

Input vs Output Length Comparison of LLaMA-3-8B for Batch Size = 1 on One GPU using TensorRT-LLM

