**Lab Exercise - Deploy a LLM in Docker**

**Objective**

Use a lightweight LLM (e.g., DistilGPT2) with FastAPI to expose a REST endpoint for text generation and deploy the app in a Docker container.

**Project Structure**

A computer code with black text

AI-generated content may be incorrect.

**Step 1: app/main.py – FastAPI Application**

from fastapi import FastAPI

from pydantic import BaseModel

from transformers import pipeline, set\_seed

app = FastAPI()

generator = pipeline("text-generation", model="distilgpt2")

set\_seed(42)

class Prompt(BaseModel):

prompt: str

@app.post("/generate/")

def generate\_text(data: Prompt):

result = generator(data.prompt, max\_length=50, num\_return\_sequences=1)

return {"generated\_text": result[0]["generated\_text"]}

**Step 2: requirements.txt**

fastapi

uvicorn

transformers

torch

**Step 3: Dockerfile**

FROM python:3.10-slim

WORKDIR /app

COPY requirements.txt .

RUN pip install --no-cache-dir -r requirements.txt

COPY app/ ./app

EXPOSE 8000

CMD ["uvicorn", "app.main:app", "--host", "0.0.0.0", "--port", "8000"]

**Step 4: Build Docker Image**

docker build -t simple-llm .

**Step 5: Run Docker Container**

docker run -p 8000:8000 simple-llm

**Step 6: Test the API**

curl -X POST "http://3.108.228.179:8000/generate/" -H "Content-Type: application/json" -d '{"prompt": "Once upon a time"}'

curl -X POST "localhost:8000/generate/" -H "Content-Type: application/json" -d '{"prompt": "Once upon a time"}'

**Expected Output**

{

"generated\_text": "Once upon a time there was a dragon that lived in a cave and..."

}