FIGURE 10: Inference Latency Comparison
LoRA: 4.7× Faster | 4-bit: 2.2× Faster