Mini Transformer Calculation Walkthrough

Input Tokens

- $x_1 = 2.0$
- $x_2 = -1.5$

Q, K, V Projections

Using $W_Q = W_K = W_V = [1.0]$:

$$Q_2 = 1.0 \cdot (-1.5) = -1.5$$
$$K_2 = 1.0 \cdot (-1.5) = -1.5$$
$$V_2 = 1.0 \cdot (-1.5) = -1.5$$

Attention Scores and Weights

Dot-product attention (scaled):

$$\text{score}_{ij} = \frac{Q_i \cdot K_j}{\sqrt{1}}$$

$$\text{Scores} = \begin{bmatrix} 2.0 \cdot 2.0 & 2.0 \cdot (-1.5) \\ -1.5 \cdot 2.0 & -1.5 \cdot (-1.5) \end{bmatrix} = \begin{bmatrix} 4.0 & -3.0 \\ -3.0 & 2.25 \end{bmatrix}$$

Apply softmax row-wise:

$$\text{softmax}([4, -3]) = \left[\frac{e^4}{e^4 + e^{-3}}, \frac{e^{-3}}{e^4 + e^{-3}} \right] \approx [0.9991, 0.0009]$$

$$\text{softmax}([-3, 2.25]) = \left[\frac{e^{-3}}{e^{-3} + e^{2.25}}, \frac{e^{2.25}}{e^{-3} + e^{2.25}} \right] \approx [0.0067, 0.9933]$$

Attention Output

$$\text{Attention output}_1 = 0.9991 \cdot 2.0 + 0.0009 \cdot (-1.5) \approx 1.9967$$

$$\text{Attention output}_2 = 0.0067 \cdot 2.0 + 0.9933 \cdot (-1.5) \approx -1.4601$$

Final Output

$$W_o = \begin{bmatrix} 1.2 \\ -0.8 \end{bmatrix}, \quad \hat{y} = 1.2 \cdot 1.9967 + (-0.8) \cdot (-1.4601)$$

$$\hat{y} = 2.396 + 1.168 = \boxed{3.56}$$