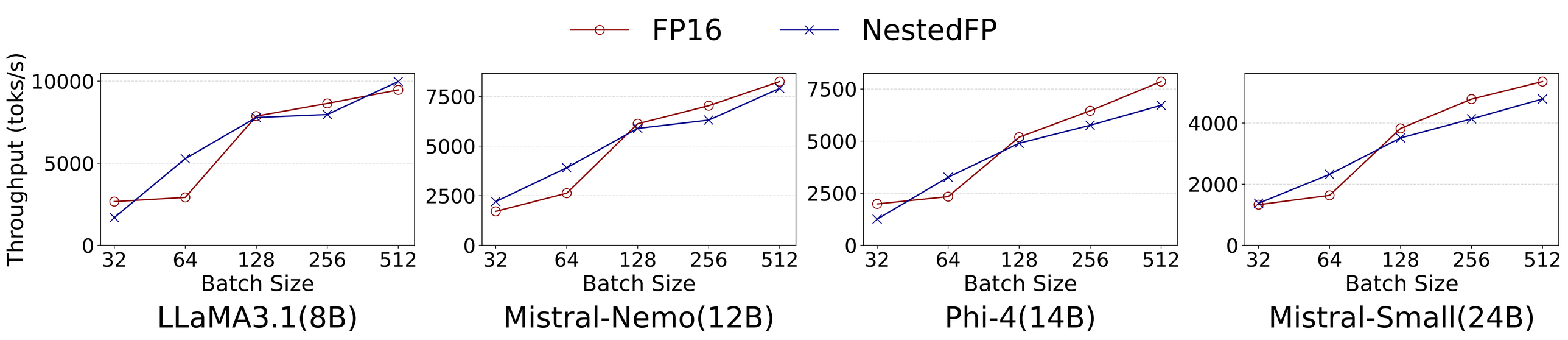
FP16
NestedFP
Throughput (toks/s)
10000
5000
0
7500
5000
2500
0
7500
5000
2500
0
4000
2000
0
32
64
128
256
512
Batch Size
LLaMA3.1(8B)
Mistral-Nemo(12B)
Phi-4(14B)
Mistral-Small(24B)