Introduction to Vector Databases

Vector databases store embeddings as high-dimensional vectors and enable efficient similarity search.

They are commonly used in modern AI systems such as recommendation engines and retrieval-augmented generation (RAG).

What is Vector Quantization?

Vector quantization is a technique used to reduce the memory footprint of embeddings.

It converts floating-point vectors into lower-precision representations such as int8 or binary.

This improves retrieval speed while trading off a small amount of accuracy.

RAG Systems Overview

A RAG system combines vector search with large language models.

Relevant documents are retrieved first and then passed as context to the language model.

Why Observability Matters

In production systems, measuring latency, cost, and retrieval quality is critical.

Observability enables adaptive systems that can balance speed and accuracy dynamically.