Reading: Summary and Highlights

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

Image captioning with Meta's Llama

- Combining computer vision with natural language processing creates powerful tools for understanding visual content.
- Three main stages of the image captioning process with a multimodal large language model (LLM) are:
  - Input processing
  - Image validation and encoding
  - Multimodal LLM processing
- Input processing receives and prepares the image and optional text prompt.
- Image validation and encoding validate and convert the image into a format (e.g., Base64) suitable for the model.
- Multimodal LLM processing combines visual and textual information to generate a descriptive caption.
- Core components of the image captioning system to produce captions tailored to prompts are:
  - Visual encoders
  - Text embedding
  - Fusion layers
  - Language generation tools
- Implementing an image captioning system using Meta's Llama 4 Maverick model via IBM watsonx involves:
  - Importing libraries and authenticating access
  - Encoding images and preparing prompts
  - Sending combined image-text messages to the model
  - Extracting descriptive text from the model's response

Text-to-video generation with OpenAI's Sora