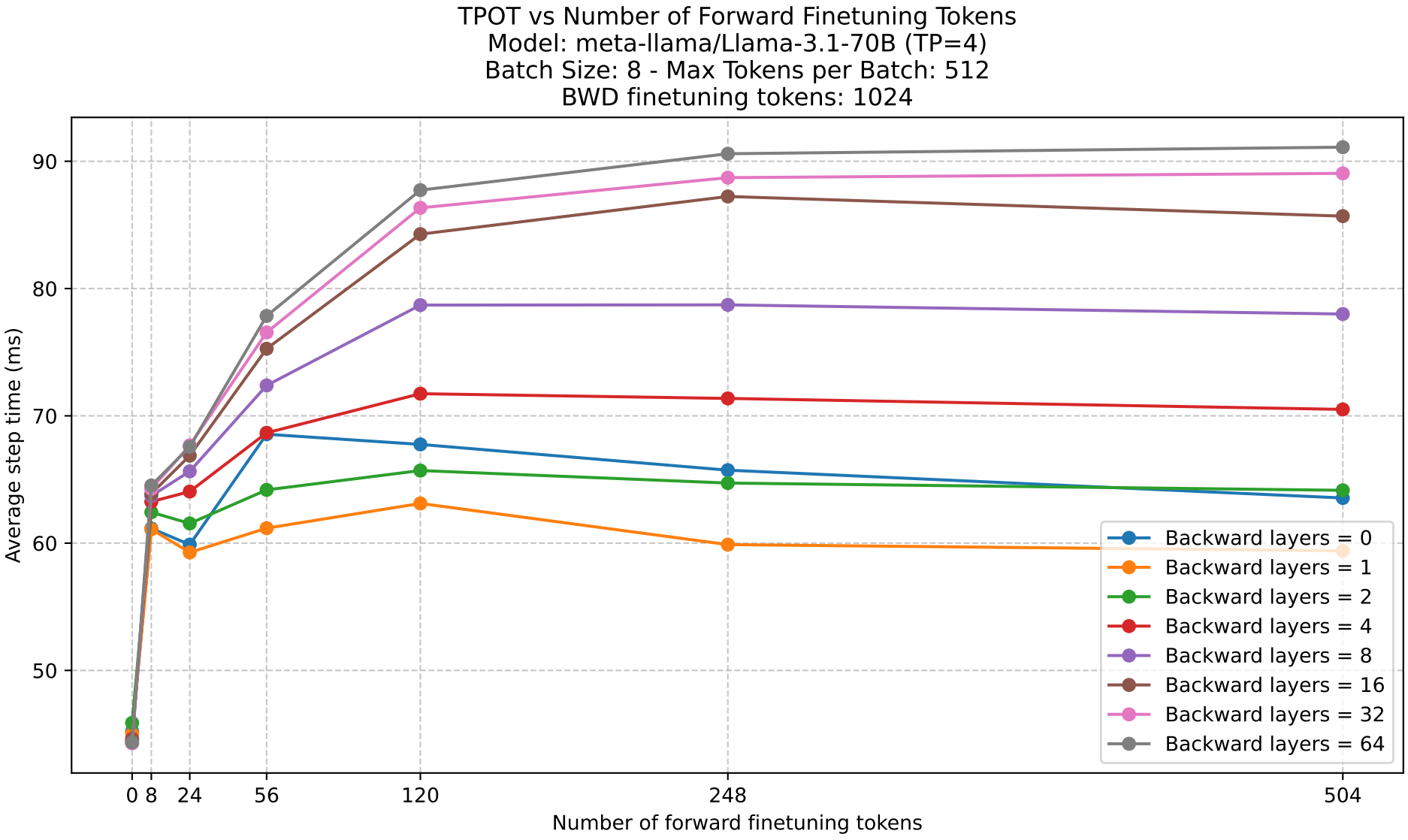
TPOT vs Number of Forward Finetuning Tokens
Model: meta-llama/Llama-3.1-70B (TP=4)
Batch Size: 8 - Max Tokens per Batch: 512
BWD finetuning tokens: 1024
Average step time (ms)
90
80
70
60
50
0
8
24
56
120
248
504
Number of forward finetuning tokens
Backward layers = 0
Backward layers = 1
Backward layers = 2
Backward layers = 4
Backward layers = 8
Backward layers = 16
Backward layers = 32
Backward layers = 64