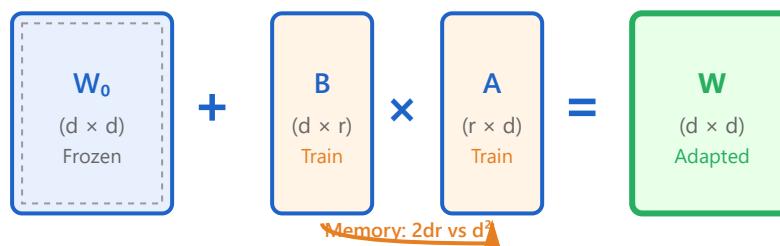


LoRA: Low-Rank Adaptation

Matrix Decomposition: $W = W_0 + BA$ where $\text{rank}(B) = \text{rank}(A) = r$



0.1%

Trainable Params

r=8-16

Typical Rank

100x

Memory Reduction



Medical NER Application

- Disease entity recognition
- Symptom identification
- Medication extraction
- Procedure coding



Configuration Guidelines

- **r=4:** Simple tasks, limited data
- **r=16:** Complex reasoning
- **r=8:** Most medical NLP tasks
- **r=32-64:** Multi-task learning

Memory Saving = 1 - (2 × r × d) / (d × d) ≈ 99.9%