Transformers

What is a Transformer ?

A **Transformer** is a deep learning model architecture designed to understand and generate human language. It uses a mechanism called **self-attention** to process all words in a sentence at once, allowing it to capture relationships between words more effectively than older models like RNNs or LSTMs.

Core Components of a Transformer

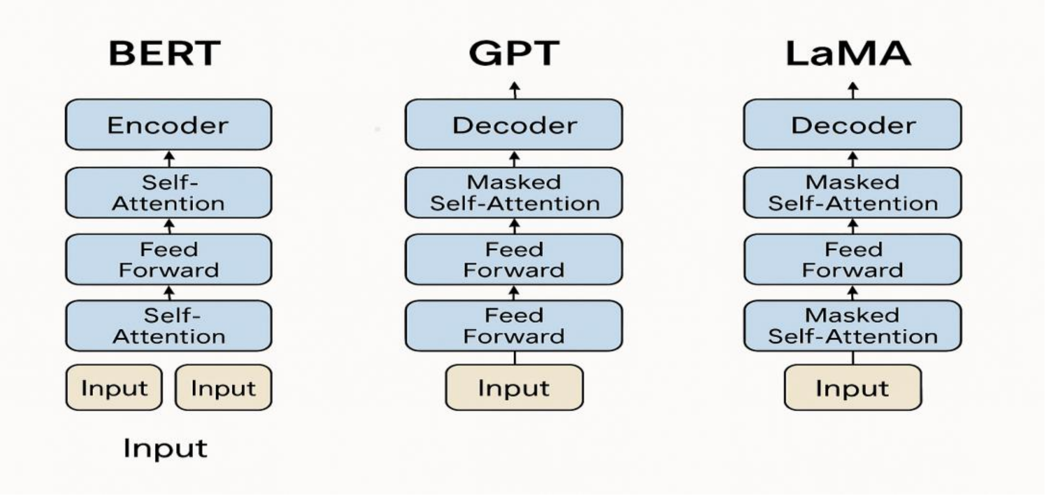
1. **Input Embedding**: Converts words into numerical vectors.
2. **Positional Encoding**: Adds information about word order.
3. **Self-Attention**: Each word looks at all other words to decide which ones to focus on.
4. **Multi-Head Attention**: Multiple attention mechanisms run in parallel to capture different relationships.
5. **Feed-Forward Network**: Processes the attention output further.
6. **Residual Connections & Layer Normalization**: Help stabilize and improve learning.

Encoder vs Decoder

- **Encoder**: Reads and understands input text.
- **Decoder**: Generates output text.
- Some models use only the encoder (like BERT), some use only the decoder (like GPT), and some use both (like T5).

Here’s a visual diagram that illustrates the Transformer architecture and how models like **BERT**, **GPT**, and **LLaMA** are built on top of it:

- **Left side**: Shows the **Transformer architecture** with its key components (Input Embedding, Positional Encoding, Self-Attention, etc.).
- **Right side**: Compares **BERT**, **GPT**, and **LLaMA** based on their use of Encoder and Decoder.



Transformer-Based Models

1. BERT (Bidirectional Encoder Representations from Transformers)

- **Architecture**: Uses only the **encoder** part of the Transformer.
- **Purpose**: Understanding language (not generating).
- **Training**: Trained to predict missing words and understand sentence relationships.
- **Use Cases**: Text classification, sentiment analysis, question answering.

Example:
Input: “The cat sat on the ___.”
BERT predicts: “mat”

---

2. GPT (Generative Pre-trained Transformer)

- **Architecture**: Uses only the **decoder** part of the Transformer.
- **Purpose**: Generating text.
- **Training**: Trained to predict the next word in a sentence.
- **Use Cases**: Text generation, chatbots, summarization.

Example:
Input: “The cat sat on the”
GPT continues: “mat and looked sleepy.”

---

3. LLaMA (Large Language Model Meta AI)

- **Architecture**: Similar to GPT, uses a **decoder-only** Transformer.
- **Purpose**: Text generation and understanding.
- **Training**: Trained on a large dataset with efficient architecture.
- **Use Cases**: Research, open-source alternatives to GPT.

Key Feature:
LLaMA is designed to be efficient and accessible for researchers, often requiring less computing power than GPT for similar performance.