# 量子化 (Quantization)

**LLaMA 2 70Bでの実例：**

| 精度 | メモリ使用量 | 性能影響 |
|---|---|---|
| FP16 | 140GB | ベースライン |
| INT8 | 70GB | 最小限の低下 |
| INT4 | 35GB | 許容範囲内 |

- 推論コストの大幅削減
- デバイスの要件緩和
- レイテンシの改善