

TPOT vs Number of Forward Finetuning Tokens
Model: meta-llama/Llama-3.1-70B (TP=4)
Batch Size: 8 - Max Tokens per Batch: 256
BWD finetuning tokens: 1024

