Introduction to RAG!-
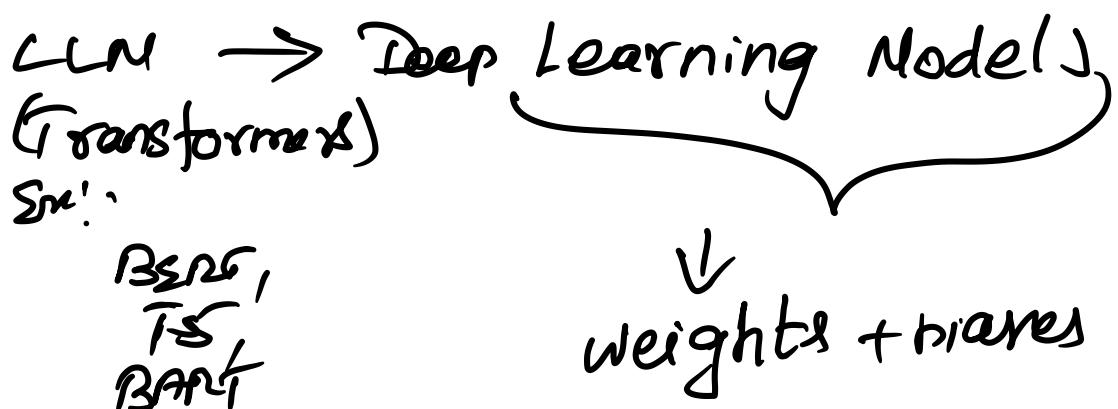
Large Language Models (LLM) follows Transformer architecture. These are Deep learning models.

LLM $\longrightarrow$ Deep Learning Model(s)
(Transformers)
Ex:-
     BERT,
      T5,
      BART

$\bigvee$
weights + biases

GPT - Decoder
BERT - Encoder

<u>Note</u> :-

          NLP terms
          upstream : Anything
          pretraining
                    Ex:- Masking

Downstream: Anything we do
past training
Sul.
fine tuning Text
classification, POS,
NER, Summarization,
QA, MT
& Text Generation

We do have some **problems** with
the **LLM** (or) even **Small Language**
Models (SLM)

Base Model (or) Foundational
Model
↓
(doesn't have your
organization
+
← data)

{ Custom Data }
↓
Fine Tuning Task (Expensive Task)

① knowledge cut off
    ↳ Cutoff Training ↓
        Date
    LLM would have Some cut off in Training data i.e not upto date In that context LLM tend to give wrong Ans (Hallucinate)

② Lack of Info in Domain specific Tasks

③ Lack of Private Data

④ Loses Trusted Source