Attention (Query, Key, Value)

- What it is?

A mechanism that lets each token decide which other tokens are important.

- Goal:

To compute a context-aware representation for each token.

- Math:

-Queries, keys, and values are linear projections:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V$$

- Attention scores:

$$\text{scores} = \frac{QK^T}{\sqrt{d_K}}$$

- Attention weighs:

$$A = \text{softmax}(\text{scores})$$

- Output:

$$\text{Output} = AV$$

- Conclusion:

Attention allows each token to dynamically focus on relevant parts of the sequence.