1. Explain the architecture of BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is based on Transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection.

2. Explain Masked Language Modeling (MLM)

MLM is a language model trained to predict the missing words in a sentence based on the context provided by the surrounding words. This is done by masking some of the words in the input text and training the model to predict the masked words based on the context of the non-masked words.

3. Explain Next Sentence Prediction (NSP)

Next sentence prediction (NSP): the models concatenates two masked sentences as inputs during pretraining. Sometimes they correspond to sentences that were next to each other in the original text, sometimes not. The model then has to predict if the two sentences were following each other or not

4. What is Matthews evaluation?

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

5. 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 tabl

6. 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. It serves to find the meaning of the sentence

7. Why Fine-tuning a BERT model takes less time than pretraining

During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks

8. Recognizing Textual Entailment (RTE)

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. This task captures generically a broad range of inferences that are relevant for multiple applications.

9. Explain the decoder stack of GPT models.

GPT is an autoregressive transformer decoder, which means that each token is predicted and conditioned on the previous token. We don't need an encoder, because the previous tokens are received by the decoder itself. This makes these models really good at tasks like language generation, but not good at classification.