1. Explain the architecture of BERT

**Ans : 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. BERTBASE has 12 layers in the Encoder stack while BERTLARGE has 24 layers in the Encoder stack.**

1. Explain Masked Language Modeling (MLM)

**Ans : MLM consists of giving BERT a sentence and optimizing the weights inside BERT to output the same sentence on the other side. So we input a sentence and ask that BERT outputs the same sentence. However, before we actually give BERT that input sentence — we mask a few tokens.**

1. Explain Next Sentence Prediction (NSP)

**Ans : Next sentence prediction (NSP) is one-half of the training process behind the BERT model (the other being masked-language modeling — MLM). ... So, in this article, we'll cover exactly how we take an unstructured body of text, and use it to fine-tune a BERT model using NSP.**

1. What is Matthews evaluation?

**Ans : 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. What is Matthews Correlation Coefficient (MCC)?

**Ans : A significant percentage of the volume of protein crystals is occupied by solvent. ... Matthews defined VM, known as the Matthews coefficient, as the crystal volume per unit of protein molecular weight, and showed that VM bears a straightforward relationship to the fractional volume of solvent in the crystal.**

1. Explain Semantic Role Labeling

**Ans : 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

**Ans : In fact, the authors recommend only 2-4 epochs of training for fine-tuning BERT on a specific NLP task (compared to the hundreds of GPU hours needed to train the original BERT model or a LSTM from scratch!).**

1. Recognizing Textual Entailment (RTE)

**Ans : Textual entailment recognition is the task of deciding, given two text fragments, whether the meaning of one text is entailed (can be inferred) from another text (see the Instructions tab for the specific operational definition of textual entailment assumed in the challenge).**

1. Explain the decoder stack of GPT models.

**Ans : GPT-2 consists of solely stacked decoder blocks from the transformer architecture. In the standard transformer architecture, the decoder is fed a word embedding concatenated with a context vector, both generated by the encoder. In GPT-2 the context vector is zero-initialized for the first word embedding**