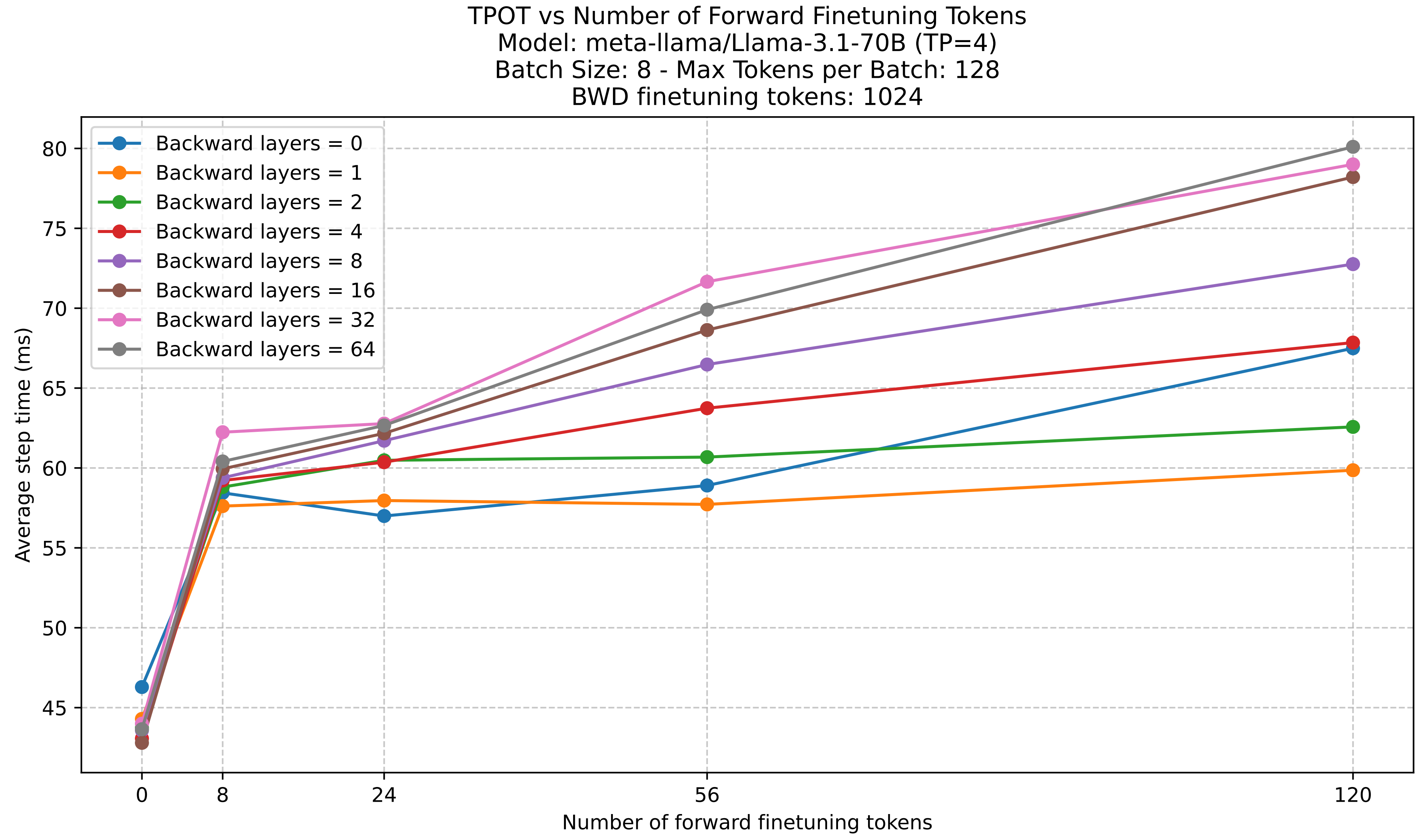
TPOT vs Number of Forward Finetuning Tokens
Model: meta-llama/Llama-3.1-70B (TP=4)
Batch Size: 8 - Max Tokens per Batch: 128
BWD finetuning tokens: 1024
Backward layers = 0
Backward layers = 1
Backward layers = 2
Backward layers = 4
Backward layers = 8
Backward layers = 16
Backward layers = 32
Backward layers = 64
Average step time (ms)
80
75
70
65
60
55
50
45
0
8
24
56
120
Number of forward finetuning tokens