**GPT**

**Generative Pretrained Transformers**  
GPT is a natural language model widely used for generating, summarizing, and translating text.

**How It Works**

GPT is built on the **Transformer architecture**, as its name indicates. It is a **pretrained transformer**, meaning it has been trained on a vast amount of data to understand language.

**Workflow of Transformers**

The process of how GPT works can be broken down into several steps:

1. **Tokenization**
   * The input text is processed and split into smaller units called tokens.
   * For example, the sentence is divided into individual words or subwords.
2. **Embedding**
   * The tokens are converted into vector representations. Think of these vectors like coordinates in a high-dimensional space.
   * For instance, the words "queen" and "king" may be close in one dimension (because both represent leaders) but far apart in another dimension (based on gender). These dimensions represent features such as meaning, context, or relationships.
3. **Positional Encoding**
   * Unlike RNNs, GPT looks at the entire sequence as a whole. However, it still needs to understand the position of each token in the sequence.
   * To achieve this, a **positional vector** is added to the embeddings to encode the position of each token.

**Main Processing Steps**

1. **Self-Attention**
   * The model determines which tokens are most relevant to each other.
   * This is done by comparing three elements: **Query (Q)**, **Key (K)**, and **Value (V)**.
   * Attention scores are calculated by comparing Queries and Keys across all tokens. These scores determine the importance of each token in the context of others.
2. **Multi-Head Self-Attention**
   * Instead of performing attention calculations once, the process is repeated across multiple "heads," each focusing on different aspects of the relationships between tokens.
   * This enhances the model's ability to capture diverse patterns and relationships.
3. **Feed-Forward Network**
   * Each token's output from the attention mechanism is passed through a **feed-forward network**, which applies:
     + **Linear transformation** (a weighted adjustment to refine the data).
     + **ReLU activation** (to introduce non-linearity and improve learning).