

7B Models on One MI250 GPU using vLLM  
with Input & Output Length = 1024

