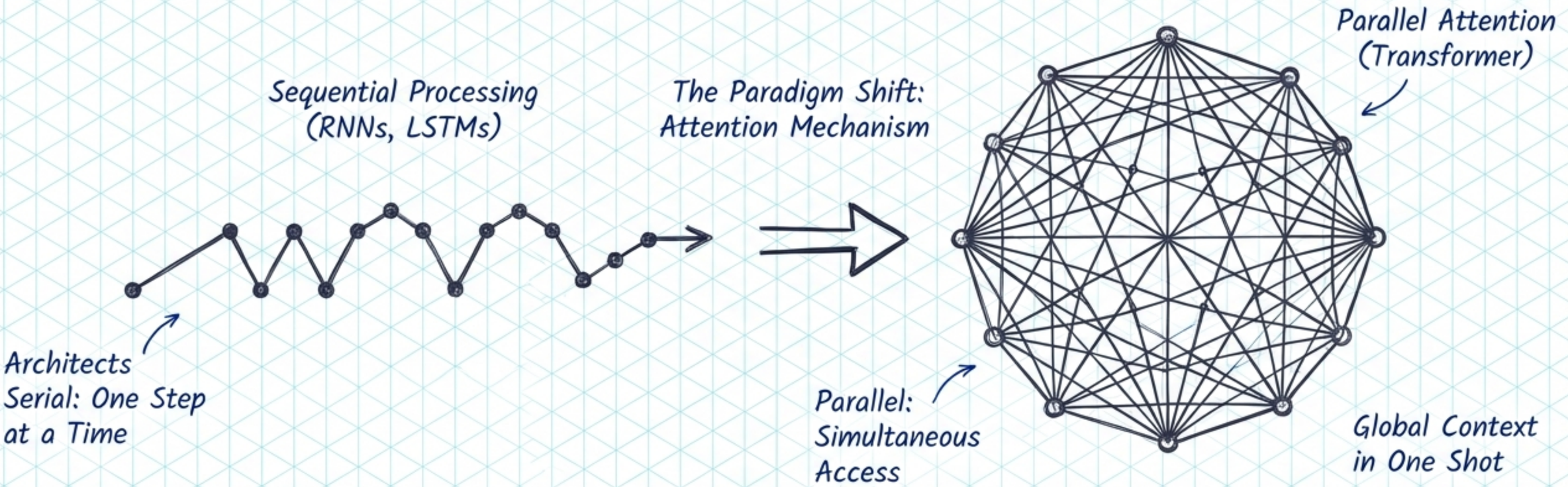
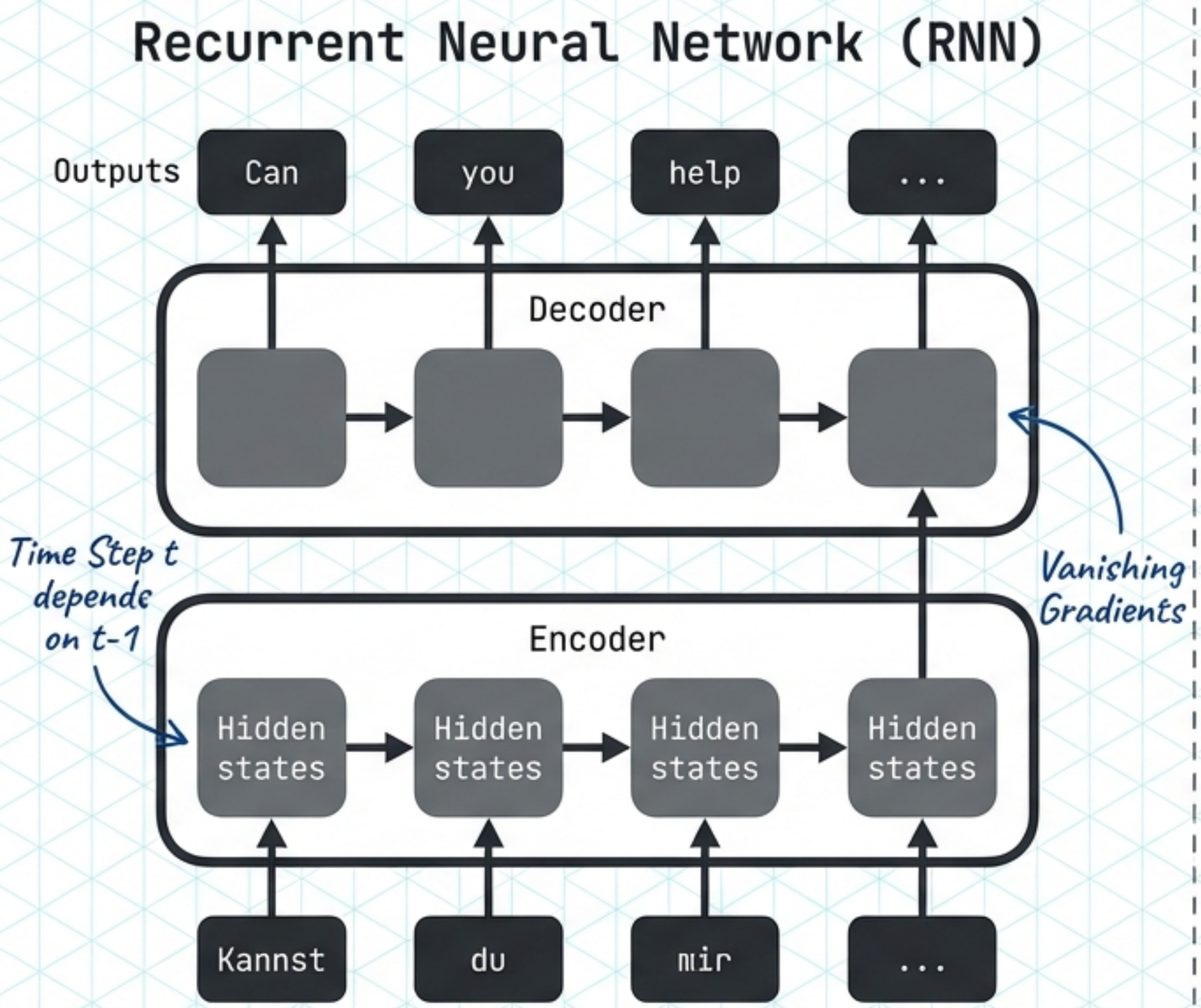
THE TRANSFORMER ARCHITECTURE

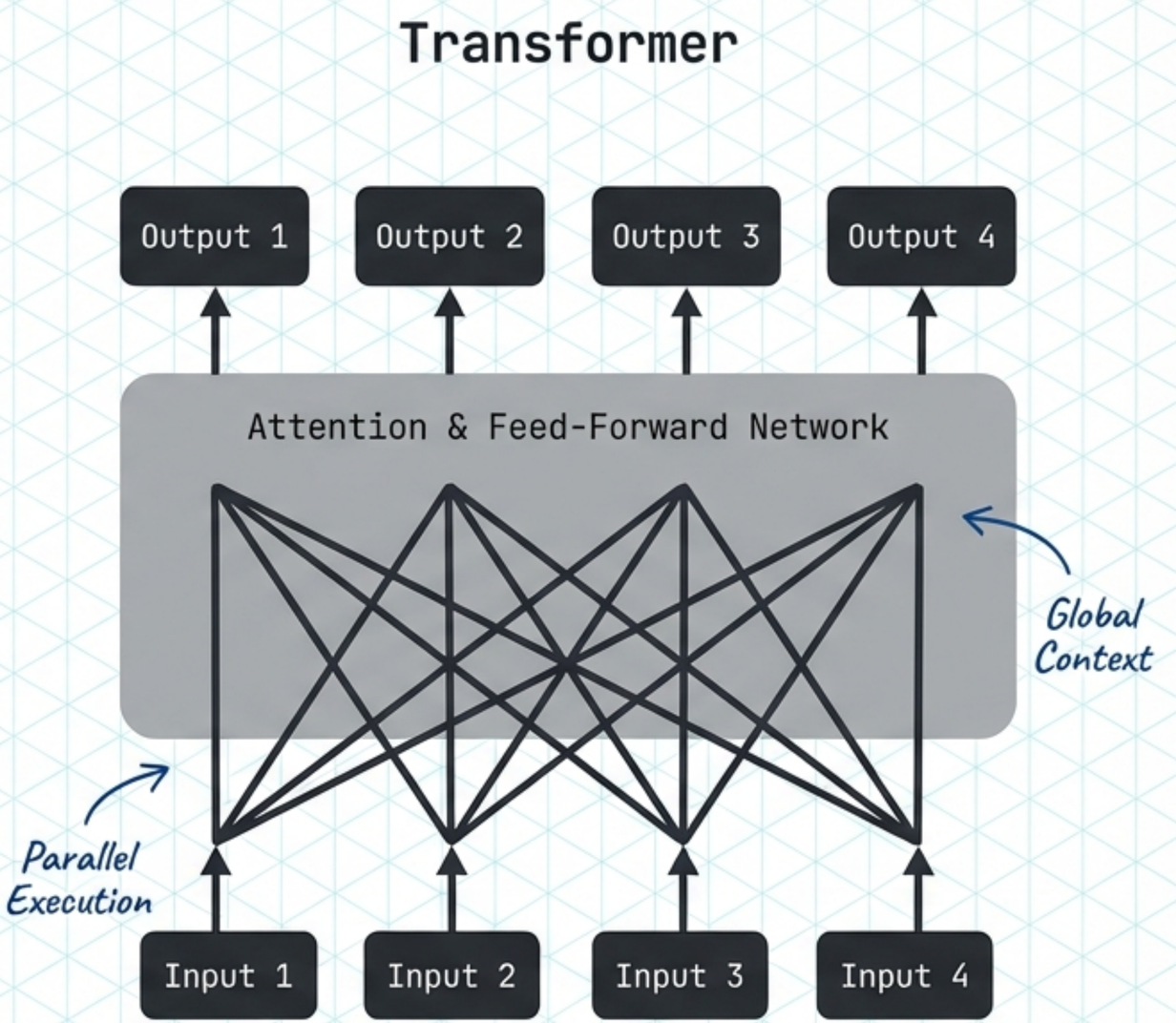
From Sequential Processing to Parallel Attention



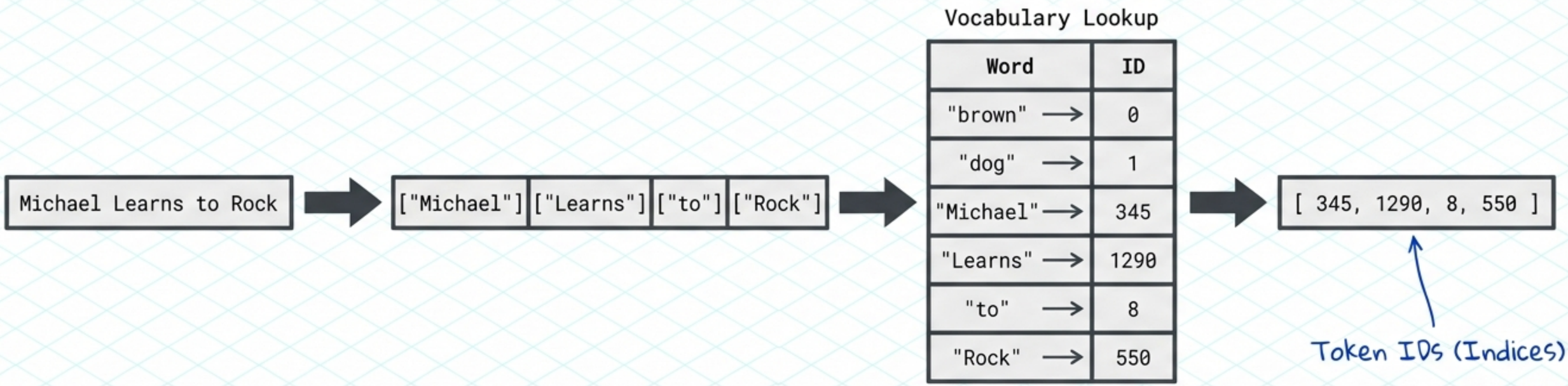
A Deep Dive for Engineering Graduate Students

THE SEQUENTIAL BOTTLENECK

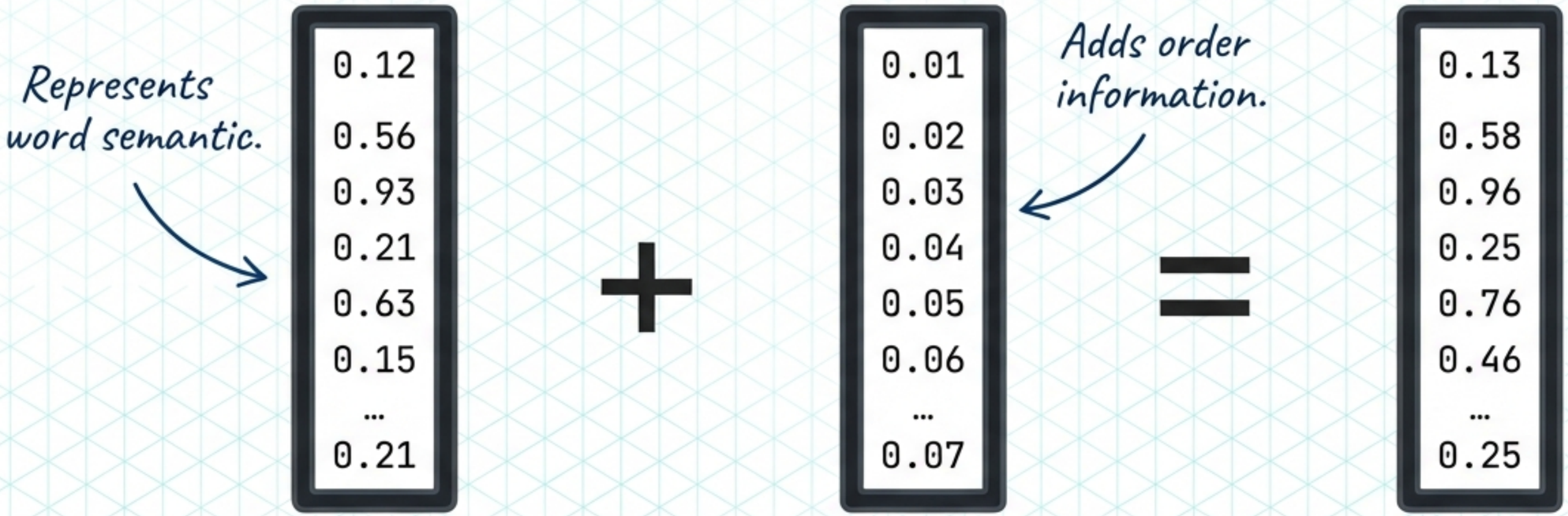




INPUT PIPELINE: TOKENIZATION



EMBEDDINGS & POSITIONAL ENCODING



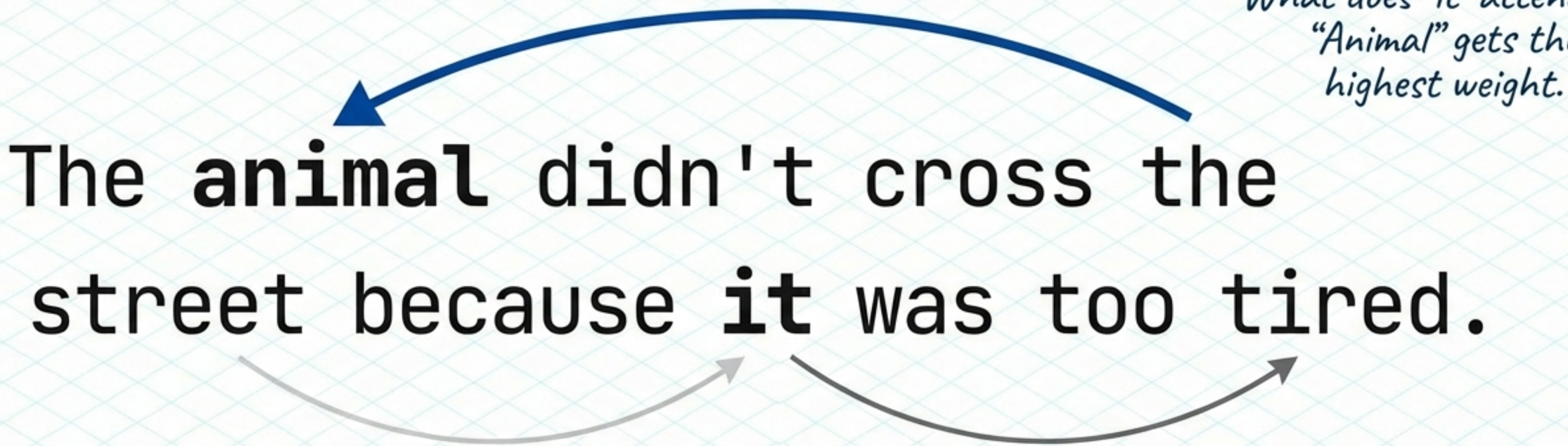
Problem: Sets have no order. {A, B} = {B, A}.

Solution: Inject geometric signatures into the vectors.

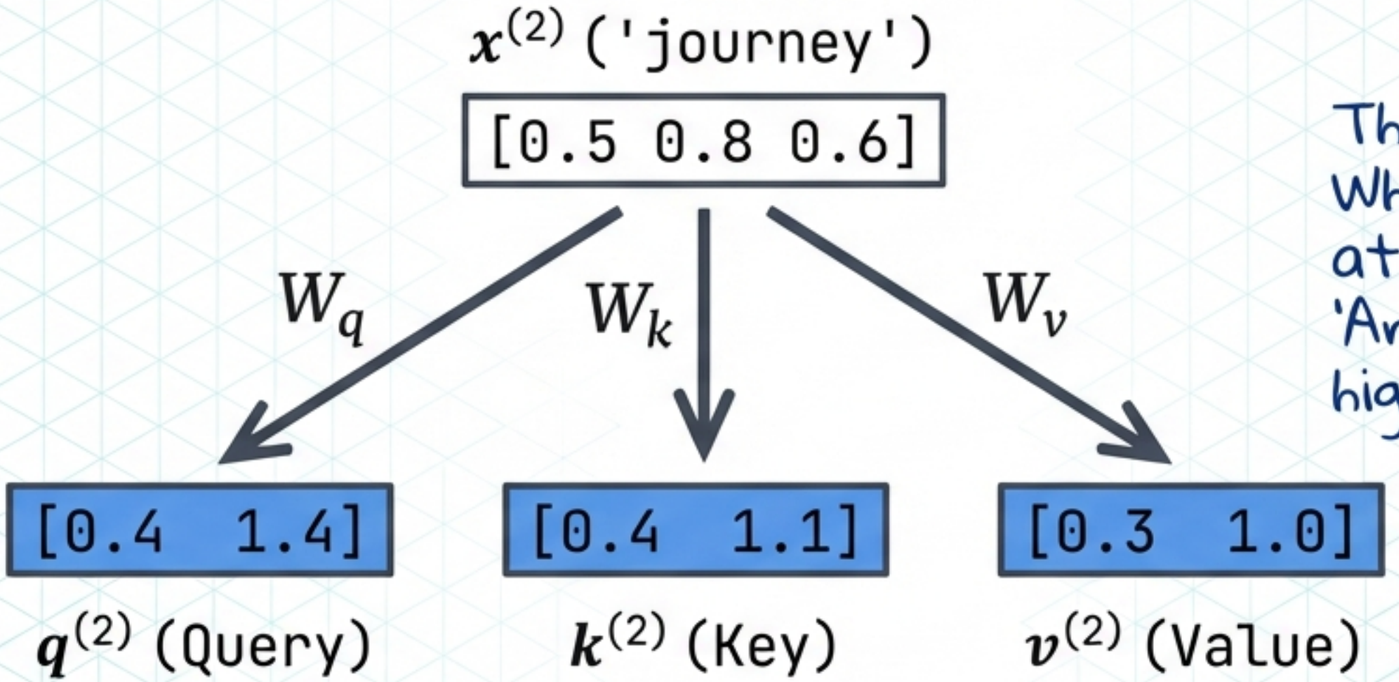
CORE MECHANISM: SELF-ATTENTION

Solving ambiguity through context.

The model asks: What does "it" attend to? "Animal" gets the highest weight.



THE MECHANICS: Q, K, V PROJECTIONS



The model asks: What does 'it' attend to? 'Animal' gets the highest weight.

Crucial: W matrices are learned. Q, K, V are derived per token.

ANALOGY 1: THE CASTING DIRECTOR

Dot Product
(Similarity Check)

Output

QUERY (Q)

The Search:
“I need an action hero.”

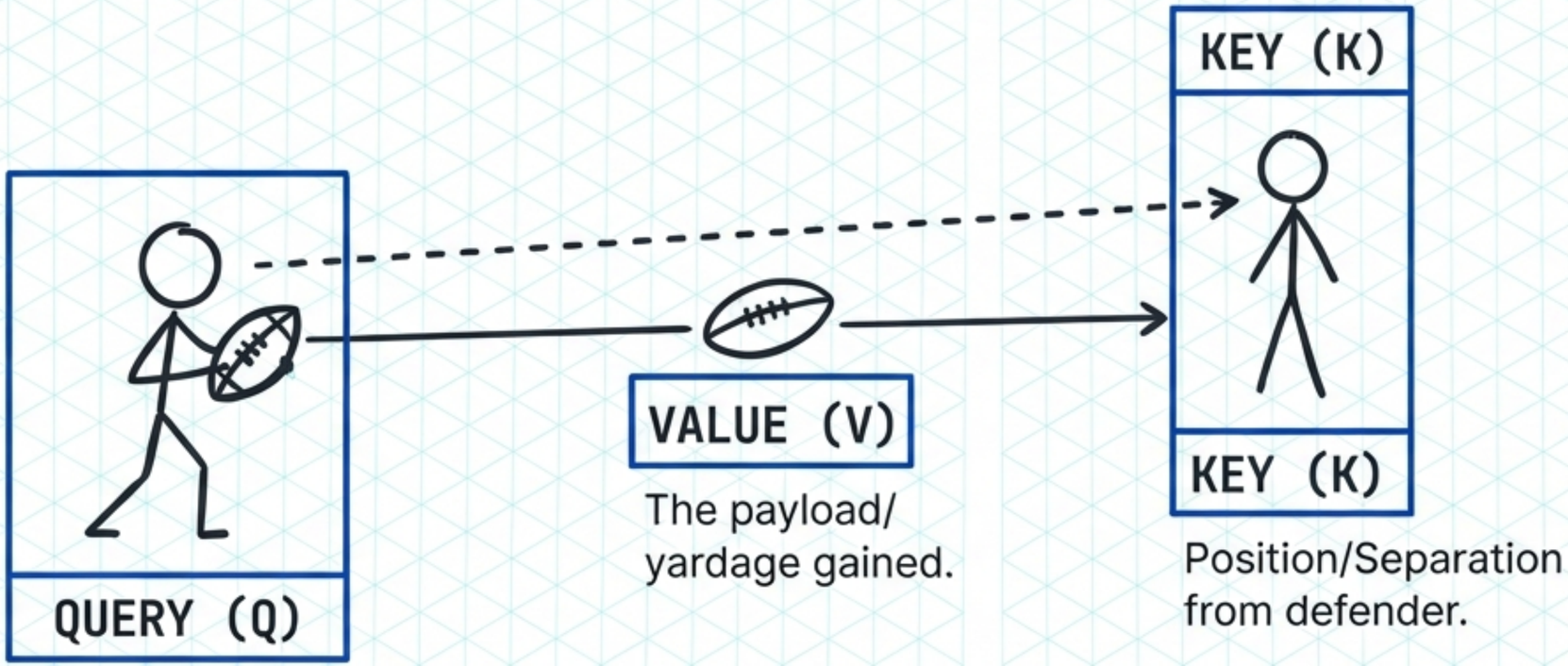
KEY (K)

The Attributes:
“I do stunts & combat.”

VALUE (V)

The Content:
The actual performance.

ANALOGY 2: THE FOOTBALL PASS



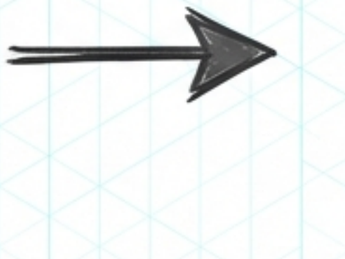
Scanning for an open player.

The payload/ yardage gained.

Position/Separation from defender.

Key Insight: The QB (Q) selects the Receiver (K) based on fit, and delivers the Ball (V) as the actual payload.

ANALOGY 3: THE VOTER



QUERY (Q)

Voter Interests: “Economy & Education”.

KEY (K)

Platform: “Lower Taxes, New Schools”.

VALUE (V)

Policies Enacted.

Key Insight: The Voter (Q) matches their interests with a Platform (K) to elect the Policies (V) that will be enacted.

THE ALGORITHM: SCALED DOT-PRODUCT ATTENTION

Convert to Probabilities (0 to 1)

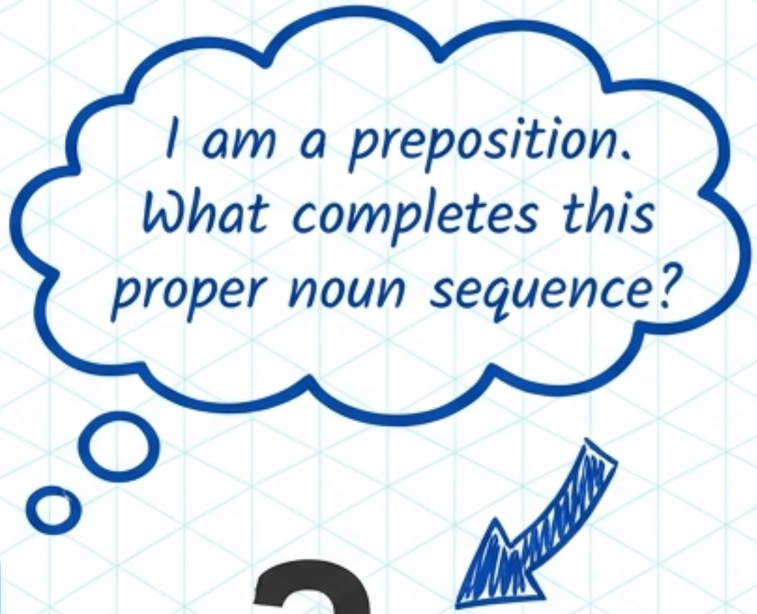
Similarity Score (Alignment)

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

Scaling Factor: Prevents vanishing gradients in Softmax

Weighted Aggregation of Information

EXAMPLE: THE QUERY



["Michael"] ["Learns"] ["to"]

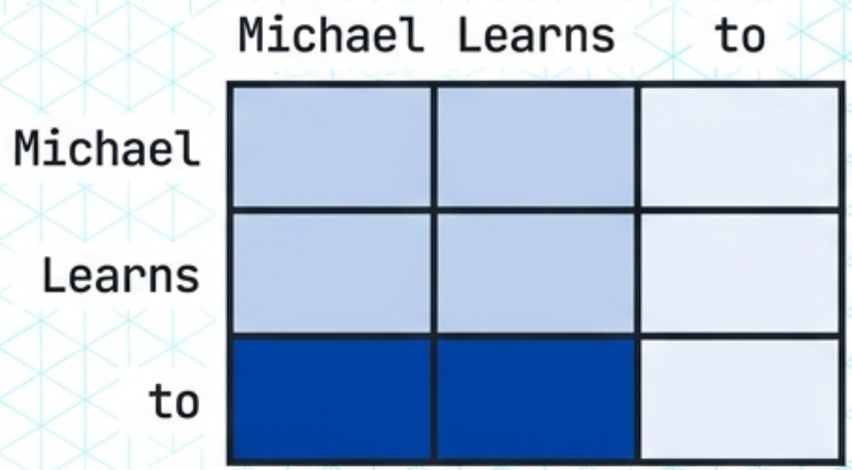
?

Q_{to}

EXAMPLE: ATTENTION & PREDICTION

Based on the "Band Name Entity" context, the most likely next word is "Rock"

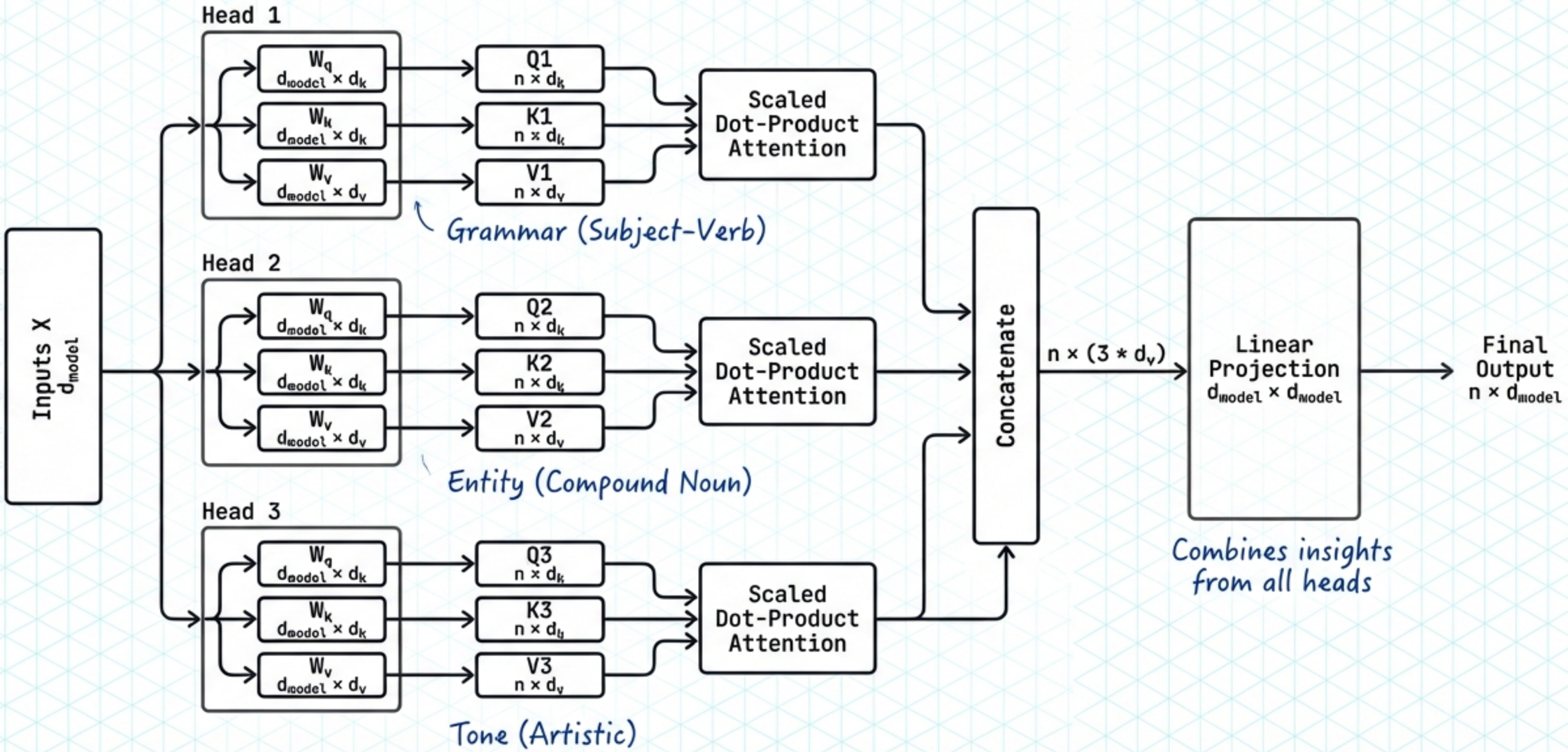
(High Probability)



Context: Band Name Entity

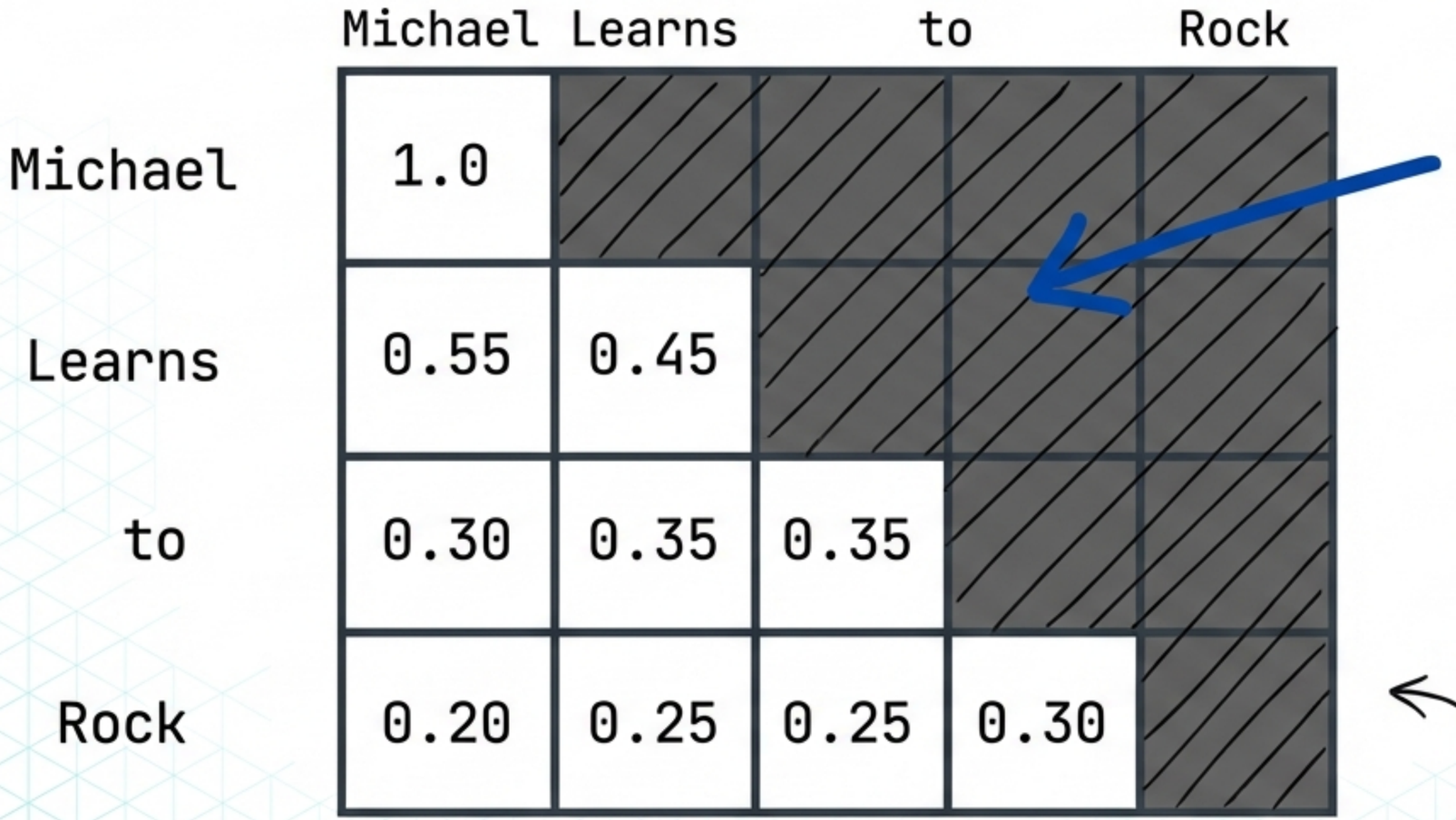
MULTI-HEAD ATTENTION

Capturing different types of relationships simultaneously.



MASKED (CAUSAL) ATTENTION

Preventing the model from seeing the future.



During training, “Michael” cannot attend to “Learns”. Future is unknown.

Masked (Causal) Attention: Upper triangle zeros out future positions.

THE TRANSFORMER BLOCK

The Fundamental Building Block of GPT.