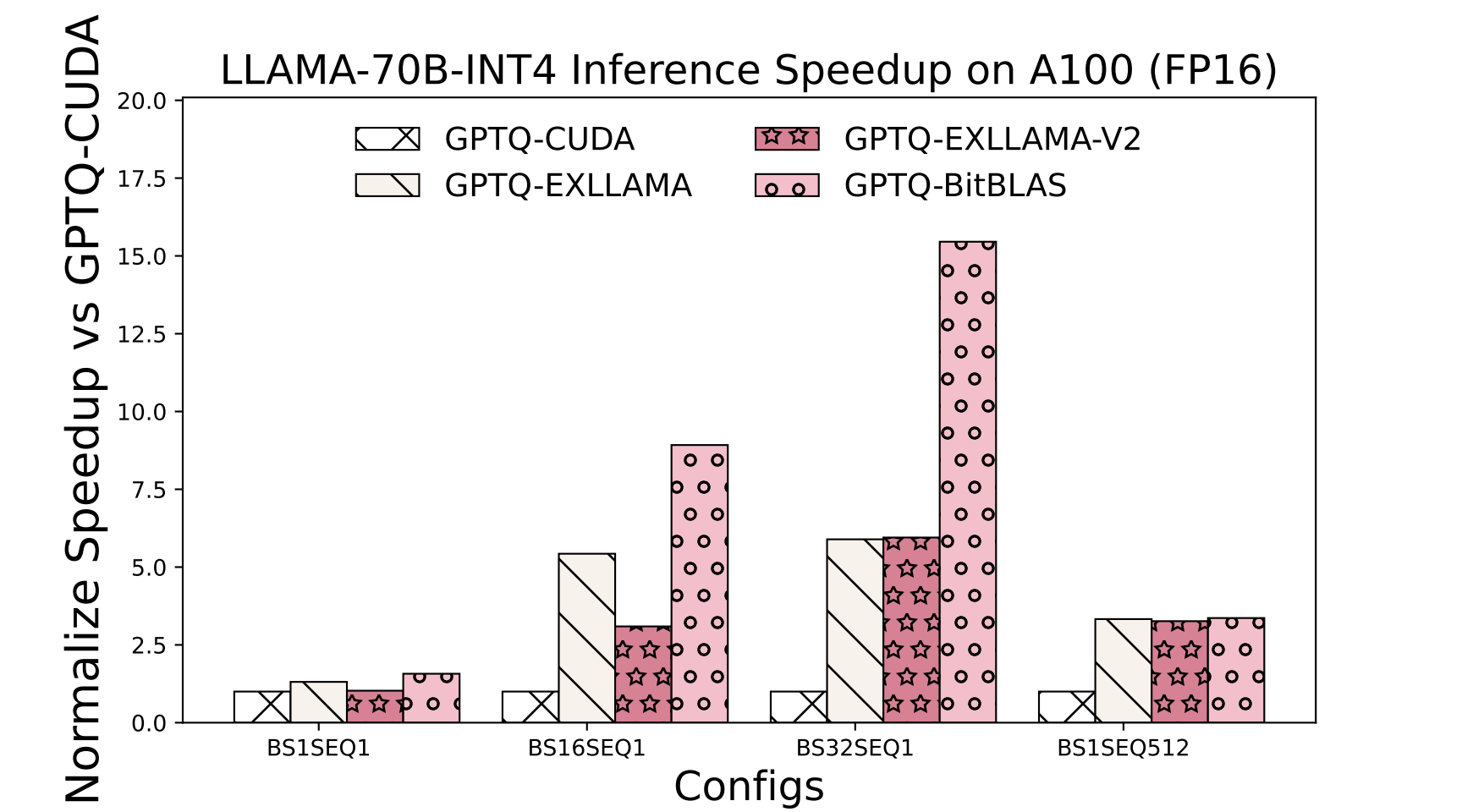
LLAMA-70B-INT4 Inference Speedup on A100 (FP16)
GPTQ-CUDA
GPTQ-EXLLAMA
GPTQ-EXLLAMA-V2
GPTQ-BitBLAS
Normalize Speedup vs GPTQ-CUDA
20.0
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
BS1SEQ1
BS16SEQ1
BS32SEQ1
BS1SEQ512
Configs