1 .Wh i ch ofthefo ll owingbe s texplainshowmulti-headattenti onrimproves npdfual understanding in Transformers?
A) By reducing the total number of parameters through parallelization
B) By enforcing uniform attention over the sequence to prevent bias
C) By increasing computation speed through batch-wise attention
D) By enabling different heads to attend to diverse relational patterns across positions

2. Which component of the Transformer architecture is exclusively utilized in GPT, making it more suited for generative tasks?
A) Decoder layers with masked self-attention
B) Encoder layers for input sequence modeling
C) A hybrid encoder-decoder combination
D) A purely feed-forward architecture

3 .Wh atdes ign c ho iceinGPT r est rictsitf rom leveragingfullbidirectionalco npdf,a nd what consequence does this have?
A)Encoder-b asedde sign;restrictso utput generation
B)Unidirectionalleft-to-ri ghtow;limi t sful lco npdfu nderstanding
C)Bidirectionalmasking ;leadst ocon p dfove rfitting
D) Cross-attention dependencies; increase inference latency

4. Which of the following best characterizes the training objectives that enable BERT to capturebothdeeptoken- le vel conpdfan dint er -sentencese mantics?
A)Predicti ng the nex t tokeninaleft-to-rightfashi onusin guni directionalc onpdf
B) Learning to generate a target sequence from an input sequence in an encoder-decoder setup
C) Jointly optimizing masked token reconstruction and inter-sentence coherence discrimination
D)Aligni ngim agefeat ure swithp dfualdescriptionsthrou ghcros s-modalsu pervision

Short Answer Questions

5. What are the potential drawbacks of the two-stage process of pretraining on large corpora followed by fine-tuning on specific tasks in Transformer models?

6. What are the potential drawbacks of GPT's autoregressive training objective when applie d tota sksrequi ringhol isticunderstandingo f pdf?

7. BERT utilizes a masked language model (MLM) during pretraining. What is the primary challenge associated with the MLM approach, and how does it affect the model's downstream performance?

8. GPT models are known for their unidirectional (left-to-right) processing. How does thi sdesi gnch oicei mpac ttheirperformanceontas k slik etext generationmparedt o task sl iketextclassification?