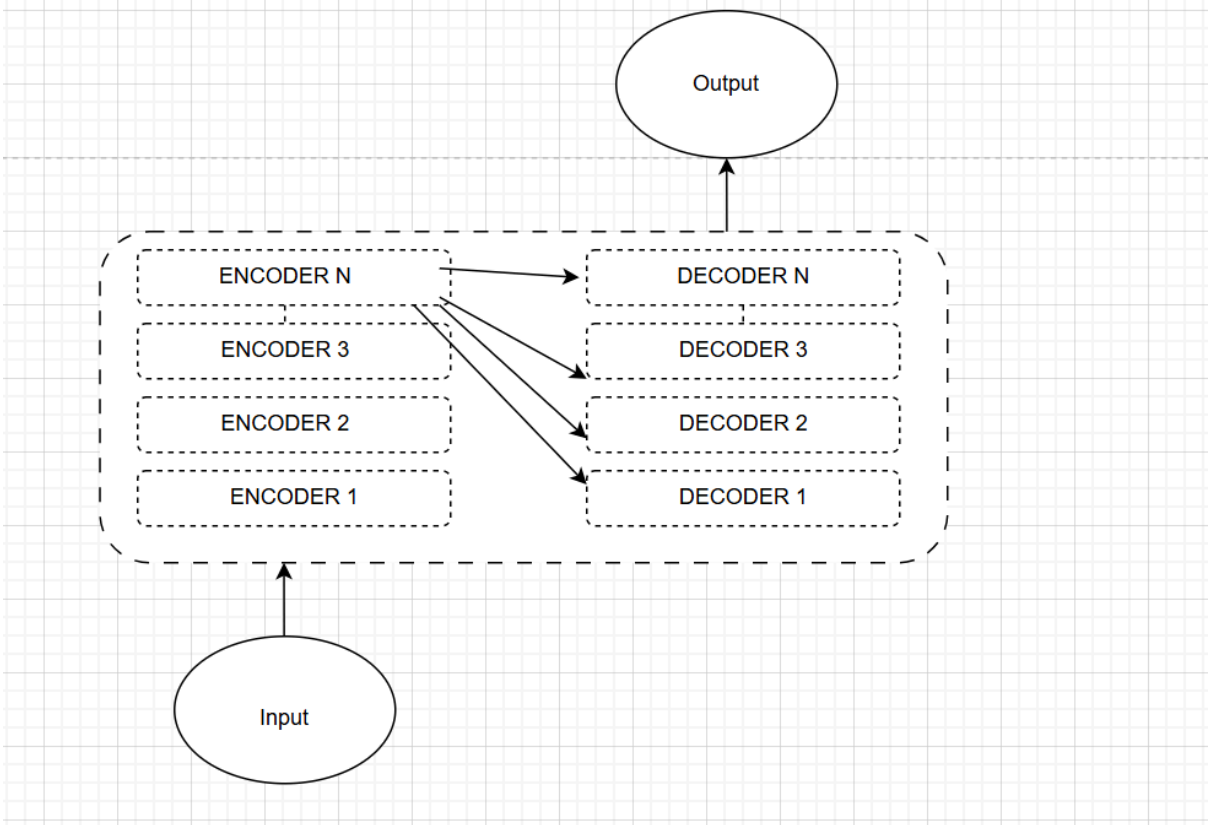
AI ENGINEERING TASK 2

A **token** is the most fundamental data unit in a text document, allowing AI to understand and process information. Tokenisation is the process of breaking down huge texts into smaller, more manageable chunks known as tokens. Depending on the approach, these tokens may be words, fragments of words, or phrases. Before an AI model processes any input, it separates it into discrete units to make it easier to analyze and generate responses. **Attention** enables the model to determine how much importance to assign to each token relative to others in the same input sequence. For every token, the model computes weighted relationships with all other tokens, allowing it to capture context and long-range dependencies effectively.

A **Transformer block** is a component of the transformer model architecture, which comprises multi-head self-attention and a Multi-Layer Perceptron layer. Most models consist of multiple such blocks that are stacked sequentially one after the other. The token representations evolve through layers, from the first block to the last one, allowing the model to build up an intricate understanding of each token.



TRANSFORMER ARCHITECTURE