Back to Cats: Tracing Attention for “sat”

Name: _____ Trace how the last token, “sat”, attends to the sentence “The cat sat”. Show your work in the margin and round every result to three decimals.

Given

Token	Embedding x
1. The	[1.0, 0.5, 0.2, 0.1]
2. cat	[0.5, 1.0, 0.3, 0.2]
3. sat (query)	[0.3, 0.2, 1.0, 0.5]

W_Q (what am I looking for?)	W_K (what do I offer?)	W_V (what will I contribute?)
[1.0, 0.0, 0.5, 0.0]	[1.0, 0.2, 0.0, 0.0]	[0.8, 0.0, 0.0, 0.1]
[0.0, 1.0, 0.0, 0.5]	[0.2, 1.0, 0.0, 0.0]	[0.0, 0.9, 0.1, 0.0]
[0.5, 0.0, 1.0, 0.0]	[0.0, 0.0, 1.0, 0.3]	[0.0, 0.1, 0.8, 0.0]
[0.0, 0.5, 0.0, 1.0]	[0.0, 0.0, 0.3, 1.0]	[0.1, 0.0, 0.0, 0.9]

Step 1: Queries, keys and values

Multiply each embedding by a weight matrix: q = x × W_Q, k = x × W_K, v = x × W_V. Three entries are left for you; the rest are given.

Token	q (only the query token needs one)	k	v
The	not needed	[1.1, 0.7, 0.23, 0.16]	[0.81, 0.47, 0.21, 0.19]
cat	not needed	[____ , ____ , ____ , ____]	[____ , ____ , ____ , ____]
sat	[____ , ____ , ____ , ____]	[0.34, 0.26, 1.15, 0.8]	[0.29, 0.28, 0.82, 0.48]

Steps 2–3: Scores and scaling

Score = q(sat) · k(token), the sum of the four products. Then divide by √d = √4 = 2.

Token	Score = q · k	Scaled = score ÷ 2
The	0.8×1.1 + 0.45×0.7 + 1.15×0.23 + 0.6×0.16 = 1.556	0.778

Token	Score = q · k	Scaled = score ÷ 2
cat	_____	_____
sat	_____	_____

Steps 4–5: Softmax

Exponentiate each scaled score, add them up, then divide each exponent by the total.

Token	exp(scaled)	Weight = exp ÷ total
The	2.177	2.177 ÷ total = 0.292
cat	_____	_____
sat	_____	_____
Total	Sum of the three exps = _____	Sum of the three weights = _____

Step 6: Weigh the values and add them up

Token	Weight (Step 5)	Value v (Step 1)	Weight × v
The	0.292	[0.81, 0.47, 0.21, 0.19]	[0.237, 0.137, 0.061, 0.055]
cat	_____	your v for “cat”	[____ , ____ , ____ , ____]
sat	_____	[0.29, 0.28, 0.82, 0.48]	[____ , ____ , ____ , ____]
Output for “sat”		Add the three rows	[____ , ____ , ____ , ____]

Questions

1. Which token gets the most attention from “sat”, and what share goes to “cat”?
2. Why must the three weights add up to 1, and what does the output vector now contain that the embedding of “sat” did not?