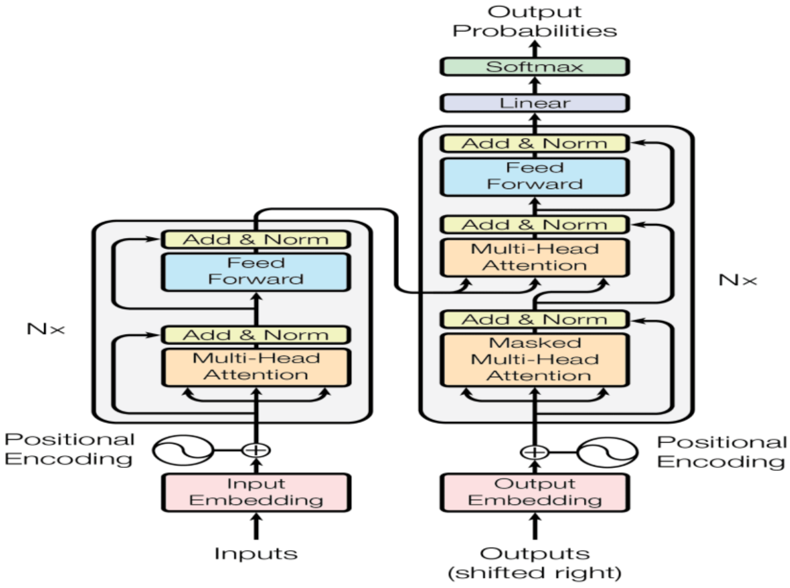
Transformers: Revolutionizing Deep Learning and AI

Transformers are a type of deep learning model architecture that has fundamentally reshaped the field of natural language processing (NLP), and more recently, computer vision, speech processing, and even protein folding. Introduced by Vaswani et al. in the landmark 2017 paper **"Attention is All You Need"**, transformers broke away from the sequential nature of earlier models like RNNs and LSTMs, enabling much faster and more scalable training on large datasets.

Key Concepts Behind Transformers

1. **Attention Mechanism**
   At the heart of transformers is the **self-attention mechanism**, which allows the model to weigh the importance of different words (or tokens) in a sequence when making predictions. For example, when translating a sentence, the model can focus on the most relevant parts of the input sentence regardless of their distance from each other in the sequence.
2. **Positional Encoding**
   Since transformers don't process data sequentially like RNNs, they use **positional encoding** to retain information about the order of tokens in the input sequence.
3. **Encoder-Decoder Structure**
   The original transformer architecture consists of two main parts:
   a. **Encoder:** Processes the input sequence and generates a contextualized representation.
   b. **Decoder:** Uses this representation to generate the output sequence, one token at a time.



1.