

Wednesday, 16 July 2025

Hi, my name is Deep, in this pdf I have written notes on how I created this project step by step.

Medical chatbot project

- The gale encyclopedia of medicine second edition volume one
- Book se knowledge base that means here we are using pine core vector database
- Pine cone DB will be knowledge database
- Cant use local vector as dataset is huge
- Hence using cloud base vector DB
- From medical book extract all docs
- LLM has I/P specific length ..try to see I/p length
- After getting chunks
- Embedding models
- Then diff diff vector embedding
- From these build sementic index
- From this to knowledge base (medicine vector db)
- This was all backend
- Now frontend
- User asks query then this to query embedding ..it will go to knowledge base
- Then we will use LLM
- We will use open ai model as its already hosted and we can access this through api requests
- But if we use any open source LLM then we have to download this in our local machine and local machine doesn't have good instance/configuration gpu cpu then we cant execute it and time will be a lot
- LLM oops platforms we will be using

- Open source LLM is using as of large size and cloud platform cost increases hence open ai model
- Based on query , knowledge base the LLM will process it and give answer to the user
- This was the frontend
- Now technology wise : open AI LLM , lang chain generative ai frame work, pinecone for DB , flask for user interface, git hub , was cloud simple deployment
- Make repo on git medical ai chatbot
- Copy it's link and clone this repo and set up it in local folder
- Now open vs code in the same folder
- Create a virtual environment in this
- Installed conda 25.5.1
- Then conda create -n medibot python=3.10 -y
- Now environment is created so now activating it
- Didn't accepted the terms and condition so accepting them via terminal
- (medibot) deeplatiyan@deeps-MacBook-Air-2 medical-ai-chatbot %
- Creating a txt file requirement.txt : sentence-transformers it will use hugging face platform to download open source model and with the help of that we will generate vector embedding (==2.2.2)

langchain

flask : for user interface

pypdf

python-dotenv

pinecone[grpc]


langchain-pinecone

langchain_community

langchain_openai

langchain_experimental

- Downloaded all these files and installed everything successfully

- In creating an end to end project the first thing is to create the project folder structure
- Creating template.py file : it will have a constructor file `__init__` Vala this is called modular approach then it will considered as local package run this file and it automatically created files n folder for me
- Now if we want to add another file just si
- -e . Run it and now project has been setup in local system
- Now pushed this local folder structure to git
- First we will write our project in Jupyter notebook then modular coding
- Selected python medibot kernel then installed some packages and tested if Jupyter notebook is working or not
- `%pwd` to check our path
- `os.chdir("../")` one folder back
- To load pdf we need `pypdfloader` and it's present inside directory so directory loader and to perform chunking operation `recursivecharacter textsplitter`
- Now extracting data from pdf file
- Since book has 700 pages so it will take time
- Now from the back end part we have extracted the data now chunking operations and to do that we will use `recursivecharacter textsplitter`
- Now embedding model to perform vector embedding
- Downloading one model from hugging face all minilm l6 v2
- Before it let's import a package
- 384 dimension whenever using pine core vector db always remember this ki what dimensional vector it's returning
- Now faced some problems while doing above things downloaded some packages
- 3 specific versions install karne ki koshish ki, lekin unmein dependency conflict aa gaya hai:
-
-  Problem:
- - `transformers==4.35.0` needs `huggingface-hub>=0.16.4`

- - huggingface-hub==0.14.1 is too old for it
 - But sentence-transformers==2.2.2 supports it
- Yo finally resolved my error ..its a relief
- Created a pinecone account copied api and now database and creating index by aws cloud provider
- Now let's create this by py code not manually by name medical-ai-chatbot
- Now chunks embedded in pinecore vector database
- Now the vector is stored in pinecore
- Here we can also visualise our vector which couldn't be possible in local system
- Now we will use these index records as our knowledge base connecting my LLM with this knowledge base and we will be asking the query to the semantic search operation
- Now loading existing index
- Now I have tested weather it's working or not
- Now we have to integrate our LLM
- Let's initialise our model by open ai model
- Faced a new problem in open ai api key
- Now using hugging face api instead of open ai api keys as it's free
- Could setup hugging api casuse of llm error
- successfully run your local LLM (FLAN-T5-Small) and it's generating answers properly
- Deleted mediabot as doing modular coding
- Dowloaded html css free template
- Created template folder inside it is chat.html
- Then a static folder inside it is style.css
- Now imported all files in app.py
- After coding in app.py i simply run this py file in my terminal n then in my localhost:8080 and took some screenshots
- Thankkkuuuuuuuuuu...yoooooooo