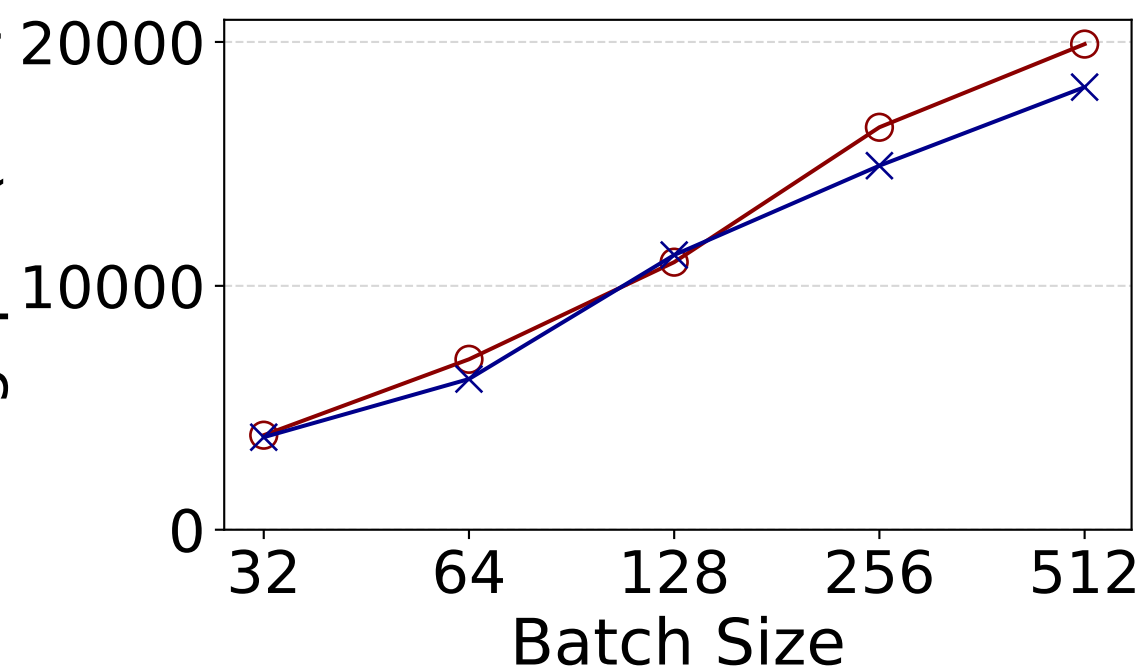
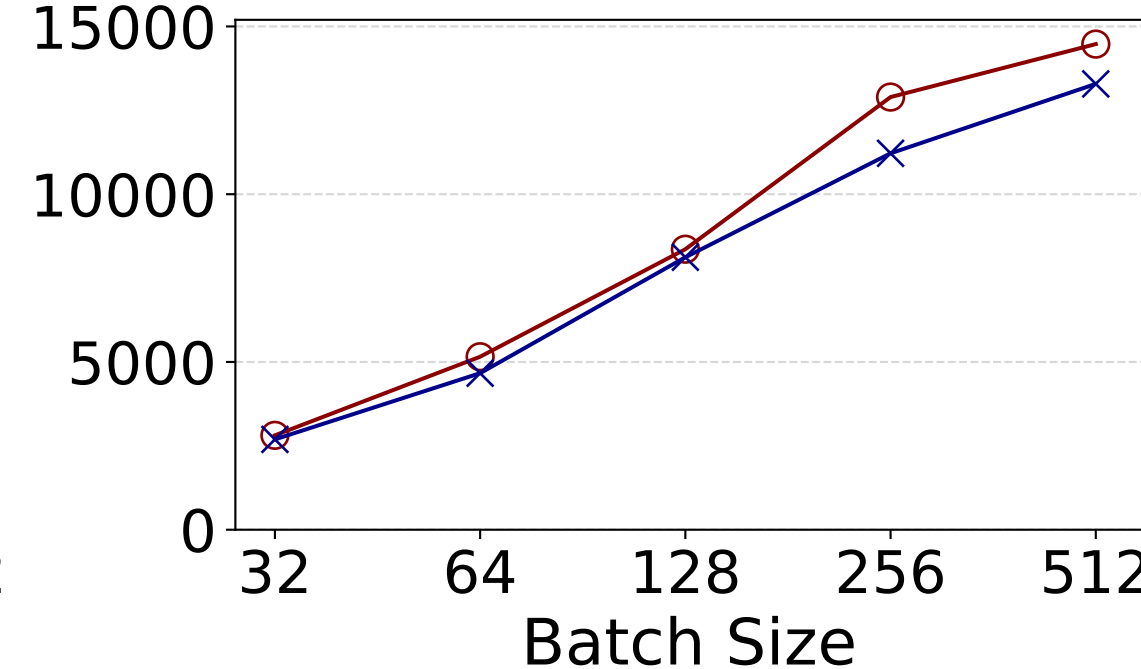


Throughput (toks/s)

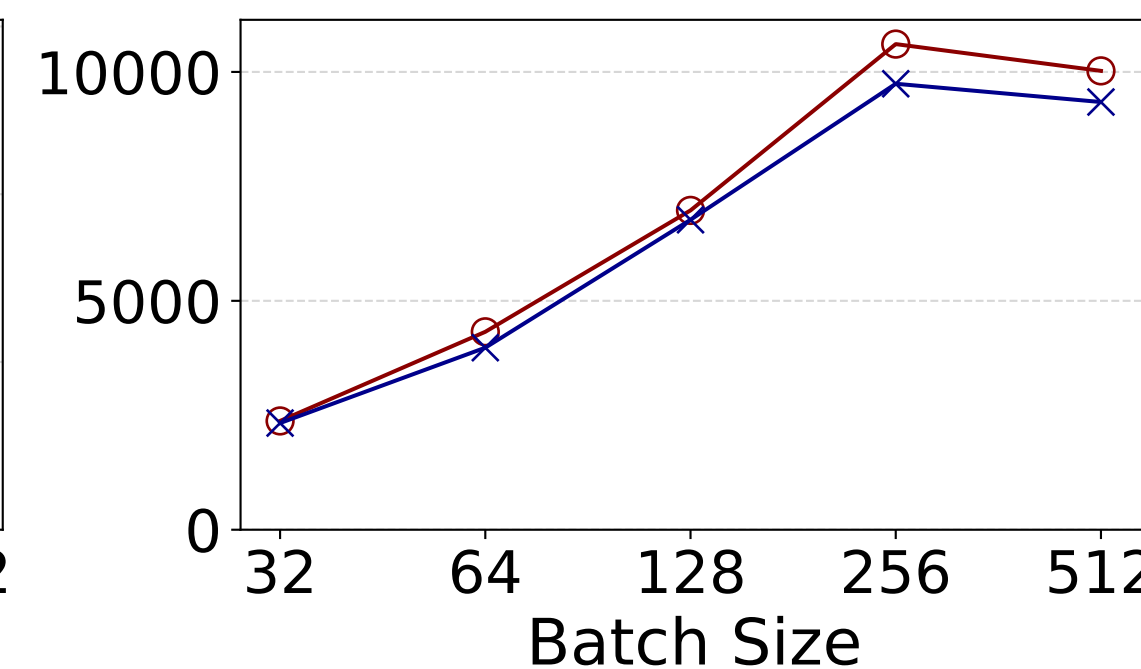
FP16 NestedFP



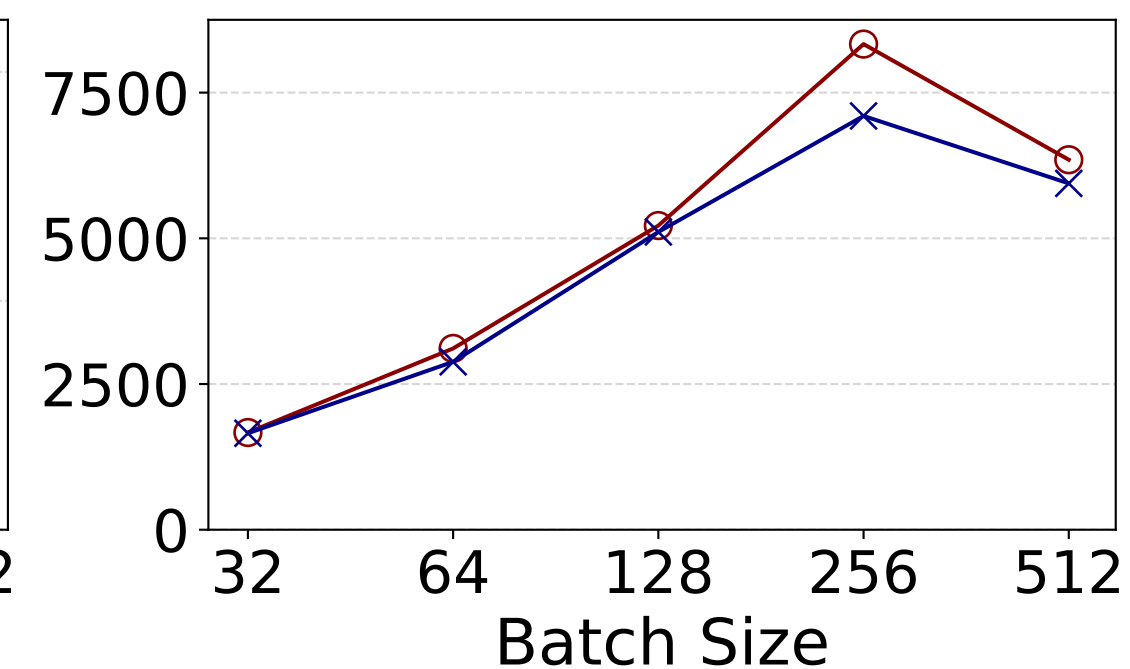
LLaMA3.1(8B)



Mistral-Nemo(12B)



Phi-4(14B)



Mistral-Small(24B)