WHAT ARE LARGE LANGUAGE MODELS ?

LARGE : means it has learned from a massive amount of data and has billions of connections (parameters) inside it.

LANGUAGE : It focuses on text and words, reading, writing, summarizing, answering and even translating.

MODEL : Its a mathematical system trained to predict what word (or idea) should come next in a sentence.

It doesn't think like humans but recognizes patterns and gives outputs based on what it has learned.

It can chat, answer questions, write stories, code or explain concepts almost like a smart assistant.

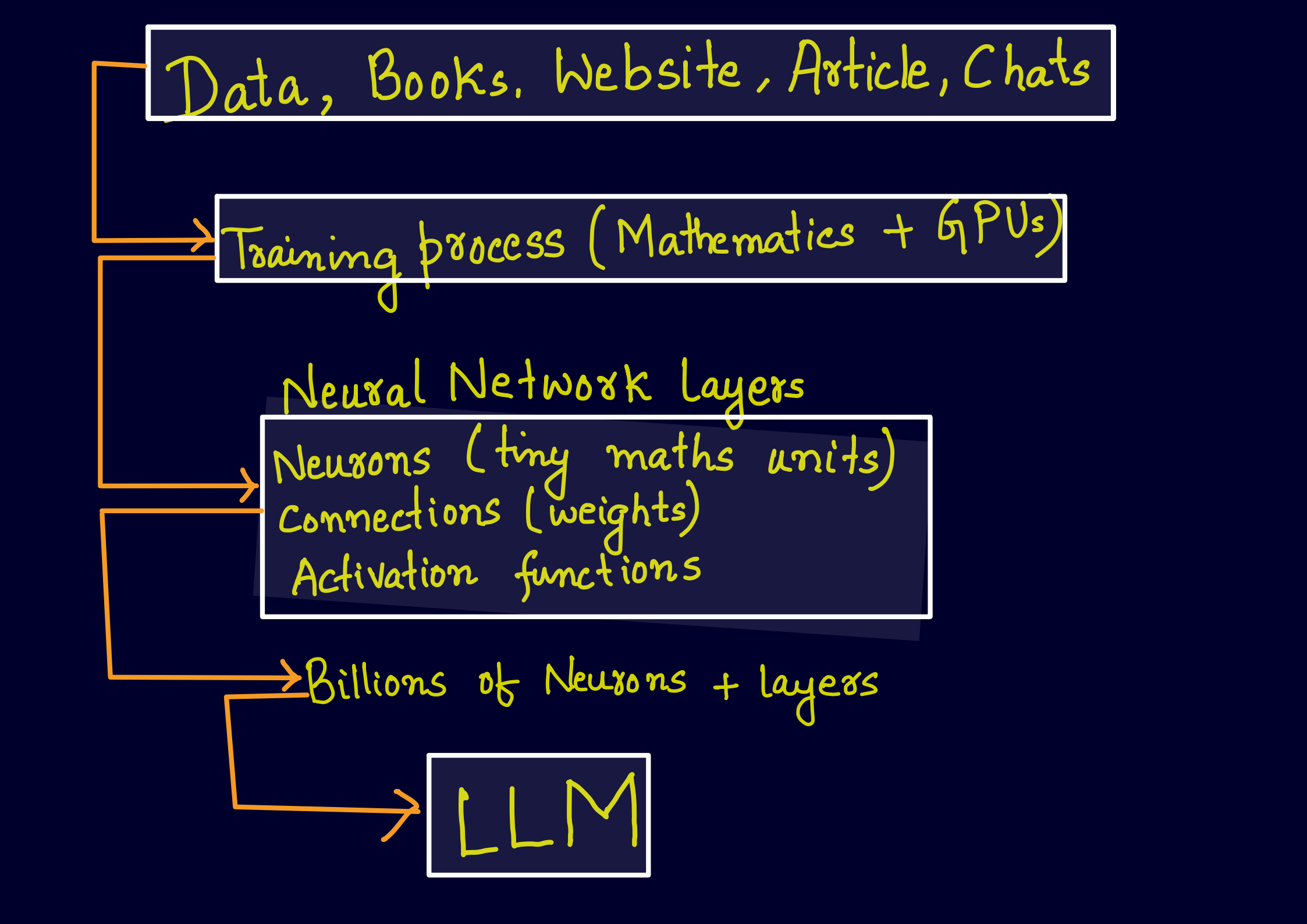
Examples: GPT, Google Gemini, Claude, Meta LLaMA are all large language models.

Large Language Models, commonly known as LLMs are sophisticated type of neural network. These models are characterized by their large number of parameters, often in billions, that make them proficient at processing, understanding and generating texts.
The primary goal of LLM is to predict the next word based on previous word. They are trained on extensive textual data, enabling them to grasp various language patterns and structure.

How Neural Networks & LLMs are Connected?

Neural Network = One small brain cell

LLM = a whole giant brain made of billions of brain cells working together.


Data, Books, Website, Article, Chats
Training process (Mathematics + GPUs)
Neural Network Layers
Neurons (tiny maths units)
Connections (weights)
Activation functions
Billions of Neurons + layers
LLM

Building Blocks of LLMs

- The Transformer
- Language Model
- Tokenization
- Embeddings
- Training / Fine-Tunning
- Prediction
- Context Size
- Scaling Laws
- Prompting

1) The Transformer

- It is a special type of Neural network architecture created in 2017 (paper: "Attention Is All You need").
- Its main idea is the attention mechanism, which helps the model focus on the most important words in a sentence.
- Transformers = the engine that makes LLMs smart and efficient.

2) Language Model

- A language model is a program that learns how words usually follow each other in sentences.
- It can predict the next word based on the word before it.
- Early LMs used n-grams (fixed window of words) and later evolved into RNNs and LSTMs, which could capture longer dependencies.

3) Tokenization

Tokenization is the initial phase of interacting with LLMs. It involves breaking the input text into smaller pieces known as tokens.

This process involves scanning the entire text to identify unique tokens, which are then indexed to create a dictionary.

4) Embeddings

It is a way to translate tokens, which are words or pieces of words (or rather their numerical IDs), into numbers that the computer can manipulate. An embedding give each token a unique numerical ID that captures its meaning.