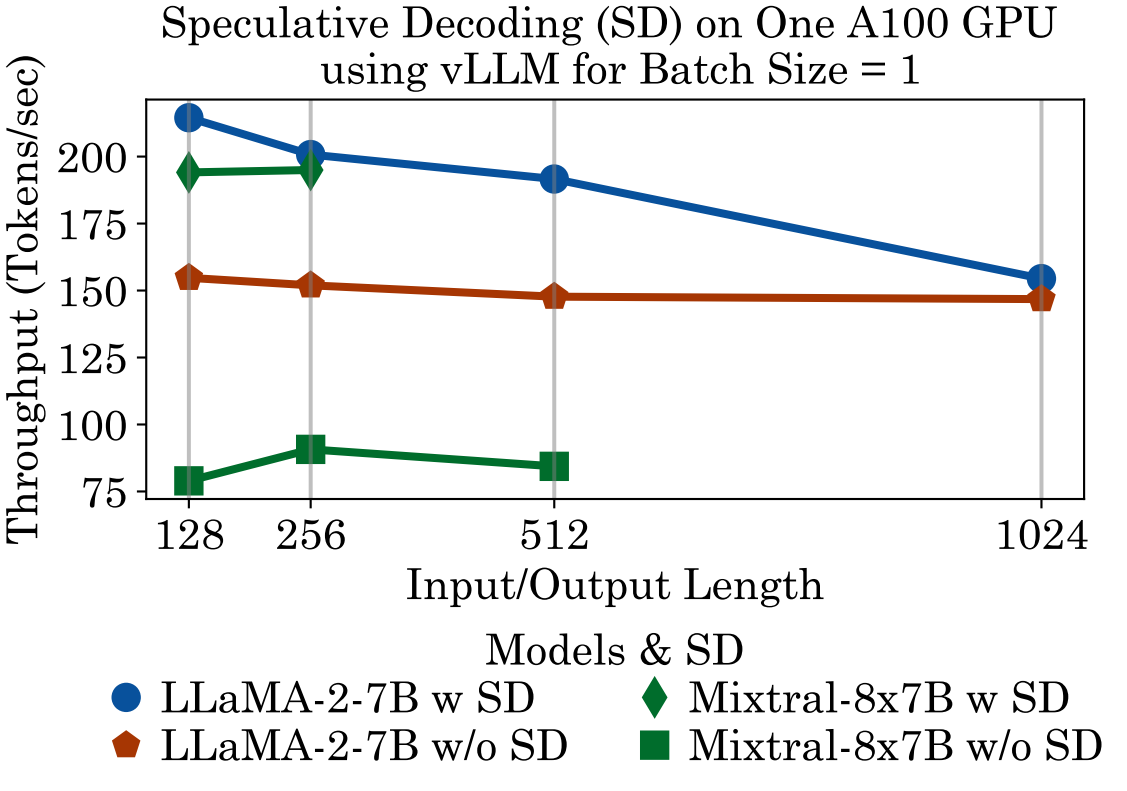
Speculative Decoding (SD) on One A100 GPU
using vLLM for Batch Size = 1
Throughput (Tokens/sec)
200
175
150
125
100
75
128
256
512
1024
Input/Output Length
Models & SD
LLaMA-2-7B w SD
LLaMA-2-7B w/o SD
Mixtral-8x7B w SD
Mixtral-8x7B w/o SD