← Back Practice Quiz: Encoder Models
Practice Assignment • 12 min

English

Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 75%. We keep your highest score.

Next item →

1. Why does BERT use an encoder-only architecture, that is, only the encoder part of the transformer model? 1 / 1 point

○ Because in encoder models, causal attention is visually represented by an 'X' for the masked attention and 'O's' for the active attention units.

● It allows BERT to process entire sequences of text simultaneously.

○ It allows BERT to be used for text-generation tasks.

○ Because encoder models possess a unidirectional training method.

✓ **Correct**
BERT uses an encoder-only architecture because this design allows BERT to process entire sequences of text simultaneously, theoretically enhancing its understanding of the context and nuances within the text.

2. In the next sentence prediction (NSP) task, which of the following determines whether the second sentence logically follows the first for a given pair of sentences? 1 / 1 point

○ Positional encoding

○ Separate token

● CLS token's contextual embedding

○ Segment embeddings

✓ **Correct**
The CLS token corresponds to the NSP classification token, which encapsulates information from a given sequence. In the NSP task, the input consists of word embeddings processed by the encoder to generate contextual embeddings. These embeddings are used to determine whether, for a given pair of sentences, the second sentence logically follows the first, just like a standard classification task.

3. How many classes are there in the output layer of the neural network for mask language modeling (MLM)? 1 / 1 point

○ The number of classes will be equal to the number of special tokens.

● The number of classes will be equal to the size of the vocabulary.

○ The number of classes will be equal to the size of NSP.

○ There will be two classes.

✓ **Correct**
As the neural network is going to predict the possible words, the output must equal the vocabulary size.

4. Identify one of the most common pretraining objectives useful for training the BERT model in PyTorch. 1 / 1 point

○ You can use semi-supervised learning with limited labeled data.

○ You can supervise the learning with labeled data.

● You can use unsupervised learning with masked language modeling (MLM).

○ You can initialize the model's parameters randomly.

✓ **Correct**
The pretraining BERT model uses unsupervised learning with MLM to mask the portion of the input tokens so that the model learns to predict them.