1.00

b. Masked self-attention
c. Layer normalization
d. Positional encoding

Question 2
Complete
Mark 1.00 out of 1.00

What is the purpose of the feedforward layer in the decoder?

a. To perform token embedding
b. To normalize the input sequence
c. To apply non-linearity and enhance feature representation
d. To compute attention scores

Question 3
Complete
Mark 1.00 out of 1.00

What role does the layer normalization play in the Transformer decoder?

a. It applies dropout to prevent overfitting
b. It normalizes the output of each sub-layer to stabilize training
c. It removes redundant information from input embeddings
d. It directly maps token embeddings to output probabilities

Question 4
Complete
Mark 1.00 out of 1.00

What happens if masking is not applied in the decoder's self-attention?

a. The model will generate incorrect predictions by seeing future tokens
b. The model will run faster and more efficiently
c. The decoder will fail to use the encoder's output
d. The model will only use the previous token for prediction

Question 5
Complete
Mark 1.00 out of 1.00

What are the inputs to the decoder in a Transformer model?

a. The encoder's output and a positional embedding
b. The raw input sequence and the encoder's output
c. The output of the previous decoder layer only
d. The final output from the encoder and the ground-truth output sequence shifted by one position

Question 6
Complete
Mark 1.00 out of 1.00

In the Transformer model, what is the main function of the decoder?

a. Encode input sequences into fixed representations
b. Generate position embeddings
c. Compute attention weights for input tokens
d. Convert encoded representations into output sequences

Question 7
Complete
Mark 1.00 out of 1.00

In the decoder's multi-head self-attention, why do we use multiple attention heads?

a. To ensure each token attends to only one other token
b. To reduce memory consumption
c. To increase computational speed
d. To capture different aspects of the input representation

Question 8
Complete
Mark 1.00 out of 1.00

Why is masking applied in the decoder's self-attention layer?

a. To reduce computational complexity
b. To allow bidirectional context understanding
c. To prevent the decoder from seeing future words during training
d. To improve performance on long sequences

Question 9
Complete
Mark 1.00 out of 1.00

How does the self-attention mechanism in the decoder differ from that in the encoder?

a. The decoder uses bidirectional attention, while the encoder does not
b. The decoder's self-attention is computed after the feedforward layer
c. The decoder does not use self-attention, only cross-attention
d. The decoder has masked self-attention, preventing attention to future tokens

Question 10
Complete
Mark 1.00 out of 1.00

What is the output of the final decoder layer before softmax activation?

a. The hidden state representations
b. A sequence of token embeddings
c. The original input sequence
d. A probability distribution over the vocabulary