Summary and Highlights

Congratulations! You have completed this module. At this point, you know that:

Multimodal Retrieval-Augmented Generation (MM-RAG)

- Combines multimodal inputs, such as text + images or videos, with retrieval-augmented generation, fetching relevant data to enhance LLM responses
- Pattern follows three steps:
  - Multimodal data retrieval
  - Contrastive learning for embeddings
  - Generative models informed by multimodal context
- Pipeline has four steps.
  - Data indexing: Diverse data, such as text, images, audio, and video, is converted into embeddings and stored in a vector database for efficient retrieval
  - Data retrieval: User query is embedded, and semantically relevant multimodal data is fetched from the vector database
  - Augmentation: Retrieved data is combined with the original query to enrich the context for generation
  - Response generation: Multimodal response is generated using the augmented input, blending information from all modalities

Multimodal Chatbots and QA Systems

- Advanced AI systems that process and respond to multiple data types, such as text, images, audio, and video.
- Can see, read, and understand the world more like humans do
- Key features:
  - Multiple input modalities
  - Integrated understanding
  - Contextual response generation
- Basic implementation steps: