NOTE: THESE ARE ROUGH NOTES, HERE IS THE LINK TO THE FULL ARTICLE: <https://medium.com/@habibahmed918131/understanding-large-language-models-the-technology-reshaping-our-digital-world-4a1190d740e6>

**What are LLMs?**

Large Language Models are Models trained on tons of data, which uses Transformer Architecture and Deep Learning Techniques.

**How they work?**

These models are based on transformer models and leverage deep learning algorithms. They consist of multiple layers of neural networks each with specific parameters, and those layers are enhanced by the “attention mechanism” layer. These models are highly efficient in capturing patterns or relationships in entities.

The workflow begins with tokenization i.e. the process through which text is broken down in to tokens or smaller units (and then converted into embeddings i.e. their numerical representation), then “positional encoding” is added to these tokens, which tells about the positions of the token (because generally transformers can’t preserve the order of tokens).

Now these tokens are sent to the transformer where they meet the “Encoder”, which based on the neural network techniques add, some hidden states which preserve the context and meaning of the text.

“The **encoder** further down, consists of multiple layers, of which ‘self-attention mechanism’ and ‘Feed Forward’ layers are sub components of each layer.”

* **Attention mechanism:** this mechanism helps in weighing the tokens and their attention scores, helping it in identifying relationships between the tokens in a context-aware manner.
* **Feed forward Neural Network:** After attention step, this network is applied to each token, these networks include non-linear connected layers of functions, enabling the model to capture complex token interactions.

Transformers also often employ **MULTI HEAD ATTENTION** mechanism, where self- attention is performed simultaneously on different learned or extracted weights. This also allows the model to capture further on the details and relations between entities.

After all the passing through the above layers, **LAYER NORMALIZATION** is applied to these components, which helps in stabilizing the learning process and improves model’s generalization across different inputs.

And at last the **OUTPUT LAYER,** come in action after the passing through all the transformer layers, the linear projection, is followed by a **SOFTMAX ACTIVATION** (which assigns probabilities of the next word).