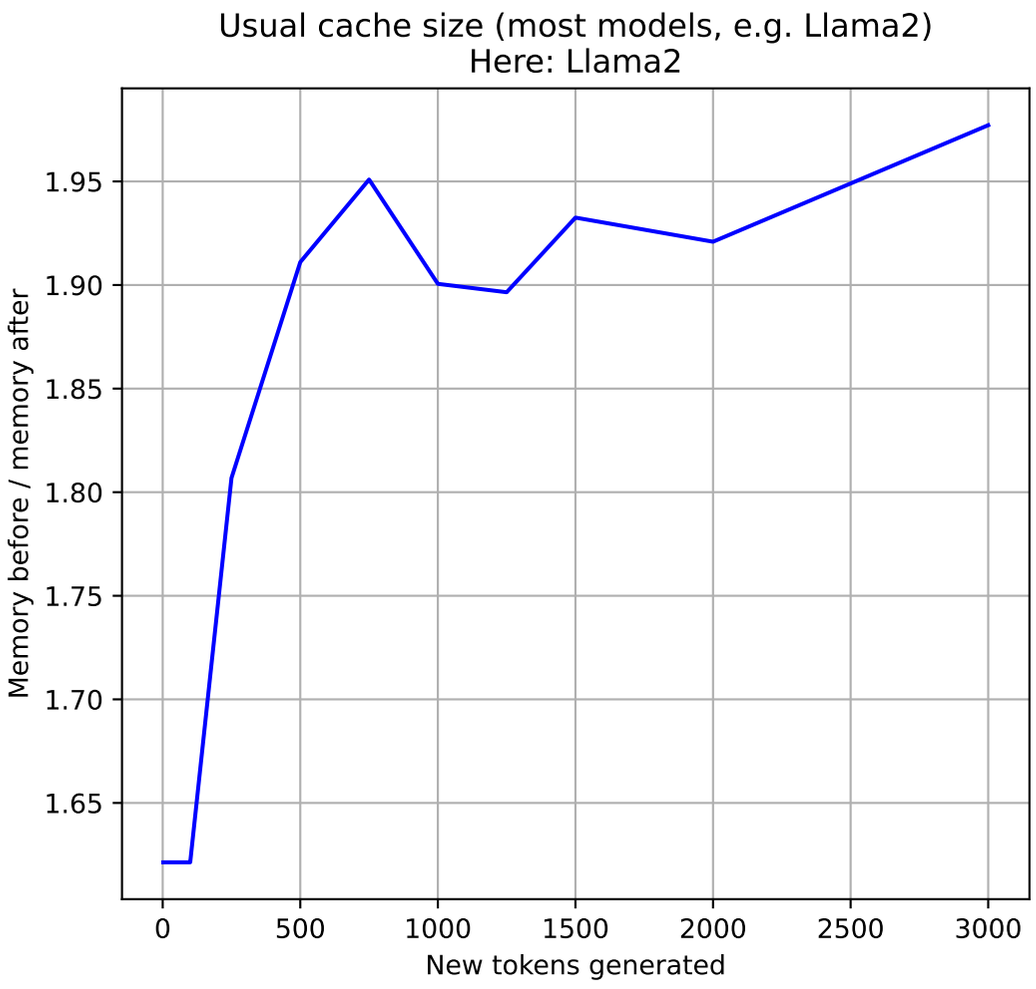
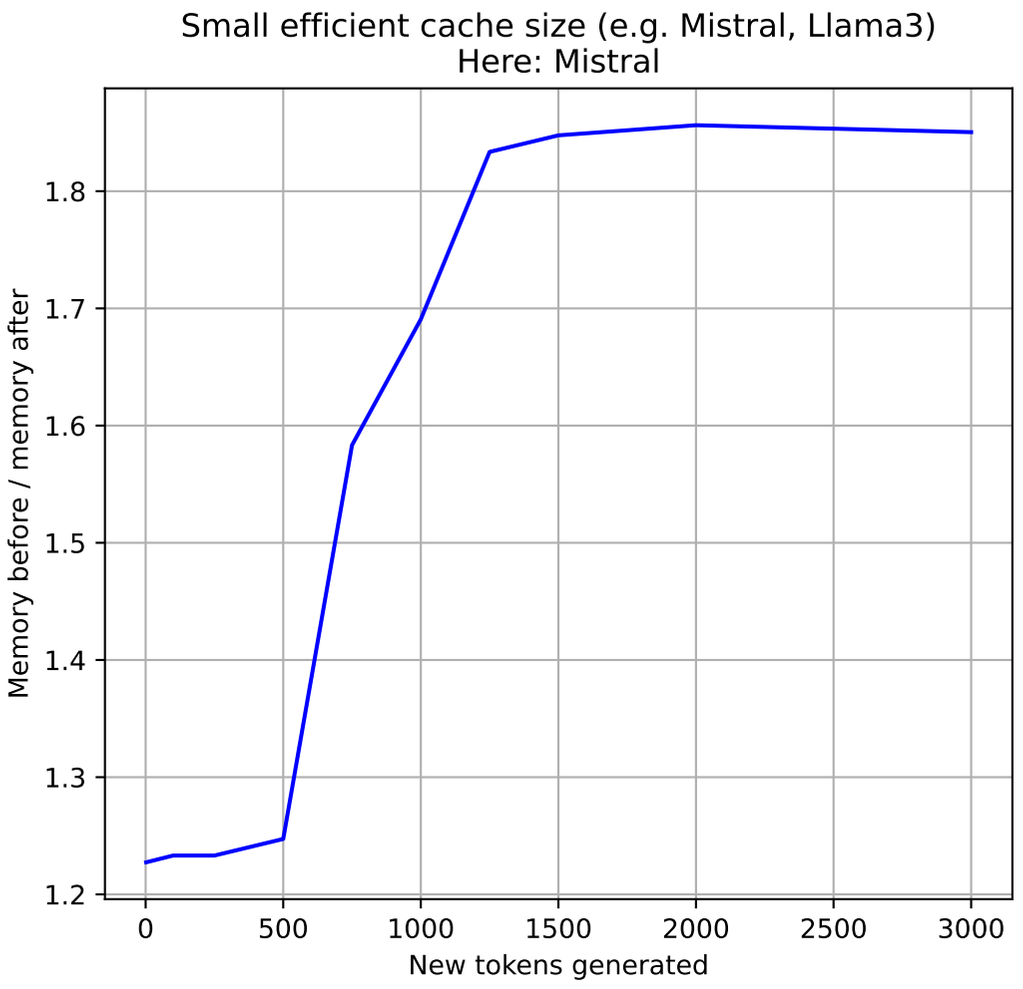

Usual cache size (most models, e.g. Llama2)
Here: Llama2
Memory before / memory after
1.95
1.90
1.85
1.80
1.75
1.70
1.65
0
500
1000
1500
2000
2500
3000
New tokens generated


Small efficient cache size (e.g. Mistral, Llama3)
Here: Mistral
Memory before / memory after
1.8
1.7
1.6
1.5
1.4
1.3
1.2
0
500
1000
1500
2000
2500
3000
New tokens generated