Inference Latency vs Output Token Count
NVIDIA A10 GPU