1 point

1. What is retrieval augmented generation (RAG)?

- ○ A deep learning technique for training large language models
- ◉ A method for improving language model outputs by using a search database
- ○ A vector search algorithm used in information retrieval
- ○ A type of few-shot learning method for natural language processing

1 point

2. What are embeddings in the context of retrieval augmented generation?

- ○ Compressed versions of the original text data stored in a vector database
- ○ Pre-trained language models used for generating text embeddings
- ◉ Numerical representations of text in a high-dimensional space that enable similarity comparisons
- ○ Optimization algorithms used for training language models on vector databases

1 point

3. What is the role of the Pandas library in the context of retrieval augmented generation with a CSV file?

- ○ To create and manage the vector database for storing embeddings
- ◉ To load and preprocess the CSV file data for creating embeddings and populating the vector database
- ○ To fine-tune the pre-trained language model on the CSV data
- ○ To implement the retrieval augmented generation algorithm itself

1 point

4. What is the purpose of cosine distance in the context of retrieval augmented generation?

- ◉ To measure the similarity between the input query and the stored document embeddings in the vector database
- ○ To optimize the training process of the language model used for creating embeddings
- ○ To compress the size of the vector database for efficient storage
- ○ To enhance the interpretability of the language model's outputs

1 point

5. What is the purpose of Azure AI Search (formerly Azure Cognitive Services Search) in the context of retrieval augmented generation?

- ○ To host and serve the large language model used for text generation
- ○ To perform data preprocessing and cleaning before creating embeddings
- ◉ To provide a cloud-based search service with retrieval augmentation capabilities for language models
- ○ To fine-tune the pre-trained language model on domain-specific data