**🔎 MLP = Multi-Layer Perceptron**

It’s just a **feed-forward neural network** made of fully connected layers.

**Structure**

Input vector (x)

│

▼

Linear Layer (W1x + b1)

│

▼

Activation (e.g., ReLU, GELU)

│

▼

Linear Layer (W2h + b2)

│

▼

Output vector (y)

* **Linear layers**: do matrix multiplications (W·x + b).
* **Activation functions**: introduce non-linearity (so the model learns complex patterns).
* **Stacked layers**: “multi-layer” = more than one dense layer in sequence.

**🔎 MLP inside a Transformer**

Each encoder/decoder block has:

1. **Self-attention** → mixes information between tokens.
2. **MLP (feed-forward network)** → processes each token’s vector *individually*.

**Example in BERT / GPT:**

* Input token hidden size: 768
* MLP expands to 3072 (4× bigger), applies activation, then projects back to 768.

So, per token:

h (768) → Linear (768→3072) → GELU → Linear (3072→768) → h'

This lets the model “re-shape” information at each layer.

✅ **Key intuition:**

* **Self-attention** = “Who should I listen to in the sentence?”
* **MLP** = “Now that I’ve listened, how should I transform my own representation?”