Attention (p-4) (LLM notes)

- from before to compare key Queries in Attention blocks we use dot product here is a visualization:

☆ Attention 8x8 pattern
• = ignore (zeroed out)
! has normalized values

	a ↓ E_1 ↓ W_K K_1	Fluffy ↓ E_2 ↓ W_K K_2	blue ↓ E_3 ↓ W_K K_3	creature ↓ E_4 ↓ W_K K_4	roamed ↓ E_5 ↓ W_K K_5	the ↓ E_6 ↓ W_K K_6	verdant ↓ E_7 ↓ W_K K_7	forest ↓ E_8 ↓ W_K K_8
	Q_1	Q_2	Q_3	Q_4	Q_5	Q_6	Q_7	Q_8
a → E_1 → K_1	$K_1 \cdot Q_1$	$K_1 \cdot Q_2$	$K_1 \cdot Q_3$	$K_1 \cdot Q_4$	$K_1 \cdot Q_5$	$K_1 \cdot Q_6$	$K_1 \cdot Q_7$	$K_1 \cdot Q_8$
fluffy → E_2 → K_2	$K_2 \cdot Q_1$ •	$K_2 \cdot Q_2$	$K_2 \cdot Q_3$	$K_2 \cdot Q_4$	$K_2 \cdot Q_5$	$K_2 \cdot Q_6$	$K_2 \cdot Q_7$	$K_2 \cdot Q_8$
blue → E_3 → K_3	$K_3 \cdot Q_1$ •	$K_3 \cdot Q_2$ •	$K_3 \cdot Q_3$	$K_3 \cdot Q_4$	$K_3 \cdot Q_5$	$K_3 \cdot Q_6$	$K_3 \cdot Q_7$	$K_3 \cdot Q_8$
creature → E_4 → K_4	$K_4 \cdot Q_1$ •	$K_4 \cdot Q_2$ •	$K_4 \cdot Q_3$ •	$K_4 \cdot Q_4$	$K_4 \cdot Q_5$	$K_4 \cdot Q_6$	$K_4 \cdot Q_7$	$K_4 \cdot Q_8$
roamed → E_5 → K_5	$K_5 \cdot Q_1$ •	$K_5 \cdot Q_2$ •	$K_5 \cdot Q_3$ •	$K_5 \cdot Q_4$ •	$K_5 \cdot Q_5$	$K_5 \cdot Q_6$	$K_5 \cdot Q_7$	$K_5 \cdot Q_8$
the → E_6 → K_6	$K_6 \cdot Q_1$ •	$K_6 \cdot Q_2$ •	$K_6 \cdot Q_3$ •	$K_6 \cdot Q_4$ •	$K_6 \cdot Q_5$ •	$K_6 \cdot Q_6$	$K_6 \cdot Q_7$	$K_6 \cdot Q_8$
verdant → E_7 → K_7	$K_7 \cdot Q_1$ •	$K_7 \cdot Q_2$ •	$K_7 \cdot Q_3$ •	$K_7 \cdot Q_4$ •	$K_7 \cdot Q_5$ •	$K_7 \cdot Q_6$ •	$K_7 \cdot Q_7$	$K_7 \cdot Q_8$
forest → E_8 → K_8	$K_8 \cdot Q_1$ •	$K_8 \cdot Q_2$ •	$K_8 \cdot Q_3$ •	$K_8 \cdot Q_4$ •	$K_8 \cdot Q_5$ •	$K_8 \cdot Q_6$ •	$K_8 \cdot Q_7$ •	$K_8 \cdot Q_8$

Attends to

	Q_1 a	Q_2 Fluffy	Q_3 blue	Q_4 creature ...
a K_1	• 0.7	• -93	• -24	• -5
Fluffy K_2	• -73	• 2.9	• -5	• +93
blue K_3	• -53	• -5.7	• 1.9	• +94
creature K_4 ⋮	• -21	• -29	• -56	• 4.9

• = size of Dot product range $(-\infty, \infty)$

→ to measure the dot product we compute it between each key Query pair. The larger Dot products mean the key and Queries align and the dot product would be some large num. meaning the embedding of keys (Fluffy, blue) attend to the embedding of Query (creature). and for neg dot products like "the" and "creature" represents that these are unrelated to each other.

- So we have numbers from $-\infty \to \infty$ that give us a score for how relevant each word is to updating the meaning of every other word

- The way we will use this score is to take a weighted sum across each column, weighted by relevance (size •)
So we need normalization specifically range $(-\infty, \infty) \to$ range $(0,1)$ sum(vals) = 1
- here we use softmax for each col

Ex

	Q_4
K_1	-5
K_2	93
K_3	94
K_4	4.9

→ range (0,1), sum(col) = 1 → softmax →

K_1	0
K_2	0.42
K_3	0.58
K_4	0
⋮	

→ sum = 1, range (0,1)