**ASSIGNMENT\_1: GEN AI**

The paper "Attention Is All You Need" introduces the Transformer, a revolutionary neural network model by Google Brain and Google Research, designed for sequence transduction tasks such as machine translation. Traditional models have relied on recurrent (RNNs) or convolutional neural networks (CNNs) combined with attention mechanisms within an encoder-decoder framework. These models, while effective, suffer from limitations in parallelization due to their sequential nature. The Transformer, however, completely abandons recurrence and convolutions in favour of self-attention mechanisms, allowing it to achieve superior performance while significantly reducing training time. Its architecture comprises an encoder and a decoder, each made up of six identical layers. These layers utilize multi-head self-attention and fully connected feed-forward networks, augmented with residual connections and layer normalization. This design enables the model to process all tokens in a sequence simultaneously, enhancing computational efficiency and performance.

Central to the Transformer's innovation is the multi-head attention mechanism, which computes attention scores across multiple subspaces to capture various aspects of the input sequence. By applying scaled dot-product attention, the model can attend to different positions in the input, effectively learning long-range dependencies without the constraints of sequential processing. Positional encodings are introduced to maintain the order of tokens, using sine and cosine functions to encode positional information into the input embeddings. This combination of self-attention and positional encoding allows the Transformer to handle long sequences more effectively than traditional models. The Transformer's ability to parallelize computations and its effectiveness in capturing relationships across sequences have set new benchmarks in machine translation tasks, marking a significant advancement in neural network design and performance.