

Attention Computation: Step-by-Step Flow

STEP 1:

Query from "mat" meets all Keys

Q("mat")
[0.8, 0.6, 0.4]

K("The")

K("cat")

K("sat")

K("on")

K("the")

STEP 2:

Calculate Q · K (dot products)

0.1

0.3

0.4

0.8

0.6

Higher score
= more relevant

STEP 3:

Apply Softmax (convert to percentages)

14%

17%

19%

28%

23%

Sum = 100%

$$\text{softmax}(x)_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

STEP 4:

Multiply weights with Values

V("The")

V("cat")

V("sat")

V("on")

V("the")

× 14%

× 17%

× 19%

× 28%

× 23%

Each Value
contributes
proportionally

STEP 5:

Sum all weighted values

14%×V("The") + 17%×V("cat") + 19%×V("sat") + ...

OUTPUT: Context-aware representation of "mat"

Knows it follows "on the" (location pattern)

Key Insight: Attention learns to identify and combine relevant information automatically