Lab 16. Getting started with LocalChat

(To be Executed on your own Machine)

You get to decide the setup for these 3 main components:

1. **LLM**: the large language model provider used for inference. It can be local, or remote, or even OpenAI.
2. **Embeddings**: the embeddings provider used to encode the input, the documents and the users' queries. Same as the LLM, it can be local, or remote, or even OpenAI.
3. **Vector store**: the store used to index and retrieve the documents.

Sequence 1. Installing Ollama

a) Installing on Linux

i. Login to your linux machine as root user and launch the terminal.
ii. Run the following command from terminal

```
curl -fsSL https://ollama.com/install.sh | sh
```

b) Installing on Windows

i. Download the following installer for windows

```
https://ollama.com/download/OllamaSetup.exe
```

ii. Run the Installer and follow the instructions

c) Installing on MacBook (M1/M2/M3 Chipset)

i. Download the Installer

```
https://ollama.com/download/Ollama-darwin.zip
```

ii. Extract the Installer and Run.

Sequence 2. Installing LocalChat

1. Create a Python Environment with Python version 3.11 with pyenv or conda:

   a. **Using** `pyenv`

   - Check the default version set by ***pyenv***. The * indicates that the system Python version is active currently.

     ```
     $ pyenv versions

       * system (set by /home/opc/.pyenv/version)
         3.11.9
     ```

   - If you see "**system**", this means that, by default, you are still using your system Python:

     ```
     $ python -V
     ```

- Change the default global version to python 3.11.9 with:

```
$ pyenv global 3.11.9

$ pyenv versions

      system
    * 3.11.9 (set by /home/opc/.pyenv/version)

$ python -V

  Python 3.11.9
```

b. **With** conda

- Check default Python Version (No Active Environment):

```
$ Python -V

  Python 3.12.6
```

- Activate base environment and check Python Version:

```
info@sangwan ~/labs$ conda activate base

(base) info@sangwan ~/labs$ Python -V

Python 3.12.8
```

- Create an environment with Python 3.11 version:

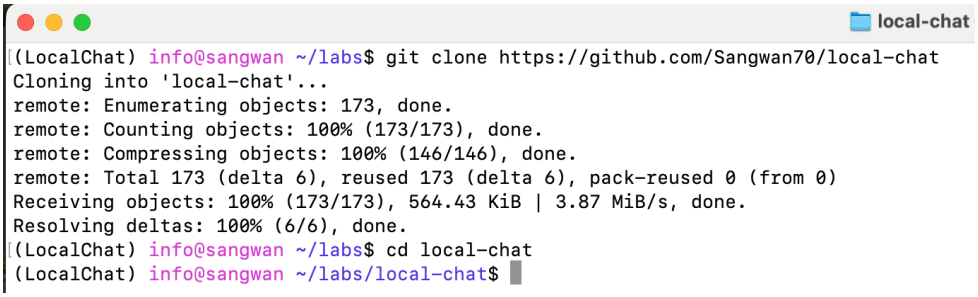
```
$ conda create -n LocalChat Python=3.11.9
```

- Activate New environment and check Python Version:

```
$ conda activate LocalChat

(LocalChat) info@sangwan ~/labs$ python -V

Python 3.11.9
```

2. Clone the LocalChat Repository and change to the downloaded directory:

```
git clone https://github.com/Sangwan70/local-chat

cd local-chat
```



3. Install Poetry for dependency management.

   ```
   pip install -qU poetry
   ```

4. Setup for Windows (Skip to step 5 for Linux/Mac)

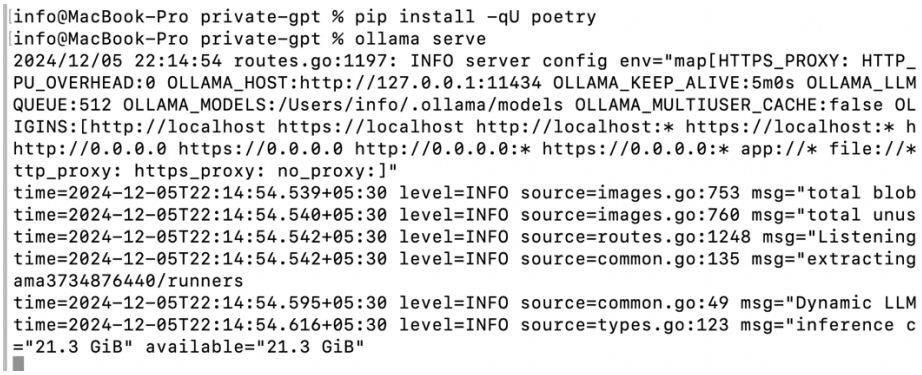
   ```
   set PGPT_PROFILES=ollama make run
   ```

5. Open another Tab of your terminal and **Start Ollama to** start a local inference server, serving both the LLM and the Embeddings:

   ```
   ollama serve
   ```

   If you get the following error, that would mean ollama is already running.

   Error: listen tcp 127.0.0.1:11434: bind: address already in use

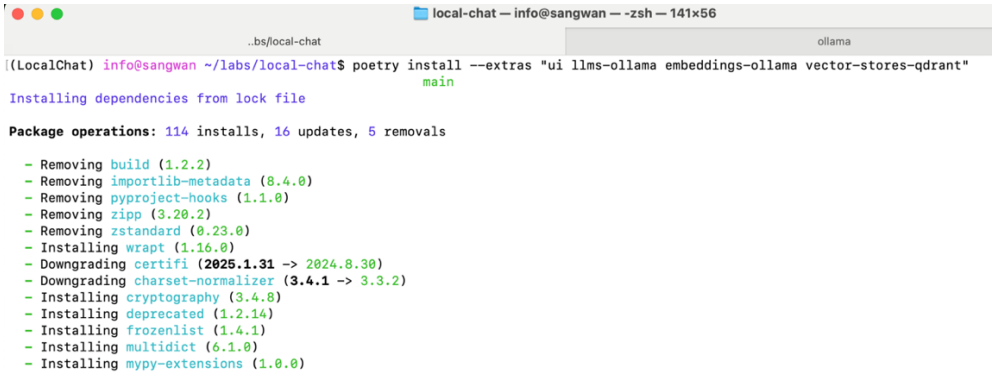


6. **The default `settings-ollama.yaml` is configured to user llama3.2 LLM (~2GB) and nomic-embed-text Embeddings (~275MB)**

   By default, PGPT will automatically pull models as needed.

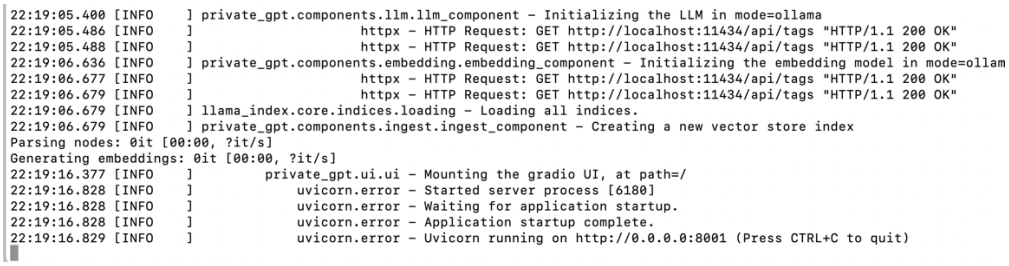
7. Back to previous Tab of your terminal, install LocalChat with:

   ```
   poetry install --extras "ui llms-ollama embeddings-ollama vector-stores-qdrant"
   ```

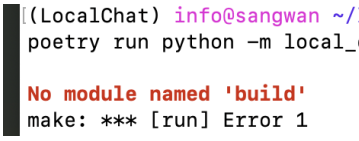


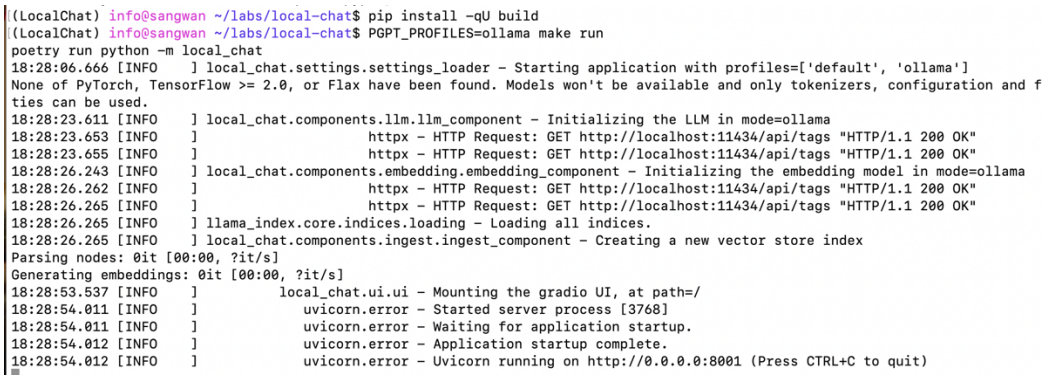
8. Now run LocalChat. (Make sure you have a working Ollama running locally.)

   `PGPT_PROFILES=ollama make run`



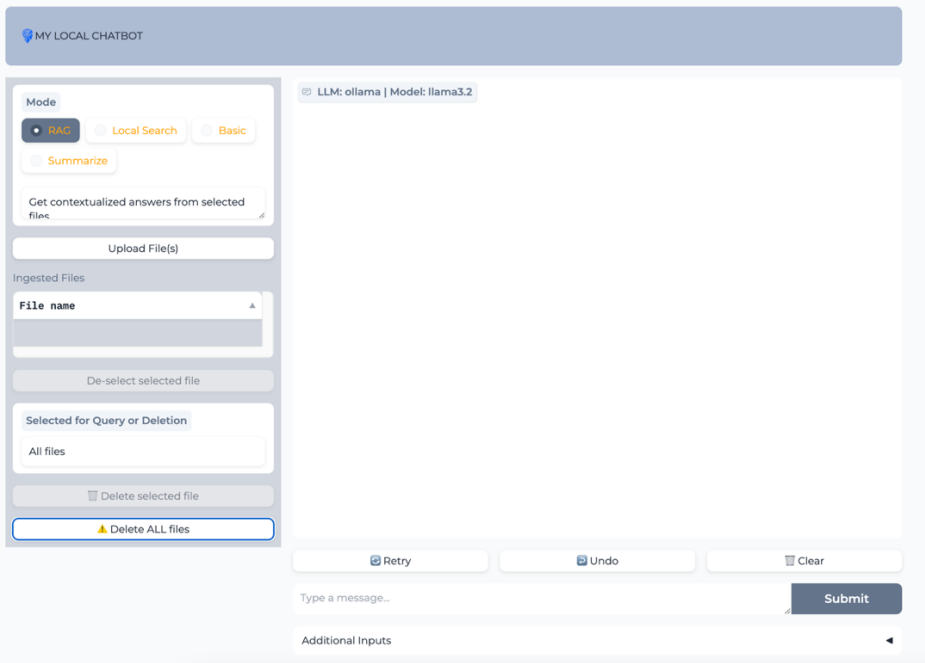
   If you get any error related to `Module Not found.` Install the module with pip. For example "`pip install -qU build`" and run the above command again.

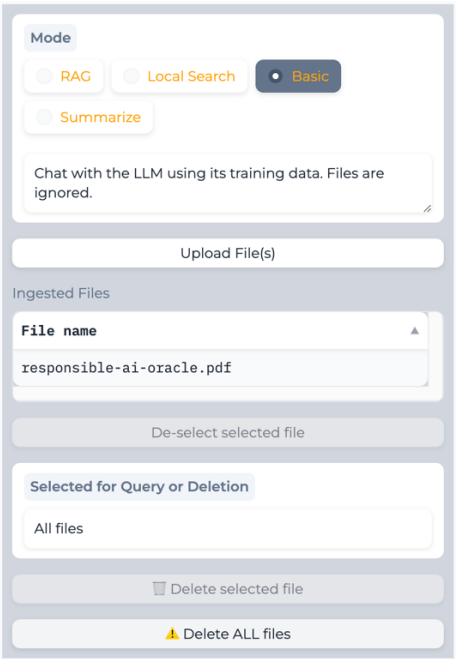




9. LocalChat will use `settings-ollama.yaml` settings file, which is already configured to use Ollama LLM and Embeddings, and Qdrant.

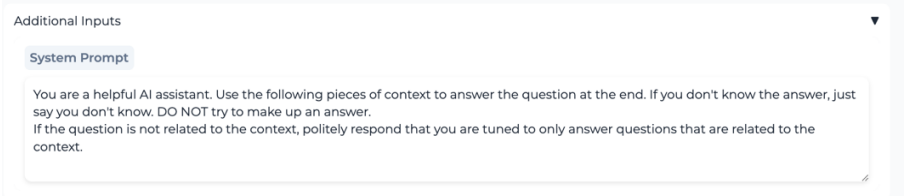
   Launch the browser and launch the UI (available at http://localhost:8001)



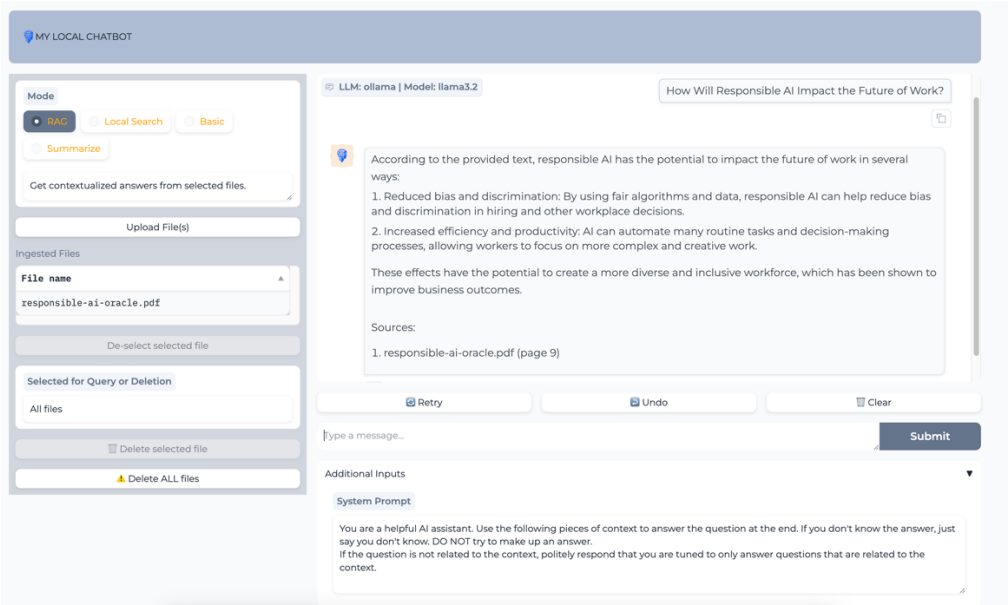


10. Select RAG and set the System Prompt to:

"*You are a helpful AI assistant. Use the following pieces of context to answer the question at the end. If you don't know the answer, just say you don't know. DO NOT try to make up an answer. If the question is not related to the context, politely respond that you are tuned to only answer questions that are related to the context.*"



11. Ask a Question "*How Will Responsible AI Impact the Future of Work?*"



12. Try asking a question not in the context. For example ask about yourself (I asked Tell me something about Ram N Sangwan)