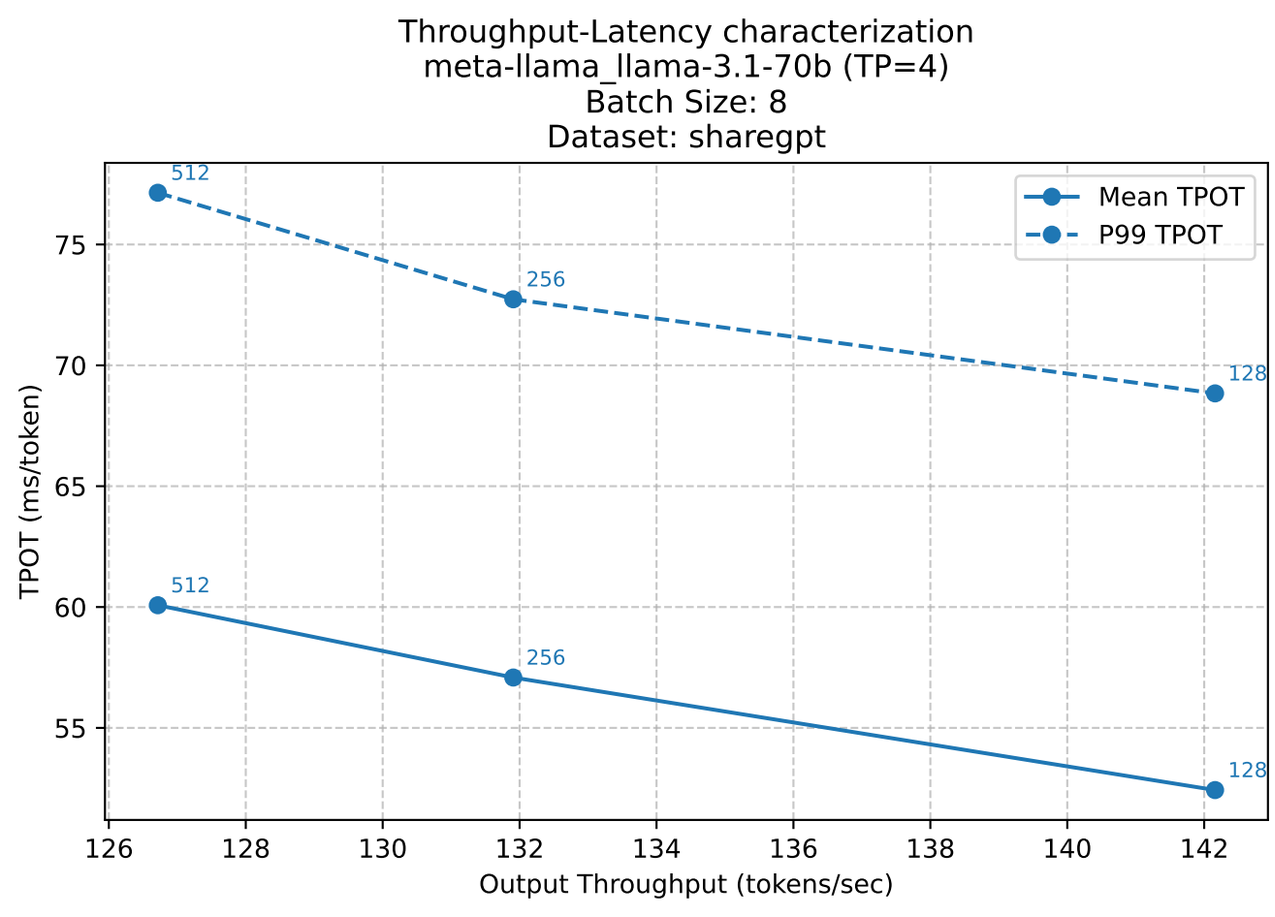
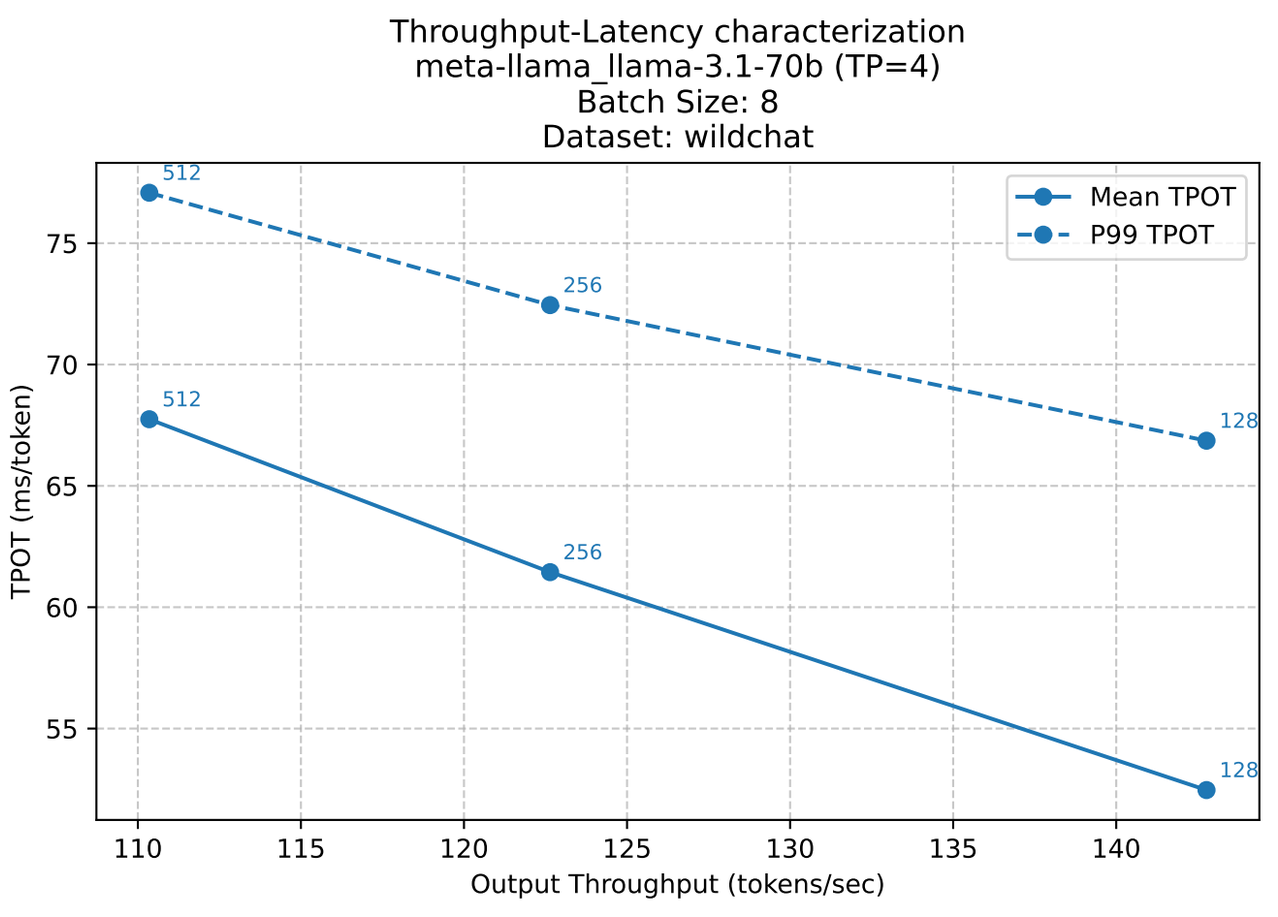

Throughput-Latency characterization
meta-llama_llama-3.1-70b (TP=4)
Batch Size: 8
Dataset: sharegpt
Mean TPOT
P99 TPOT
TPOT (ms/token)
Output Throughput (tokens/sec)
512
256
128
512
256
128
75
70
65
60
55
126
128
130
132
134
136
138
140
142


Throughput-Latency characterization
meta-llama_llama-3.1-70b (TP=4)
Batch Size: 8
Dataset: wildchat
Mean TPOT
P99 TPOT
TPOT (ms/token)
Output Throughput (tokens/sec)
512
256
128
512
256
128
75
70
65
60
55
110
115
120
125
130
135
140