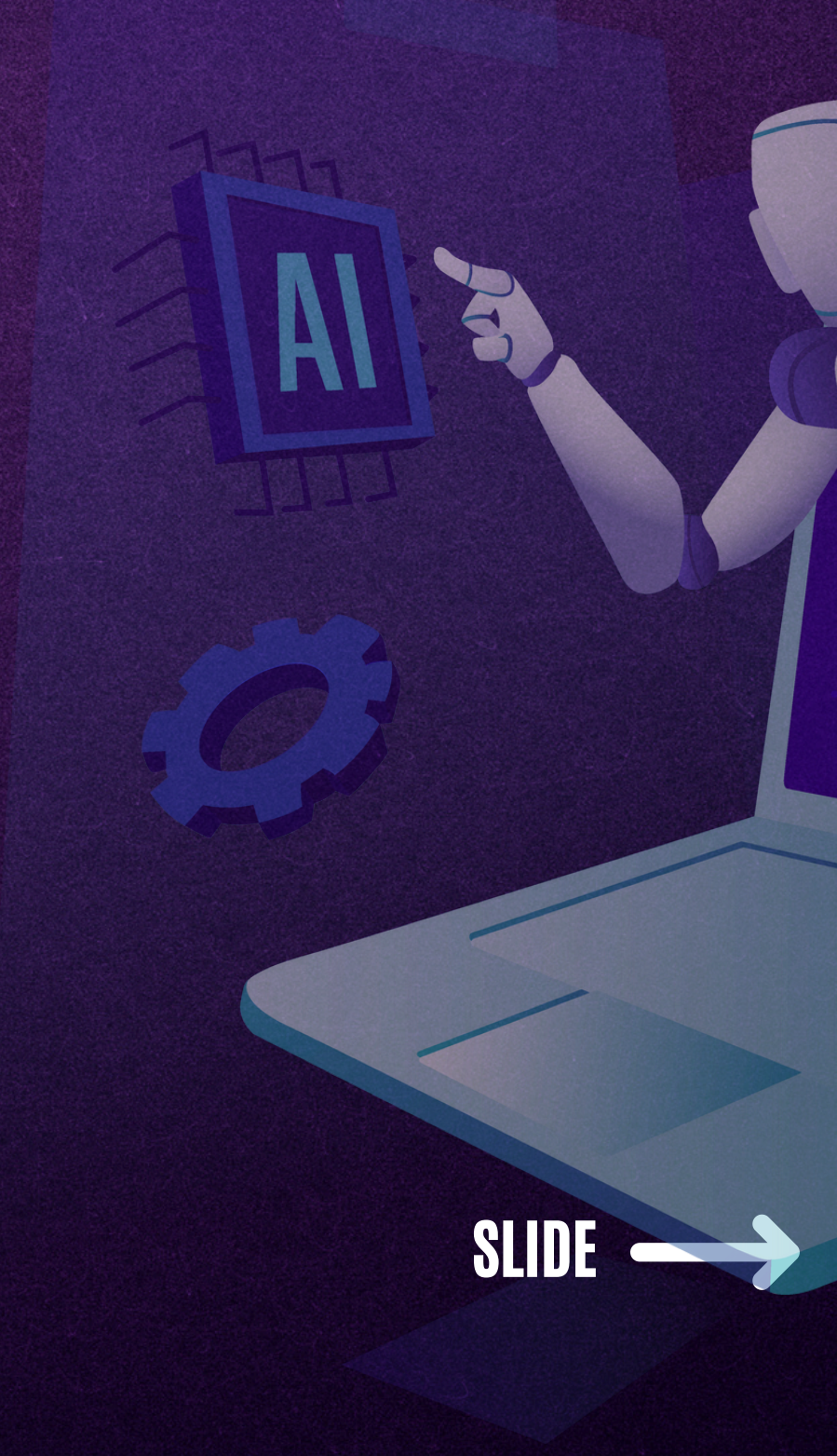
TRANSFORMER ARCHITECTURE

Attention is all you need



MINIs

With ♥ by **Mohit Kanwar**

#CXO Incubator

TRANSFORMER MODEL

The Transformer model is a deep learning architecture introduced in the paper "Attention is All You Need". It revolutionized the field of natural language processing (NLP) by introducing a novel mechanism based on attention rather than recurrent or convolutional layers.

Scan here
To Read the paper

MINIs

With ♥ by **Mohit Kanwar**

TEXT GENERATION BEFORE TRANSFORMER

Before transformer architecture, text generation was done using RNN (Recurrent Neural Network).

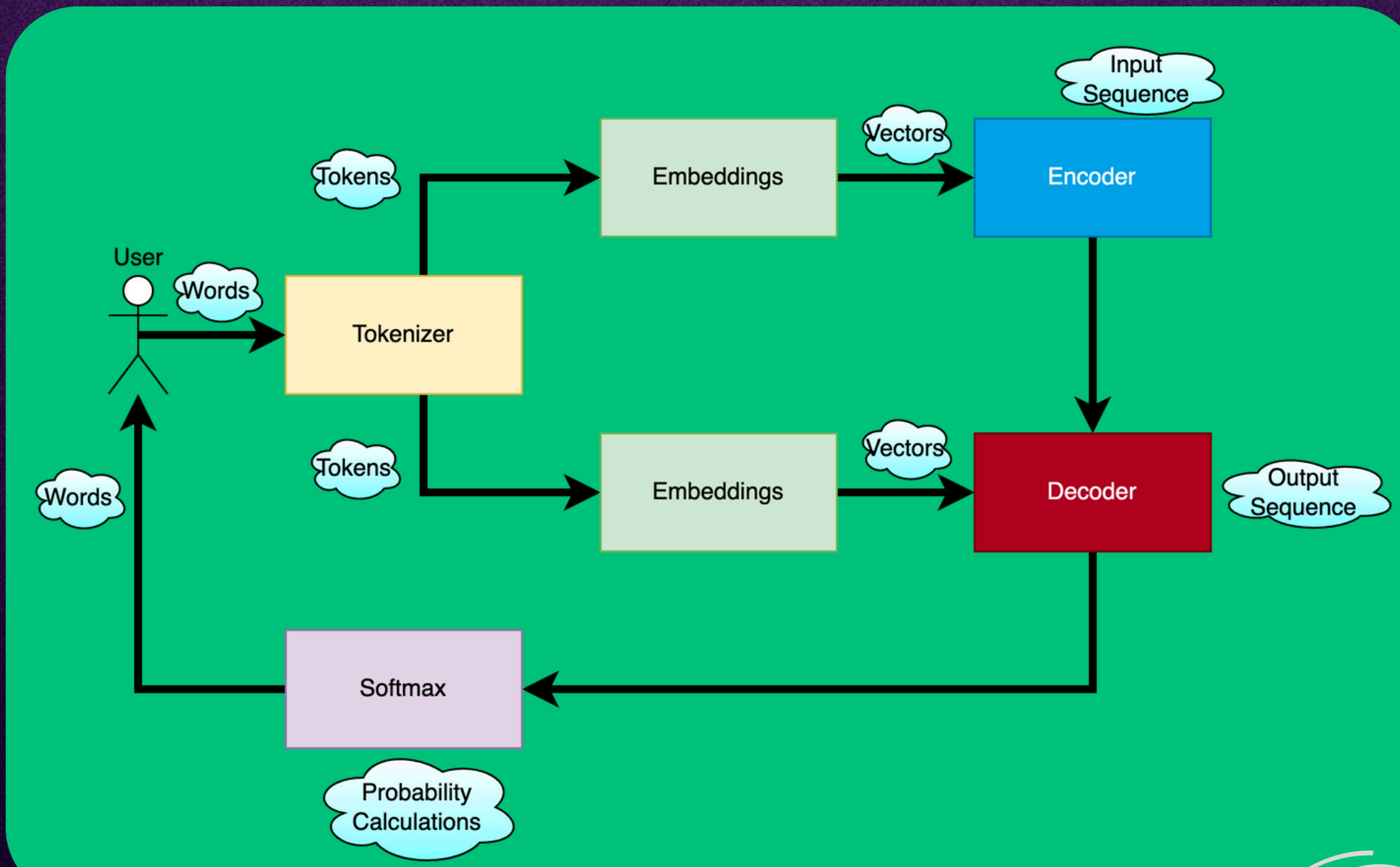
In this technique, the next word was predicted based upon past few words.

RNN worked, but there were a few limitations.

- Vanishing and Exploding Gradients
- Difficulty in Capturing Long-Range Dependencies
- Limited Context Understanding

With ♥ by **Mohit Kanwar**

TRANSFORMER ARCHITECTURE



MINIs

With ♥ by **Mohit Kanwar**

TOKENIZER

Tokenizer breaks down the given text into smaller units called as Tokens.

These tokens can be understood by the LLM, and can process them effectively.

Tokens generally are from a pre-defined vocabulary that the LLM understands.

Some additional tokens required for processing may be added by the tokenizer.

MINIs

With ♥ by **Mohit Kanwar**

EMBEDDINGS

Embeddings are vector representations of tokens that encode semantic and syntactic information about the language. These embeddings are learned during the training process of the language model.

Embeddings map the tokens in a vector space, such that tokens with similar meanings are closer to each other.

In advanced models, the embeddings are contextual. They derive meaning from the surroundings.

With ♥ by **Mohit Kanwar**

ENCODER

An encoder is a neural network component that processes the embeddings (or vector tokens) to create a context to understand the input.

Attention is all you need *

By utilizing self-attention mechanisms encoders can capture long-range dependencies and contextual information.

Encoders are generally layered, with each upcoming layer enhancing the output of previous layer.

With ♥ by **Mohit Kanwar**

DECODER

While the encoder helps the LLM to understand the context, the Decoder is another neural network component that helps the LLM to prepare a response based on the understood context.

Decoder; similar to encoder is also layered. Each layer works on the abstraction from the previous layer to deliver a meaningful result.

Different types of decoders are used to deliver different results e.g. sentences or summaries etc.

However, at this stage, the output is still not finalized.

SOFTMAX

Softmax is the final decision maker in generating the response. It takes into account the probability of tokens and normalizes them.

Softmax predicts the sequence of tokens based on this normalized probability.

SoftMax Is called so because **it normalizes (or softens) the probability of token occurrences** and **selects the result that has a maximum value** (closest to 1 probability).

MINIs

With ♥ by **Mohit Kanwar**

#CXO Incubator

Hi, I'm Mohit Kanwar

Architect | Xebia

I help Fintechs and Banks develop their software solutions

Scan me