What can “cat” read?

Each row is one token. Use the query at “cat”. The key width is 2 and the score temperature is 1. A calculator is enough. `q(cat) = [1, 1]`

> $s_i = (q \cdot k_i) / \sqrt{2}$
> $w_i = \exp(s_i) / \Sigma_j \exp(s_j)$

1. Fill in two attention rows

First allow every key. Then apply a causal mask: “cat” can read itself and earlier positions. Sum over visible keys only. Write 0 for a blocked key.

Token	K	V	q · k	All keys	Causal
The	[1, 0]	[1, 0]			
cat	[1, 1]	[0, 2]			
sat	[0, 1]	[2, 1]			
here	[-1, 0]	[-1, 1]			

2. Mix the values

Multiply each V by its weight and add the four vectors. Write the two output vectors, one for each mask setting.

3. Change one future value

Add [0, 8] to V(here). How much does each output move? Use the weights you already calculated.

4. Find the masking bug

A teammate applies softmax to all four scores, then zeros the last two weights. What quick check reveals the error?

© 2026 Seongeun So · github.com/sungeuns/foundation-model-engineering · MIT license

Worked solution

Divide each raw dot product by √2. In the causal run the first two tokens are visible. Their scores are about 0.7071 and 1.4142. Compute each weight using only the visible terms in the denominator.

```
w(The) = exp(1/√2) / (exp(1/√2) + exp(2/√2)) ≈ 0.3302
```

Token	K	V	q · k	All keys	Causal
The	[1, 0]	[1, 0]	1	0.2341	0.3302
cat	[1, 1]	[0, 2]	2	0.4748	0.6698
sat	[0, 1]	[2, 1]	1	0.2341	0.0000
here	[-1, 0]	[-1, 1]	-1	0.0569	0.0000

Displayed weights are rounded. Keep full precision until the final output.

2. Mix the values

For each weight row w, the first coordinate is w(The) + 2w(sat) − w(here). The second coordinate is 2w(cat) + w(sat) + w(here).

```
All keys → [0.6455, 1.2407]
Causal → [0.3302, 1.3395]
```

3. Change one future value

Q and K stay fixed, so the weights stay fixed. Only the second output coordinate can change. With all keys its increase is 8 × w(here) ≈ 0.4554. With the causal mask that weight is zero, so both output coordinates stay the same.

4. Find the masking bug

After the late mask the surviving weights sum to 0.7090. A normalized row should sum to 1. Mask the scores before softmax, or divide the surviving weights by their sum when that sum is positive.

Try every query in the browser · Download the matching PyTorch script

© 2026 Seongeun So · github.com/sungeuns/foundation-model-engineering · MIT license