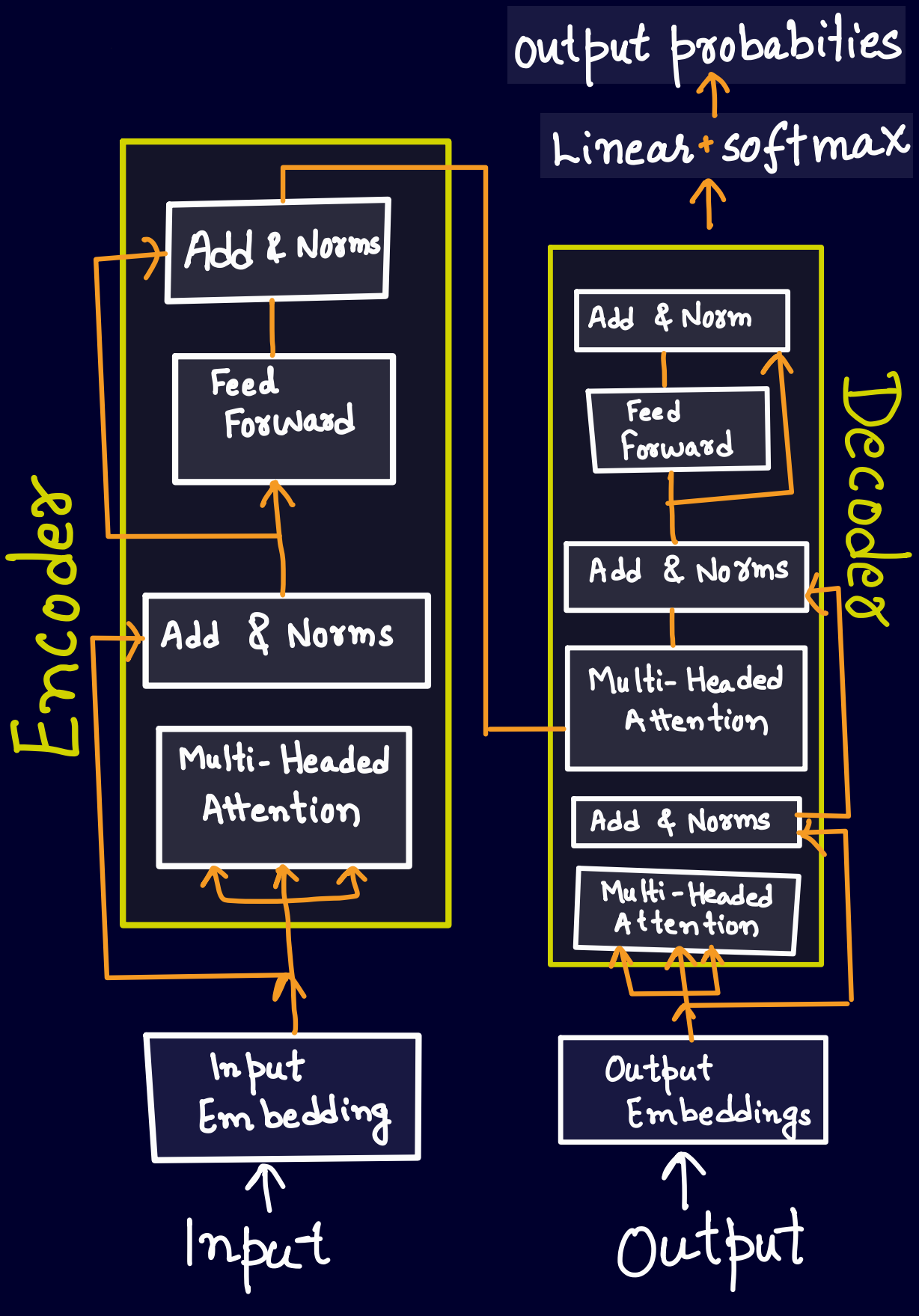
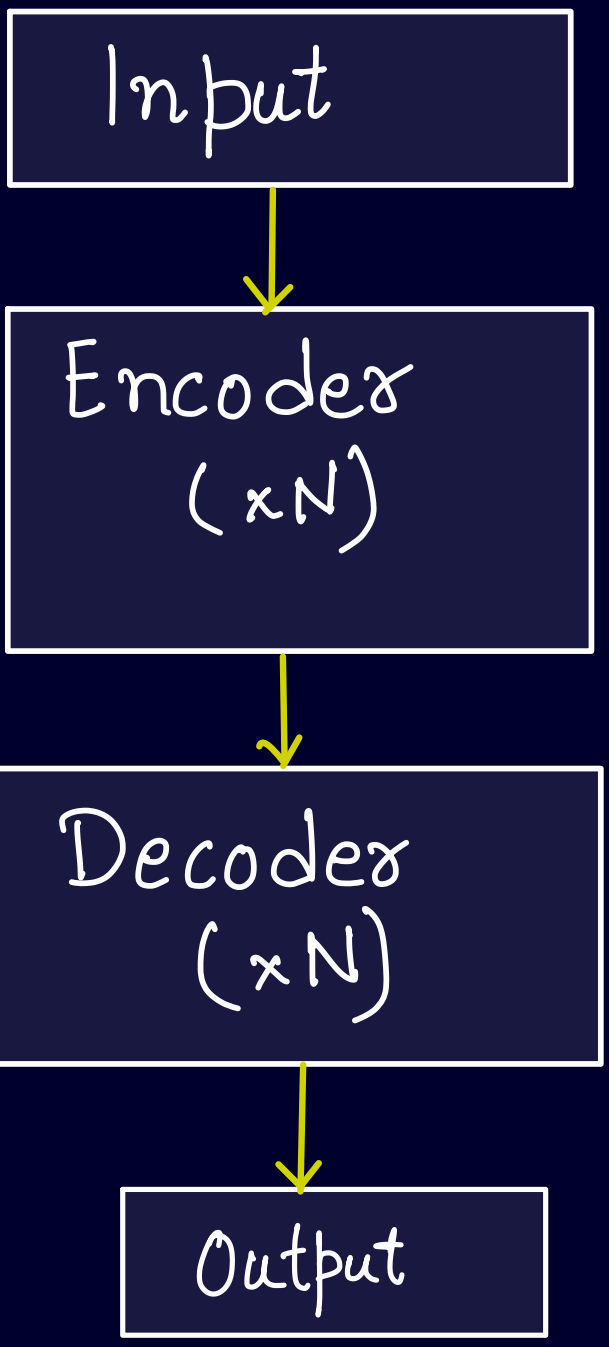
Building Block Of LLM

Transformer Architecture



In Simple :

Transformer



Why the transformer changed AI forever ?

In 2017, a research paper titled "Attention Is All You Need" introduced something that completely reshaped AI : The Transformer.

Parallel Processing – Unlike RNNs / LSTMs, transformers don't read text word by word. They process the entire sequence at once → making training lightning fast.

Attention Mechanism – Instead of treating all words equally, the model focuses on the most

relevant ones

Scalability – Transformer scale beautifully Add more data + more parameters which is strong model performance.

Versatality – They are not just for text. Transformers power vision models, speech recognition, protein folding and even recommendation systems.

The encoder only architecture is typically dedicated to extracting context-aware representations from input data. In encodes data into a dense but rich representation of its meaning. Encode only models are employed for tasks like classification and sentiment analysis, among others. A represent model from this category is BERT, which can be useful for classification tasks

The encoder-decoder architecture facilates sequence-to-sequence tasks such as translation, summarization and training multimodels like caption generator.

The decoder Only architecture is specially designed to produce outputs by following the intructions provided, as demonstrated in recent LLMs. It is specifically designed for next token prediction. The mean the model is not trained for a specific task (such as translation for encoder decoder). Rather they are trained for all tasks at the same time based on their training data. They will learn to predict words after an input sequence.