

# Questions

 [Bookmark this page](#)

## Q1

5.0/5.0 points (graded)

What is a "word embedding" in the context of NLP deep learning models? (Choose the best answer)

- ☐ A word that is found in a paragraph
- ☒ A vector that encodes a representation of a word such that similar words will have similar representations
- ☐ An algorithm that turns words into numbers
- ☐ A model that can translate one language into another



Submit

## Q2

5.0/5.0 points (graded)

The "Attention is All You Need!" paper introduced the Transformer architecture in 2017. What are some of the significant features of the architecture? (Check all that apply)

- ☒ RNNs are no longer required for sequence modeling in tasks such as NMT
- ☒ Fewer sequential operations result in greater parallelization when processing the model
- ☒ Relationships between words that are sequentially farther from each other can be encoded
- ☐ LSTM layers combined with self-attention layers create a robust parallel encoder



Submit

 Correct (5/5 points)

## Q3

5/5 points (graded)

What is the purpose of an "attention mechanism" in the Transformer architecture? (Choose the best answer)

- ☐ It sets an interrupt in the system when something important happens in an input sequence
- ☐ It is not used in the Transformer architecture
- ☐ It is a secondary model that periodically gets the "attention" of the primary model to provide additional information
- ☒ It looks at an input sequence and decides, at each step, which other parts of the sequence are important and encodes that information



Submit

 Correct (5/5 points)

## Q4

5/5 points (graded)

Which of the following are goals of deploying an NLP model to production on a system such as NVIDIA Triton Inference Server? (Check all that apply)

- ☐ Improve the accuracy of the deployed inference model
- ☒ Faster throughput of inference via various optimizations
- ☐ Quickly training a new model in real time
- ☒ Ability to run inference on multiple models simultaneously



Submit

 Correct (5/5 points)

## Q5

5/5 points (graded)

Which of the following is an example of post-training optimization? (Choose the best one)

- ☐ self-attention
- ☐ embedding
- ☒ quantization
- ☐ self-supervision



Submit

 Correct (5/5 points)

## Q6

5/5 points (graded)

Which of the following model types are supported by Triton Server? (Check all that apply)

- ☒ TensorRT
- ☒ PyTorch
- ☒ ONNX
- ☒ TensorFlow
- ☒ Custom Backends



Submit

 Correct (5/5 points)