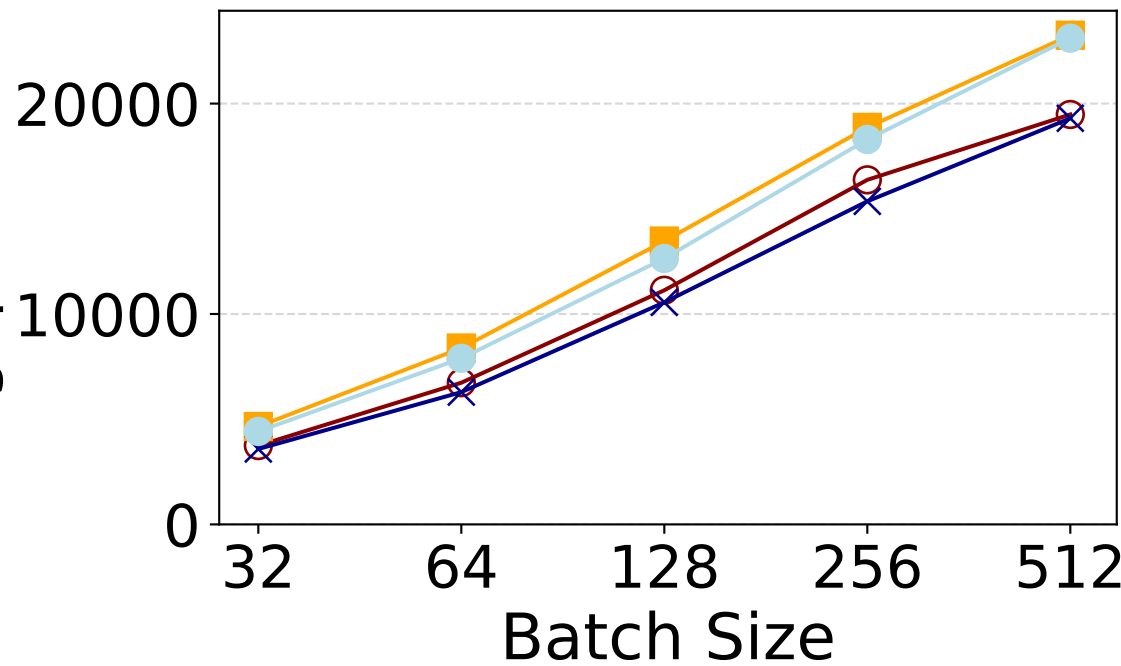
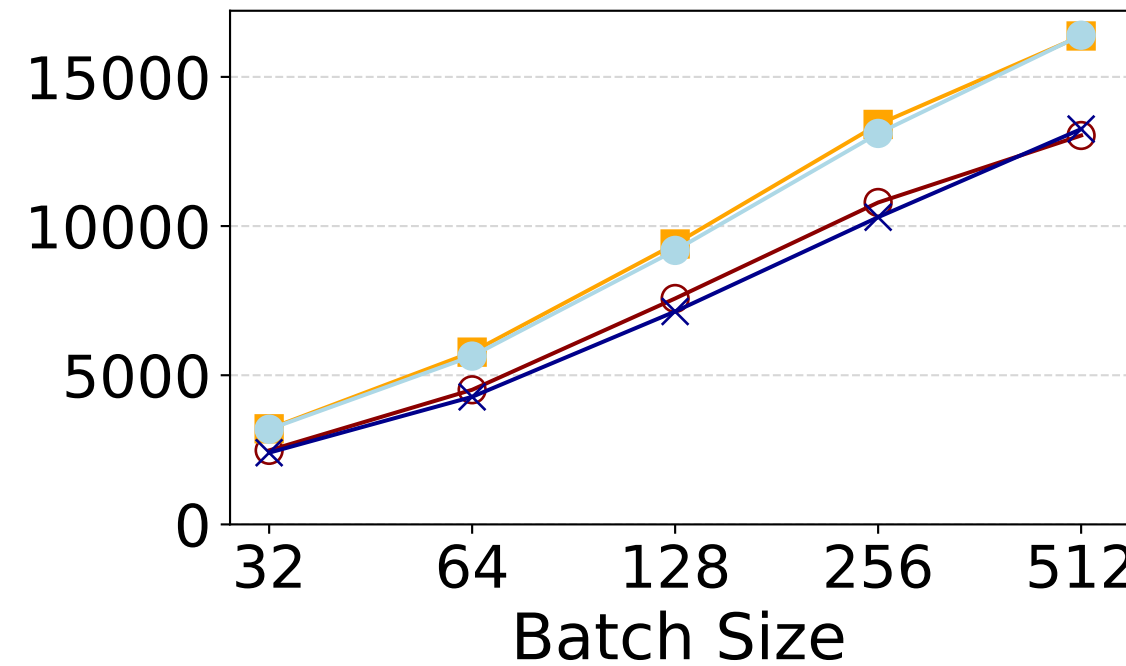


—○— Torch FP16    —■— Torch FP8    —×— NestedFP16    —●— NestedFP8

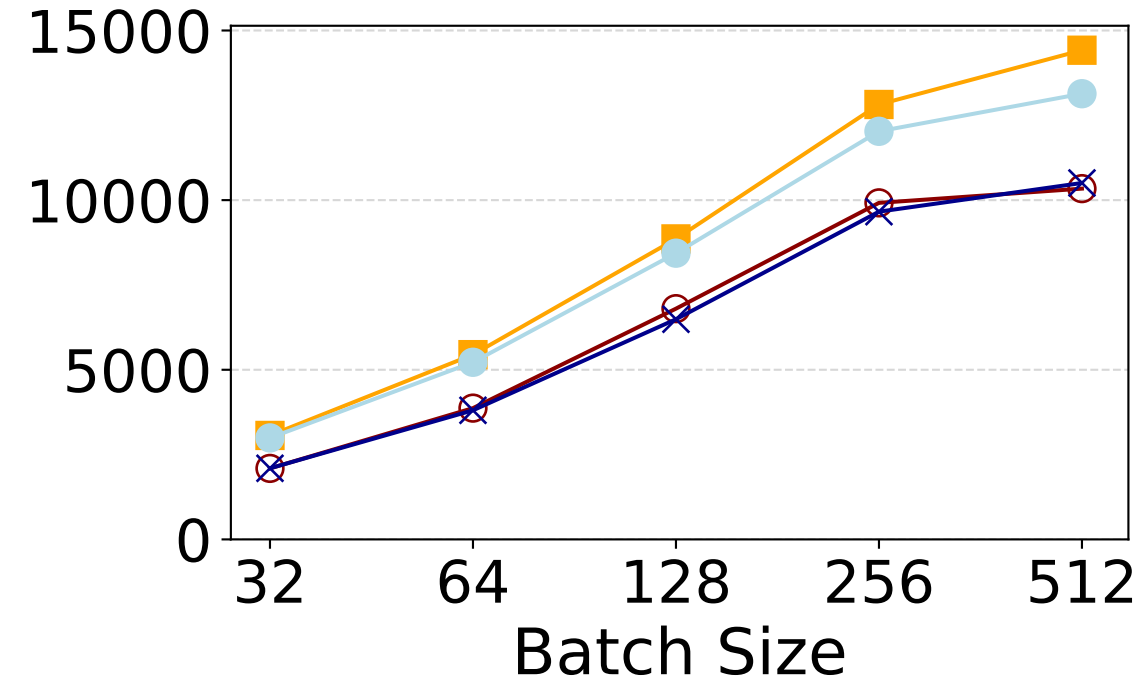
Throughput (toks/s)



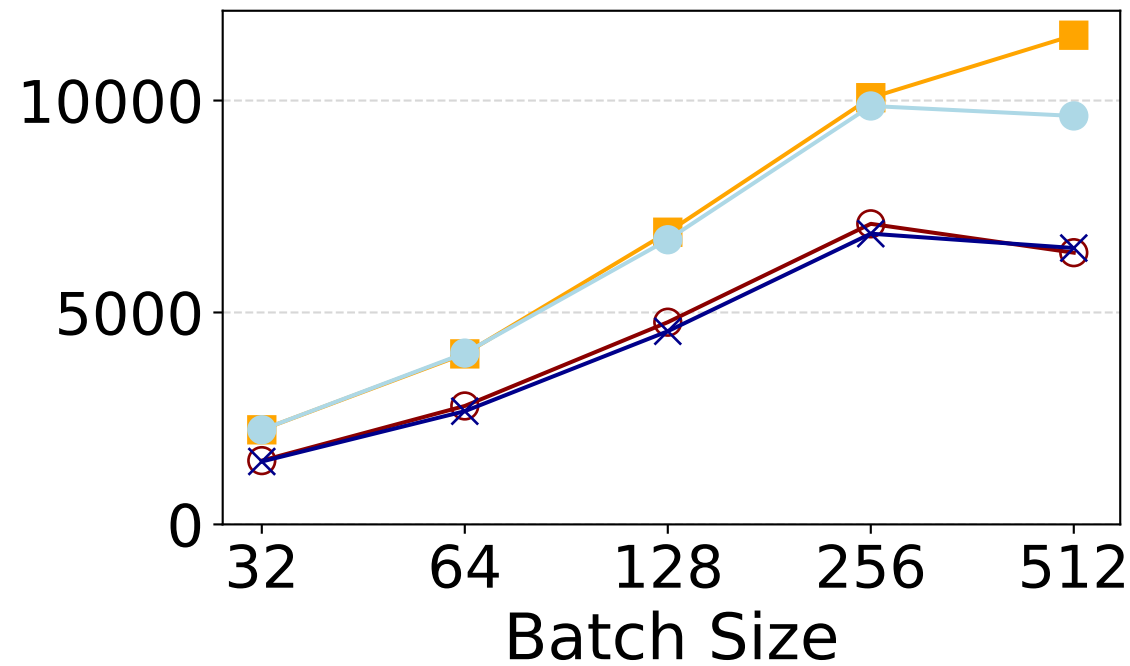
LLaMA3.1(8B)



Mistral-Nemo(12B)



Phi-4(14B)



Mistral-Small(24B)