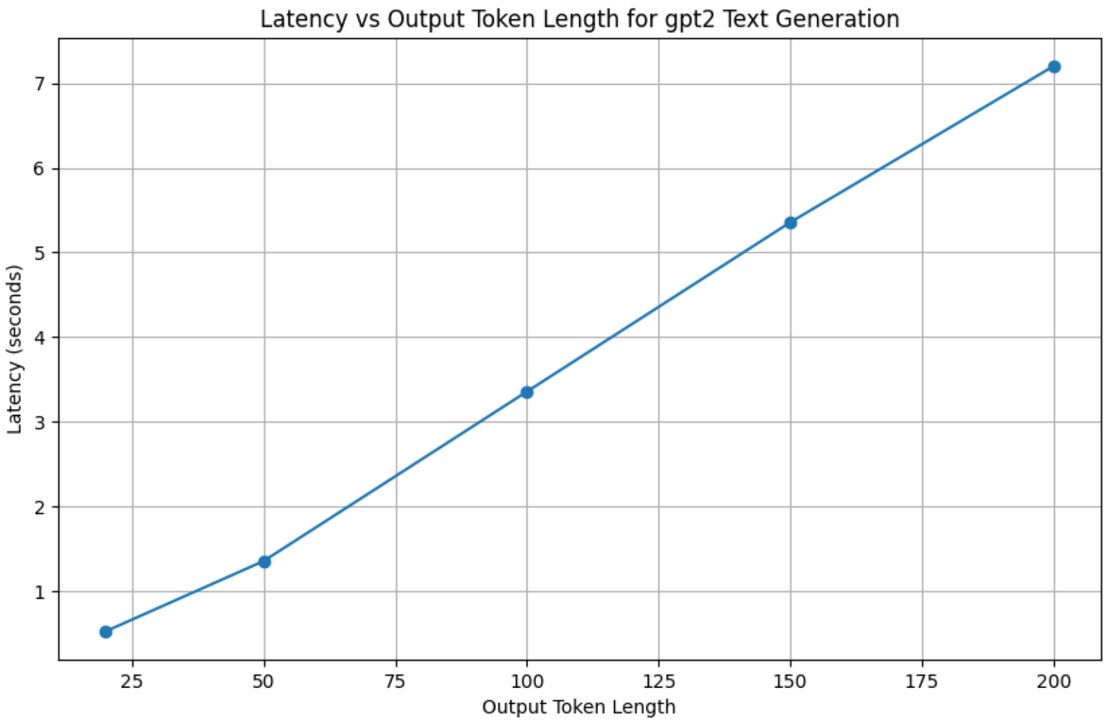
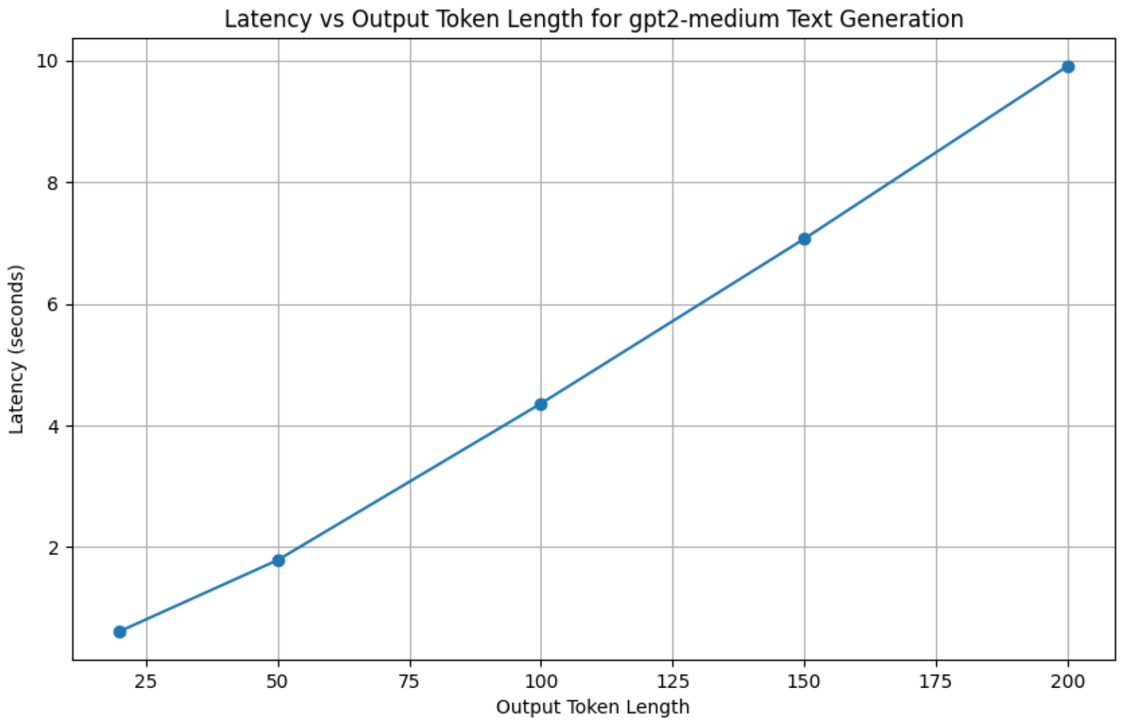
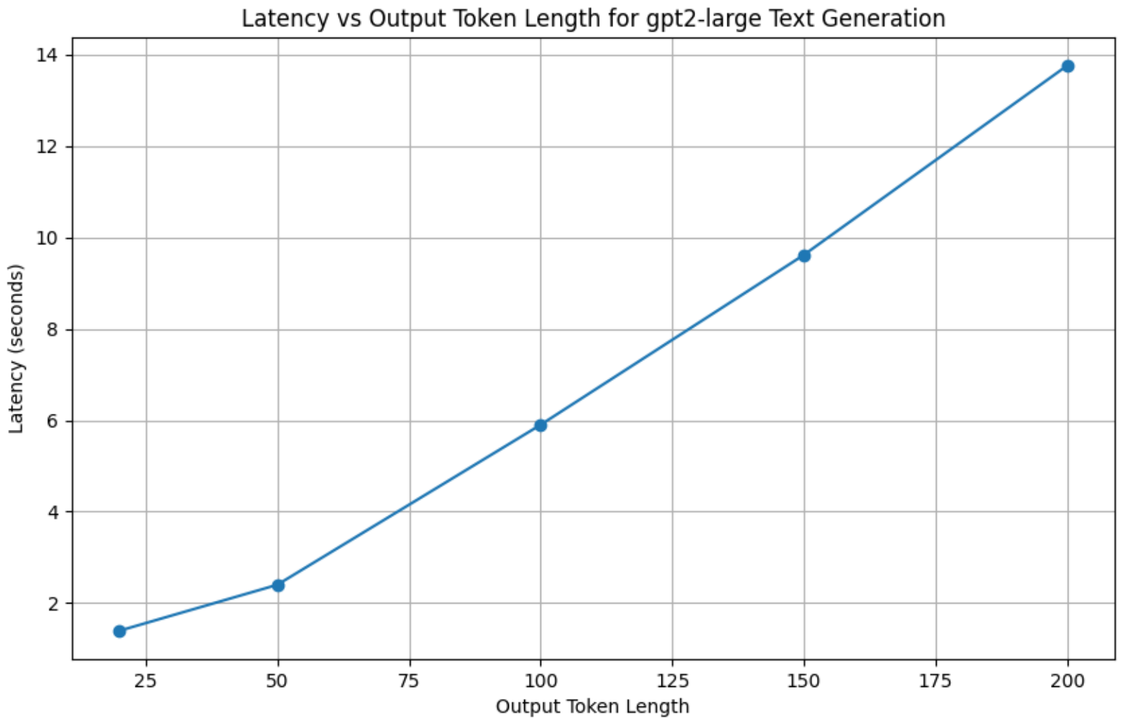
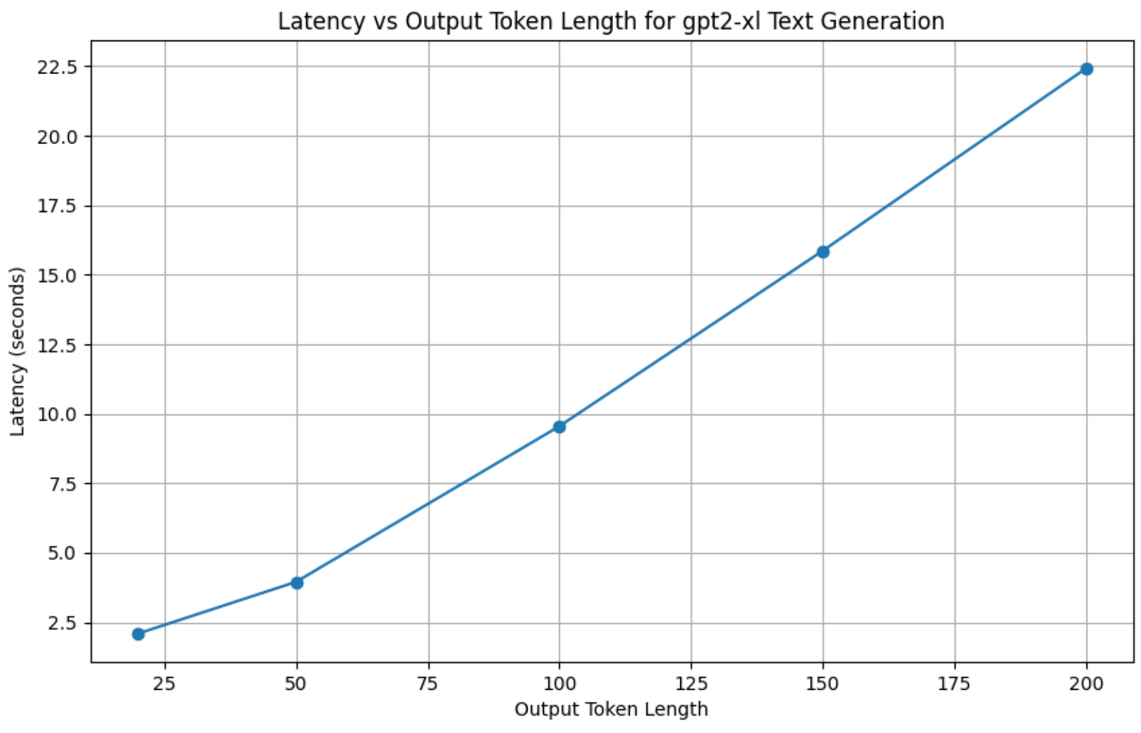
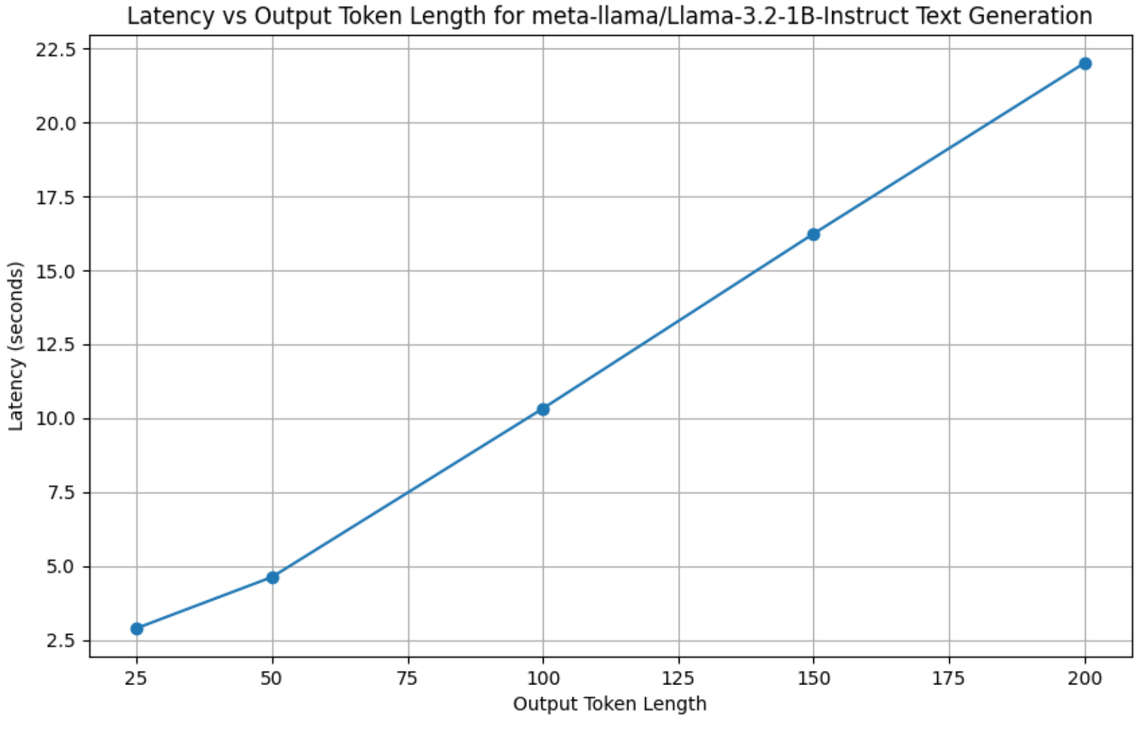
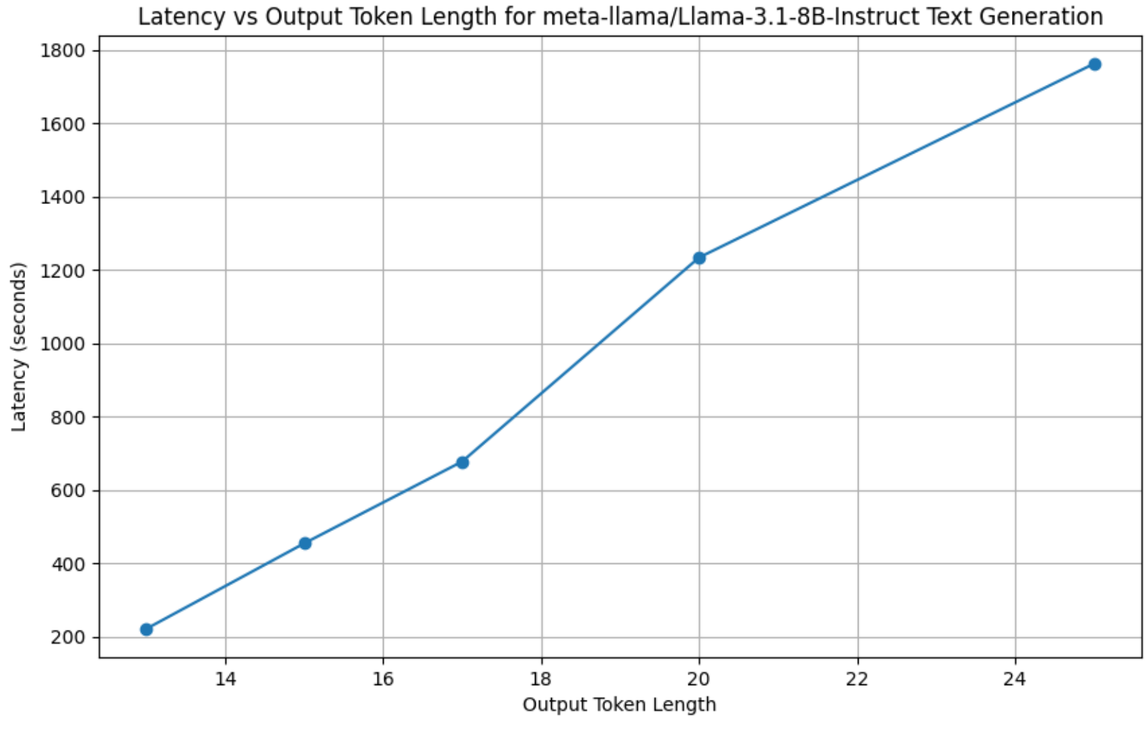

Latency vs Output Token Length for gpt2 Text Generation
Latency (seconds)
Output Token Length


Latency vs Output Token Length for gpt2-medium Text Generation
Latency (seconds)
Output Token Length


Latency vs Output Token Length for gpt2-large Text Generation
Latency (seconds)
Output Token Length


Latency vs Output Token Length for gpt2-xl Text Generation
Latency (seconds)
Output Token Length


Latency vs Output Token Length for meta-llama/Llama-3.2-1B-Instruct Text Generation
Latency (seconds)
Output Token Length


Latency vs Output Token Length for meta-llama/Llama-3.1-8B-Instruct Text Generation
Latency (seconds)
Output Token Length