# AirNir – Generative AI Platform

## Phase 3.2 Documentation – Embeddings & Vector Storage

### 📂 Project Structure

/AirNir  
├── Library  
│ ├── ArNir.Core → Entities, DTOs, Config  
│ ├── ArNir.Data → DbContexts (SQL Server + Postgres), EF Migrations  
│ └── ArNir.Service → Business logic (EmbeddingService, RetrievalService)  
├── Presentation  
│ ├── ArNir.Admin → ASP.NET Core MVC (AdminLTE UI, embeddings test page)  
│ └── ArNir.Frontend → ASP.NET Core MVC (User-facing, RAG search – future)  
├── sql  
│ ├── create\_tables.sql → SQL Server (Documents, Chunks)  
│ ├── update\_documents\_chunks.sql → SQL Server updates  
│ └── update\_embeddings.sql → Postgres schema for embeddings (pgvector)  
└── docs  
 └── Phase3.2\_Architecture.png → Embedding & Vector Storage Architecture

## Phase 3.2 – Embeddings & Vector Storage

### 🎯 Goal

Enable semantic representation of document chunks using **OpenAI embeddings** and store them in a **vector database (Postgres + pgvector)** to support similarity search.

### 🔑 Key Features

* **pgvector Setup**:
  + Installed Postgres in Docker with ankane/pgvector image.
  + Enabled CREATE EXTENSION vector;.
* **DbContexts**:
  + SQL Server → Documents & Chunks.
  + Postgres → Embeddings (vector(1536)).
* **Entities**:
  + Embedding entity with Pgvector.Vector type.
* **EF Core Migrations**:
  + Created embeddings table in Postgres.
* **EmbeddingService**:
  + Retrieves chunks from SQL Server.
  + Calls OpenAI text-embedding-ada-002.
  + Saves embeddings into Postgres.
* **Admin Test UI**:
  + Page /Embedding/Test to input custom text.
  + Generates embedding + runs similarity search.
  + Displays top-k nearest chunks.
* **Validation**:
  + Verified embeddings stored in Postgres.
  + Confirmed similarity search queries via EF Core and raw SQL.

### 🖼️ Architecture Diagram

|  |
| --- |
| Phase 3.2 Architecture |

Phase 3.2 Architecture

### ✅ Outcome

A fully functional **embedding pipeline**: - Documents + Chunks → SQL Server. - Embeddings → Postgres with pgvector. - Admin UI enables quick validation and semantic search. - System ready for **Phase 3.3 – Retrieval Service**.