1. Which of the following best explains how multi-head attention improves contextual 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. What design choice in GPT restricts it from leveraging full bidirectional context, and what consequence does this have?
   A) Encoder-based design; restricts output generation
   B) Unidirectional left-to-right flow; limits full context understanding
   C) Bidirectional masking; leads to context overfitting
   D) Cross-attention dependencies; increase inference latency

4. Which of the following best characterizes the training objectives that enable BERT to capture both deep token-level context and inter-sentence semantics?
   A) Predicting the next token in a left-to-right fashion using unidirectional context
   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) Aligning image features with textual descriptions through cross-modal supervision

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 applied to tasks requiring holistic understanding of text?

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 this design choice impact their performance on tasks like text generation compared to tasks like text classification?