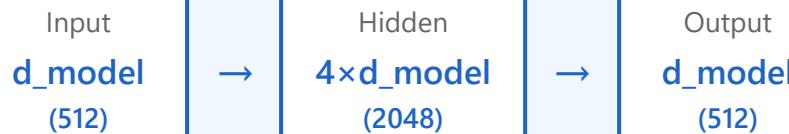


# Feed-Forward Network & Layer Normalization

## ◆ Feed-Forward Network (FFN)

$$\text{FFN}(\mathbf{x}) = \max(0, \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$$



🔧 Two linear transformations with ReLU activation

📐 Hidden dimension typically  $4 \times d_{\text{model}}$  (e.g., 2048)

🎯 Applied identically to each position separately (position-wise)

## 📊 Layer Normalization



### Normalization

Normalizes across feature dimension



### Training Benefits

Stabilizes training and speeds convergence



### Application

Applied after each sub-layer with residual connections



### Placement Strategy

Post-LN vs Pre-LN

affects training dynamics