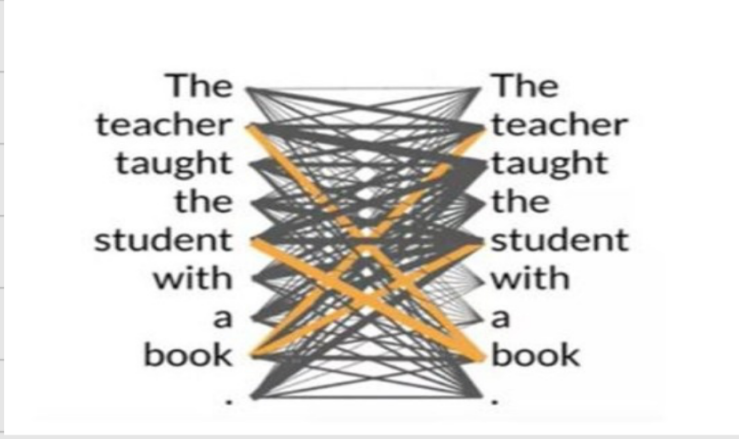
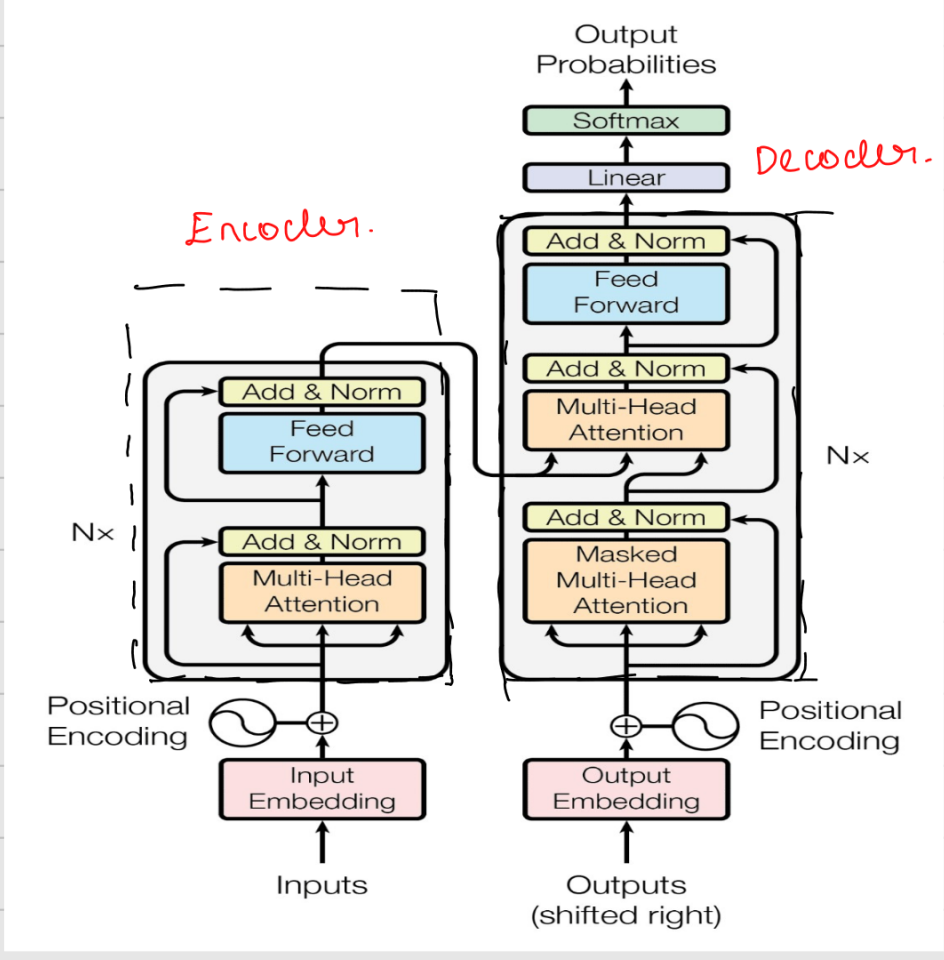
Transformer Model.

→ Backbone of Large Language Model.

→ The advantage of the transformer model is in its ability to understand the significance and context of every word in sentence.



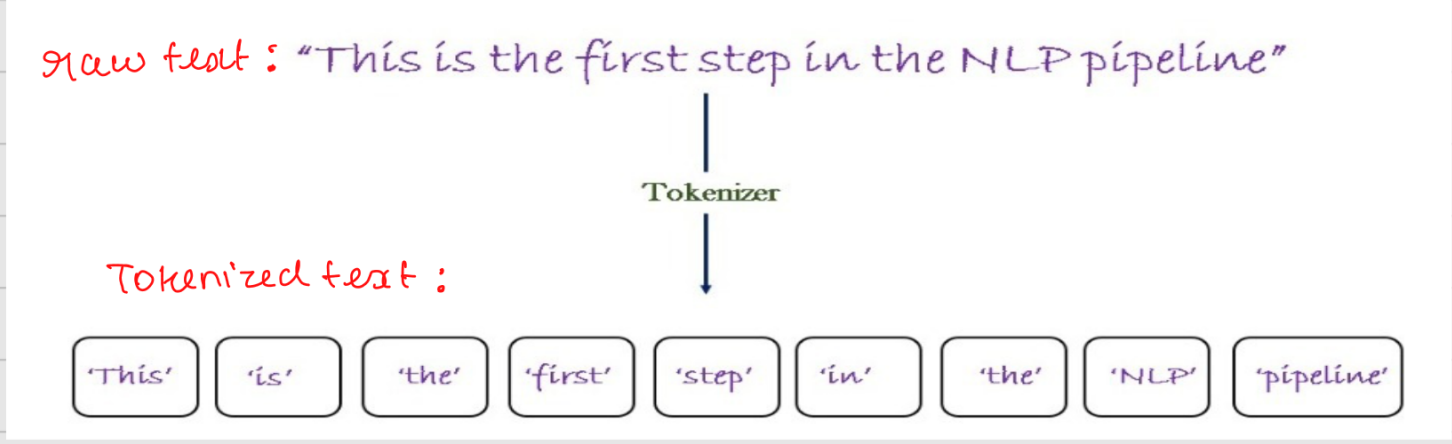
Working of transformer Model!



A transformer model is a neural network that learns context and thus meaning by tracking relationship in sequential data like the words in the sentences.

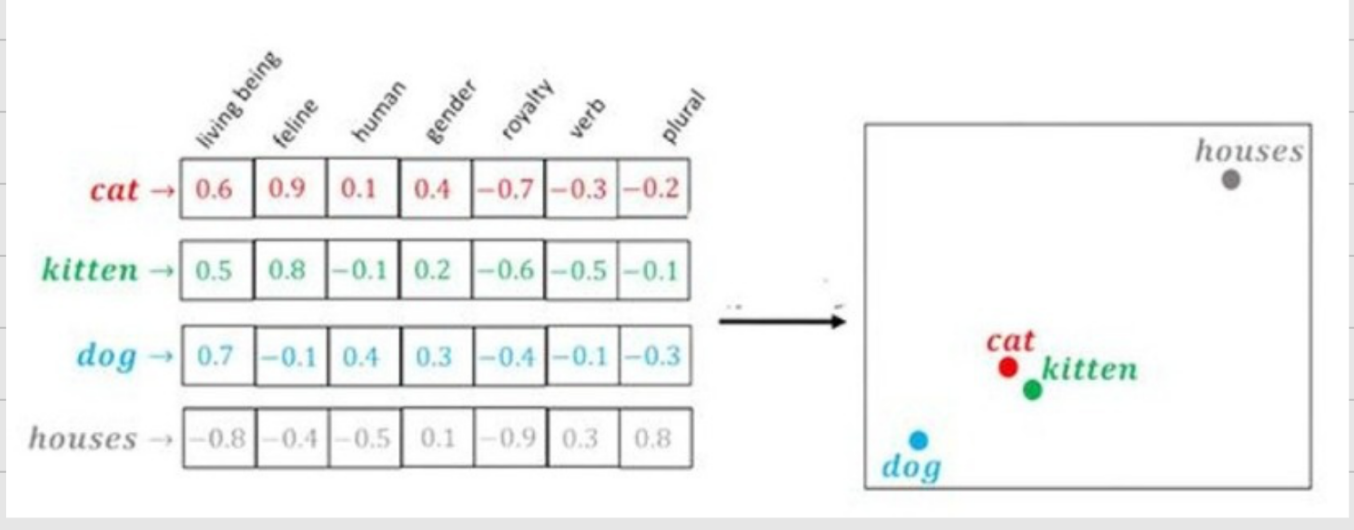
① Tokenization.

Tokenization is the process of breaking down text into smaller units, such as words / phrases for easier processing and analysis.



② Embeddings

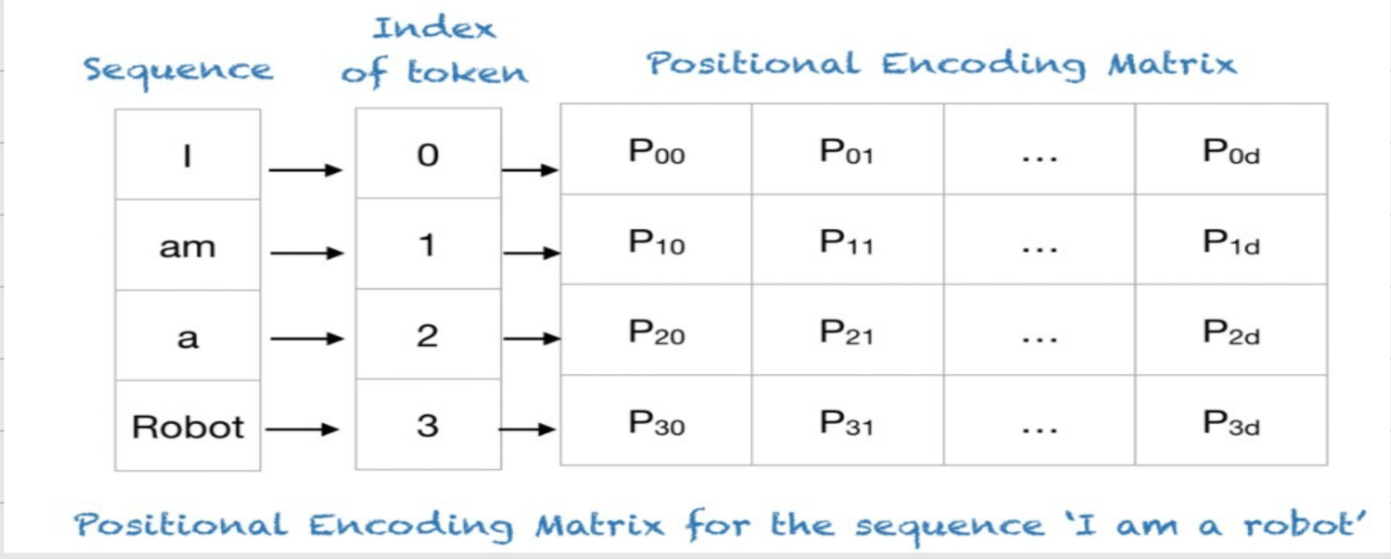
Embedding is the process where each token is then transformed into a vector in a high-dimensional space. This embedding captures the meaning and context of each word.



③ Positional Encoding

Since transformers do not process text sequentially like RNN, they need a way to understand the order of words

Positional encoding is the process of adding information to a model about the position of elements in a sequences.



Positional Encoding Matrix for the sequence 'I am a robot'

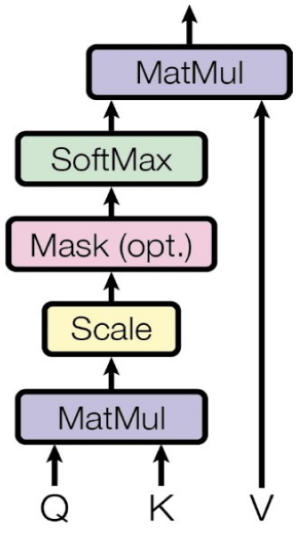
④ Self Attention

The model calculates attention scores for each word, determining how much focus it should put on other words in the sentences when trying to understand a particular word.

This helps the model capture relationship and context within the text.

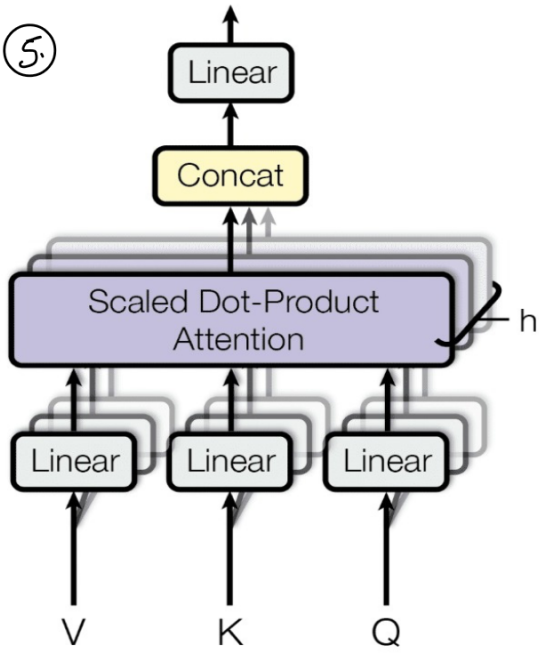
Scaled Dot-Product Attention

(4.)



Multi-Head Attention

(5.)



(5.) Multi-Headed attention.

Multi-Headed attention is a mechanism in transformer that runs several self-attention processes in parallel, allowing the model to focus on different parts of the input sequences from different perspective at the same time.

(6.) output

The final layers of the transformer convert the processed data into an output format suitable for the task at hand, such as classifying the text or generating new text.