

LLMs, or Large Language Models, are based on the Transformer architecture. This architecture uses self-attention layers to understand the relationships between words in a sentence or text. The pipeline of an LLM typically starts with a tokenizer, which breaks down the input text into individual words or tokens. These tokens are then converted into embeddings, which are numerical representations of the words that can be processed by the model.

The model then passes these embeddings through multiple Transformer layers, which are composed of self-attention mechanisms and feed-forward neural networks. The self-attention mechanism allows the model to weigh the importance of different words in the input text and focus on the most relevant ones. The feed-forward neural networks then process the weighted input and produce an output.

The final output of the LLM is a prediction of the next word in the sequence, which is generated by a next-token prediction mechanism. This mechanism takes the output of the Transformer layers and produces a probability distribution over all possible next words in the sequence.

Overall, the architecture of LLMs is designed to handle long context and complex instructions, making them well-suited for a wide range of natural language processing tasks, including language translation, text summarization, and question answering.