# Assignment 6 LLMOPs with AWS

## Steven Kok

## Context of the work and report

During this case I had a number of structural limitations because I am behind Zscaler in my corporate laptop. This required changes throughout the codebase. – especially on the pipelines. But there were also several changes I had to make throughout the code to make it work. The details of this are in : REPO\_CHANGES\_ZASCALER.md

For the AWS setup section, I aimed to fully automate the setup to learn how to use CLI as much as possible. The information is in AWS\_README.md and the scripts are setup-aws-bedrock.sh

For Bedrock support, I wanted to also learn the difference in use of different models, and how to best encapsulate configurability, so I generalised the code to support multiple model options including OpenAI+SentenceTransformer (initial implementation), OpenAI full, Bedrock and Google, and has placeholder implementations for Huggingface, Ollama and. Mistral. This also required to make enhancements on the client app and the introduction of new services, primarily the need to rebuild the vector database, as the vector size changed by model (this was covered in the exercises, and I found that different models needed different vector sizes to work – maybe this is a limitation in my understanding that I need to explore further) The summary of changes and adjustments to the repo can be found in: REPO\_MULTI\_MODEL\_DIFF\_REPORT.md.

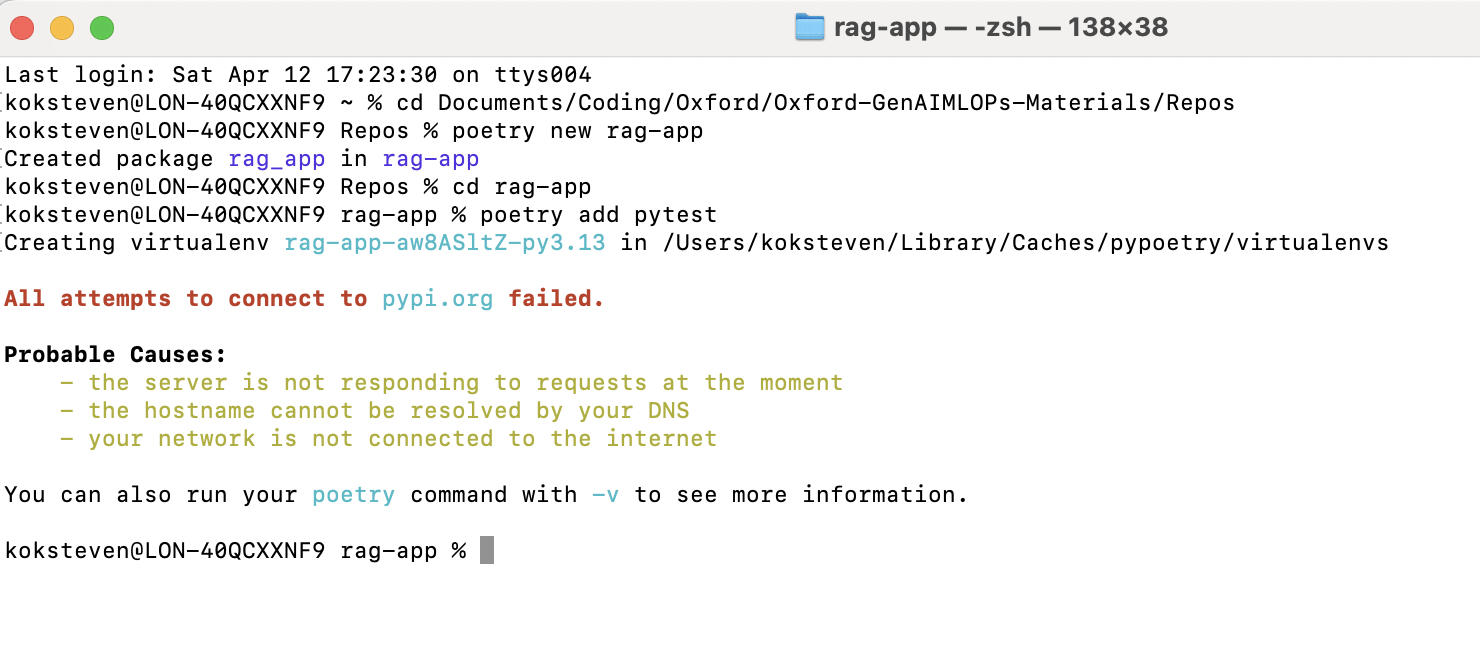
The full project is in the following repo: https://github.com/skok007/MLOPs\_with\_AWS

The repo main folder has a REPO\_CHANGES.md with all the details of the changes I had to do across the codebase

## OpenAI implementation

### 00 – Python Fundamentals – A Warmp Up

#### 1. Create a directory called ‘rag-app’ and use the Python dependency management solution poetry to initialize your project. Use poetry to install the pytest.



As you can see, poetry does not work because I have a limitation on my corporate laptop. I can make it work inside the devcontainer but not outside. This is being addressed; but going forward everything that I need to install outside the devcointainer I need to do with creating and environment and using pip.

#### 2. Write a small python script with functions within it that will add two numbers, subtract two numbers and multiply two numbers. You can call the script ‘basic\_functions.py’.

def add(*x*: int, *y*: int) -> int:

return *x* + *y*

def multiply(*x*: int, *y*: int) -> int:

return *x* \* *y*

#### 3. Write unit tests for all three of these functions inside a folder called tests (create this under your ‘rag-app’ folder) and put these inside a Python script called test\_basic\_functions.py.

import pytest

from basic\_functions import add, multiply

def test\_add\_positive\_numbers():

assert add(1, 2) == 3

def test\_add\_negative\_numbers():

assert add(-1, -2) == -3

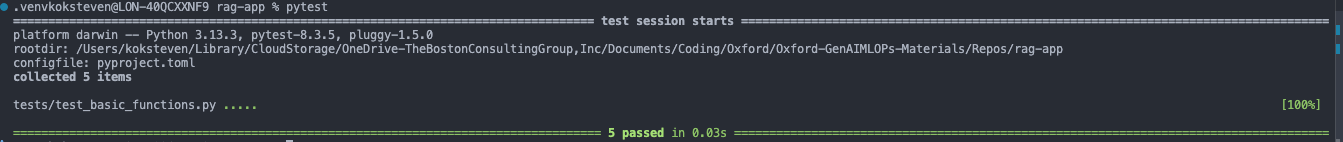
def test\_multiply\_positive\_numbers():

assert multiply(1, 2) == 2

def test\_multiply\_negative\_numbers():

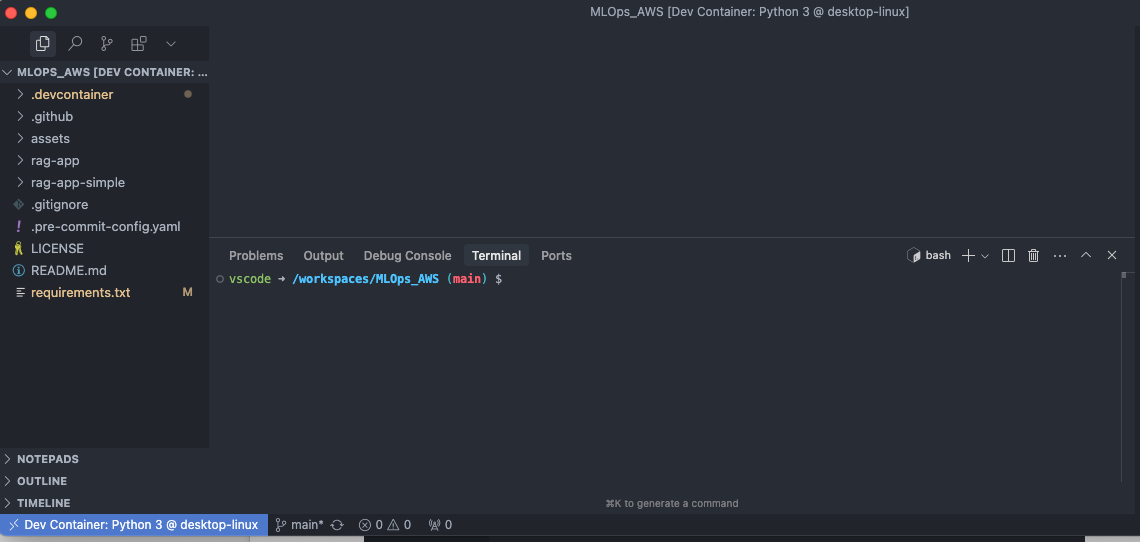
assert multiply(-1, -2) == 2

#### 4. Run pytest to return a unit test completion report.

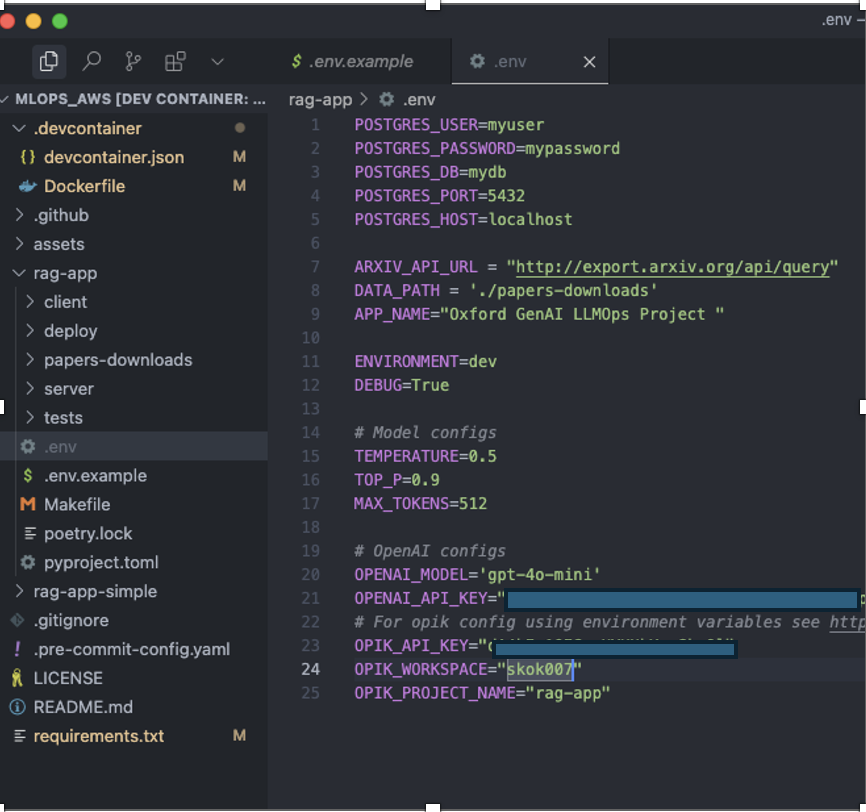


### 01 – Project Setup, FastAPI & Docker

#### 1. Clone the project repository, 2. Open the repository in IDE and 3. Rebuild and reopen container



#### 4. create .env, 5. Fill missing variables add Opik details



#### 7. make install

I had multiple issues with this as I have a corporate laptop behind Zscaler. Had to re-train on the use of certificates for npm and docker. dockerfile here

Adjusted

*# Use dev container base image*

FROM mcr.microsoft.com/devcontainers/python:1-3.12-bookworm

*# ✅ Trust Zscaler cert early (use .crt extension for update-ca-certificates)*

COPY ZscalerRootCertificate-2048-SHA256.pem /usr/local/share/ca-certificates/zscaler.crt

RUN apt-get update && apt-get install -y ca-certificates && update-ca-certificates

*# ✅ Make pip, curl, and Node respect custom certs*

ENV REQUESTS\_CA\_BUNDLE=/etc/ssl/certs/ca-certificates.crt

ENV NODE\_EXTRA\_CA\_CERTS=/etc/ssl/certs/ca-certificates.crt

ENV SSL\_CERT\_FILE=/etc/ssl/certs/ca-certificates.crt

ENV PIP\_CERT=/etc/ssl/certs/ca-certificates.crt

*# ✅ Install AWS CLI (auto-detect architecture)*

RUN apt-get update && \

apt-get install -y --no-install-recommends curl unzip && \

ARCH=$(uname -m) && \

if [ "$ARCH" = "x86\_64" ]; then \

CLI\_URL="https://awscli.amazonaws.com/awscli-exe-linux-x86\_64.zip"; \

else \

CLI\_URL="https://awscli.amazonaws.com/awscli-exe-linux-aarch64.zip"; \

fi && \

curl "$CLI\_URL" -o "awscliv2.zip" && \

unzip -o awscliv2.zip && \

./aws/install && \

rm -rf awscliv2.zip aws && \

apt-get clean && rm -rf /var/lib/apt/lists/\*

*# ✅ Install git and configure safe directory*

RUN apt-get update && apt-get install -y git && \

git config --global --add safe.directory /workspaces/oxford-genai-capstone

*# ✅ Install Python dependencies (using trusted cert explicitly)*

COPY requirements.txt requirements.txt

RUN pip3 install --cert /etc/ssl/certs/ca-certificates.crt -r requirements.txt

*# ✅ Final cleanup*

RUN rm -rf /var/lib/apt/lists/\*

#### 8 make run-app

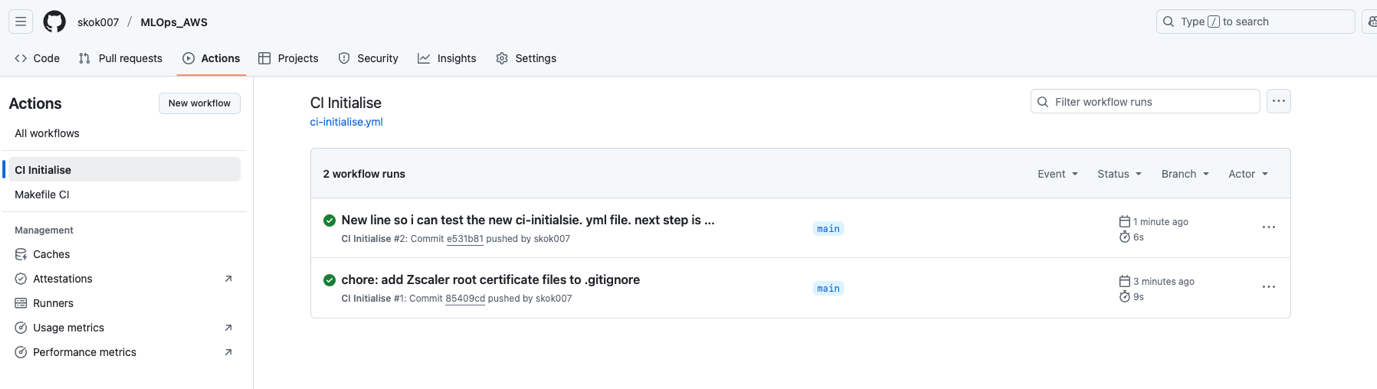
### 02 GitHub Actions

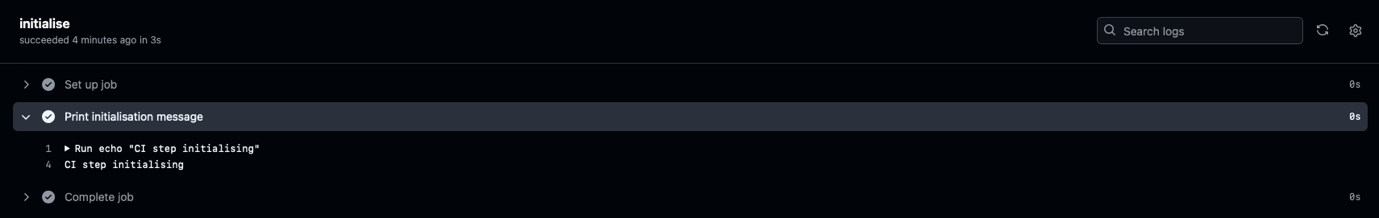
#### 1 Create a ci-initialise.yml and build a simple workflow that will print the message CI step initalising

I had to add several adjustments to makethe workflow work. I also added a test stage eventually but couldn’t make it work. See details below



#### 3 ,4 make small change and commit and push this change





#### Adapt the yml file so that it triggers a run of pytest

By trying to do this I realised a lot of the tests failed. And the reason for this is that there were some aspects of the workflow missing; including the installation of dependencies, and needed to setup secrets in Github to store all the necessary api keys and global variables.

So first I proceeded to setup secrets. I wanted to do this as securely as possible, so created a script that extracted the secrets from the .env files called set\_github\_secrets.sh in the root folder. Code here:

*#!/bin/bash*

*# Exit on any error*

set -e

*#!/bin/bash*

*# Exit on any error*

set -e

*# CONFIGURABLE VARIABLES*

*# Path to the .env file (adjust this if needed)*

ENV\_FILE\_PATH="rag-app/.env" *# <-- CHANGE this path if needed*

*# Check if gh CLI is installed*

if ! command -v gh &> /dev/null; then

echo "Error: GitHub CLI (gh) is not installed"

echo "Please install it with: brew install gh (macOS) or apt install gh (Ubuntu)"

exit 1

fi

*# Check if user is authenticated with GitHub*

if ! gh auth status &> /dev/null; then

echo "Error: Not authenticated with GitHub"

echo "Please run 'gh auth login' first"

exit 1

fi

*# Check if .env file exists*

if [ ! -f "$ENV\_FILE\_PATH" ]; then

echo "Error: .env file not found at $ENV\_FILE\_PATH"

exit 1

fi

*# Get the repository from git remote*

REPO=$(git config --get remote.origin.url | sed 's/.\*github.com[:/]//' | sed 's/\.git$//')

if [ -z "$REPO" ]; then

echo "Error: Could not determine repository from git remote"

echo "Please set the repository manually:"

read -p "Enter repository (format: owner/repo): " REPO

if [ -z "$REPO" ]; then

echo "Error: No repository provided"

exit 1

fi

fi

echo "Setting GitHub secrets for repository: $REPO"

*# Clean up the .env file for sourcing (remove spaces around '=')*

CLEANED\_ENV=$(mktemp)

sed 's/ \*= \*/=/' "$ENV\_FILE\_PATH" > "$CLEANED\_ENV"

*# Load secrets from cleaned .env*

echo "Loading secrets from $ENV\_FILE\_PATH (cleaned)..."

set -a

source "$CLEANED\_ENV"

set +a

rm "$CLEANED\_ENV"

*# Function to set a secret if it exists*

set\_secret\_from\_env() {

local secret\_name=*$1*

local secret\_value=${!secret\_name}

*# Remove leading and trailing double quotes if present*

secret\_value=$(echo "$secret\_value" | sed 's/^"\(.\*\)"$/\1/')

if [ ! -z "$secret\_value" ]; then

echo "Setting ${secret\_name}..."

printf "%s" "$secret\_value" | gh secret set "$secret\_name" -R "$REPO" -b -

if [ $? -ne 0 ]; then

echo "Error: Failed to set ${secret\_name}"

return 1

fi

echo "Successfully set ${secret\_name}"

return 0

else

echo "Warning: No value found for ${secret\_name}, skipping"

return 1

fi

}

*# List of secrets to set (add more here if needed)*

SECRETS\_LIST=(

"POSTGRES\_USER"

"POSTGRES\_PASSWORD"

"OPIK\_API\_KEY"

"OPIK\_WORKSPACE"

"OPIK\_PROJECT\_NAME"

"OPENAI\_API\_KEY"

"ARXIV\_API\_URL"

)

*# Set all secrets*

SECRETS\_SET=0

for secret\_name in "${SECRETS\_LIST[@]}"; do

set\_secret\_from\_env "$secret\_name" && SECRETS\_SET=$((SECRETS\_SET + 1))

done

if [ $SECRETS\_SET -eq 0 ]; then

echo "Error: No secrets were set."

exit 1

else

echo "Successfully set $SECRETS\_SET secrets!"

fi

*# Optionally, list the secrets just set*

echo "Listing current secrets in repository $REPO:"

gh secret list -R "$REPO"

The adjusted workflow has the following requirements:

I'll add detailed comments to the workflow and then provide a summary of the requirements.

1. **Trigger Conditions**:

* Runs on pushes to main and develop branches
* Runs on pull requests to main
* Can be manually triggered

1. **Environment Setup**:

* Sets up various environment variables for testing, including:
* Database configurations
* API endpoints
* Model parameters
* Opik API settings
* Poetry version (1.8.2)

1. **Main Job Steps**:

* Runs on Ubuntu latest
* Sets up a PostgreSQL database service using pgvector
* Checks out the code
* Sets up Python 3.12
* Caches Poetry and dependencies
* Installs Poetry package manager
* Installs project dependencies
* Initializes the database and runs tests with pytest
* Uploads test results and coverage reports as artifacts

1. **Disabled Features** (commented out):

* Linting tools installation and checks (flake8, black, isort)
* Docker image builds (disabled due to certificate access limitations)

The testing process found a multitude of errors in the code. Especially with the Opik setup. So I eventually commented out the tests for this stage. Will try to fix later

The final ci-initialise.yml including breaking it down into a sequence of jobs here:

*# This workflow runs automated tests whenever code is pushed to any branch*

name: CI Pipeline

*# Define when the workflow should run*

on:

push:

branches:

- 'main' *# Run on push to main branch only*

- 'develop' *# Run on push to develop branch*

pull\_request:

branches:

- 'main' *# Run on pull requests to main branch*

workflow\_dispatch: *# Manual trigger*

*# Define environment variables needed for the application*

env:

*# Basic application settings*

ENVIRONMENT: test

APP\_NAME: rag-app

DEBUG: "true"

*# Database connection settings (for testing only)*

POSTGRES\_HOST: localhost

POSTGRES\_DB: test\_db

POSTGRES\_USER: test\_user

POSTGRES\_PASSWORD: test\_password

POSTGRES\_PORT: 5432

*# API endpoints and paths*

ARXIV\_API\_URL: https://export.arxiv.org/api/query

DATA\_PATH: ./data

*# Model generation parameters*

TEMPERATURE: "0.7"

TOP\_P: "0.9"

MAX\_TOKENS: "1000"

*# Opik API settings (for AI model evaluation)*

OPIK\_API\_KEY: ${{ secrets.OPIK\_API\_KEY }}

OPIK\_WORKSPACE: ${{ secrets.OPIK\_WORKSPACE }}

OPIK\_ENVIRONMENT: test

OPIK\_PROJECT\_NAME: rag-app-test

*# OpenAI settings*

*#OPENAI\_MODEL: gpt-3.5-turbo*

*#OPENAI\_API\_KEY: ${{ secrets.OPENAI\_API\_KEY }}*

*# Poetry version to use*

POETRY\_VERSION: "1.8.2"

jobs:

ci:

name: CI Pipeline

runs-on: ubuntu-latest

services:

postgres:

image: ankane/pgvector:latest

env:

POSTGRES\_DB: test\_db

POSTGRES\_USER: test\_user

POSTGRES\_PASSWORD: test\_password

ports:

- 5432:5432

options: >-

--health-cmd pg\_isready

--health-interval 10s

--health-timeout 5s

--health-retries 5

steps:

*# Step 1: Check out the code*

- name: Checkout code

uses: actions/checkout@v4

*# Step 2: Set up Python environment*

- name: Set up Python

uses: actions/setup-python@v5

with:

python-version: '3.12'

cache: 'pip'

*# Step 3: Cache Poetry installation and dependencies*

- name: Cache Poetry and dependencies

uses: actions/cache@v3

with:

path: |

~/.local

~/.cache/pypoetry

./rag-app/.venv

key: poetry-${{ runner.os }}-${{ env.POETRY\_VERSION }}-${{ hashFiles('\*\*/poetry.lock') }}

restore-keys: |

poetry-${{ runner.os }}-${{ env.POETRY\_VERSION }}-

*# Step 4: Install Poetry package manager*

- name: Install Poetry

run: |

curl -sSL https://install.python-poetry.org | python3 - --version ${{ env.POETRY\_VERSION }}

poetry config virtualenvs.create true

poetry config virtualenvs.in-project true

*# Step 5: Install project dependencies*

- name: Install dependencies

working-directory: ./rag-app

run: |

echo "Installing dependencies..."

poetry install --no-interaction

echo "Installation complete!"

*# Step 6: Initialize database and run tests*

- name: Initialize database and run tests

working-directory: ./rag-app

run: |

# Wait for PostgreSQL to be ready

echo "Waiting for PostgreSQL to be ready..."

until poetry run python -c "import psycopg2; conn = psycopg2.connect(dbname='test\_db', user='test\_user', password='test\_password', host='localhost'); conn.close()" 2>/dev/null; do

echo "PostgreSQL is unavailable - sleeping"

sleep 1

done

echo "PostgreSQL is up - executing tests"

# Run tests with coverage

poetry run pytest tests/ -v --junitxml=test-results.xml --cov=server --cov-report=xml

*# Step 7: Upload test results and coverage as artifacts*

- name: Upload test results and coverage

uses: actions/upload-artifact@v4

if: always()

with:

name: test-results

path: |

rag-app/test-results.xml

rag-app/coverage.xml

*# Step 13: Install linting tools - TEMPORARILY DISABLED*

*# - name: Install linting tools*

*# working-directory: ./rag-app*

*# run: |*

*# echo "Installing linting tools..."*

*# poetry add --group dev flake8 black isort*

*# echo "Linting tools installation complete"*

*# Step 14: Run linting - TEMPORARILY DISABLED*

*# - name: Run linting*

*# working-directory: ./rag-app*

*# run: |*

*# poetry run flake8 server/src tests*

*# poetry run black --check server/src tests*

*# poetry run isort --check-only server/src tests*

*# Step 15: Build Docker images - TEMPORARILY DISABLED*

*# NOTE: Docker builds are disabled due to certificate access limitations.*

*# The PostgreSQL image build requires Zscaler certificates which cannot be*

*# shared in the public repository for security reasons.*

*# - name: Set up Docker Buildx*

*# uses: docker/setup-buildx-action@v3*

*# - name: Build PostgreSQL image*

*# working-directory: ./rag-app/deploy/docker/postgres*

*# run: |*

*# docker build -t rag-app-postgres:latest -f pgvector2.Dockerfile .*

*# - name: Build application image*

*# working-directory: ./rag-app*

*# run: |*

*# docker build -t rag-app:latest -f Dockerfile .*

### 03 – Database Setup

#### 1. Create a container for the database

I had to again address the issues of being behind Zscaler. So had to move to http vs. https in the dockerfile. Added this code:

*# Override repositories to use HTTP instead of HTTPS to skip SSL verification*

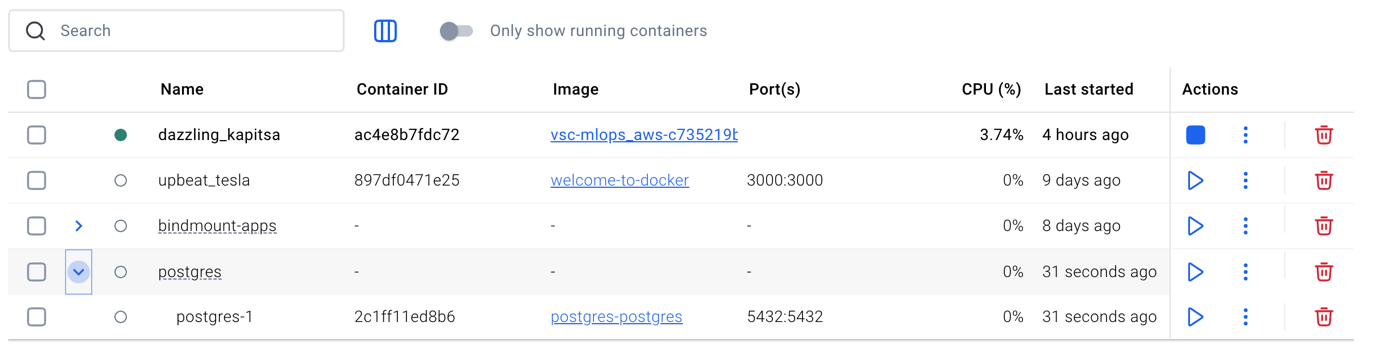
RUN echo "http://dl-cdn.alpinelinux.org/alpine/v3.21/main" > /etc/apk/repositories && \

echo "http://dl-cdn.alpinelinux.org/alpine/v3.21/community" >> /etc/apk/repositories && \

apk update && apk add --no-cache ca-certificates && update-ca-certificates

COPY ZscalerRootCertificate-2048-SHA256.pem /usr/local/share/ca-certificates/zscaler.crt

After this it worked. See here docker dashboard:



#### 2. Then I ran the test with this command with all the right credentials:

PGPASSWORD=mypassword psql -h localhost -p 5432 -U myuser -d mydb

mydb=# \dt

List of relations

Schema | Name | Type | Owner

--------+--------+-------+--------

public | papers | table | myuser

(1 row)

mydb=# \du

List of roles

Role name | Attributes

-----------+------------------------------------------------------------

myuser | Superuser, Create role, Create DB, Replication, Bypass RLS

mydb=#

#### 3. adding comments to init\_pgvector.sql

*/\* Create the vector extension \*/*

CREATE EXTENSION IF NOT EXISTS vector;

*/\**

*\* Vector dimension explanation:*

*\* The dimension 384 is commonly used with models like:*

*\* - all-MiniLM-L6-v2 (384 dimensions)*

*\* - all-mpnet-base-v2 (384 dimensions)*

*\**

*\* Different models have different embedding dimensions:*

*\* - BERT base: 768 dimensions*

*\* - BERT large: 1024 dimensions*

*\* - OpenAI text-embedding-ada-002: 1536 dimensions*

*\* - OpenAI text-embedding-3-small: 1536 dimensions*

*\* - OpenAI text-embedding-3-large: 3072 dimensions*

*\**

*\* If using a different model, adjust the vector dimension accordingly.*

*\* For example, if using OpenAI's text-embedding-ada-002, change to:*

*\* embedding vector(1536)*

*\*/*

CREATE TABLE papers (

id SERIAL PRIMARY KEY,

title TEXT NOT NULL,

summary TEXT NOT NULL,

chunk TEXT NOT NULL,

embedding vector(384)

);

### 4. Data ingestion and Embedding

##### 1- Where in the code is the search against arXiv API?

Here:

Here: if \_\_name\_\_ == "\_\_main\_\_":

*# papers = fetch\_papers(query="ti:perovskite", max\_results=10)*

papers = fetch\_papers\_paginated(

*query*="ti:perovskite", *max\_results*=20, *results\_per\_page*=5, *wait\_time*=5

)

##### 2 Try different variants of your query string using the API quickstart examples

These are the query variants I used

*# ========================*

*# Query Variant 1: Basic title-only search*

*# ========================*

*# papers = fetch\_papers\_paginated(*

*# query="ti:perovskite", max\_results=20, results\_per\_page=5, wait\_time=5*

*# )*

*# ========================*

*# Query Variant 2: Title or abstract contains "perovskite"*

*# ========================*

*# papers = fetch\_papers\_paginated(*

*# query="ti:perovskite OR abs:perovskite", max\_results=20, results\_per\_page=5, wait\_time=5*

*# )*

*# ========================*

*# Query Variant 3: Filter to category - materials science in condensed matter*

*# ========================*

*# papers = fetch\_papers\_paginated(*

*# query="(ti:perovskite OR abs:perovskite) AND cat:cond-mat.mtrl-sci", max\_results=20, results\_per\_page=5, wait\_time=5*

*# )*

*# ========================*

*# Query Variant 4: Keyword combo (solar + perovskite in title)*

*# ========================*

*# papers = fetch\_papers\_paginated(*

*# query="ti:perovskite AND ti:solar", max\_results=20, results\_per\_page=5, wait\_time=5*

*# )*

*# ========================*

*# Query Variant 5: Specific author (Michael Grätzel as example – which comes back empty)*

*# ========================*

*# papers = fetch\_papers\_paginated(*

*# query="ti:perovskite AND au:Grätzel", max\_results=20, results\_per\_page=5, wait\_time=5*

*# )*

##### 3 run ingestion results

These are the results of running ingestion.

poetry run python ./server/src/ingestion/pipeline.py

Reading JSON files from ./papers-downloads...

Succesfully processed 80 papers.

Successfully inserted 80 rows into the papers table.

Completed ingestion into database mydb

### 05 RAG Workflows

#### 1 Isolate where the retrieval process for the top\_k numer is enforced on retrieval from the database.

##### What does top\_k mean in this context (i.e. of RAG)

Here:

*# SQL query to find the top\_k chunks using cosine similarity*

*query* = """

SELECT id, title, chunk, embedding <=> %s::vector AS similarity

FROM papers

ORDER BY similarity ASC

LIMIT %s;

"""

##### What will happen if I increase or decrease top\_k?

In RAG (Retrieval-Augmented Generation), top\_k refers to the number of most relevant documents that the system should retrieve based on similarity to the input query.

Think of it as saying: “Give me the k most similar pieces of context (documents/chunks) from the knowledge base.”

These retrieved documents are typically passed into a language model to help answer the query more accurately.

Increasing top\_k:

* You retrieve more context chunks from the database.
* Can improve accuracy by giving the LLM more information — but too many might introduce noise, redundancy, or exceed input limits (token cap).
* May increase latency slightly (more data to fetch and process).

Decreasing top\_k:

* Fewer chunks retrieved.
* Can lead to faster retrieval and response time.
* Risk of missing relevant information if the top few aren’t sufficient.

##### We have used an algorithm inside the database query to retrieve the top\_k similar matches, what is this algorithm and what does it do/how does it work?

We used it here:

embedding <=> %s::vector

That <=> operator is provided by pgvector, and it computes cosine distance between two vectors:

* One vector is from your query (user input),
* The other is the document embedding stored in your papers table.

Despite looking like a custom operator, it’s part of pgvector. What it means in plain English is : “Compute the cosine distance between the embedding column (in the DB) and the query embedding vector.”

Cosine distance measures how similar two vectors are in direction, regardless of their magnitude. It’s based on the cosine of the angle between them.

* If the angle is 0° → Cosine = 1 → Distance = 0 (perfectly similar)
* If the angle is 90° → Cosine = 0 → Distance = 1 (completely dissimilar)

Therefore, pgvector’s <=> operator:

* Calculates cosine distance between embeddings,
* Sorts results in ascending order (lowest distance = highest similarity),
* Returns the top\_k most relevant document chunks.

##### Given the answer to c, why is the query to the Postgres database asking for the results in **ascending** order?

The query sorts results by ascending cosine distance because:

* Smaller distance = higher similarity
* You want the most similar documents first

So, ordering by ASC (ascending) ensures that the top\_k most relevant results are returned.

We are using this in the query:

*ORDER BY similarity ASC → sorts from lowest to highest distance*

Why is this important?

Cosine distance (as used in pgvector) ranges from 0 (most similar) to 2 (most dissimilar). Ordering by ascending means:

* The first row = closest match to the query
* Next few rows = next closest, etc.
* After sorting, you take the first k rows using LIMIT

If you used DESC instead (descending), you’d get the least relevant documents first — which defeats the purpose of retrieval.

So ASC is essential for returning the most useful chunks to the LLM for generation.

#### 2. There is an error control flow in the retrieve\_top\_k\_chunks\_endpoint that deals with a failure to retrieve chunks (i.e context). Under what scenarios could this happen?

* **No Relevant Documents in the DB:** The query’s embedding might not be close enough (in cosine space) to any document embeddings. For example, if the query is too vague, irrelevant, or outside domain, similarity scores may be poor — and if your DB contains only a few documents, it might return nothing.
* **Database Table is Empty:** The papers table might not have any rows (e.g., ingestion hasn’t run or failed). So even a good query can’t return anything.
* **All Document Embeddings are Null or Invalid:** If documents exist but the embedding column is unpopulated or corrupted, pgvector can’t compute similarity.
* **Query Embedding Failed:** If the SentenceTransformer model fails to generate a query embedding (e.g., bad input, or model not loaded correctly), then the search won’t work, and might quietly return an empty result instead of throwing.
* **Unexpected errors:** These could also occurred and that is where we have the except (500 error).
  + Invalid or missing DB credentials (POSTGRES\_USER, etc.)
  + Database is unreachable (network, port, auth issues)
  + SQL query syntax error
  + psycopg2 throws a connection or execution error
  + Model loading issue (though handled earlier in your code, could break on startup)

#### 3. Can you outline the process flow that happens when a user hits the “generate” endpoint with a query?

1. **User Sends a Request to the /generate Endpoint**:
   * The FastAPI application defines a route for the /generate endpoint, which is handled by a function in the controllers/generation.py module.
2. **Controller Receives the Request**:
   * The controller function extracts the user’s query from the request parameters.
   * It then invokes the generate\_response function from the services/generation\_service.py module, passing along the query.
3. **Service Layer Handles Business Logic**:
   * Within generation\_service.py, the generate\_response function orchestrates the core logic:
     + **Retrieval**: It calls the retrieve\_top\_k\_chunks function from services/retrieval\_service.py to fetch the most relevant document chunks related to the query.
     + **Generation**: It then uses these retrieved chunks as context to generate a response, likely by interfacing with a language model.
4. **Retrieval Service Fetches Relevant Chunks**:
   * The retrieve\_top\_k\_chunks function:
     + Generates an embedding for the user’s query using a sentence transformer model.
     + Queries the PostgreSQL database (with pgvector extension) to find the top K document chunks that are most similar to the query embedding.
     + Returns these chunks to the generate\_response function.
5. **Generation Service Produces the Final Response**:
   * Using the retrieved chunks as context, the generate\_response function:
     + Constructs a prompt or input suitable for the language model.
     + Invokes the model to generate a coherent and relevant response to the user’s query.
     + Returns this generated response back to the controller.
6. **Controller Sends Back the Response**:
   * The controller receives the generated response from the service layer.
   * It then formats this response appropriately (e.g., as JSON) and sends it back to the user as the HTTP response.

**Understanding the Structure**:

* **Controllers**: Located in server/src/controllers/, controllers are responsible for handling incoming HTTP requests, extracting necessary parameters, and invoking appropriate service functions. They act as the interface between the client and the server’s business logic.
* **Services**: Found in server/src/services/, services contain the core business logic of the application. They perform operations such as data retrieval, processing, and interfacing with machine learning models.
* **Main Application**: The server/src/main.py file initializes the FastAPI application, includes the routers from the controllers, and sets up any necessary middleware or configurations.

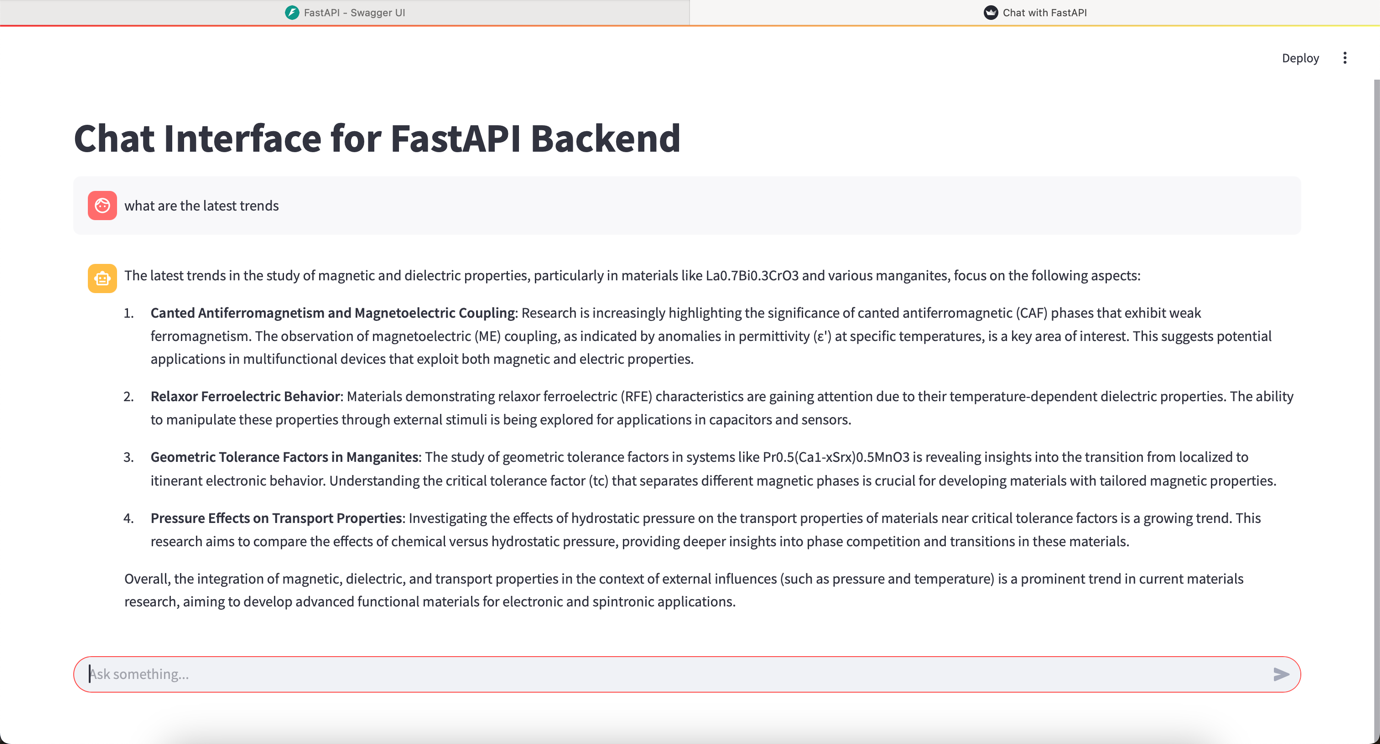
This separation of concerns ensures a modular and maintainable codebase, where each component has a clear responsibility

#### 4. Run the app again with the command:

make run-app

had to make several adjustments to the endpoints for them to work

This is the app running:



This is the client running

Fastapi here:

