Total Inference Time vs Output Token Size