Quantization Tradeoffs: Size vs Accuracy
Fewer bits = Smaller model but slight accuracy loss