1. Explain the Architecture of BERT

BERT is based on the **Transformer encoder** architecture. It uses multiple layers of self-attention and feed-forward networks to create deep bidirectional representations of text. BERT reads the entire sequence at once, allowing it to understand context from both left and right sides simultaneously.

---

2. Explain Masked Language Modeling (MLM)

MLM is a pretraining task where some tokens in the input are randomly masked, and the model tries to predict the original tokens. This helps BERT learn bidirectional context by predicting missing words based on surrounding words.

---

3. Explain Next Sentence Prediction (NSP)

NSP is another pretraining task where the model is given two sentences and learns to predict whether the second sentence follows the first one in the original text. This helps BERT understand relationships between sentences.

---

4. What is Matthews Evaluation?

It usually refers to the evaluation using the Matthews Correlation Coefficient (MCC), a metric for measuring the quality of binary classifications, especially useful for imbalanced datasets.

---

5. What is Matthews Correlation Coefficient (MCC)?

MCC is a metric that measures the correlation between predicted and true binary classifications. It returns a value between -1 and +1, where +1 indicates perfect prediction, 0 is random, and -1 is total disagreement.

---

6. Explain Semantic Role Labeling

Semantic Role Labeling (SRL) identifies the predicate-argument structure in a sentence, labeling phrases with roles such as who did what to whom, when, and where. It helps machines understand sentence meaning beyond syntax.

---

7. Why Fine-Tuning a BERT Model Takes Less Time Than Pretraining

Pretraining BERT requires learning from scratch on massive data, which is very compute-intensive. Fine-tuning starts from a pretrained model and only adjusts weights for a specific task with smaller datasets, making it much faster.

---

8. Recognizing Textual Entailment (RTE)

RTE is the task of determining whether one sentence logically follows (entails), contradicts, or is neutral with respect to another sentence. It tests a model's understanding of sentence meaning and relationship.

---

9. Explain the Decoder Stack of GPT Models

GPT uses the **Transformer decoder** architecture, which consists of layers with masked self-attention (to prevent seeing future tokens), feed-forward networks, and layer normalization. It generates text autoregressively, predicting one token at a time based on previous tokens.