1. **Explain the architecture of BERT**

BERT is basically **an Encoder stack of transformer architecture**. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side

1. **Explain Masked Language Modeling (MLM)**

Masked language modeling (MLM), **a self-supervised pretraining objective**, is widely used in natural language processing for learning text representations. MLM trains a model to predict a random sample of input tokens that have been replaced by a [MASK] placeholder in a multi-class setting over the entire vocabulary.

1. **Explain Next Sentence Prediction (NSP)**

**Next sentence prediction** (NSP) is one-half of the training process behind the BERT model (the other being masked-language modeling — MLM). Where MLM teaches BERT to understand relationships between words — NSP teaches BERT to understand longer-term dependencies across sentences.

1. **What is Matthews evaluation?**
2. **What is Matthews Correlation Coefficient (MCC)?**

Matthew's correlation coefficient, also abbreviated as MCC was invented by Brian Matthews in 1975. MCC is **a statistical tool used for model evaluation**. Its job is to gauge or measure the difference between the predicted values and actual values and is equivalent to chi-square statistics for a 2 x 2 contingency table.

1. **Explain Semantic Role Labeling**

In natural language processing, semantic role labeling (also called shallow semantic parsing or slot-filling) is **the process that assigns labels to words or phrases in a sentence that indicates their semantic role in the sentence, such as that of an agent, goal, or result**.

1. Why Fine-tuning a BERT model takes less time than pretraining
2. **Recognizing Textual Entailment (RTE)**

Recognizing Textual Entailment (RTE) was proposed as **a unified evaluation framework to compare semantic understanding of different NLP systems**

1. **Explain the decoder stack of GPT models.**

general architecture of the GPT-2 model consists of a sequence-to-sequence predictive task based on the transformer architecture [9] [1]. However, **it consists solely of the 12-layer decoder-only**