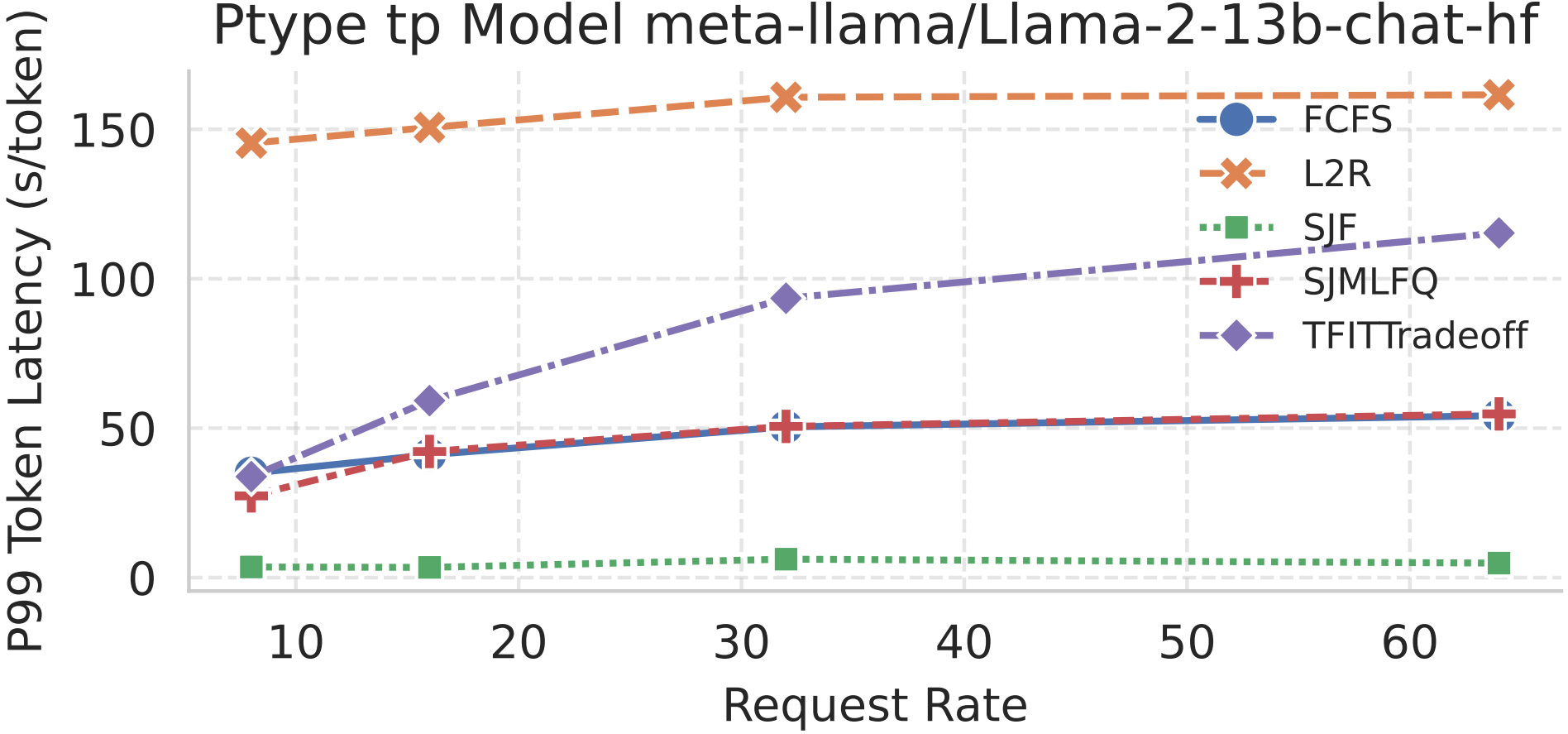
Ptype tp Model meta-llama/Llama-2-13b-chat-hf
P99 Token Latency (s/token)
150
100
50
0
10
20
30
40
50
60
Request Rate
FCFS
L2R
SJF
SJMLFQ
TFITTradeoff