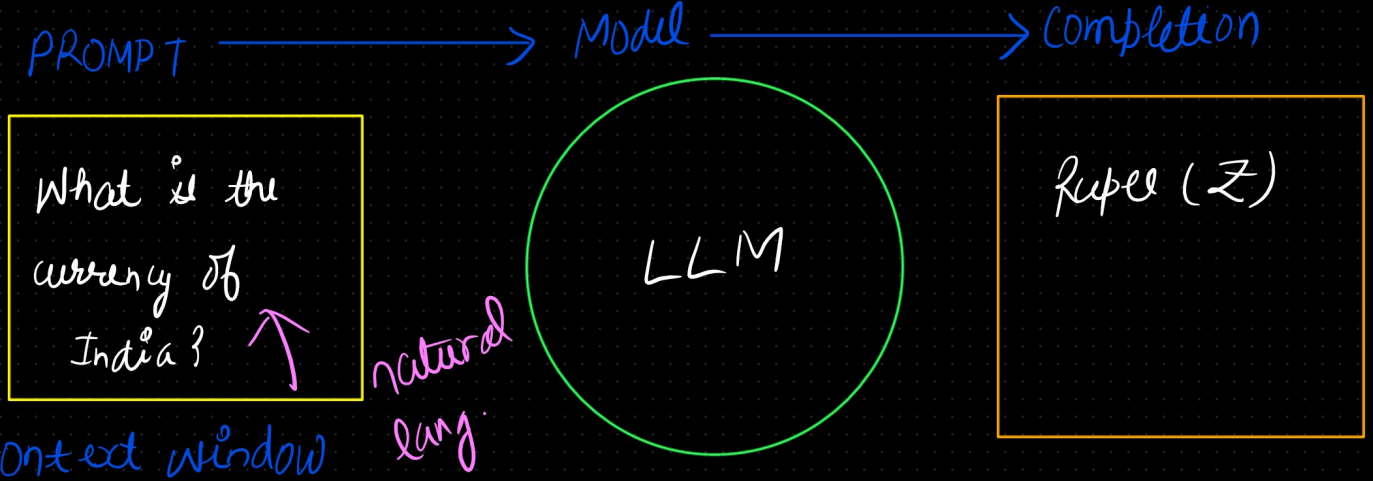
Large Language Model (LLM)

What is LLM?

LLM are machine learning models that have learned from massive datasets of human-generated content, finding statistical patterns to replicate human-like abilities.



This process is called "Inference"

ChatGPT → Application

↓

LLM Model → GPT 3, GPT 4, GPT 5

Google — Gemma, Gemini

What are the Use Cases of LLMs?

→ Chatbots

→ Writing → Essay, Emails, Report

→ Summarisation → Summarize long content into a meaningful shorter length.

→ Code

→ Language Translation

→ Information Retrieval

→ Agumented LLM

TRANSFORMERS

↓

Model Architecture

↓

2017 → Research Paper

↓

"Attention is All You Need"

↓

revolutionised generative AI

How was text generation done before Transformers ?

Recurrent Neural Network (RNN)

↓

predict next words

→ The next word was predicted looking at the previous few words

Eg → I took my money to the bank.

River Bank?

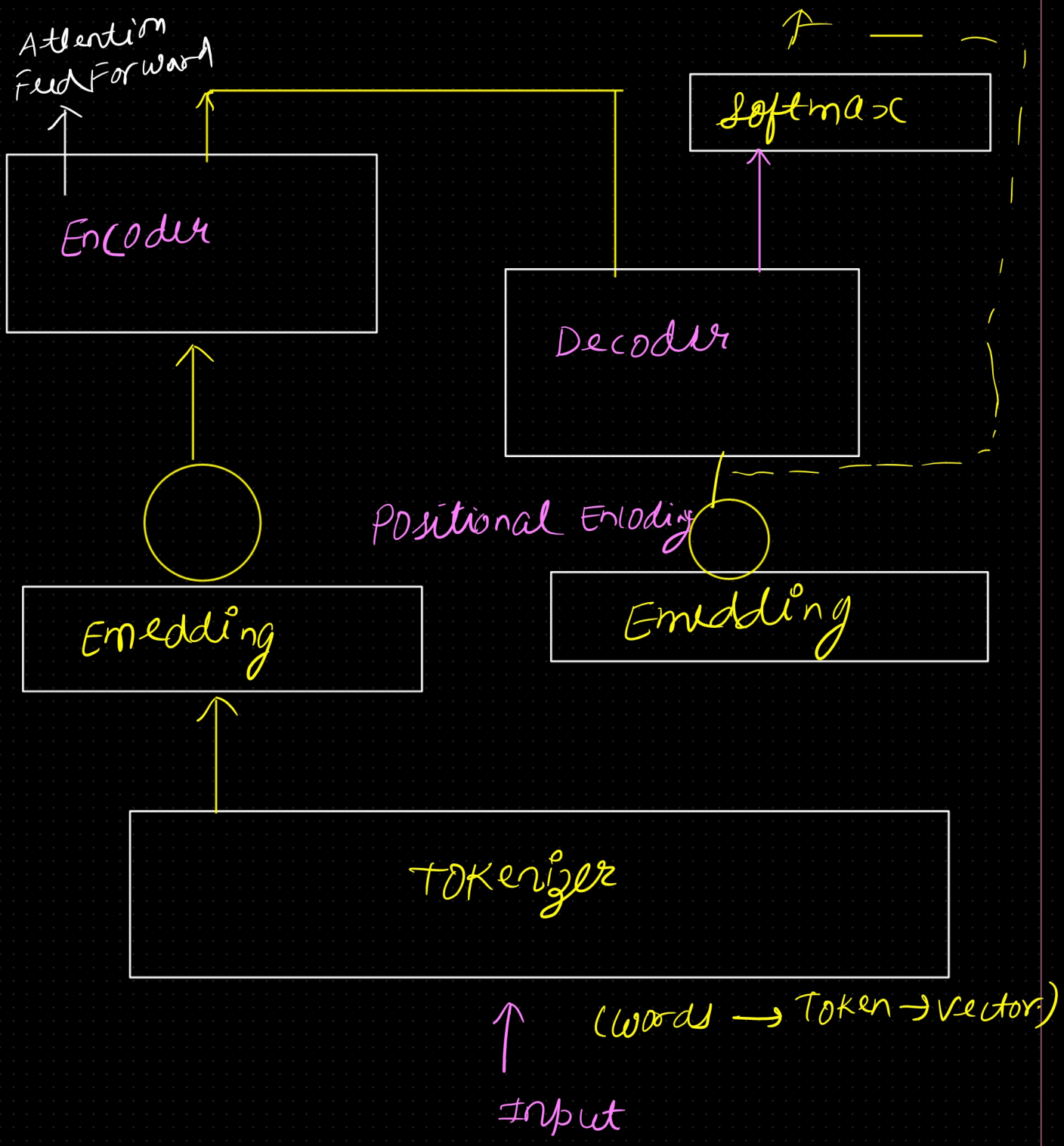
Financial Bank?

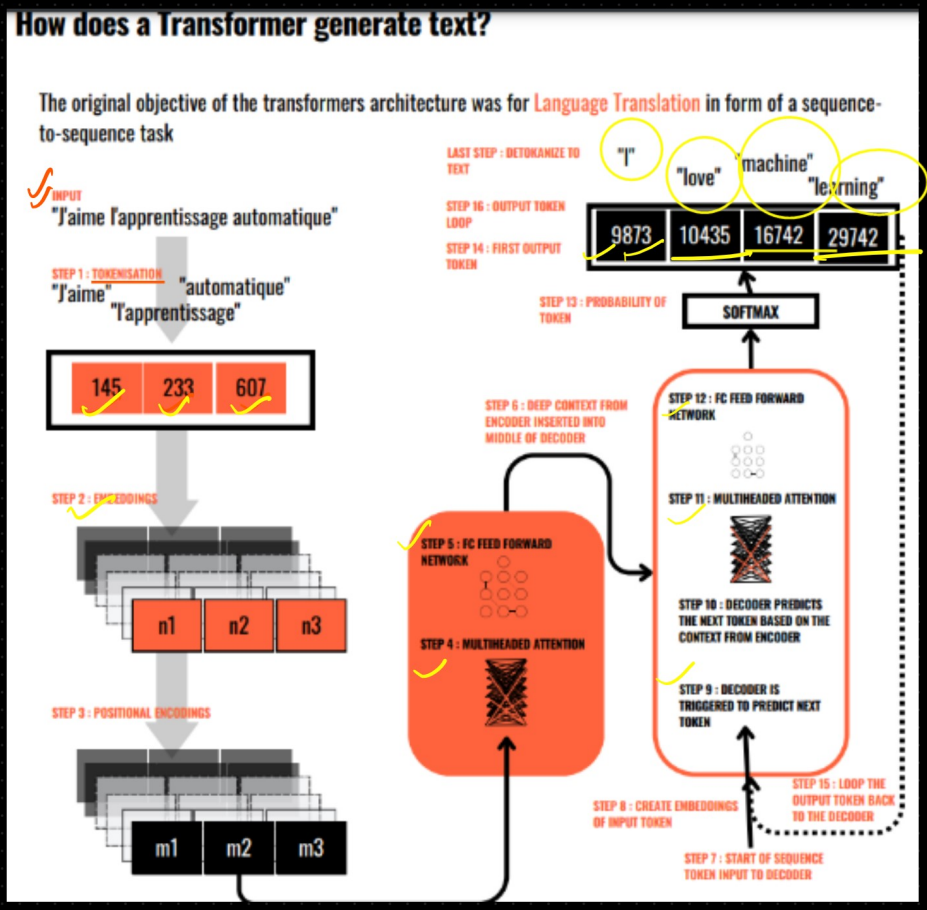
→ Transformers are able to pay Attention to the meaning of the words.

↓

Multi Head Attention

output


Attention
Feed Forward
Encoder
Softmax
Decoder
Positional Encoding
Embedding
Embedding
Tokenizer
(words → Token → vector)
Input



What is Prompt ?

→ The natural language instruction in which we interact with LLM is called Prompt.

→ The construction of prompts is called prompt Engineering

Zero shot → Without giving an example

One shot → Single example

Few shot → More than one example

The inferencing that an LLM does and completes the instruction given in the prompt is called 'in context learning'

↓ 0 – 1