Reading: Course Overview

Welcome to the **Build Multimodal Generative AI Applications** course!

This intermediate-level course is designed for AI developers, data scientists, and engineers looking to expand their expertise in building applications that leverage multimodal generative AI. You'll gain hands-on experience integrating text, speech, images, and video to develop practical, cross-modal applications using powerful models and frameworks.

Throughout this course, you'll explore how to build applications that transform and generate across modalities—such as turning speech into text, text into images, or images into captions—using models like Meta's Llama 4, OpenAI Whisper, and DALL-E. You'll work with tools such as Gradio, Flask, LangChain, and Hugging Face to prototype and deploy user-facing multimodal AI systems.

You'll begin by mastering the fundamentals of multimodal AI, gaining experience with speech-to-text and text-to-speech applications. Then, you'll learn how to generate and caption images and videos using advanced generative models. Finally, you'll build real-world multimodal applications—such as chatbots, personal assistants, and recommendation systems—using cross-modal retrieval and generation techniques.

This course is part of the **IBM RAG and Agentic AI Professional Certificate**, aimed at equipping learners with cutting-edge skills to develop next-generation AI systems that understand and generate multiple forms of media.

Prerequisites

To succeed in this course, you should have:

- Proficiency in Python
- Prior exposure to Flask, Gradio, and LangChain
- A working knowledge of generative AI fundamentals and frameworks like Hugging Face

Learning objectives

After completing this course, you will be able to:

- Explain key challenges and integration techniques in multimodal AI
- Build applications using models like Whisper, DALL-E, and Llama 4