Feed-Forward Network & Layer Normalization

Feed-Forward Network (FFN)

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

Input	Hidden	Output
d_model (512)	→ 4×d_model (2048)	→ d_model (512)

- Two **linear transformations** with **ReLU activation**
- Hidden dimension typically **4× d_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