Llama 3.1 8B Latency P95 (s)