Introduction to Generative AI

Generative AI refers to algorithms that can be used to create new content, such as text, images, or audio, often indistinguishable from human-made data. These models include GANs, VAEs, and large language models like GPT.

Applications of Generative AI

Common applications include content generation, chatbots, personalized marketing, game development, and data augmentation for training machine learning models. These applications leverage the model's ability to learn underlying patterns and generate new data accordingly.

Retrieval-Augmented Generation (RAG)

RAG is a hybrid approach that combines retrieval-based methods with generative models. It retrieves relevant documents from a knowledge base and uses them to condition the generation of more accurate and contextually aware responses. This approach helps large language models stay grounded and up-to-date.

Vector Databases

Vector databases like FAISS, Chroma, Pinecone, and Weaviate allow efficient similarity search by storing data as high-dimensional vectors. These are often used in RAG pipelines to quickly find semantically similar content from a document corpus.