Attention Mechanism

⭐ Master Rhyme

"Query meets Key to measure degree,
Value flows weighted—attention is free.
Scale the dot, softmax the plot,
Focus on tokens that matter a lot."

⭐ Self-Attention Rhyme

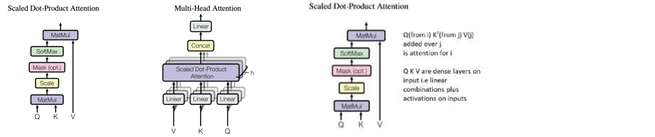
"Every word looks at every other,
Weights decide who feels like a brother."

⭐ Q–K–V Rhyme

"Keys define where memories lie,
Queries ask who should apply,
Values carry what to supply."

⭐ Scaled Dot-Product Rhyme

"Dot it, scale it, soften the score—
Attention finds what to look for."



🧠 KEY EQUATIONS (Short + Clear)

1 Compute Attention Scores

$$\text{scores} = QK^\top$$

2 Scale by $\sqrt{d_k}$

$$\text{scaled} = \frac{QK^\top}{\sqrt{d_k}}$$

3 Softmax to get Attention Weights

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

4 Weighted Sum of Values

$$\text{Attention}(Q, K, V) = AV$$

⭐ FULL EQUATION

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

Rhyme for the full formula:

"Dot with Keys, scale with ease,
Softmax breeze, Values please."

🔥 IMPORTANT POINTS

Conceptual

- Attention decides **which tokens matter most**.
- It replaces recurrence (RNN/LSTM) with **parallel processing**.
- Each word creates **Q**, **K**, **V** vectors.

Why scaling?

- Prevents large dot-products → stable softmax.

Why softmax?

- Converts scores into **probabilities** → weights.

Self-attention vs Cross-attention

- **Self-attention**: Q = K = V (same sequence)
- **Cross-attention**: Q comes from decoder, K & V from encoder

Multi-head attention

- Multiple attention heads allow the model to learn **different relations** (syntax, semantics, etc.)

Transformers

- Stack of attention + feed-forward layers.
- Gives SOTA performance in NLP, vision, speech.

📘 SMALL MATH PROBLEM

Given:

Query:

$$Q = [1, 2]$$

Keys:

$$K_1 = [1, 0], \quad K_2 = [0, 1]$$

Values:

$$V_1 = [1, 1], \quad V_2 = [2, 0]$$

Let $d_k = 2$.

Step 1: Compute dot products

$$QK_1^\top = 1(1) + 2(0) = 1$$

$$QK_2^\top = 1(0) + 2(1) = 2$$

Scores:

$$[1,\ 2]$$

Step 2: Scale

$$\sqrt{d_k} = \sqrt{2}$$

$$\text{scaled} = \left[\frac{1}{\sqrt{2}}, \frac{2}{\sqrt{2}}\right]$$

Step 3: Softmax

Let

$$a = \frac{1}{\sqrt{2}}, \quad b = \frac{2}{\sqrt{2}}$$

Softmax:

$$w_1 = \frac{e^a}{e^a + e^b}, \quad w_2 = \frac{e^b}{e^a + e^b}$$

Since $b > a$, weight 2 will be larger.

Approximate values:

$$w_1 \approx 0.27,\ w_2 \approx 0.73$$

Step 4: Weighted sum of values

$$\text{Output} = w_1 V_1 + w_2 V_2$$

$$= 0.27[1, 1] + 0.73[2, 0]$$

$$= [0.27 + 1.46,\ 0.27 + 0]$$

$$= [1.73,\ 0.27]$$

✅ Final Answer

$$\text{Attention Output} \approx [1.73,\ 0.27]$$