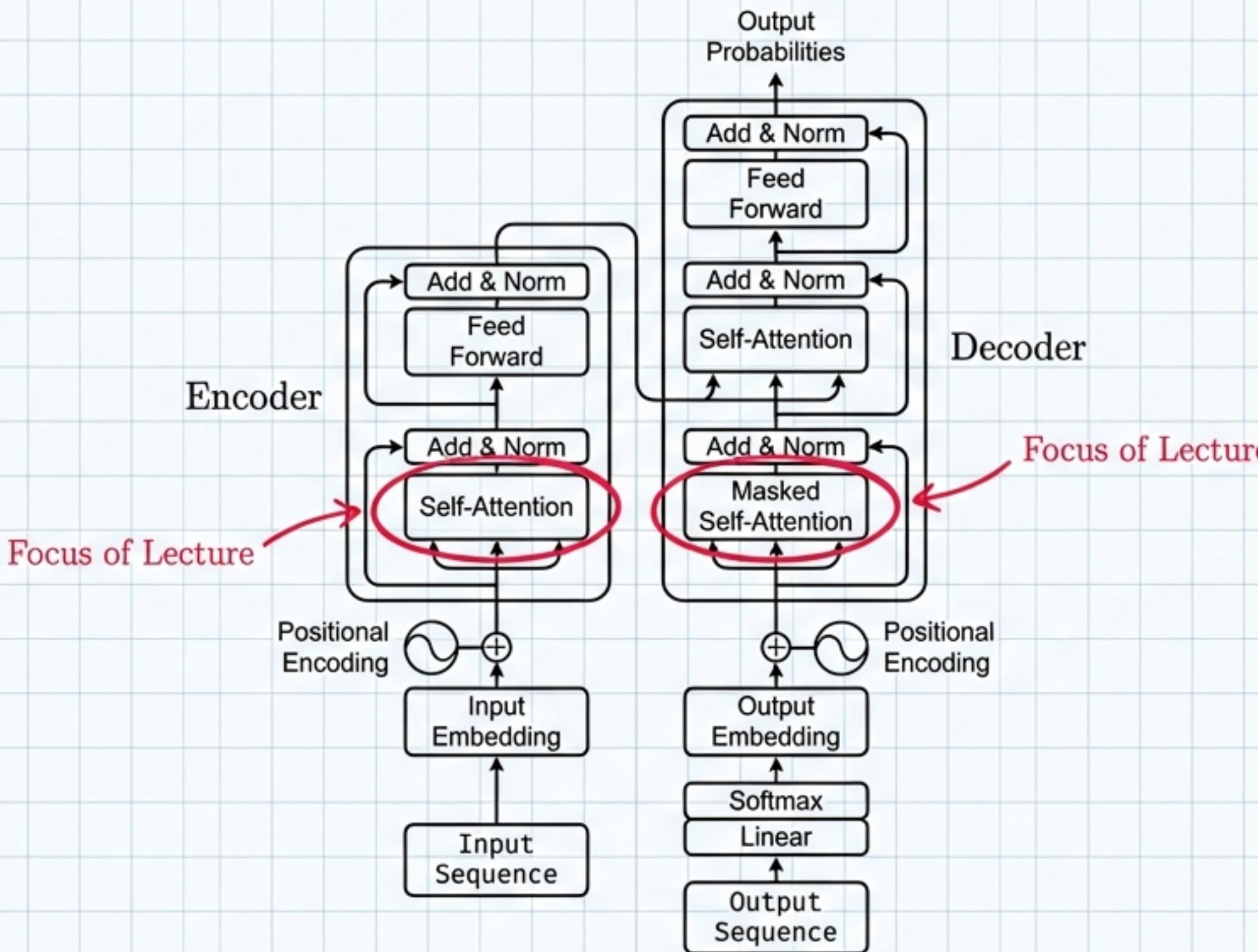


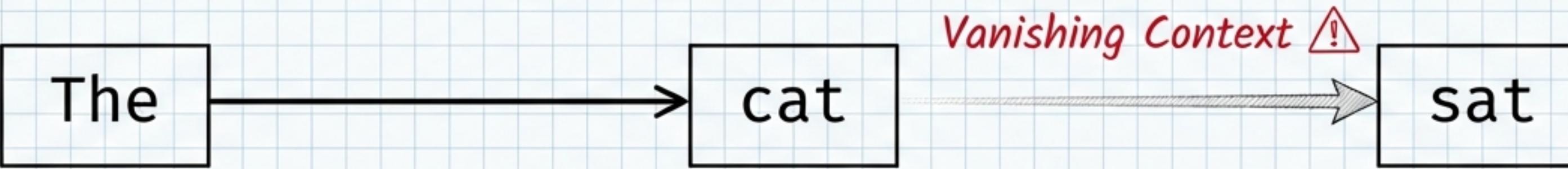
The Heart of the Transformer

Understanding the Self-Attention Mechanism

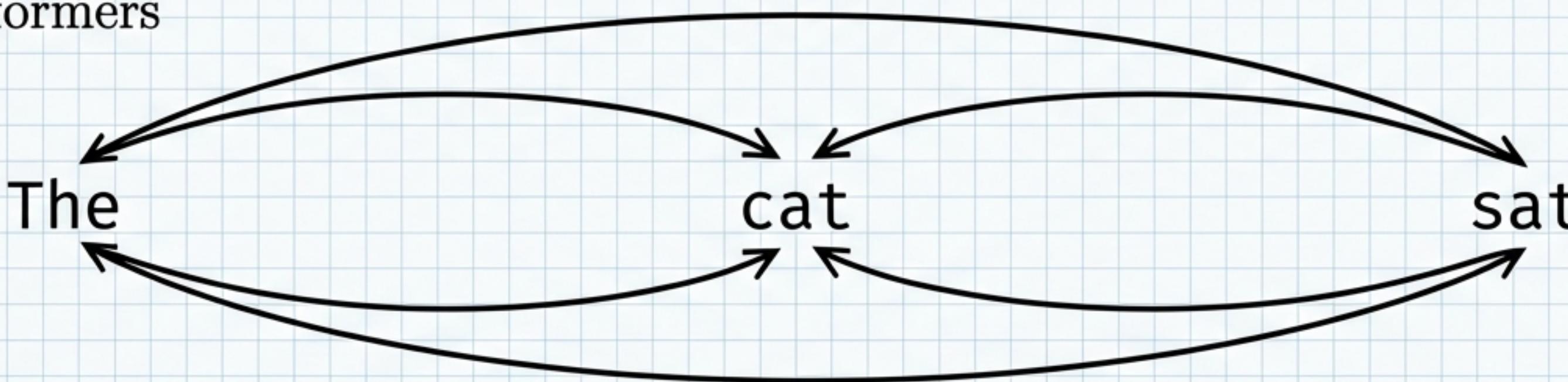


The Problem with Sequences

Recurrent Neural Networks (RNNs)

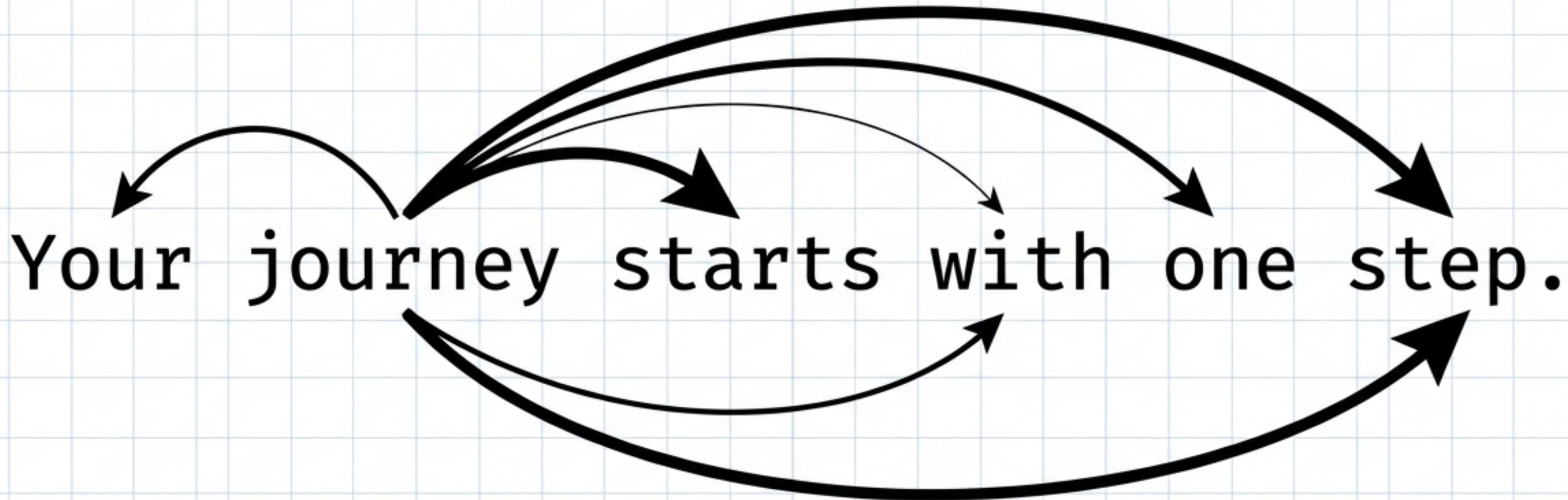


Transformers



Parallel Context: All tokens see all history instantly.

Defining Self-Attention



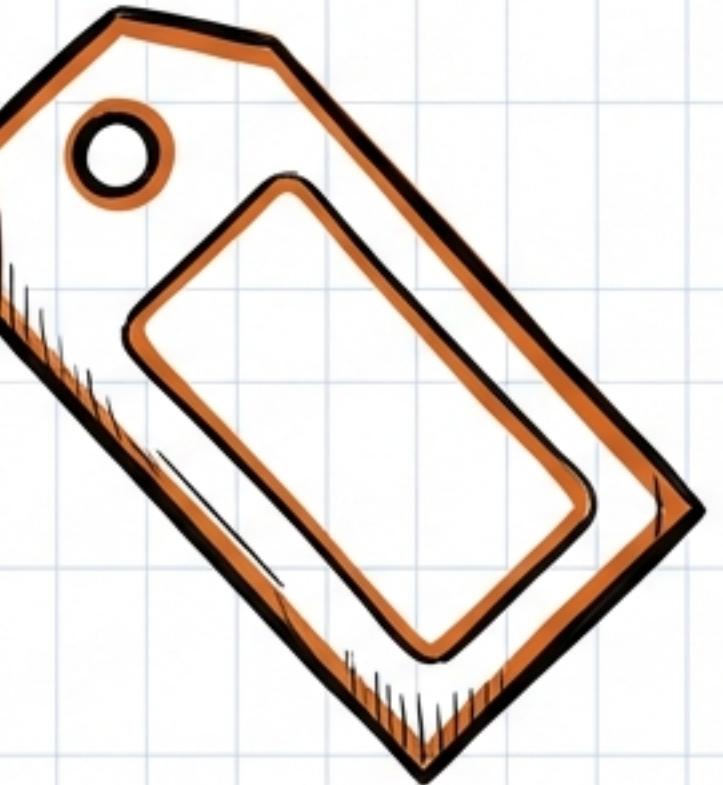
Self-Attention: Relating positions within a single sequence.

The Core Trinity: Q, K, V



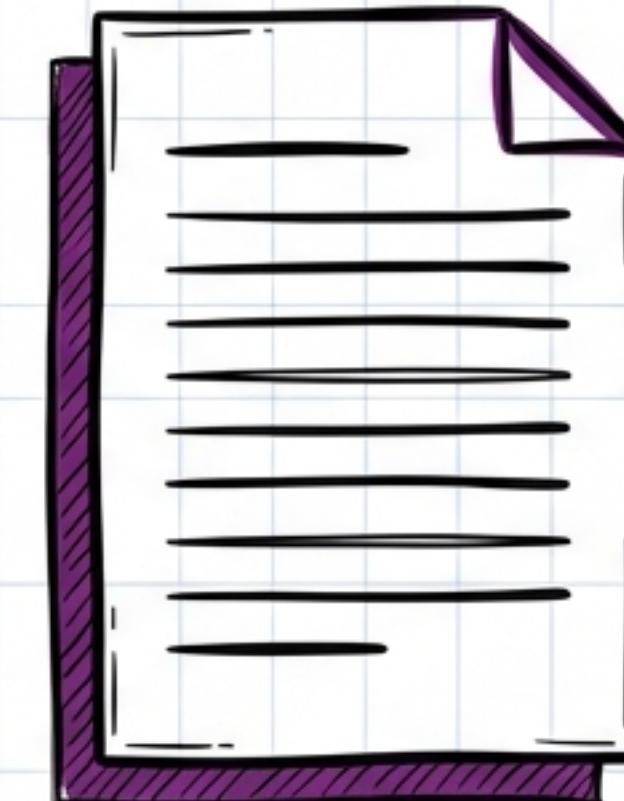
Query (Q)

What I am
looking for.



Key (K)

What defines
the content.



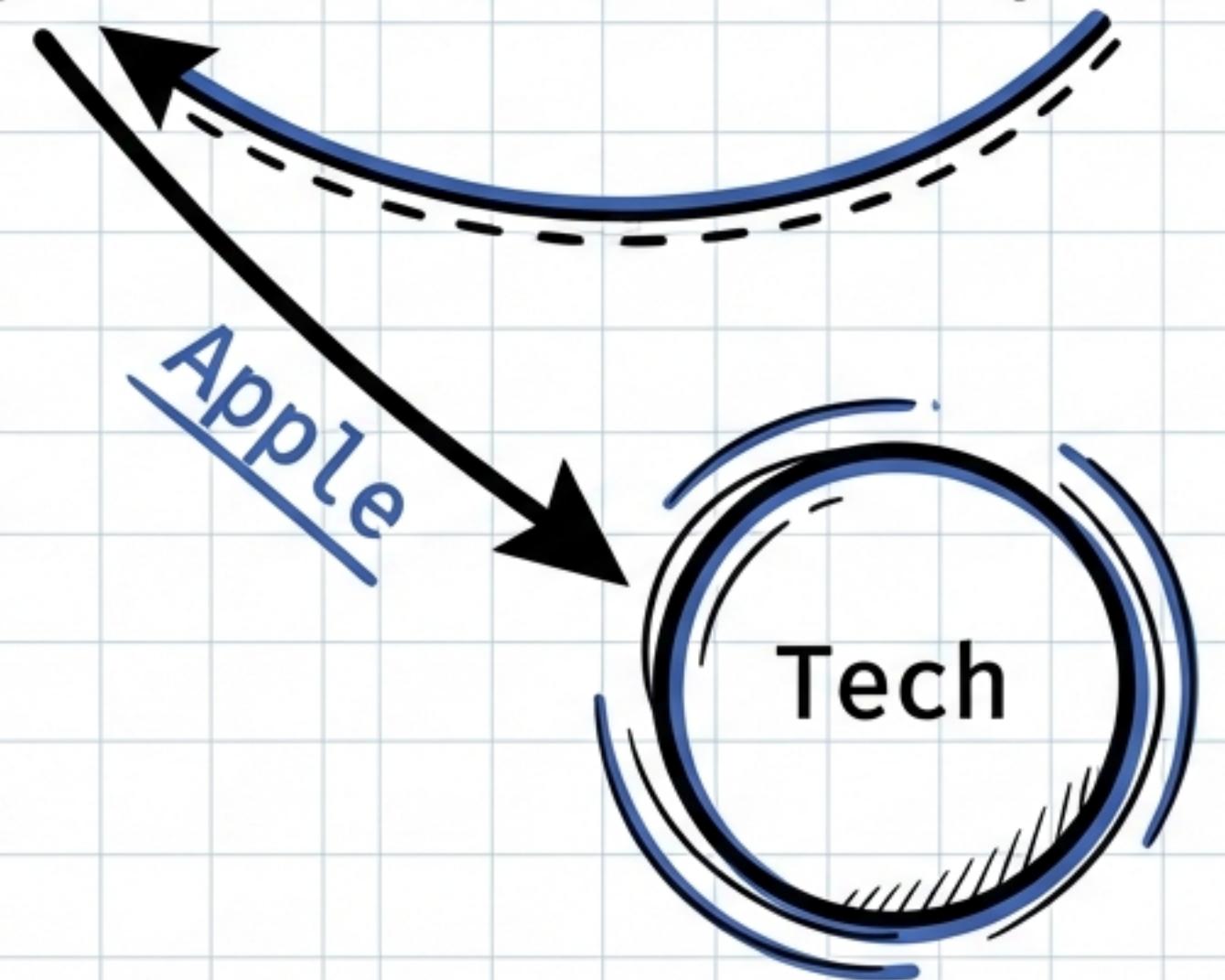
Value (V)

The actual
content.

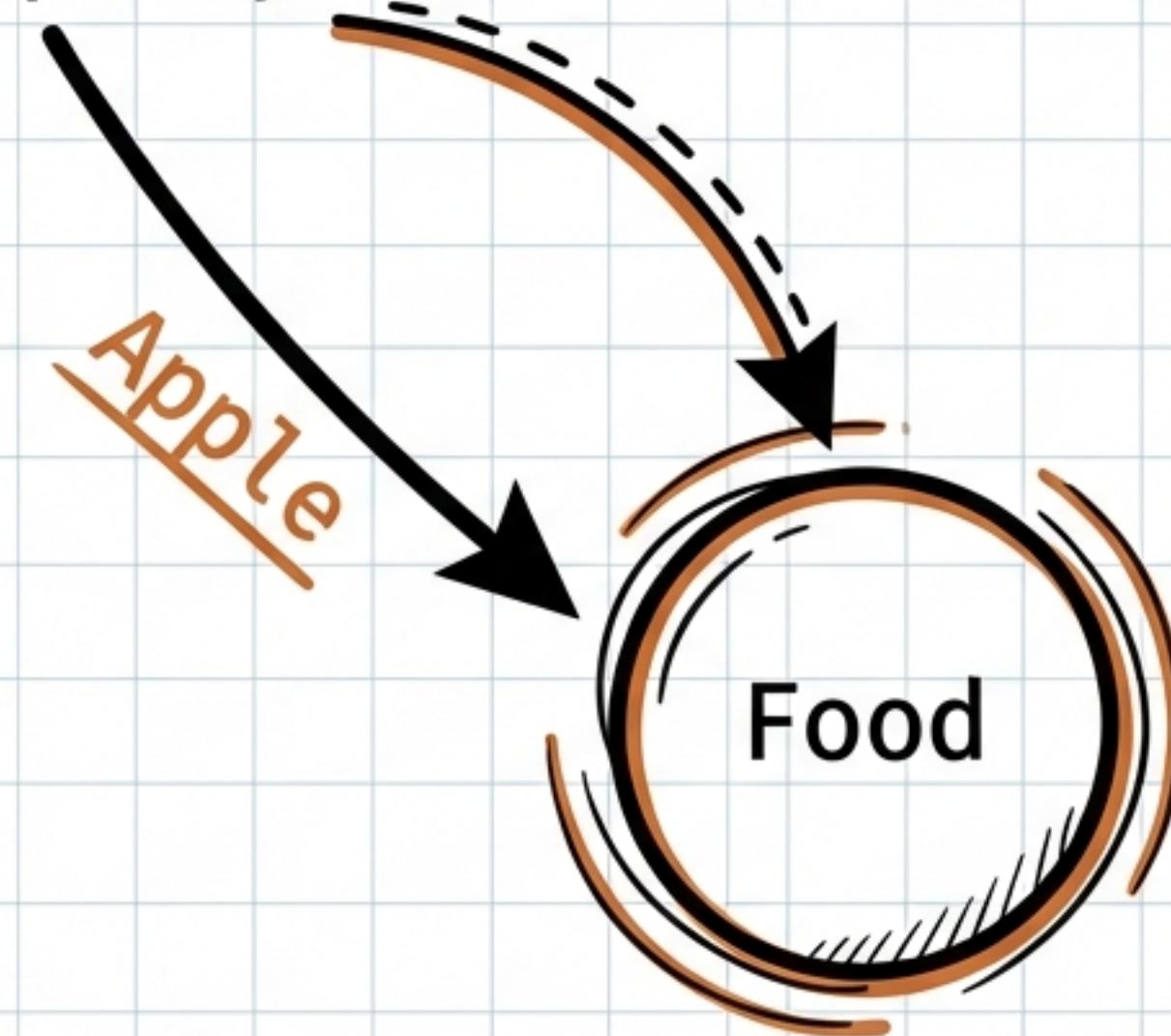
Database Analogy: Match the **Query** to the **Key** to retrieve the **Value**.

Intuition: Context Disambiguates Meaning

Apple unveiled a new phone.



Apple pie is delicious.



Intuition: Resolving Pronouns

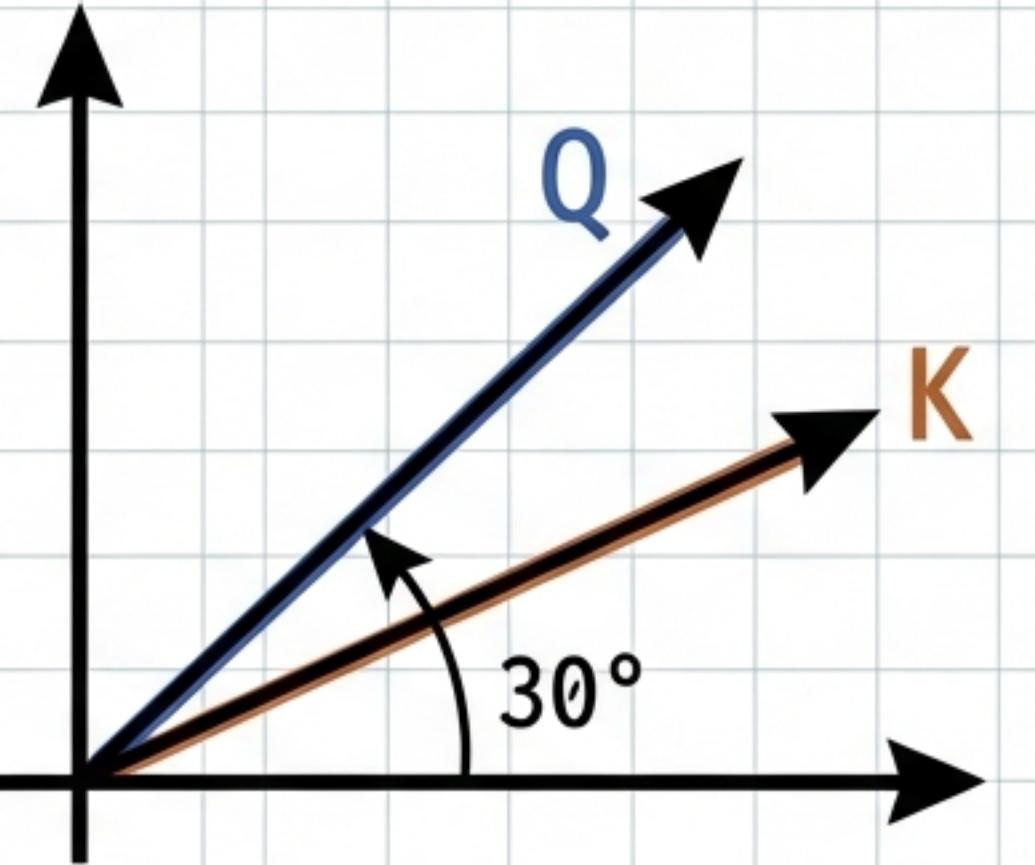
The animal didn't cross the street because **it** was too tired.

A horizontal blue bar with the text "High Attention Score" is positioned above the word "it". A thick black curved arrow points from the left side of the bar down towards the word "it". Below the word "it", there is a small square box with a dashed line extending upwards from it, pointing towards the blue bar. The word "it" is enclosed in this square box.

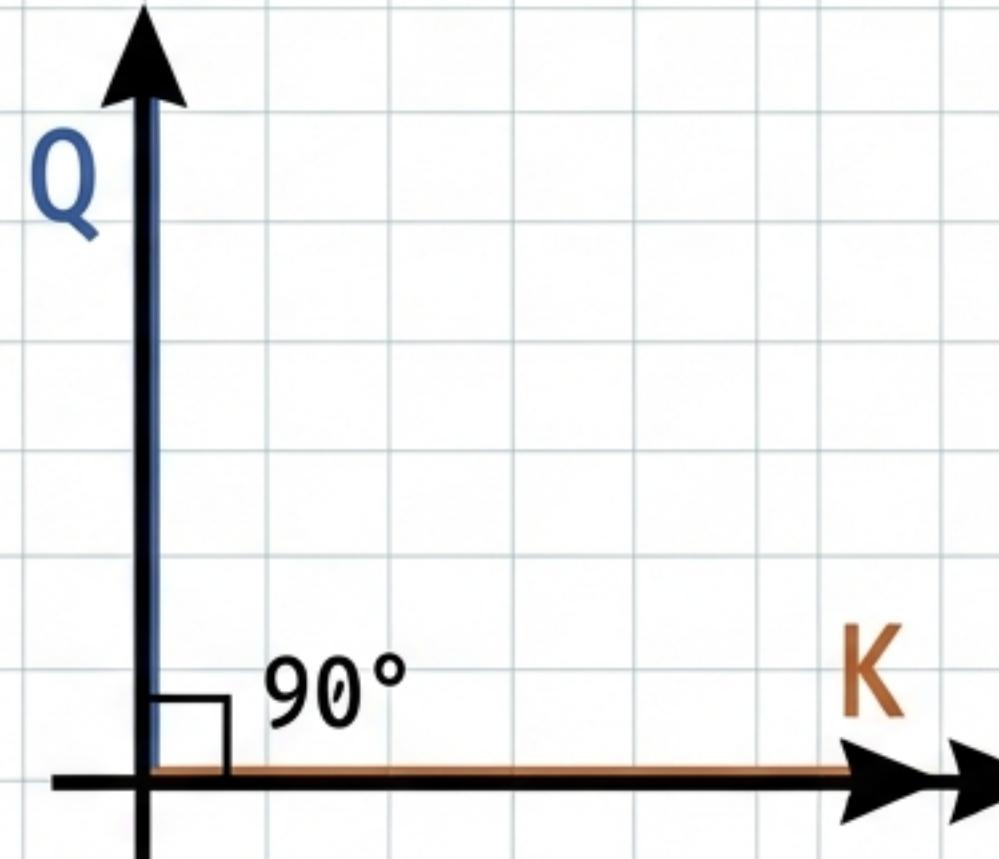
Low Attention Score

The model resolves “it” by attending to the entity capable of being “tired”.

The Math of Similarity: The Dot Product



High Dot Product = Similar



Zero Dot Product = Not Similar

Similarity = $Q \cdot K$ (Transpose)

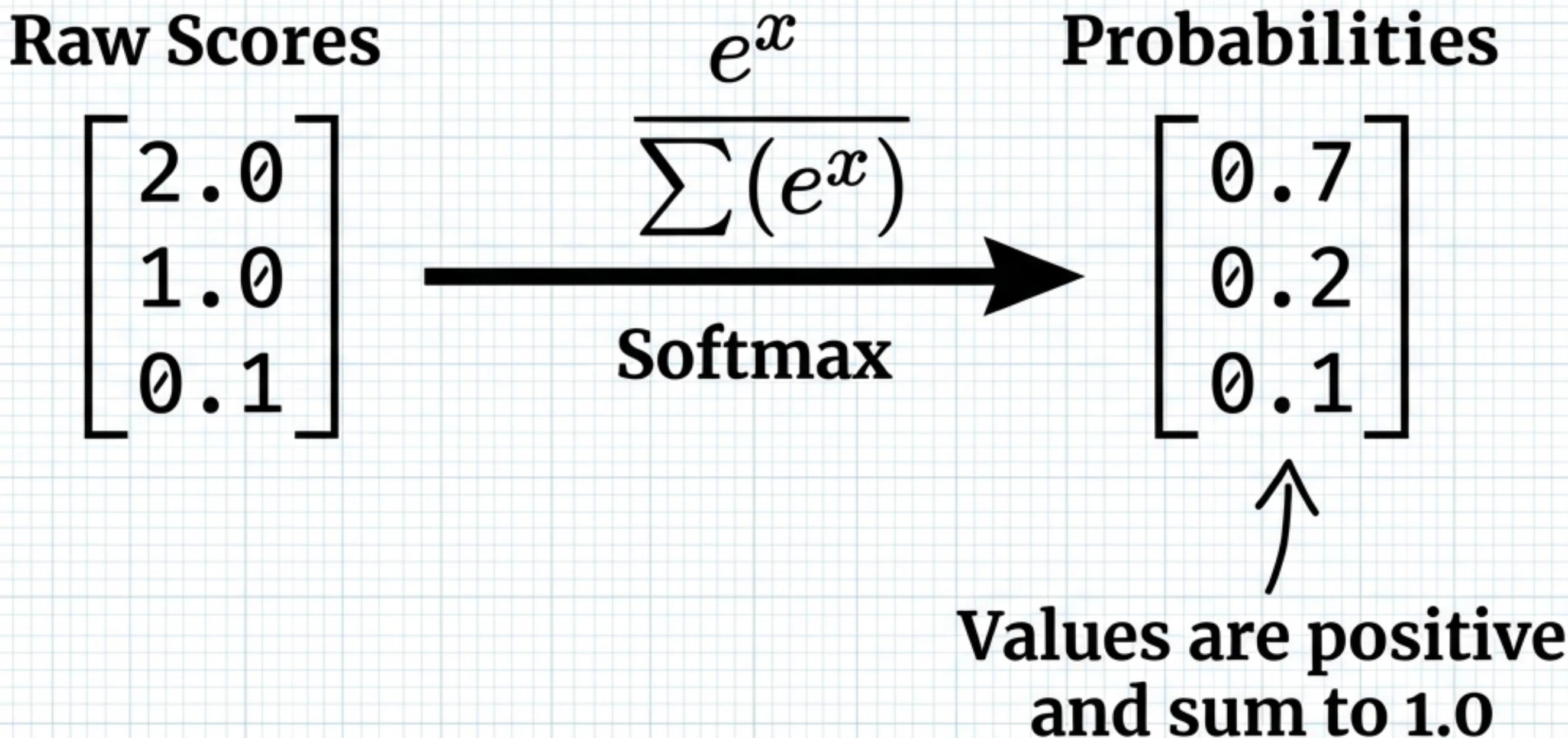
Scaling the Dot Product

$$\text{Score} = \frac{\mathbf{Q} \cdot \mathbf{K}^T}{\sqrt{d_k}}$$

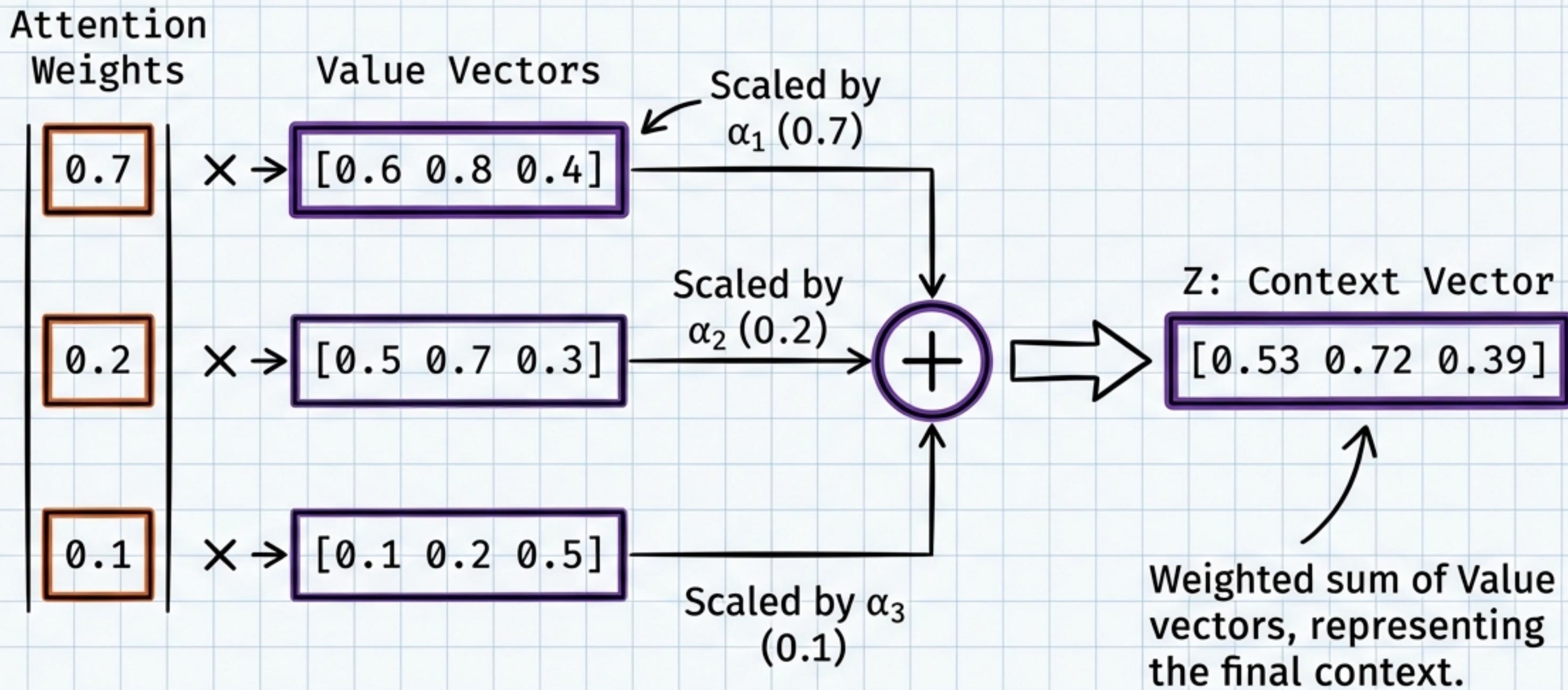
Normalization Factor

Prevents extremely large values
that would cause exploding
gradients during training.

Normalization: The Softmax Function



Computing the Context Vector



The Matrix View

$$\text{Attention}(Q, K, V) = \text{softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) V$$

Diagram illustrating the components of the Attention formula:

- Similarity**: Curly brace under Q, K, V .
- Probability**: Curly brace under the softmax term.
- Scaling**: Curly brace under the denominator $\sqrt{d_k}$.
- Content Extraction**: Curly brace under the matrix V .

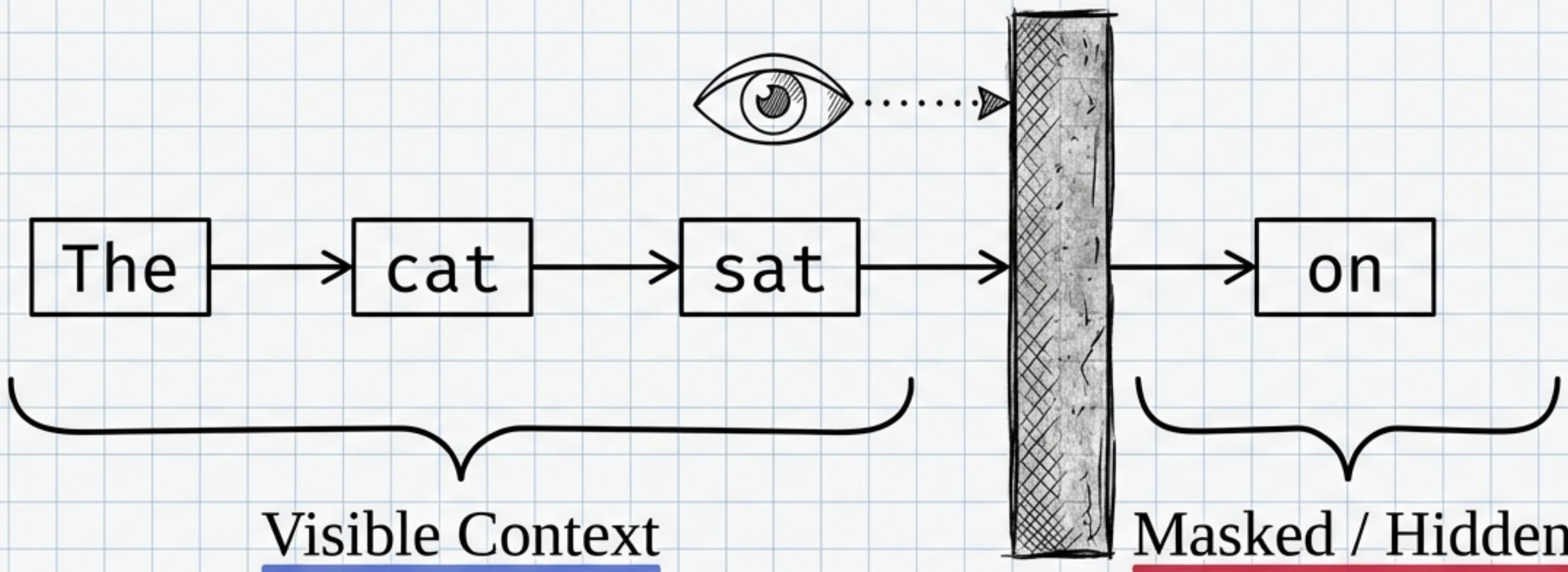
Visualizing the Matrix

Your journey starts with one					
Your	0.05	0.10	0.25	0.10	0.50
journey	0.15	0.45	0.20	0.10	0.10
starts	0.05	0.10	0.28	0.10	0.10
with	0.05	0.10	0.25	0.10	0.10
one	0.05	0.10	0.25	0.10	0.50



Simultaneous
calculation for
all word pairs.

Causal Attention: Hiding the Future



Implementing the Mask

Pre-Softmax Scores

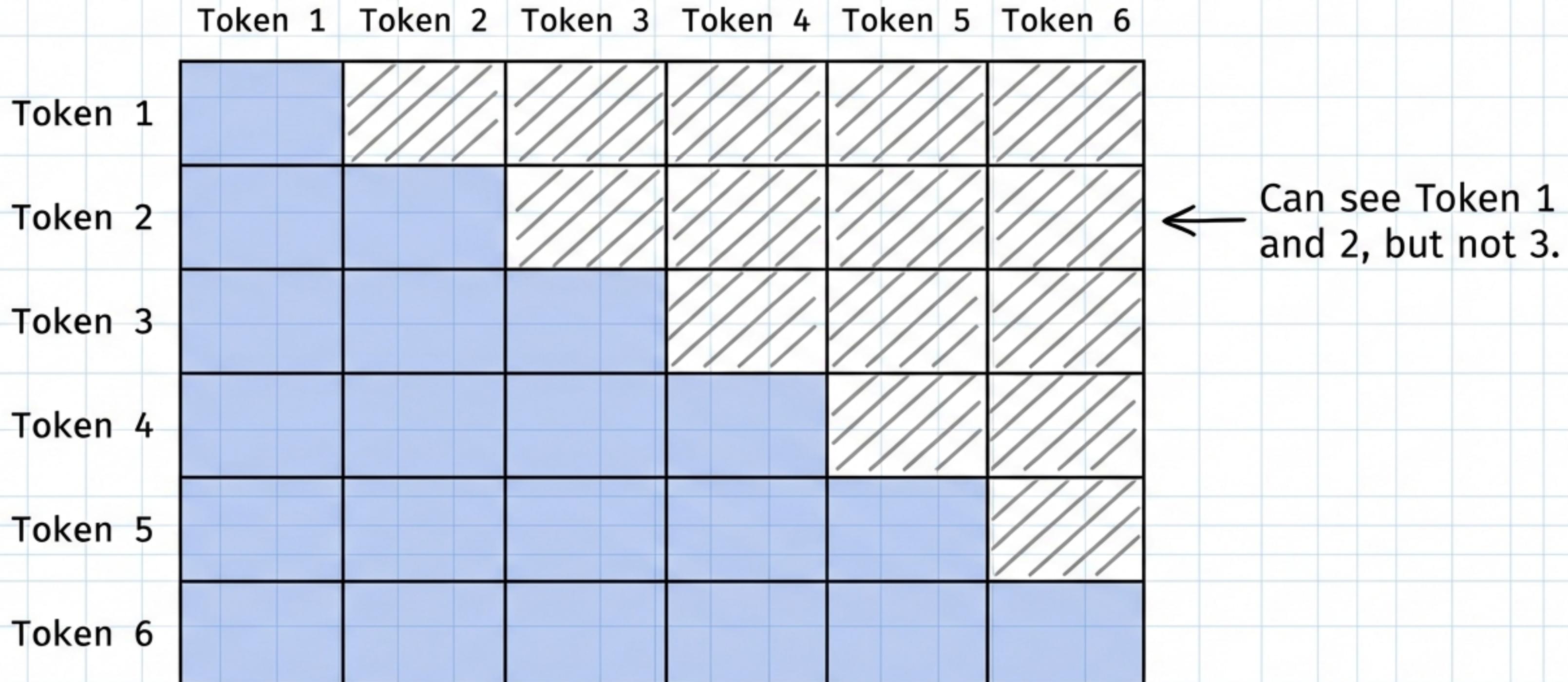
-inf	-inf	-inf	-inf
1.2	0.8	-inf	-inf
2.1	-0.5	3.0	1.1
0.2	2.5	0.9	1.4

Softmax(-inf) ≈ 0

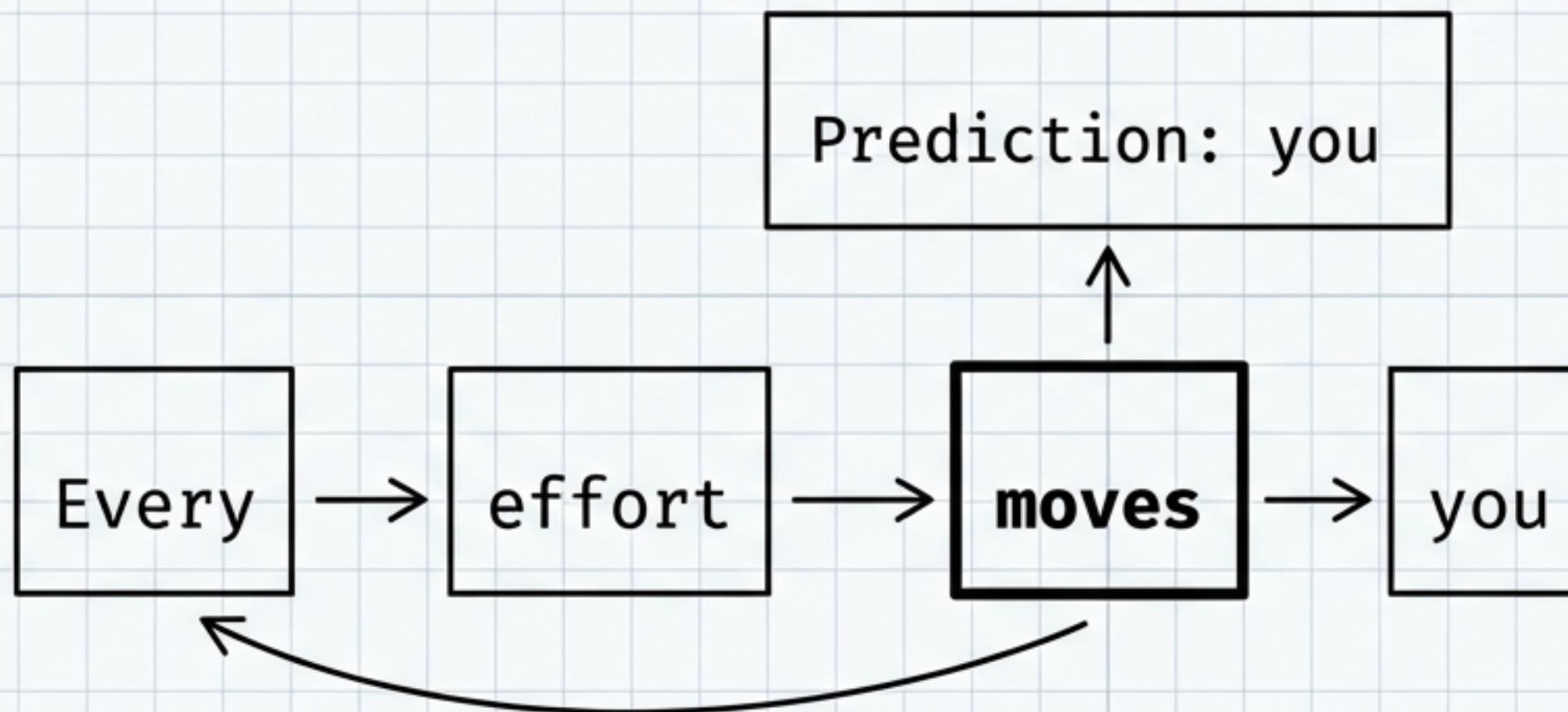
Post-Softmax Probabilities

0.0	0.0	0.0	0.0
0.40	0.32	0.70	0.0
0.10	0.85	0.85	0.45
0.08	0.65	0.35	0.40

The Look of Causal Attention



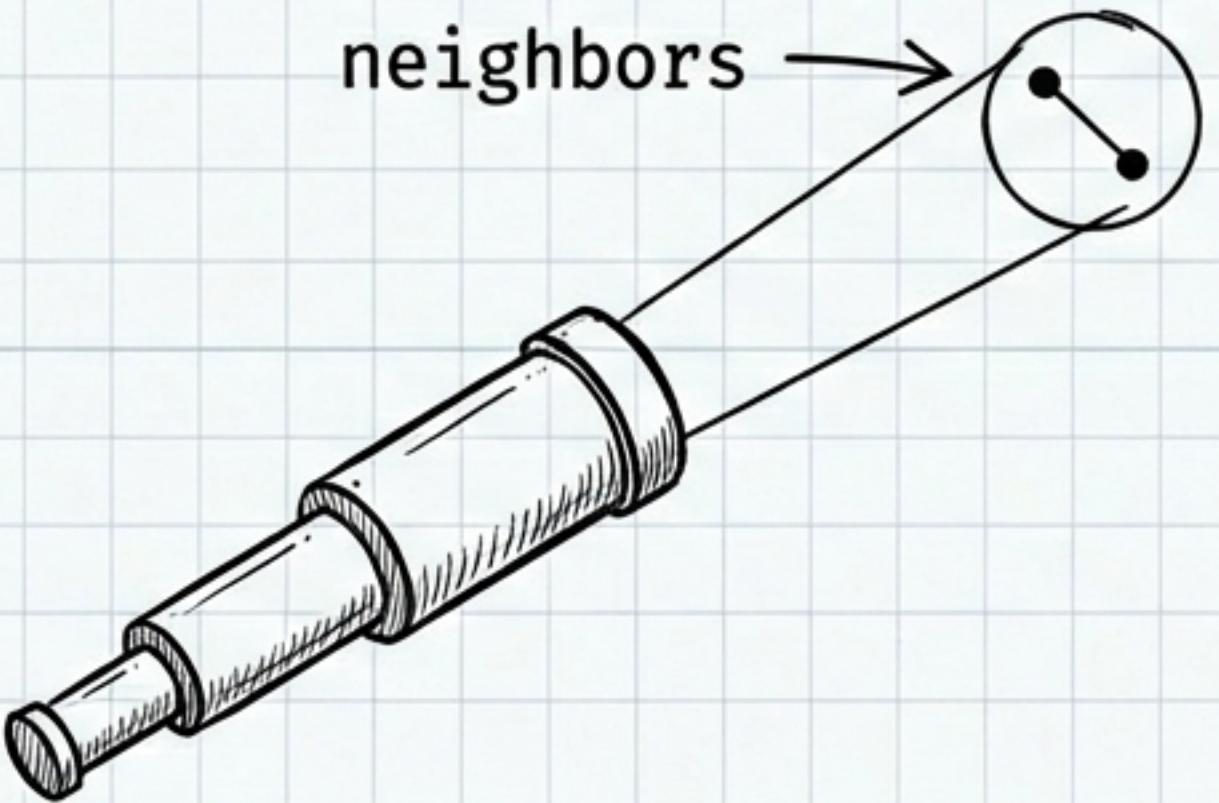
Putting It Together



The model uses the context of “Every effort moves” to predict the next token “you”.

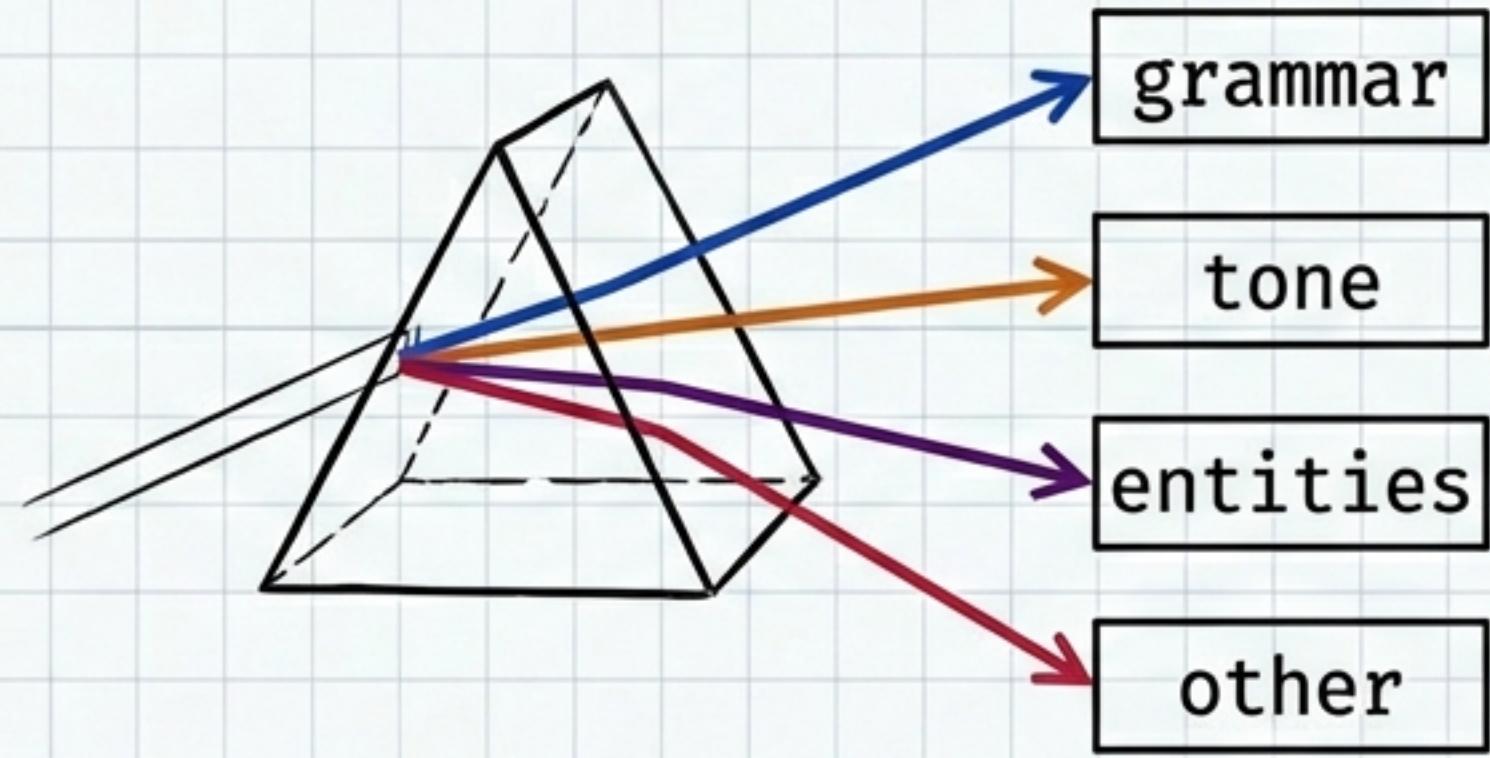
Why One Head Isn't Enough

Single Head



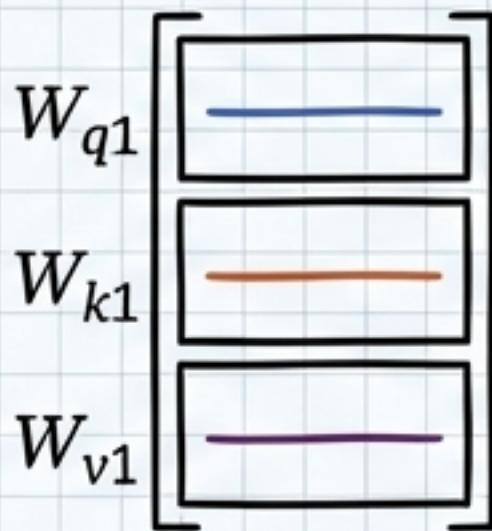
Focuses on one relationship
(e.g., neighbors).

Multi-Head

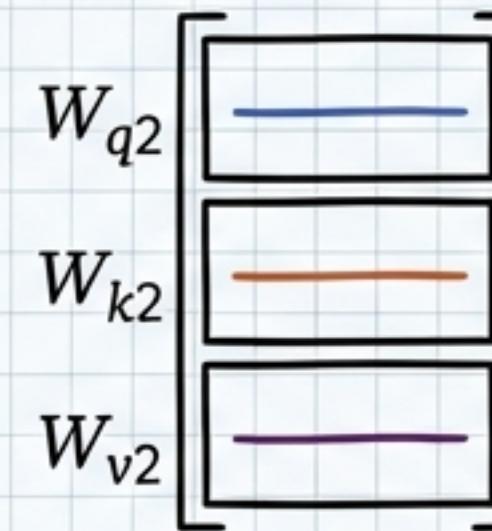


Focuses on multiple relationships
(grammar, tone, entities)
simultaneously.

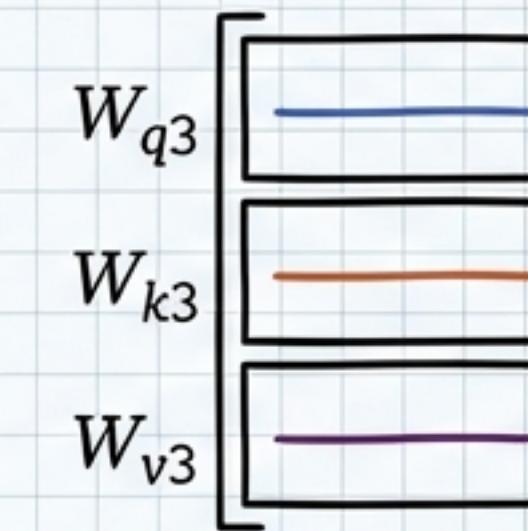
Multi-Head Attention: A Panel of Experts



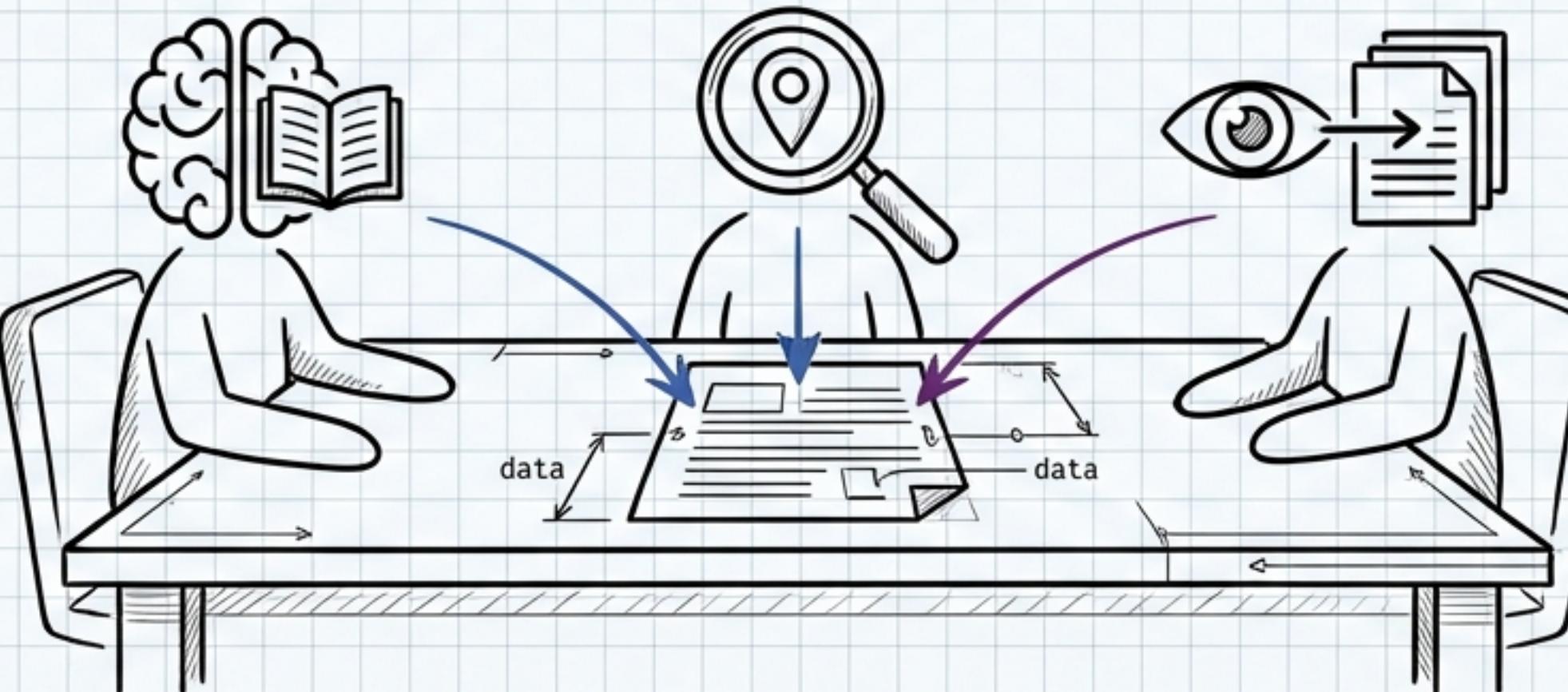
Grammar Head
(Checks verb forms)



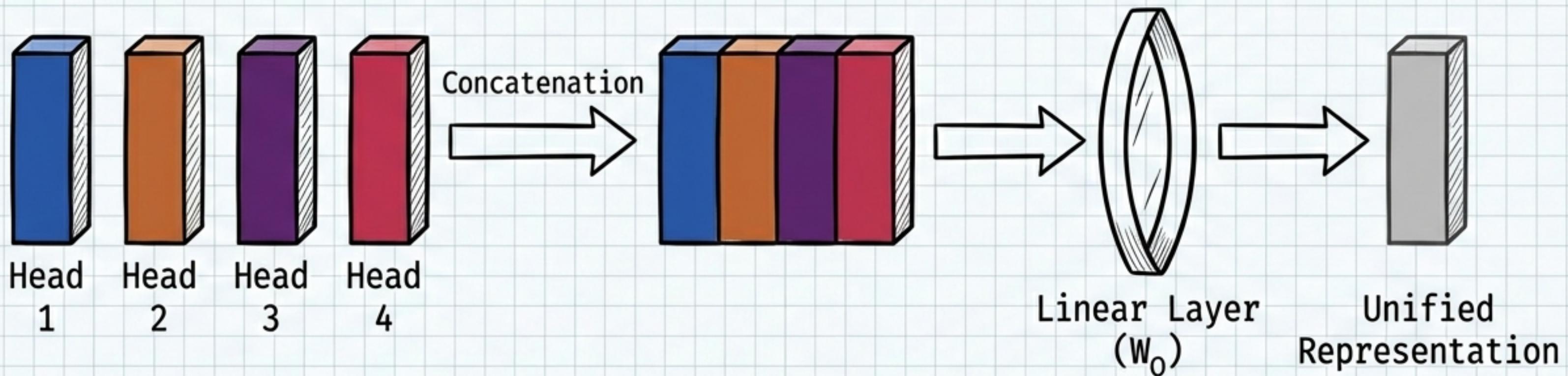
Entity Head
(Checks names/places)



Context Head
(Checks previous sentences)

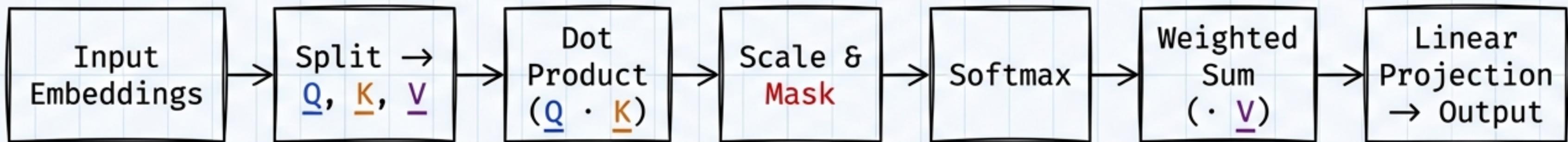


Aggregating the Insights



Combining diverse perspectives into a unified representation.

The Complete Attention Pipeline



This parallel process drives the modern AI revolution.