1. Which tokenization method generates a smaller vocabulary but increases input dimensionality and computational needs?

1 point

○ Character-based tokenization

○ WordPiece tokenization

○ SentencePiece tokenization

○ Word-based tokenization

2. Imagine you are training a sentiment analysis model where the input consists of user reviews. After tokenization, you find that the sequences have varying lengths. Which concept will you employ to address the issue of varied lengths while using data loaders?

1 point

○ Shuffling

○ Padding

○ Batching

○ Iteration

3. Fill in the blank.

1 point

In subword-based tokenization, the ________ indicates that the word should be attached to the previous word without a space.

○ <pad> token

○ Underscore symbol

○ ## symbol

○ <eos> special token

4. Identify an advantage of word-based tokenization.

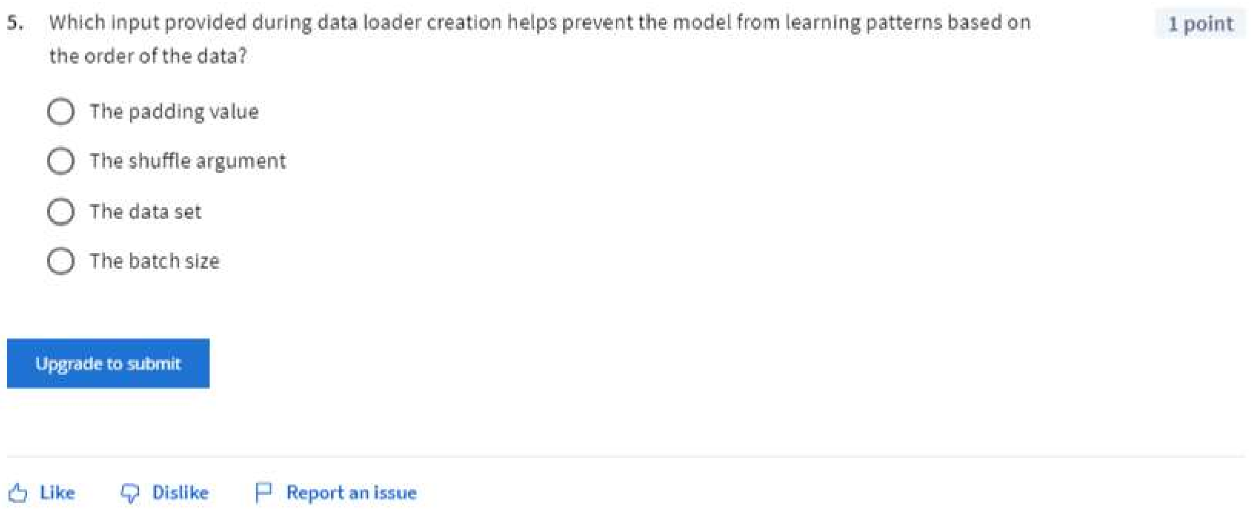
1 point

○ It preserves the semantic meaning

○ It evaluates the benefits and drawbacks of splitting and merging two symbols

○ It breaks down infrequent words to meaningful subwords

○ It creates smaller vocabulary



5. Which input provided during data loader creation helps prevent the model from learning patterns based on the order of the data?

1 point

○ The padding value

○ The shuffle argument

○ The data set

○ The batch size