Problem 3 (30 points)

We have three tokens in the sentence **"bagel with cheese"**:

- $w_1 =$ *bagel*
- $w_2 =$ *with*
- $w_3 =$ *cheese*

with query, key, value vectors

$$\begin{aligned} q_1 &= [1,2,3], & k_1 &= [1,1,1], & v_1 &= [2,0,1], \\ q_2 &= [2,3,2], & k_2 &= [0,0,0], & v_2 &= [3,0,0], \\ q_3 &= [5,6,7], & k_3 &= [2,2,0], & v_3 &= [1,2,2]. \end{aligned}$$

We want the self-attention output z_1 for token w_1 (*bagel*).

(a) Unnormalized attention scores for w_1

First compute dot products $q_1 \cdot k_j$:

$$\begin{aligned} a_{11} &= q_1 \cdot k_1 = 1\cdot 1 + 2\cdot 1 + 3\cdot 1 = 6, \\ a_{12} &= q_1 \cdot k_2 = 1\cdot 0 + 2\cdot 0 + 3\cdot 0 = 0, \\ a_{13} &= q_1 \cdot k_3 = 1\cdot 2 + 2\cdot 2 + 3\cdot 0 = 2 + 4 + 0 = 6. \end{aligned}$$

Scale by $\sqrt{|k_1|}$.
The key vectors have dimension 3, so $|k_1| = 3$, hence

$$\sqrt{|k_1|} = \sqrt{3} \approx 2 \quad (\text{given to round to } 2).$$

Thus

$$\begin{aligned} a_{11} &= \frac{6}{2} = 3, \\ a_{12} &= \frac{0}{2} = 0, \\ a_{13} &= \frac{6}{2} = 3. \end{aligned}$$

(b) Normalize attention weights for row $i = 1$

Compute the row sum

$$\sum_k a_{1k} = 3 + 0 + 3 = 6.$$

Then

$$\alpha_{11} = \frac{a_{11}}{\sum_k a_{1k}} = \frac{3}{6} = 0.5,$$
$$\alpha_{12} = \frac{a_{12}}{\sum_k a_{1k}} = \frac{0}{6} = 0,$$
$$\alpha_{13} = \frac{a_{13}}{\sum_k a_{1k}} = \frac{3}{6} = 0.5.$$

So the normalized attention weights for w_1 are

$$\alpha_{1\cdot} = [0.5,\ 0,\ 0.5].$$

(c) Weighted sum of value vectors

Now compute

$$z_1 = \sum_{j=1}^{3} \alpha_{1j} v_j = \alpha_{11} v_1 + \alpha_{12} v_2 + \alpha_{13} v_3.$$

$$\alpha_{11} v_1 = 0.5[2, 0, 1] = [1, 0, 0.5],$$
$$\alpha_{12} v_2 = 0[3, 0, 0] = [0, 0, 0],$$
$$\alpha_{13} v_3 = 0.5[1, 2, 2] = [0.5, 1, 1].$$

Add them:

$$z_1 = [1, 0, 0.5] + [0, 0, 0] + [0.5, 1, 1] = [1.5,\ 1,\ 1.5].$$

Final answer

$$\boxed{z_1 = [1.5,\ 1,\ 1.5]}$$

for the token *bagel*.

Time spend

- Q1: 4 hours
- Q2: 2 hours
- Q3: 40 mins