CSE 576 Quiz 7 Transformer Pretraining Quiz

Q1.

Which of the following best explains how multi-head attention improves contextual understanding in T

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

Q2.

Which component of the Transformer architecture is exclusively utilized in GPT, making it more suited

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

Q3.

What design choice in GPT restricts it from leveraging full bidirectional context, and what consequenc

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

Q4.

Which of the following best characterizes the training objectives that enable BERT to capture both de

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

Q5.

What are the potential drawbacks of the two-stage process of pretraining on large corpora followed b

Q6.

What are the potential drawbacks of GPT's autoregressive training objective when applied to tasks re

Q7.

BERT utilizes a masked language model (MLM) during pretraining. What is the primary challenge assoc

Q8.

GPT models are known for their unidirectional (left-to-right) processing. How does this design choice i