Summary and Highlights: Introduction to RAG

Congratulations! You have completed this lesson. At this point in the course, you know that:

- Retrieval-Augmented Generation (RAG) is a machine-learning technique that integrates information retrieval with generative AI to produce accurate and context-aware responses.
- RAG enhances Large Language Models (LLMs) by integrating external or domain-specific knowledge without retraining. This helps LLMs generate more accurate and contextually relevant responses for specialized queries, such as a company's mobile policy.
- RAG consists of two main components: the Retriever, which extracts relevant data from a knowledge base, and the Generator, which uses the retrieved information to generate responses in natural language.
- The RAG process comprises four steps: Text Embedding, Retrieval, Augmented Query Creation, and Model Generation.
- Text Embedding converts user prompts and knowledge base documents into high-dimensional vectors using AI models such as BERT or GPT.
- Retrieval matches the user query with similar vectors from the knowledge base to retrieve relevant information.
- Augmented Query Creation combines retrieved content with the user prompt.
- Model Generation uses the created augmented query to generate a response using the content from the knowledge base.
- Prompt encoding converts a text-based prompt into a numerical representation that an LLM can process. It uses Token Embedding and Vector Averaging to break documents into smaller text chunks, convert them into vectors, and index them in a vector database.
- Distance Metrics measure similarity between user queries and document vectors using methods such as dot product (magnitude-based) or cosine similarity (direction-based).
- RAG is an efficient response generation technique. It retrieves the most relevant text chunks from the knowledge base to augment the model's knowledge and produce an informed response. This ensures responses are accurate, domain-specific, and up-to-date.