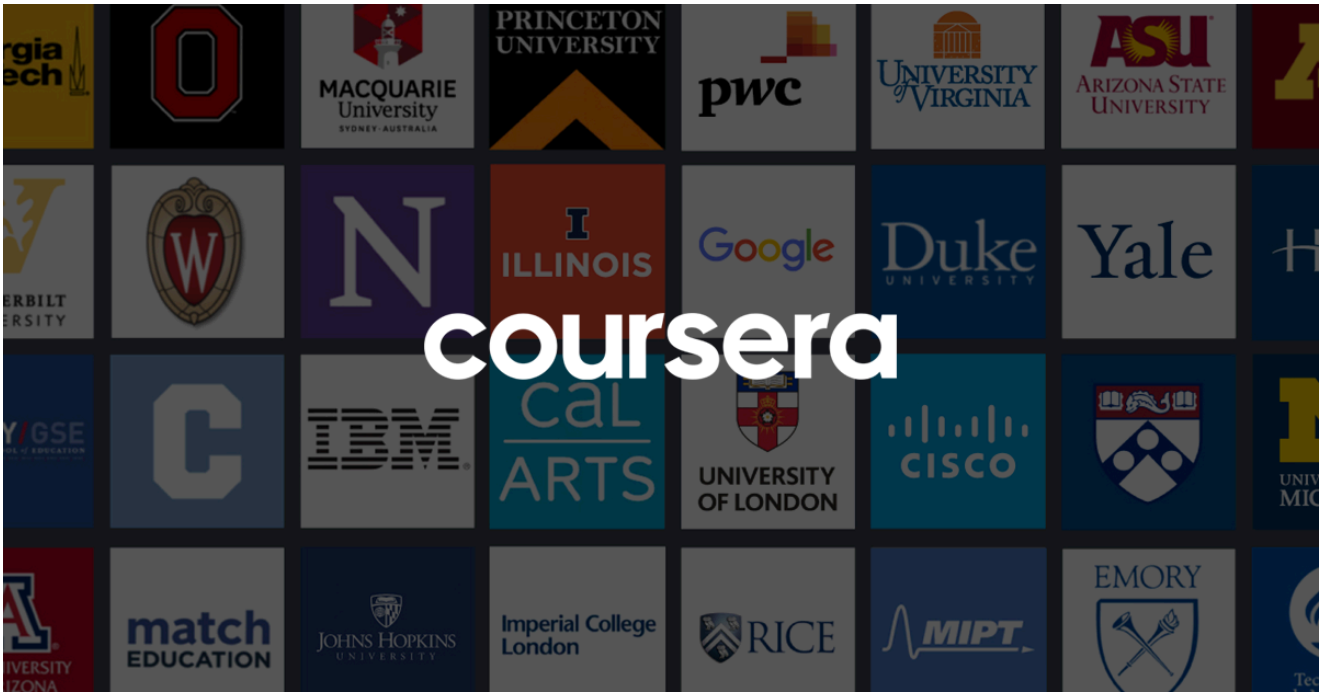
Graded Quiz: Data Preparation for LLMs

coursera.org/learn/generative-ai-llm-architecture-data-preparation/assignment-submission/Nf5tU/graded-quiz-data-preparation-for-llms/view-feedback



Ready to review what you've learned before starting the assignment? I'm here to help.

Assignment details

Due

October 4, 11:59 PM PDTOct 4, 11:59 PM PDT

Submitted

October 2, 2:57 AM PDTOct 2, 2:57 AM PDT

Attempts

2 left (3 attempts every 8 hours)

Your grade

To pass you need at least 80%. We keep your highest score.

100%

Graded Assignment • 15 min

DueOct 4, 11:59 PM PDT

Your grade: 100%

Your latest: 100%•

Your highest: 100%•

To pass you need at least 80%. We keep your highest score.

Next item→

1.

Question 1

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

✓
Correct

Character-based tokenization generates a smaller vocabulary as each character is treated as a separate token. However, since each character becomes a unique token, it increases the input dimensionality and computational needs due to the larger number of tokens.

Status: [object Object]

1 / 1 point

2.

Question 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?

Correct

You would use padding to ensure that each sample in a data loader is of the same length.

Status: [object Object]

1 / 1 point

3.

Question 3

Fill in the blank.

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

Correct

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

Status: [object Object]

1 / 1 point

4.

Question 4

Identify an advantage of word-based tokenization.

Correct

In word-based tokenization, the text is divided into individual words, each word considered a token. An advantage of word-based tokenization is that it preserves the semantic meaning.

Status: [object Object]

1 / 1 point

5.

Question 5

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

Correct

You can mention the shuffle argument as true. This shuffling is particularly useful for training deep learning models, as it prevents the model from learning patterns based on the order of the data.

Status: [object Object]

1 / 1 point