BCS JUDGE IT WELL

LLMs and Transformer Architecture

- **LLMs**

  LLMs stand for Large Language Models, which generate text based on the user's prompt. It learns patterns, grammar and context from text and can answer questions, write content, translate languages and many more. Examples- gpt, gemini, claude etc. It works on transformer architecture.

- **Transformer Architecture**

  Transformer architecture enables the model to retain long term dependencies of the data and get the contextual meaning of the text. The transformer architecture works on encoder-decoder mechanism. Encoder helps in the representation of the data whereas decoder helps to generate the output sequence based on the context understood by encoder.

  DETAILED EXPLANATION

  - **INPUT REPRESENTATION**

    First a given text is tokenized to words or further. Like "I am a good boy" is tokenized as 'I' 'am' 'a' 'good' 'boy'. Now each token is represented as a vector by word embeddings. The vector is also determined by the position of the token in the text. This is called positional embeddings.

  - **ENCODER PROCESS**

    Each token in the text attends to each other in order to get the contextual meaning and dependencies of tokens on each other. This is called self attention mechanism.

- **DECODER PROCESS**
  Now, the model predicts the next word of the sequence based on the previous words in the sequence. This is called Masked Self Attention. It also prevents the model from attending the future tokens which could affect the output.

Attention in Detail

The Attention Mechanism in Machine Learning is a technique that allows models to focus on the most important parts of input data when making predictions. It assigns different weights to different elements which captures most relevant information ensures that the contextual meaning of the text is understood.

It has 3 main elements key, value and query:

QUERY: This is a vector representing the current focus or question the model has about a specific word in the sequence. It's like a flashlight the model shines on a particular word to understand its meaning in context.

KEY: Each word has a label or reference point — the key vector acts like this label. The model compares the query vector with all the key vectors to see which words are most relevant to answer the question about the focused word.

VALUE: This vector holds the actual information associated with each word. Once the model identifies relevant words through the key comparisons, it retrieves the corresponding value vectors to get the actual details needed for understanding.