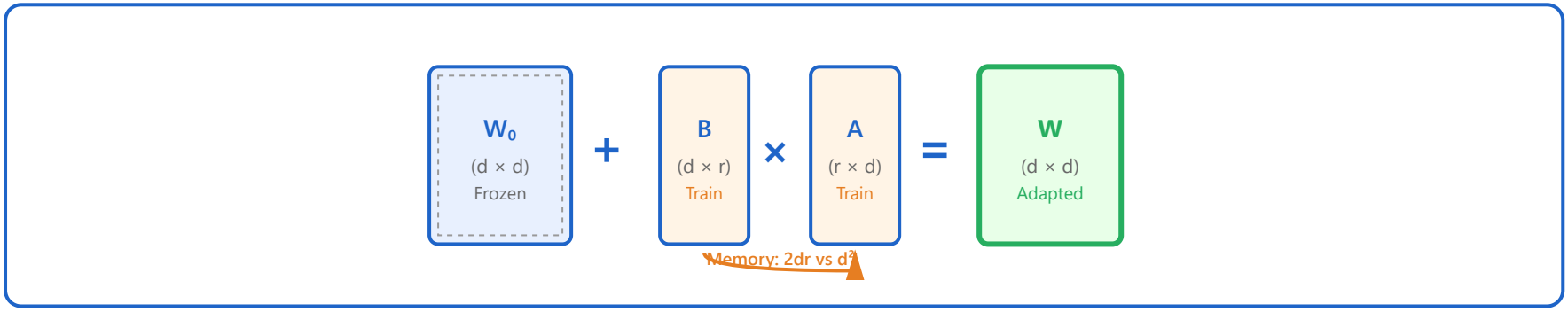
LoRA: Low-Rank Adaptation

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



0.1%	r=8-16	100x
Trainable Params	Typical Rank	Memory Reduction

Medical NER Application

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

Configuration Guidelines

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

$$\text{Memory Saving} = 1 - (2 \times r \times d) / (d \times d) \approx 99.9\%$$