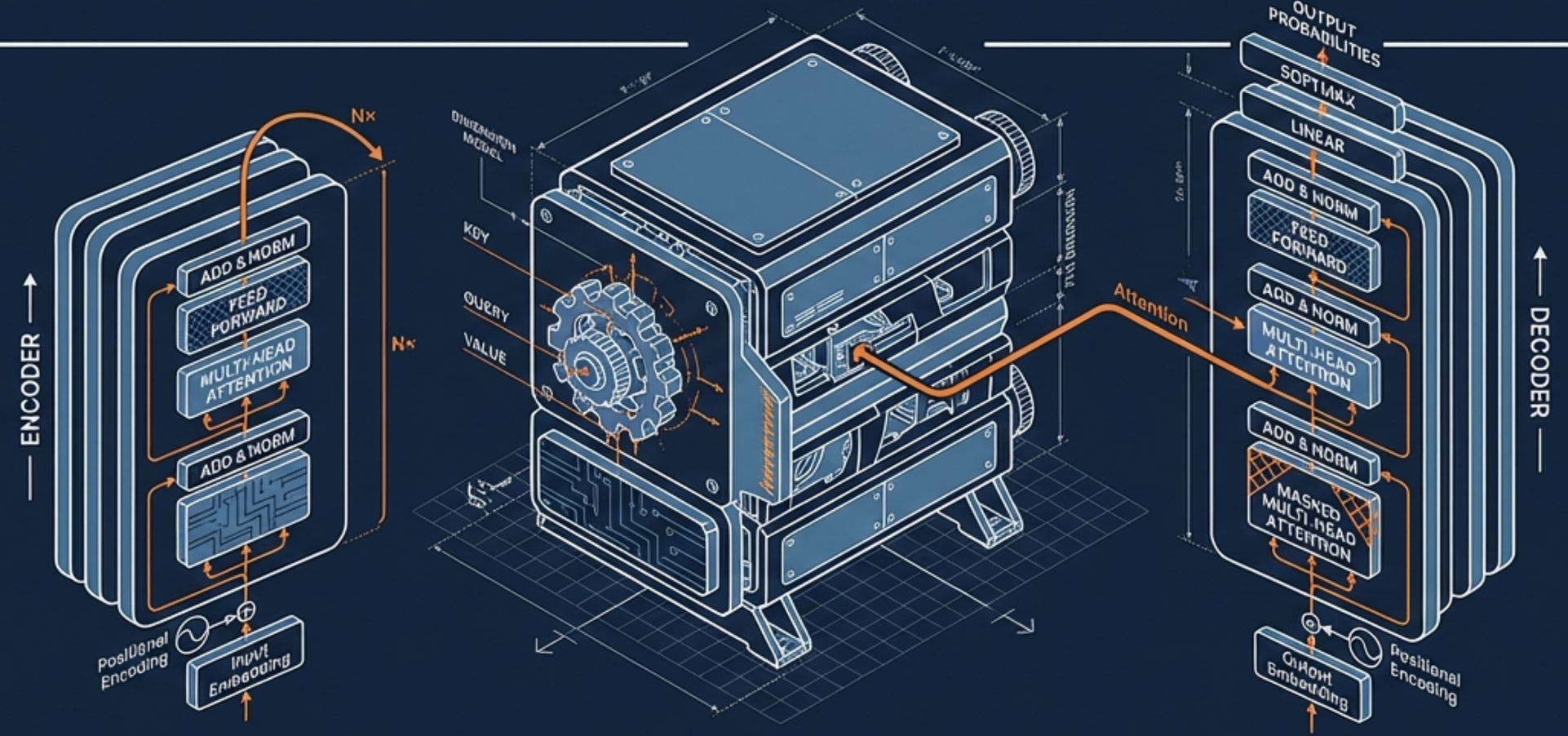
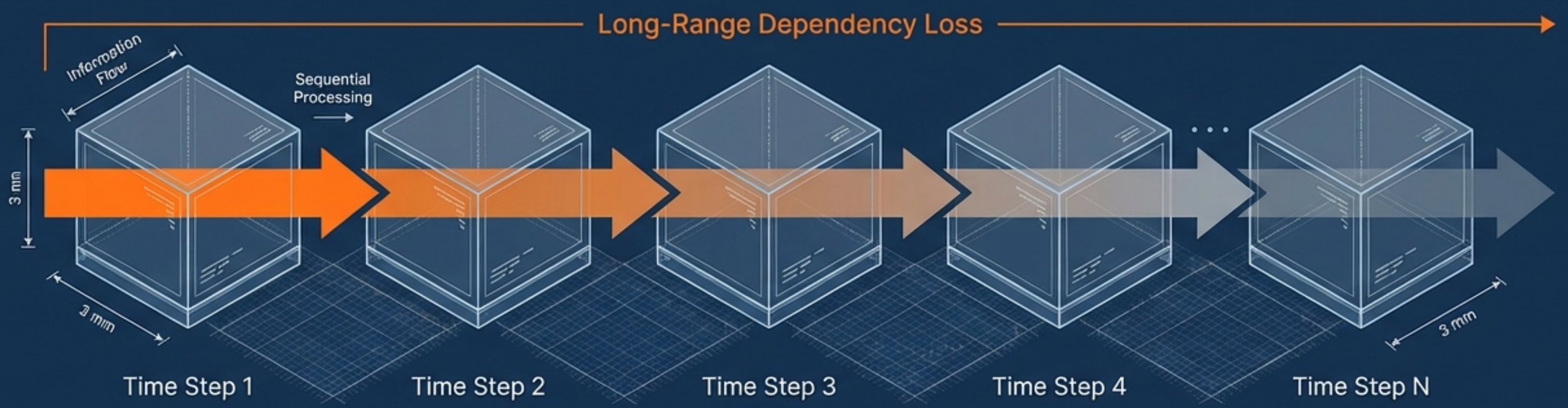
THE TRANSFORMER ARCHITECTURE

A Deep Dive into “Attention Is All You Need”



Moving from Sequential Recurrence to Parallel Attention.

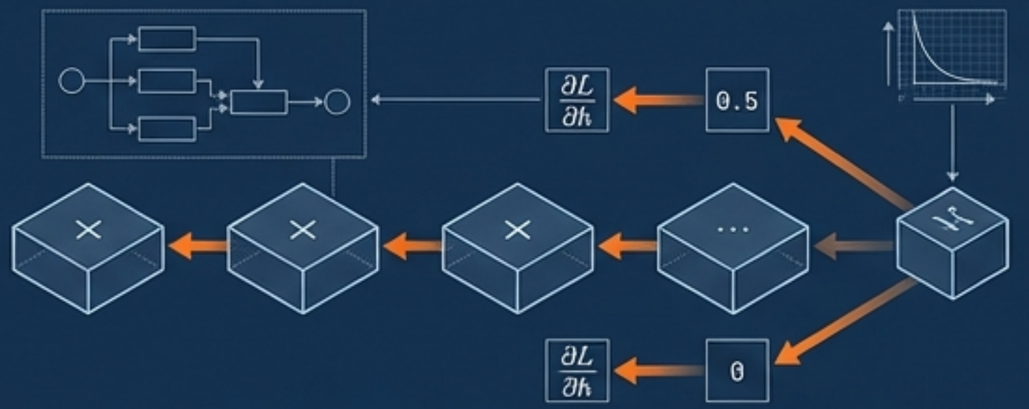
The Bottleneck of Recurrent Neural Networks (RNNs)



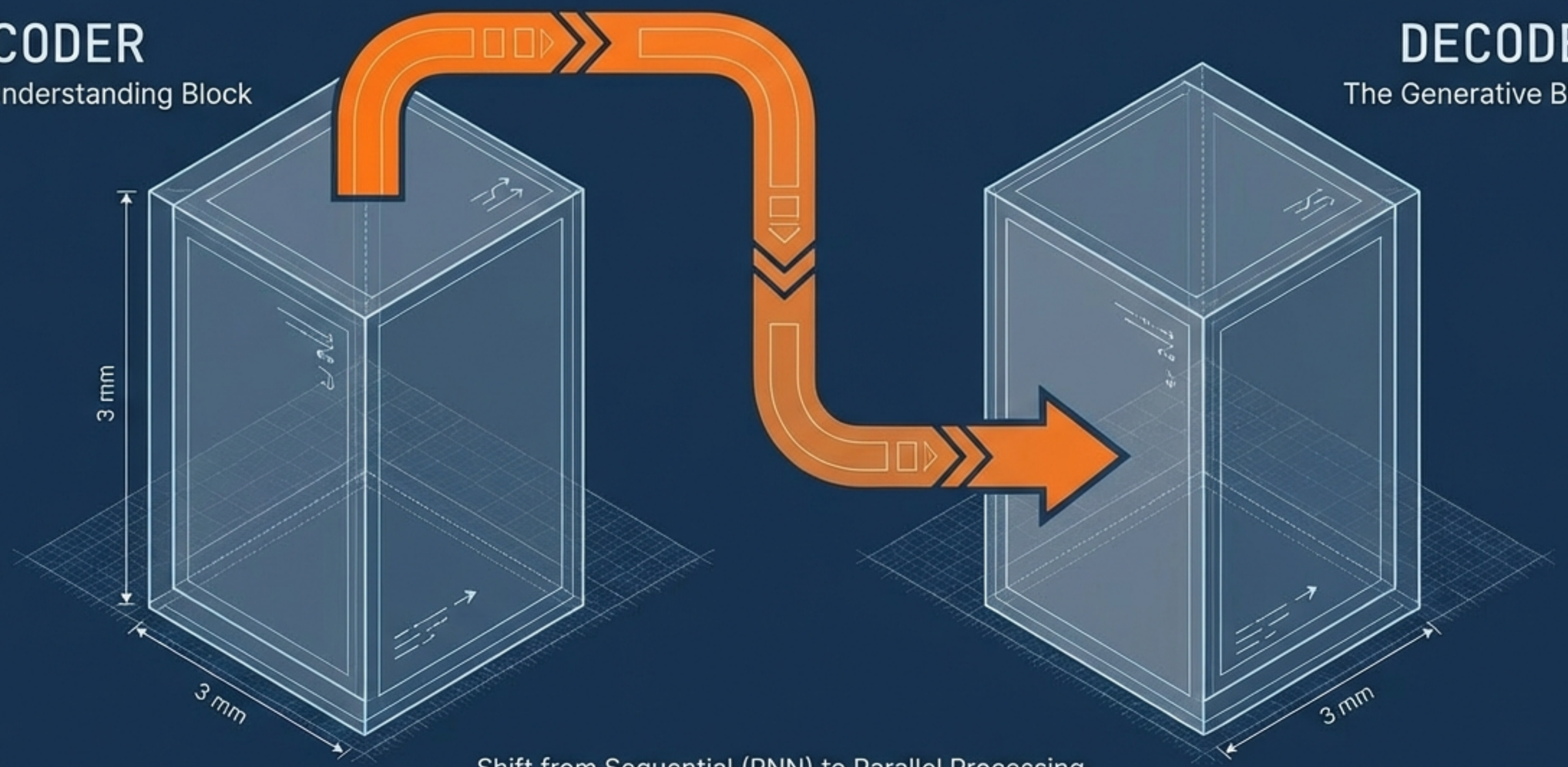
The Vanishing Gradient Problem

$$0.5 \times 0.5 \times 0.5 \times \cdots \rightarrow 0.0$$

Chain Rule: Multiplying derivatives < 1 results in zero updates.



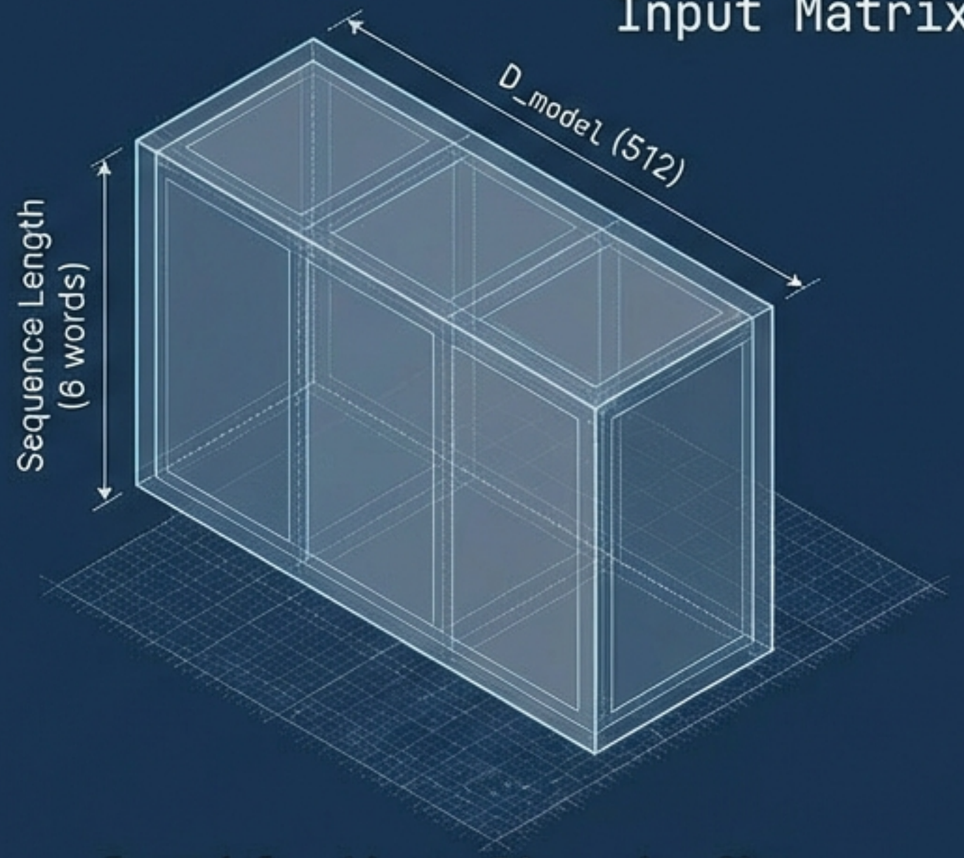
The Encoder-Decoder Architecture



Shift from Sequential (RNN) to Parallel Processing.
The Encoder extracts features; the Decoder generates output.

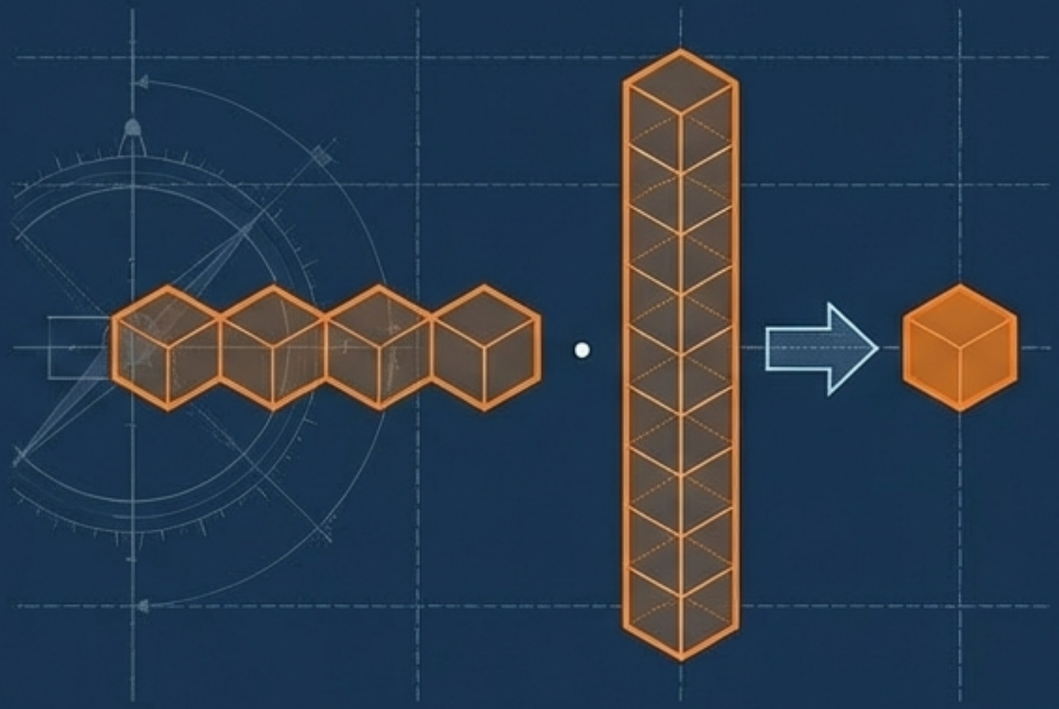
The Vocabulary of Dimensions

Matrix Shapes



`D_model` = Vector Dimension Size

Dot Product Visual



`Row · Column = Scalar Score`

Matrix Multiplication: Inner dimensions cancel, outer dimensions remain.

Inside the Encoder

ENCODER

The Understanding Block

ENCODER PROCESS:

1. Input Embeddings
2. Positional Encoding
3. Multi-Head Attention
4. Add & Norm
5. Feed Forward

Feed Forward

Add & Norm

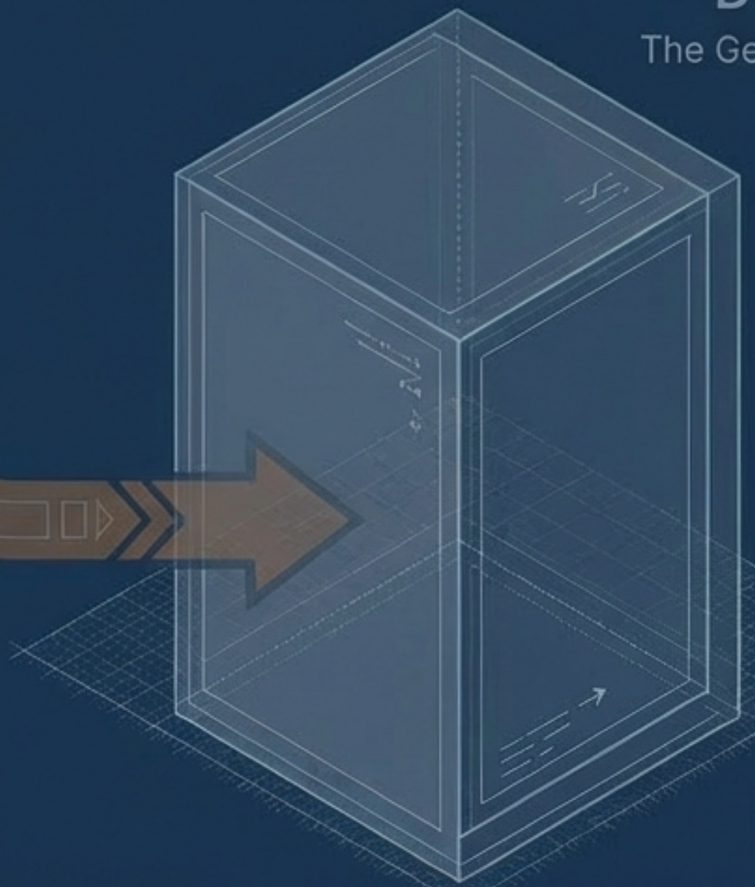
Multi-Head Attention

Positional Encoding

Input Embeddings

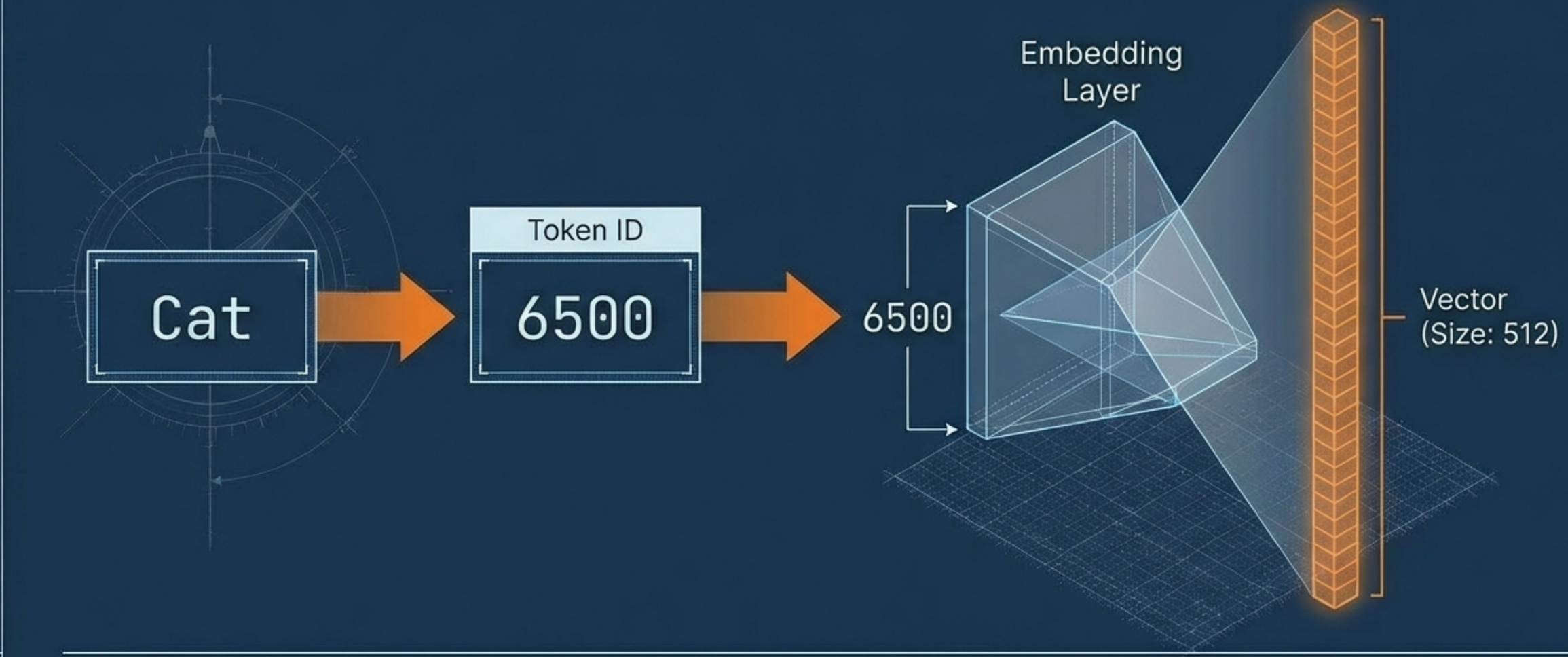
DECODER

The Generative Block (Ghosted)



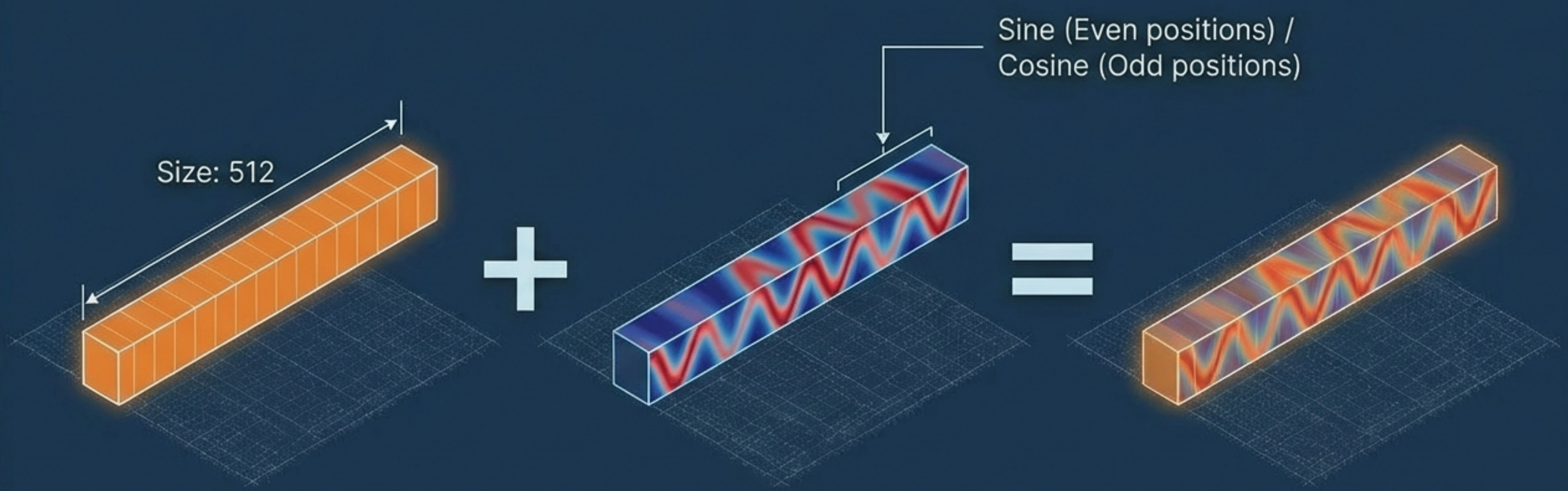
Detailed view of the internal processing layers within the Encoder block.

Input Embeddings



Learnable Parameters: Weights are updated during training to capture semantic meaning.

Positional Encoding



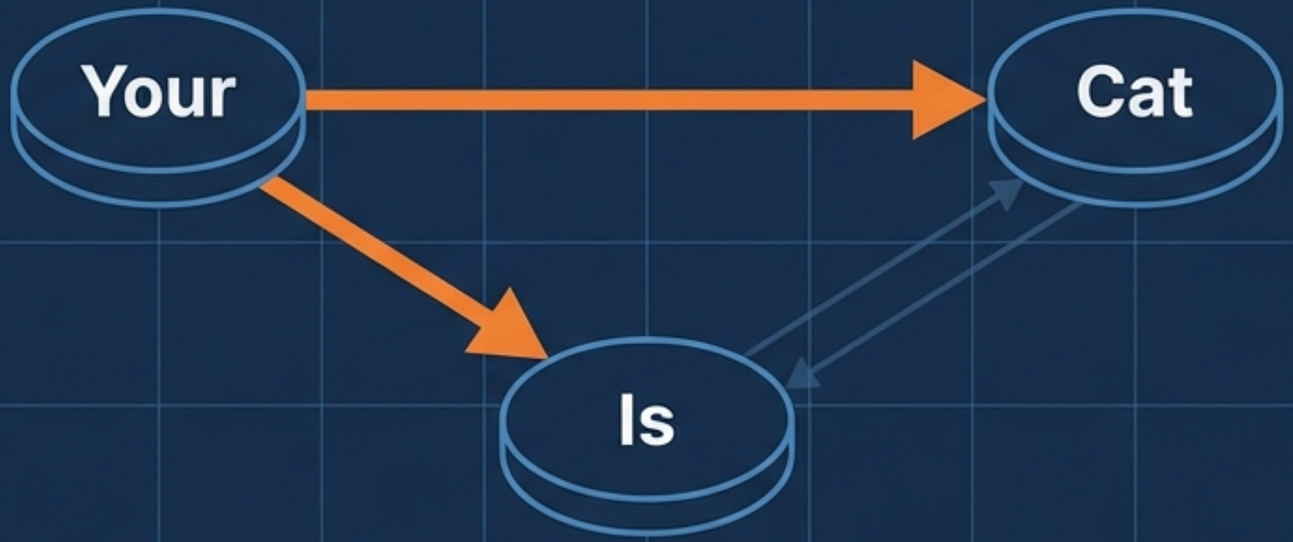
Embedding Vector

Waveform Vector

Final Input Vector

Injecting Order: Since the model processes in parallel, this pattern gives the model spatial awareness.

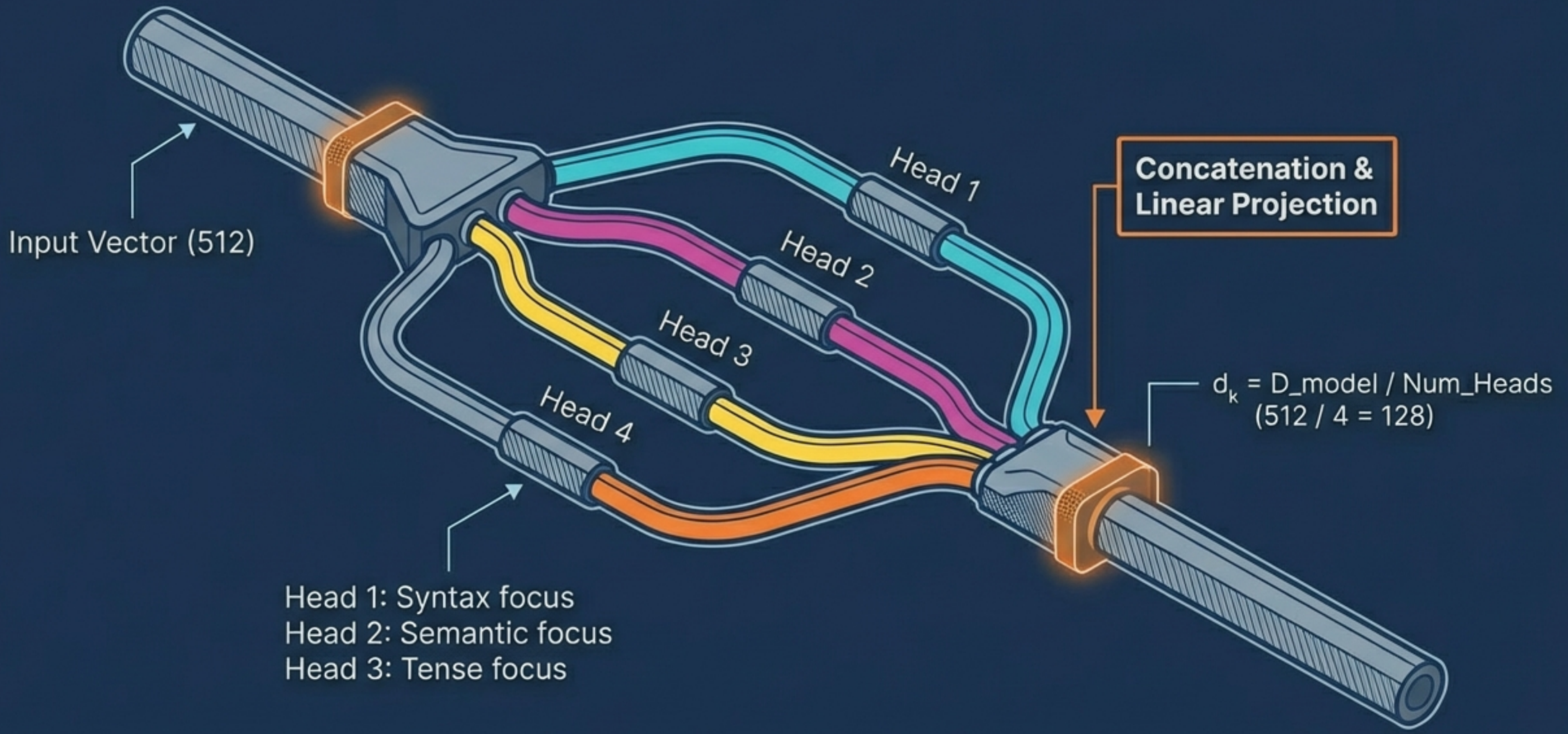
Single-Head Self-Attention



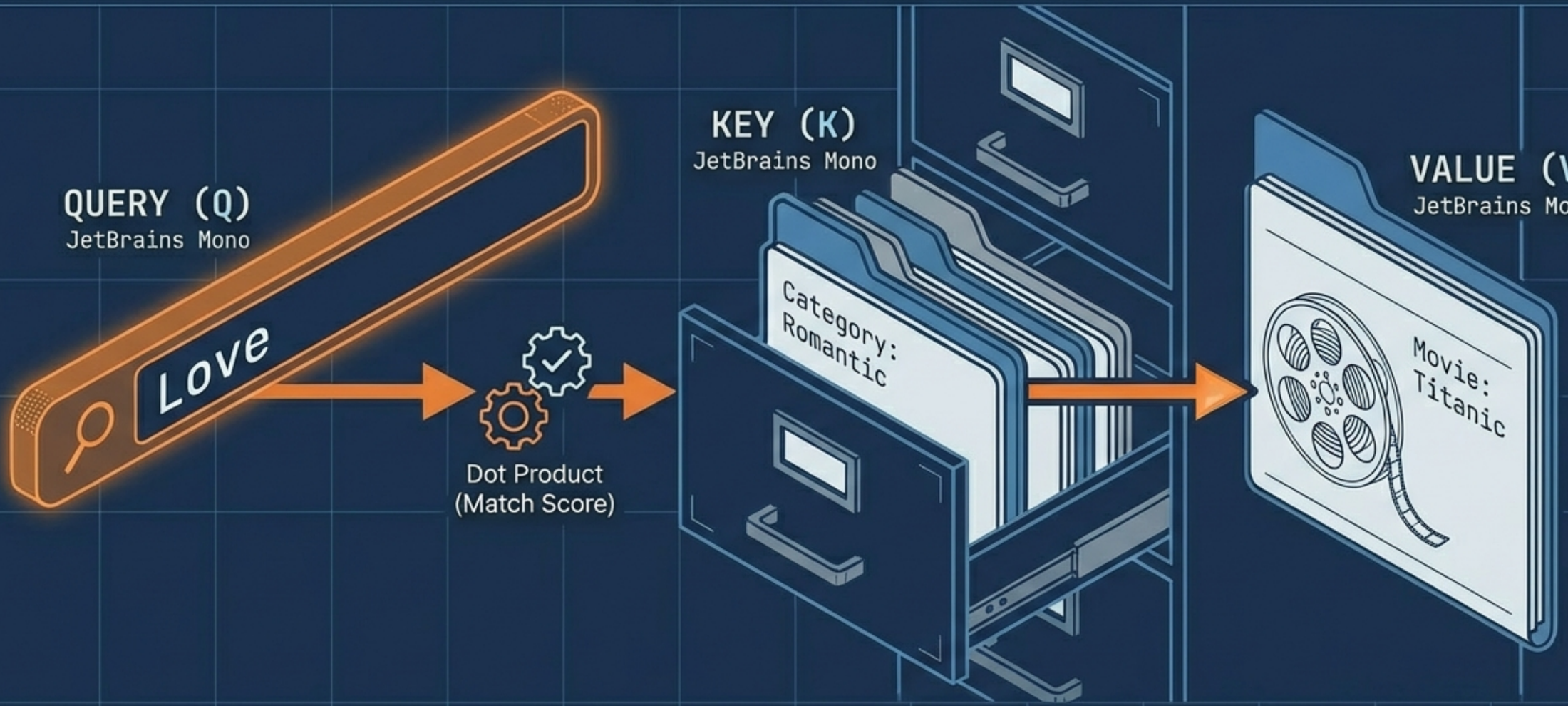
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V$$

Parameter-Free: Relies on dot products to calculate affinity scores.

Multi-Head Attention

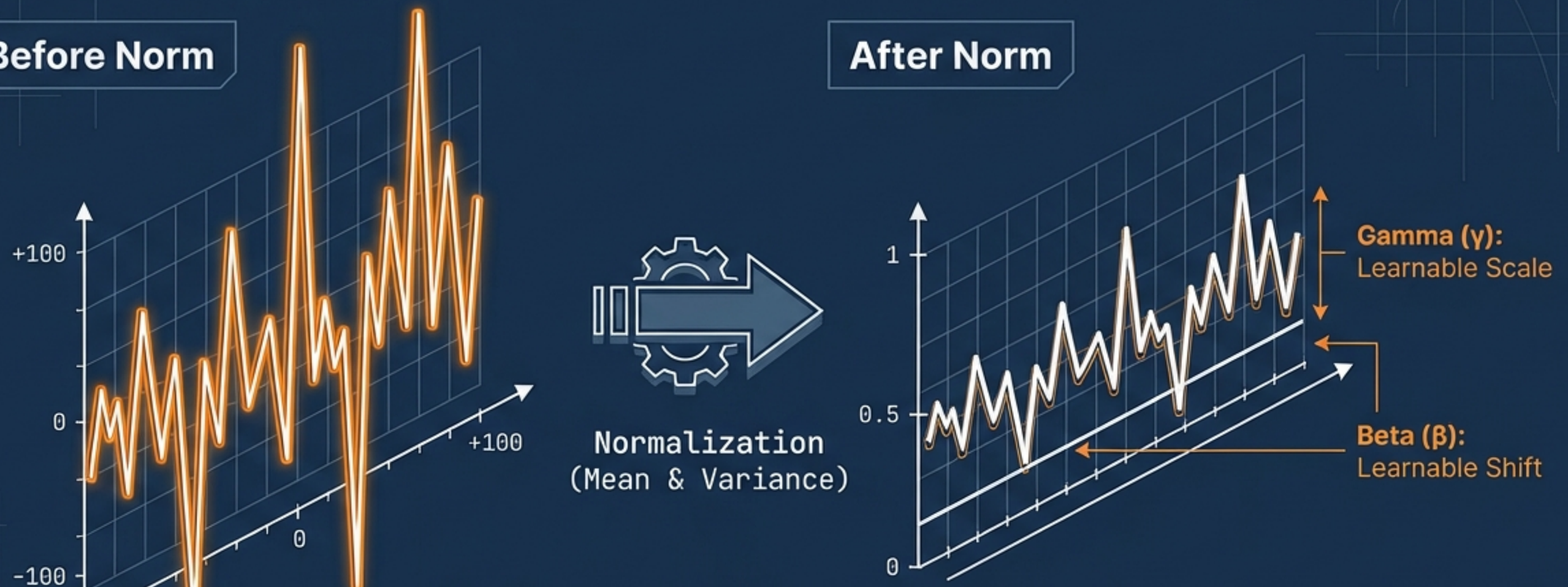


Query, Key, and Value



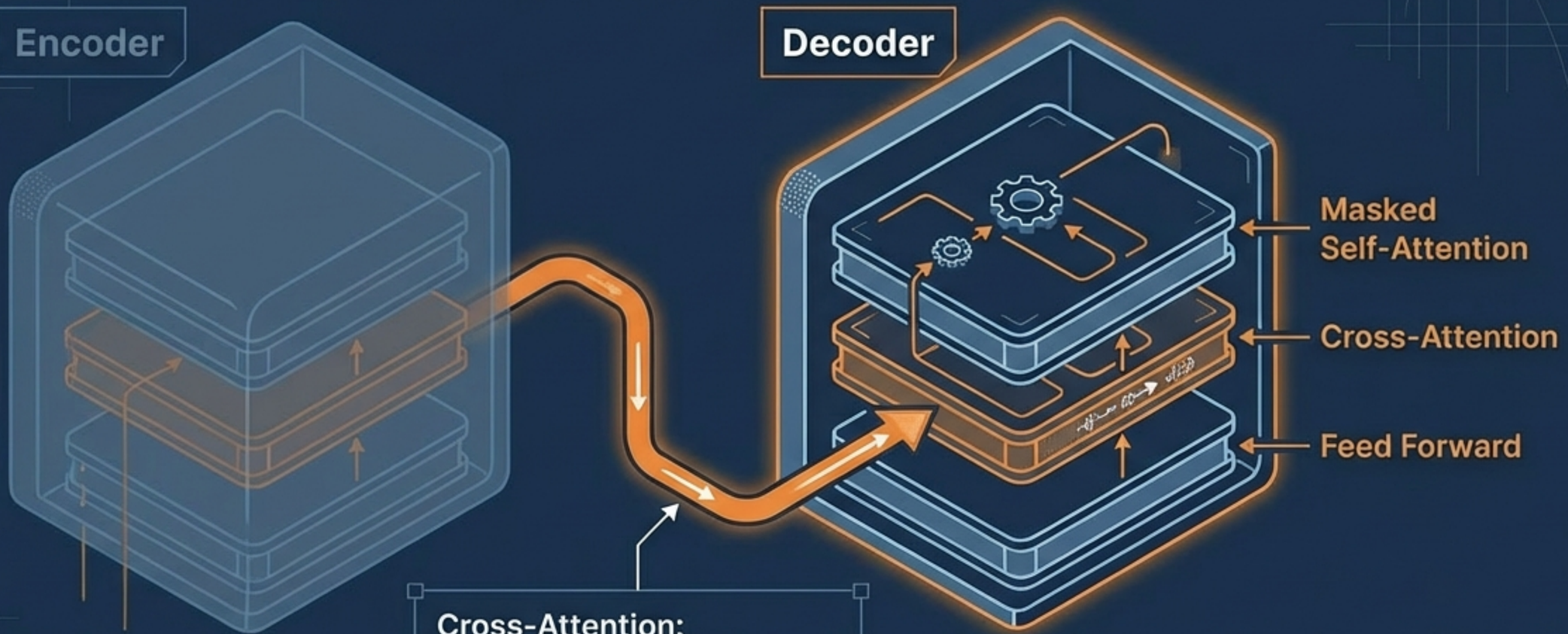
The **Database Analogy**: **Q** searches for **K** to retrieve **V**.

Layer Normalization



Independent normalization per item, distinct from Batch Norm.

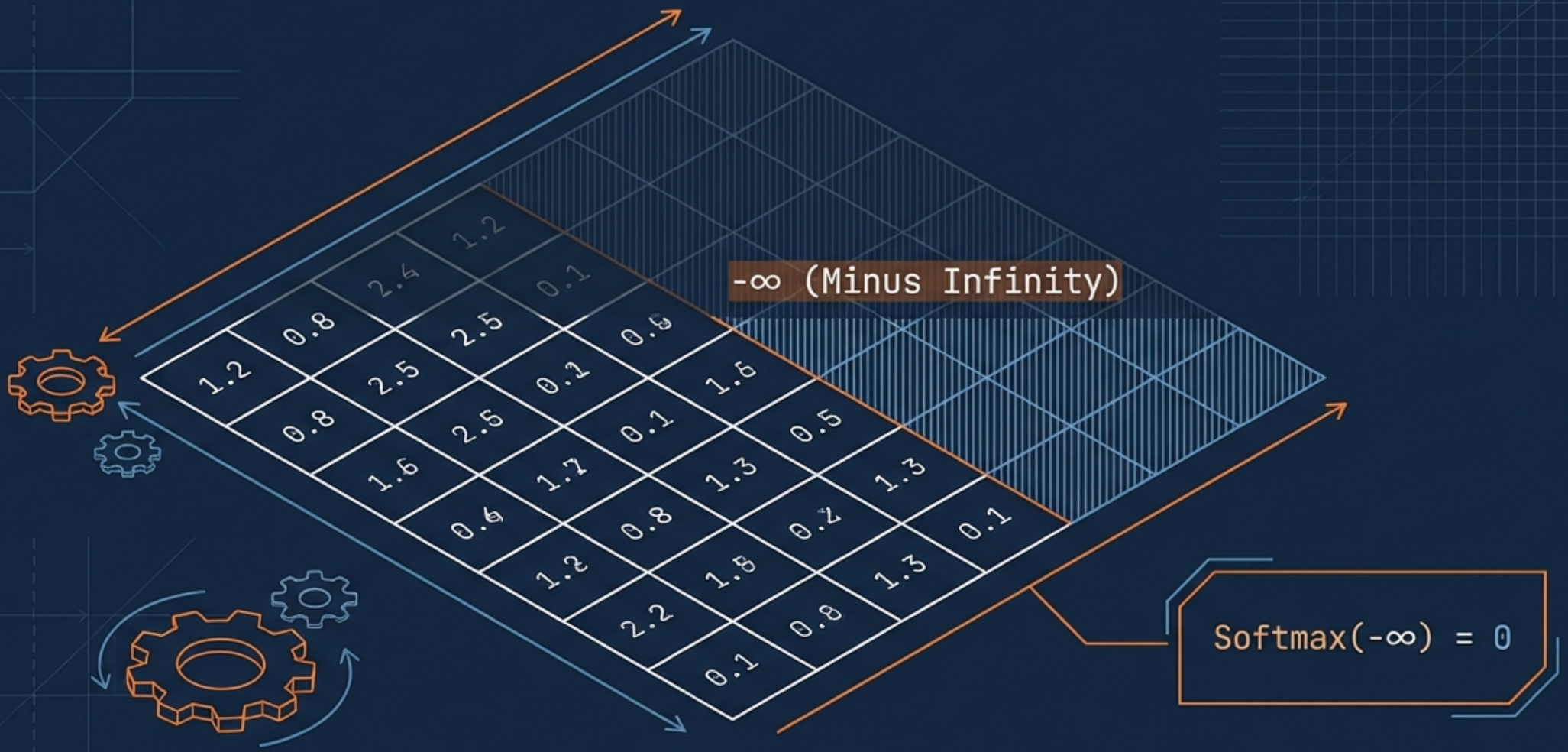
Inside the Decoder



Cross-Attention:
Uses Encoder Output (K, V) and Decoder Input (Q).

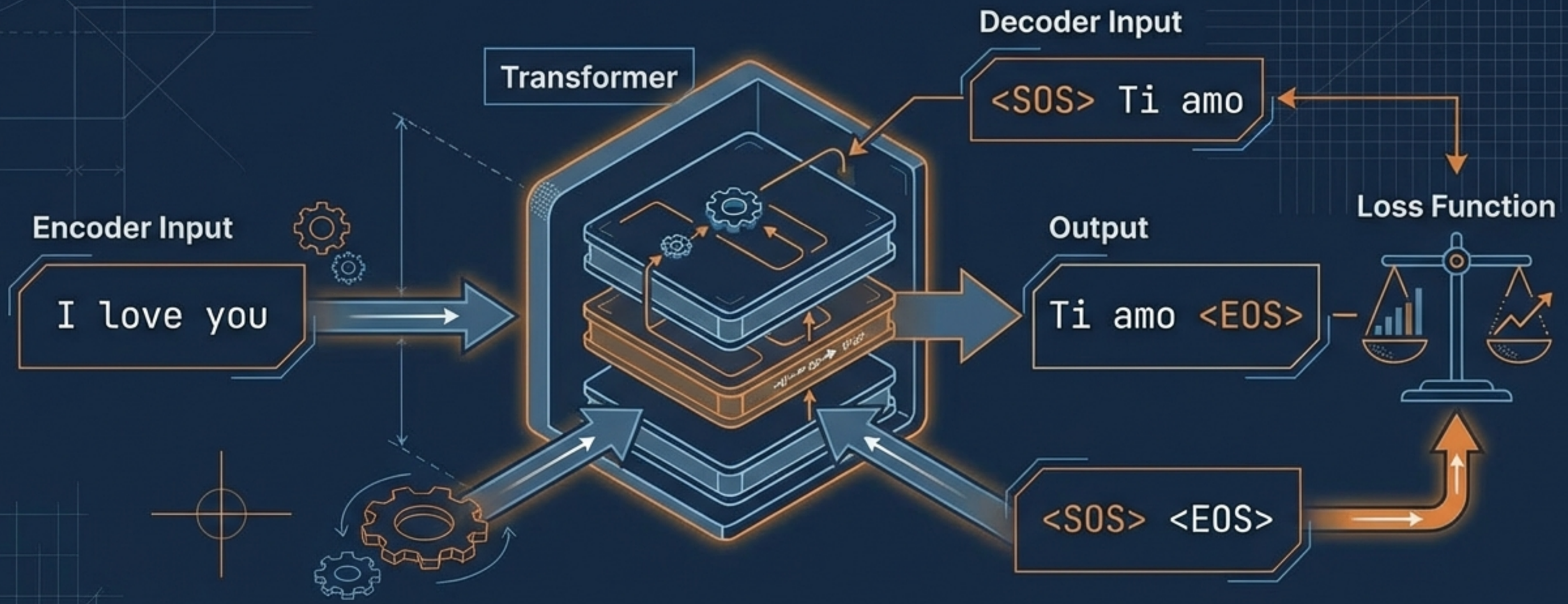
Helvetica Now Display - JetBrains Mono

Masked Multi-Head Attention



Causality Enforcement:
During training, future tokens are masked so the model cannot “cheat” by seeing ahead.

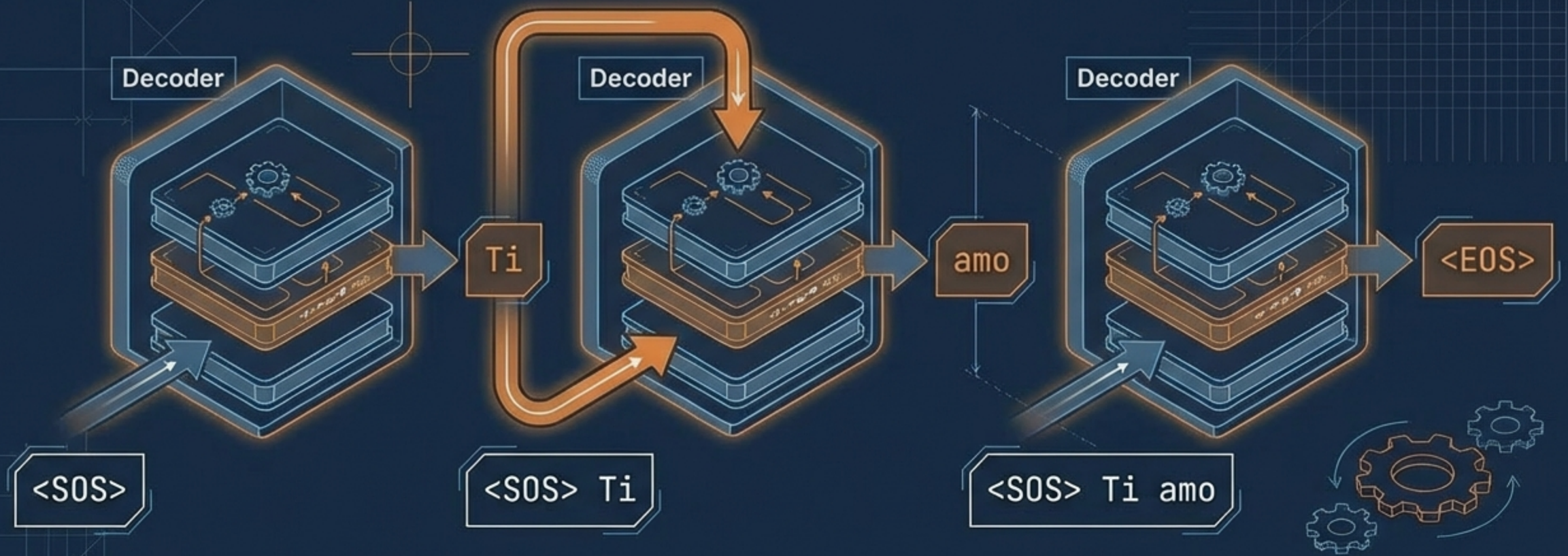
Training: Parallel Teacher Forcing



Parallel Processing:
Unlike RNNs, the entire sequence is trained in one time step.

Helvetica Now Display - JetBrains Mono

Inference: Token-by-Token Generation



Greedy Search (Top %) vs. Beam Search (Top 'B' Paths)