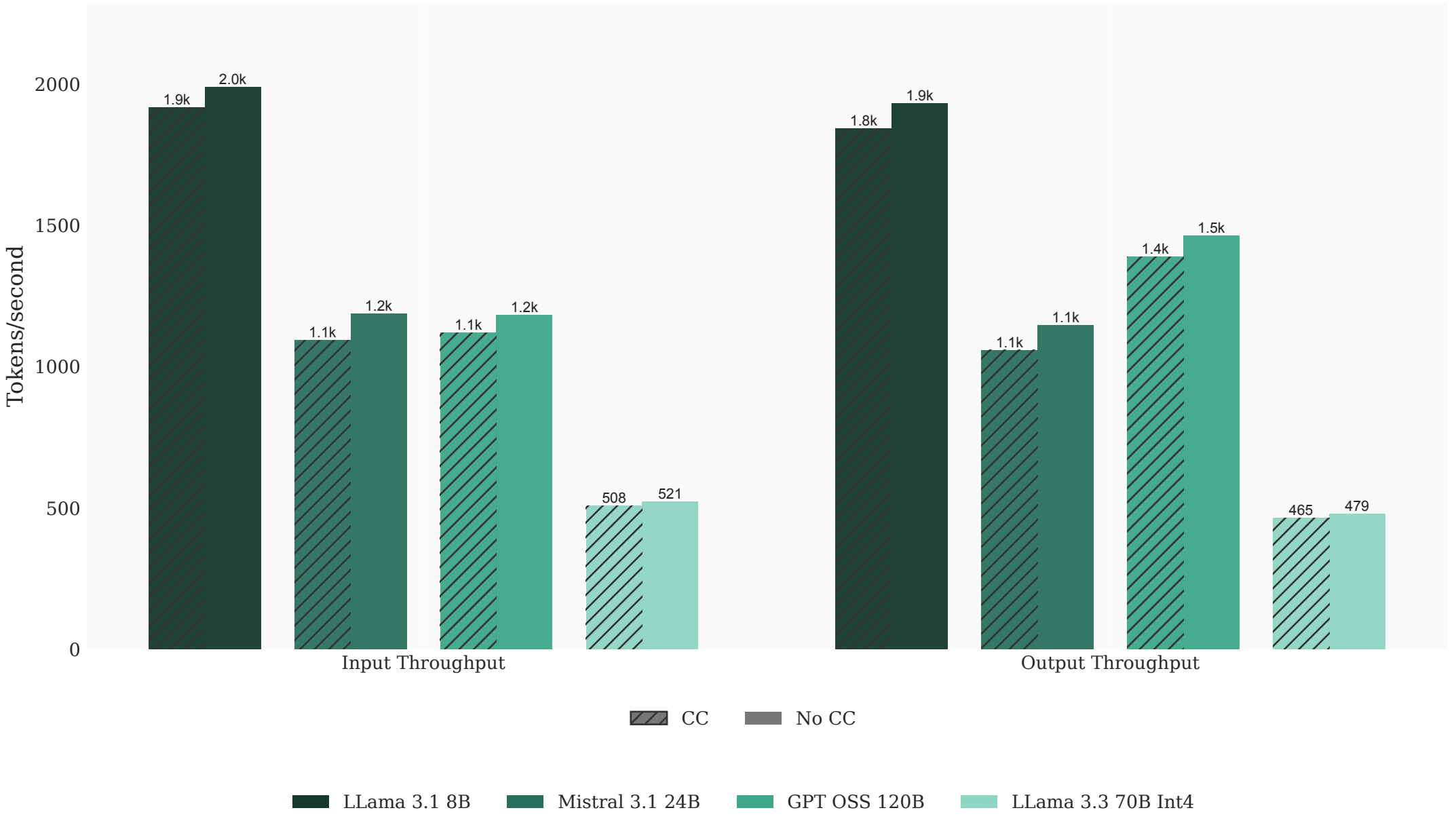
Edit 10K Characters (Rate 100)



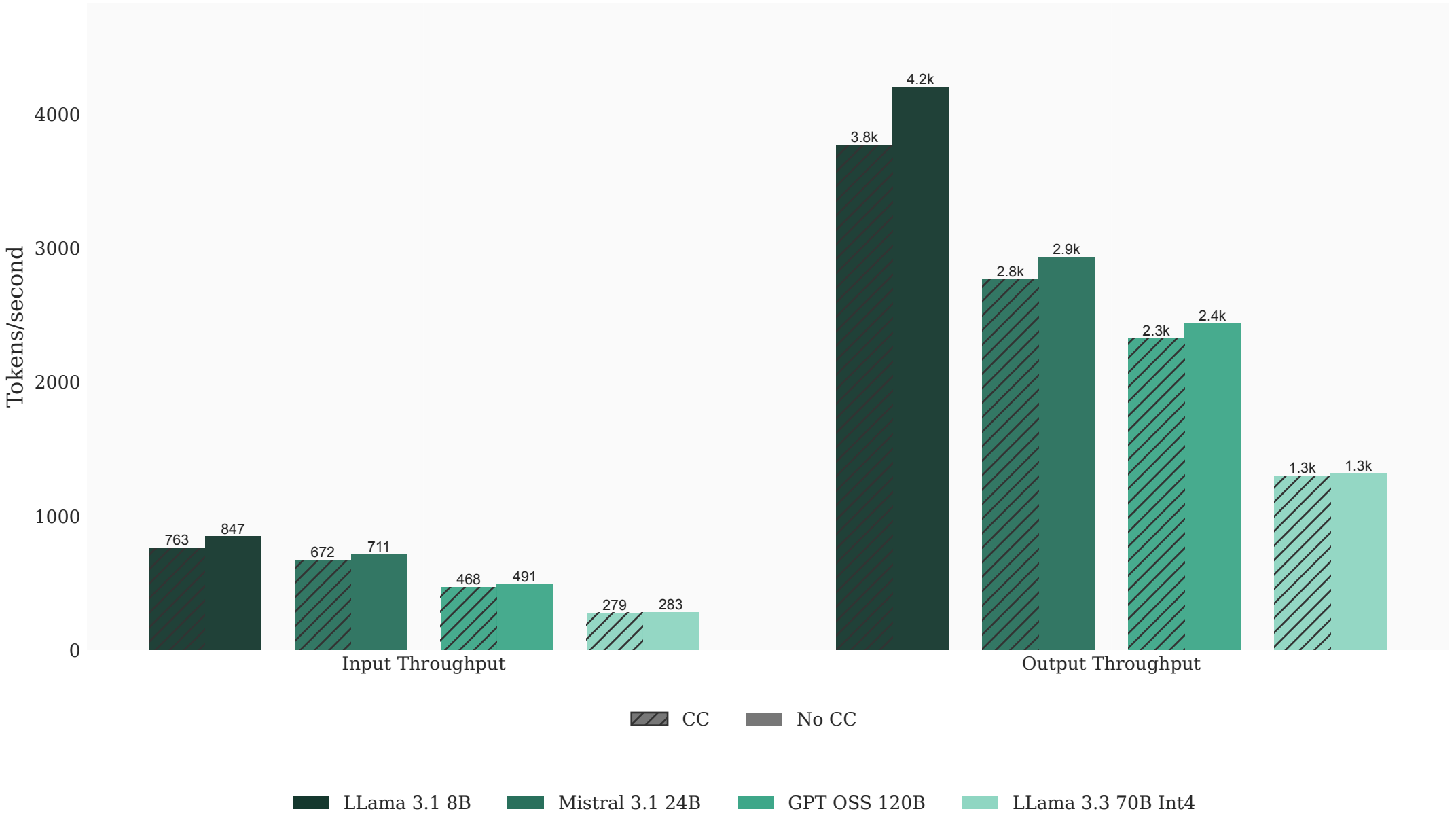
Edit 10K Characters (Rate 50)



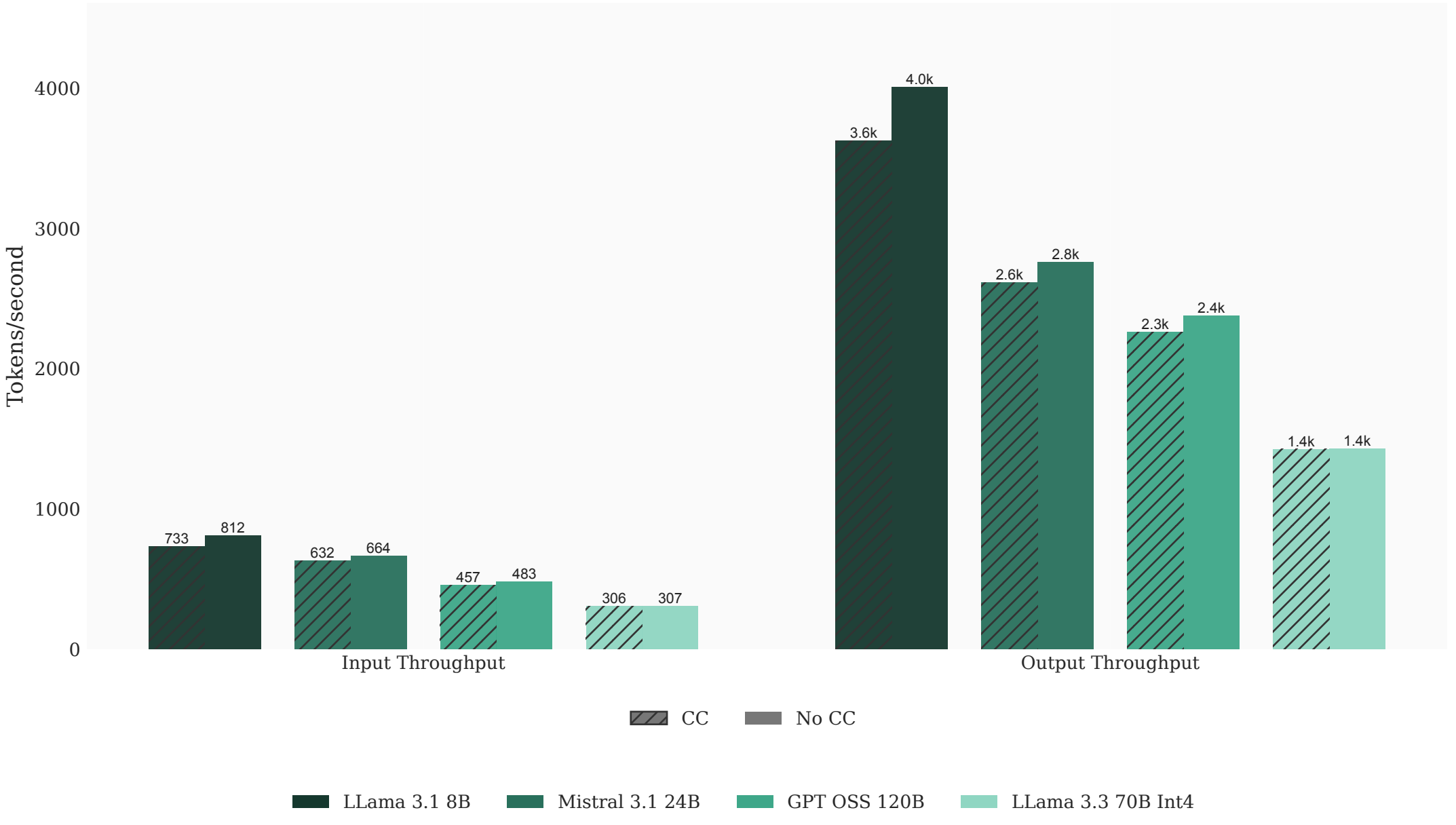
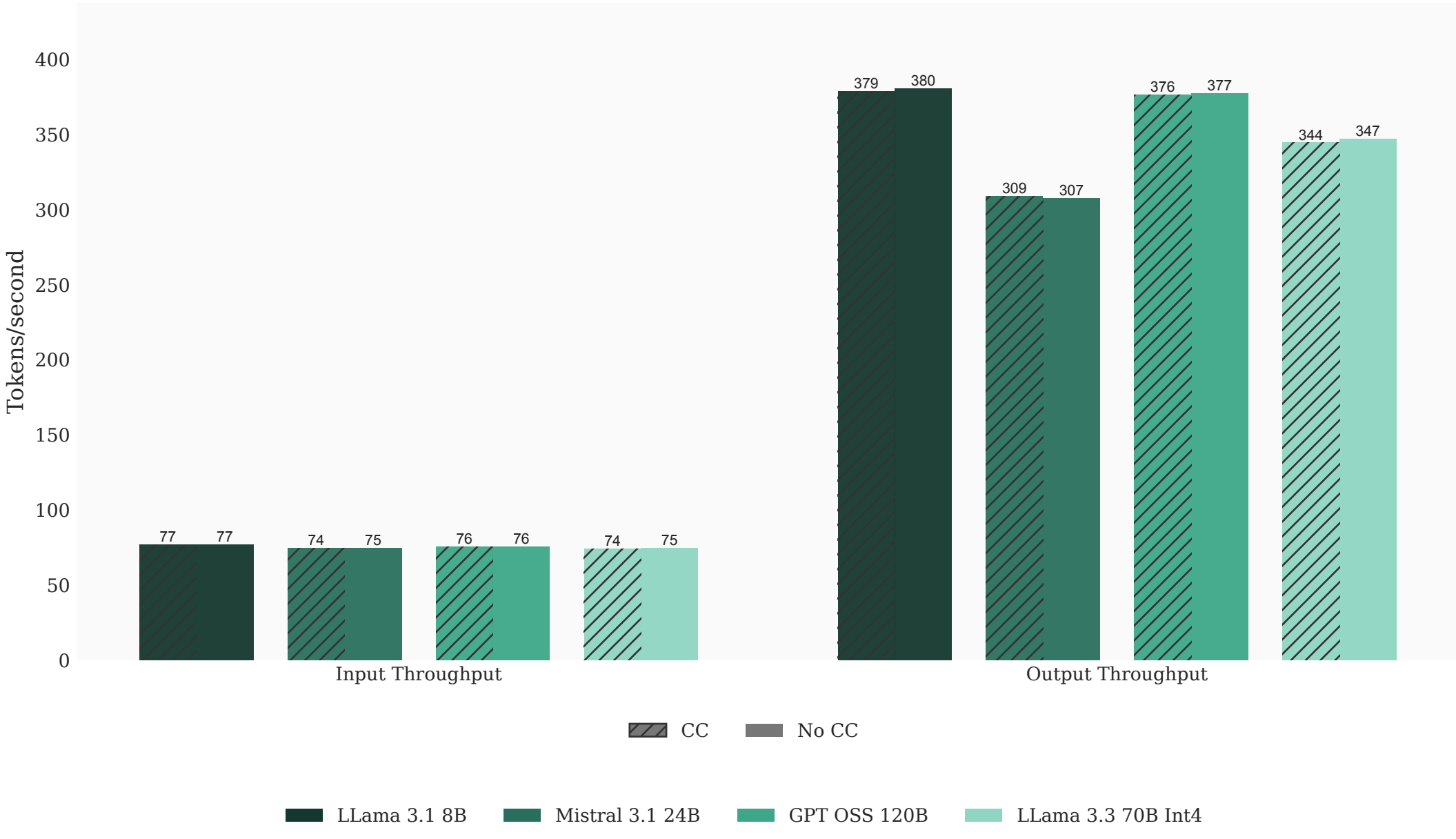
Edit 10K Characters (Rate 1)



Numina Math (Rate 100)



Numina Math (Rate 50)



Numina Math (Rate 1)