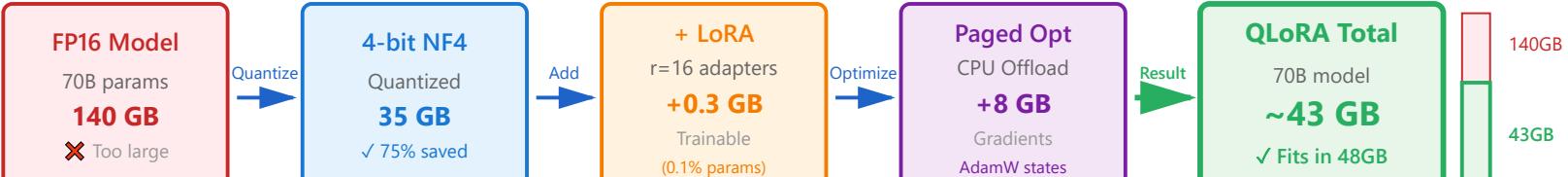


## QLoRA: Quantized Low-Rank Adaptation

**4-bit Quantization + LoRA**  
Train 70B models on single 48GB GPU



**75%**

Memory Reduction

**<1%**

Performance Loss

**24GB**

Min GPU Memory

### Key Benefits

- Enables large model fine-tuning
- Reduces training costs
- Maintains model accuracy
- Faster convergence

### Implementation Details

- NF4 (Normal Float 4-bit)
- Paged optimizers
- Double quantization
- BitsAndBytes library