

DeepSeek-R1-Distill-Llama-70B  
TP=4, PP=1, Input=2048, Output=512  
Instance=g6e.12xlarge, GPU=L40S  
Elapsed=145.6s, GPU Time=145.4s

