Module Summary: RAG Using LangChain

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

- LangChain uses text splitters to split a long document into smaller chunks.
- Text splitters operate along two axes: Method used to break the text and how the chunk is measured.
- Key parameters of a text splitter: Separator, chunk size, chunk overlap, and length function.
- Commonly used splitters: Split by Character, Recursively Split by Character, Split Code, and Markdown Header Text Splitter.
- Embeddings from data sources can be stored using a vector store.
- A vector database retrieves information based on queries using similarity search.
- Chroma DB is a vector store supported by LangChain that saves embeddings along with metadata.
- To construct the Chroma DB vector database, import the Chroma class from LangChain vector stores and call the chunks and embedding model.
- A similarity search process starts with a query, which the embedding model converts into a numerical vector format.
- The vector database compares the query vector to all the vectors in its storage to find the ones most similar to the query.
- A LangChain retriever is an interface that returns documents based on an unstructured query.
- Vector Store-Based Retriever retrieves documents from a vector database using similarity search or MMR.
- Similarity search is when the retriever accepts a query and retrieves the most similar data.
- MMR is a technique used to balance the relevance and diversity of retrieved results.