First look at the problem :

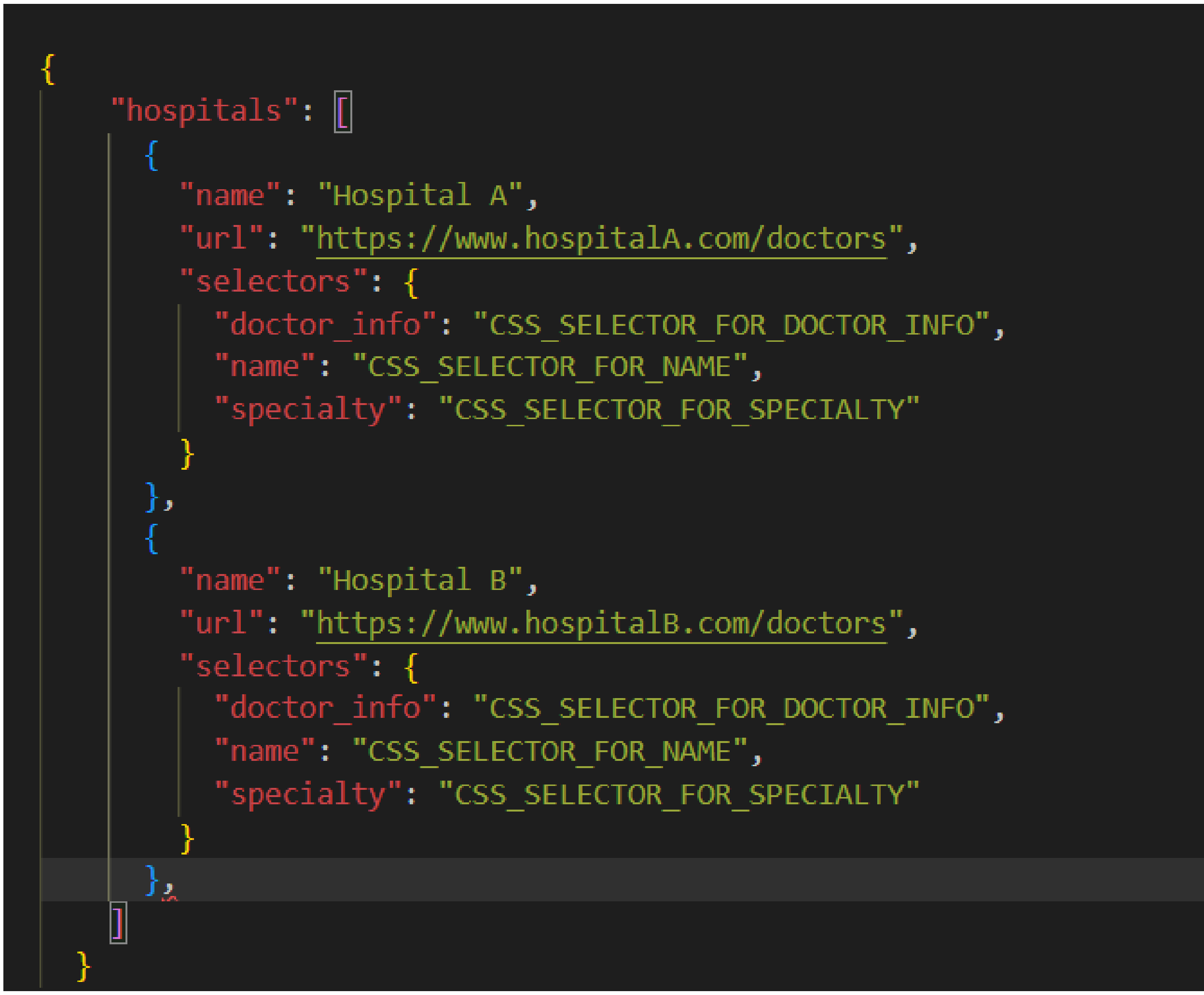
When I looked at the websites of the top few hospitals I found out that each of them have a different DOM structure and some were dynamic. So it is quite difficult to scrape all 50 websites with a single script.

Approaches to solve the problem :

I thought of two type of solutions

- Modular Approach
    - Build a config file with json containing rules of how to scrape the websites and iterate the json for hospitals.
      That way we can build a scalable solution and can scrape as many hospitals as we want by adding a json config and few changes to script if needed.

  - Example ,



```
{
    "hospitals": [
      {
        "name": "Hospital A",
        "url": "https://www.hospitalA.com/doctors",
        "selectors": {
          "doctor_info": "CSS_SELECTOR_FOR_DOCTOR_INFO",
          "name": "CSS_SELECTOR_FOR_NAME",
          "specialty": "CSS_SELECTOR_FOR_SPECIALTY"
        }
      },
      {
        "name": "Hospital B",
        "url": "https://www.hospitalB.com/doctors",
        "selectors": {
          "doctor_info": "CSS_SELECTOR_FOR_DOCTOR_INFO",
          "name": "CSS_SELECTOR_FOR_NAME",
          "specialty": "CSS_SELECTOR_FOR_SPECIALTY"
        }
      },
    ]
  }
```

This solution will take time to implement and as we have to analyze and write the selectors for each hospital.

It will be difficult to debug as some website have dynamic structures and we are scraping 50 at once.

- Scattered Solution
  In this solution I want to go to each website and write a script for

it and scrape the data. Do it for all 50 hospitals.
This approach is fast but its not scalable if we want to accomplish larger goals.

-> I chose the Scattered Solution.

This was scraping was fairly easy. I just looked at the DOM and scraped the data with relevant tags and selectors.
Some websites have the information within lists and some within div tags, some have “next” button to navigate while some have “load more”. Since we are writing custom scripts it was easy to handle.

Data preparation
-> I saved the data as csv file in each folder.
-> Found out 4 common problems

1. White spaces/Empty rows
    a. Occurred due to script not finding relevant tags/data points
2. Missing values.
3. Removing Duplicates.
4. Scraping irreverent info
    a. This happened because within tag like doctor name there was information like location enclosed within. I had to deal with it.

-> Then I got both csv to a common, clean structure I merged them.

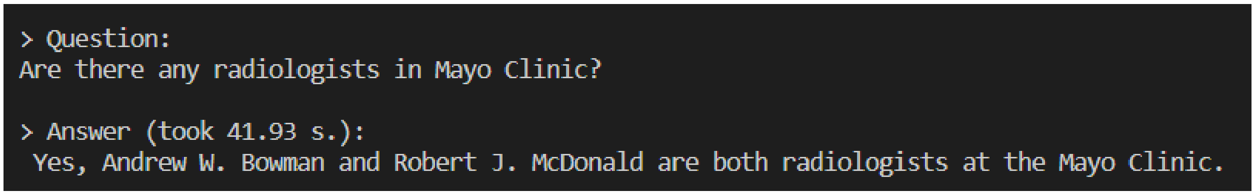
Model Training :

I followed the docs to setup the private gpt model.

Then In fed the data file to the model to train by running the ingest.py file.

Interaction :

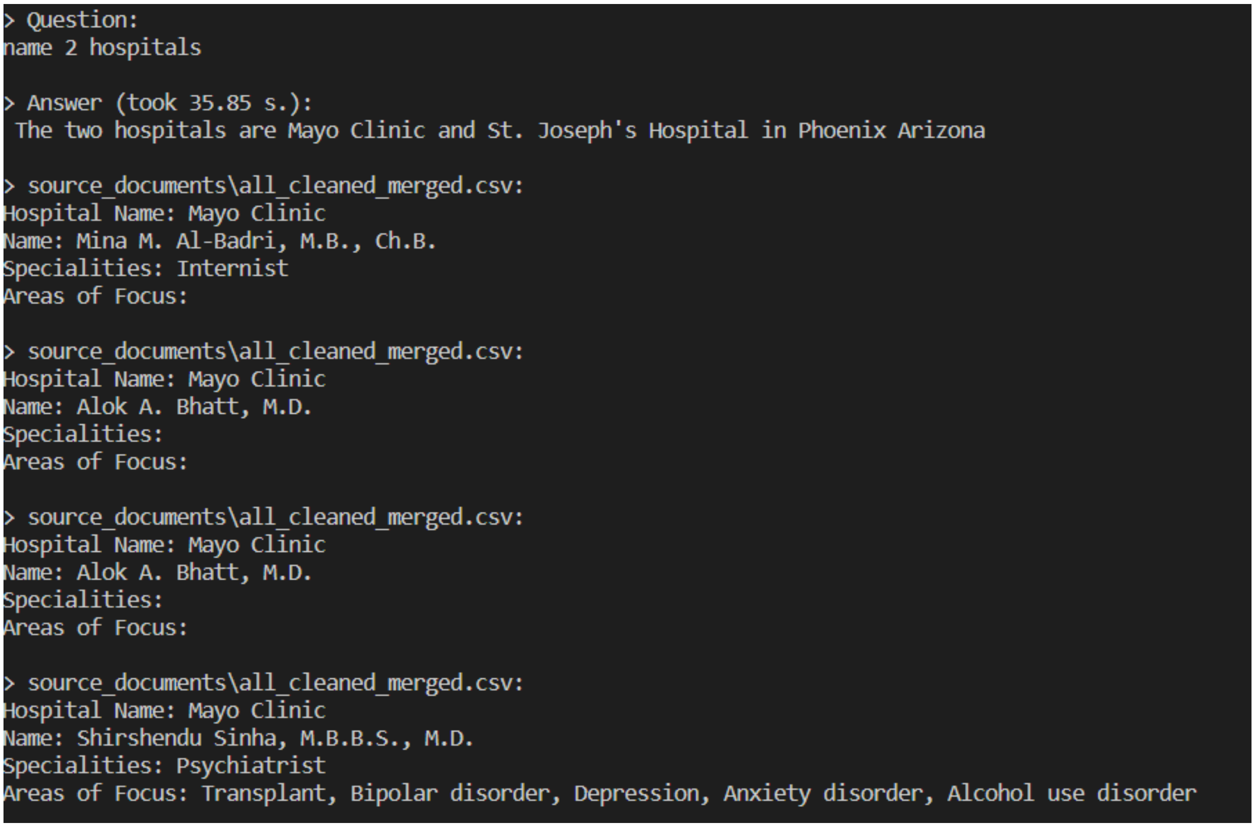
Model was able to answer some simple questions like.



```
> Question:
Are there any radiologists in Mayo Clinic?

> Answer (took 41.93 s.):
 Yes, Andrew W. Bowman and Robert J. McDonald are both radiologists at the Mayo Clinic.
```

Sometimes it was “Hallucinating.”

This is a term used when model makes up information



```
> Question:
name 2 hospitals

> Answer (took 35.85 s.):
 The two hospitals are Mayo Clinic and St. Joseph's Hospital in Phoenix Arizona

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Mina M. Al-Badri, M.B., Ch.B.
Specialities: Internist
Areas of Focus:

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Alok A. Bhatt, M.D.
Specialities:
Areas of Focus:

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Alok A. Bhatt, M.D.
Specialities:
Areas of Focus:

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Shirshendu Sinha, M.B.B.S., M.D.
Specialities: Psychiatrist
Areas of Focus: Transplant, Bipolar disorder, Depression, Anxiety disorder, Alcohol use disorder
```

As you can see St. Joseph's Hospital doesn't exists in training data and there are no sources for it.
Its made up.

-> How to counter Hallucination?
1. Improving training data
2. Fine tuning with prompt engineering

Since the data was limited, i went with second approach.

-> Prompt engineering

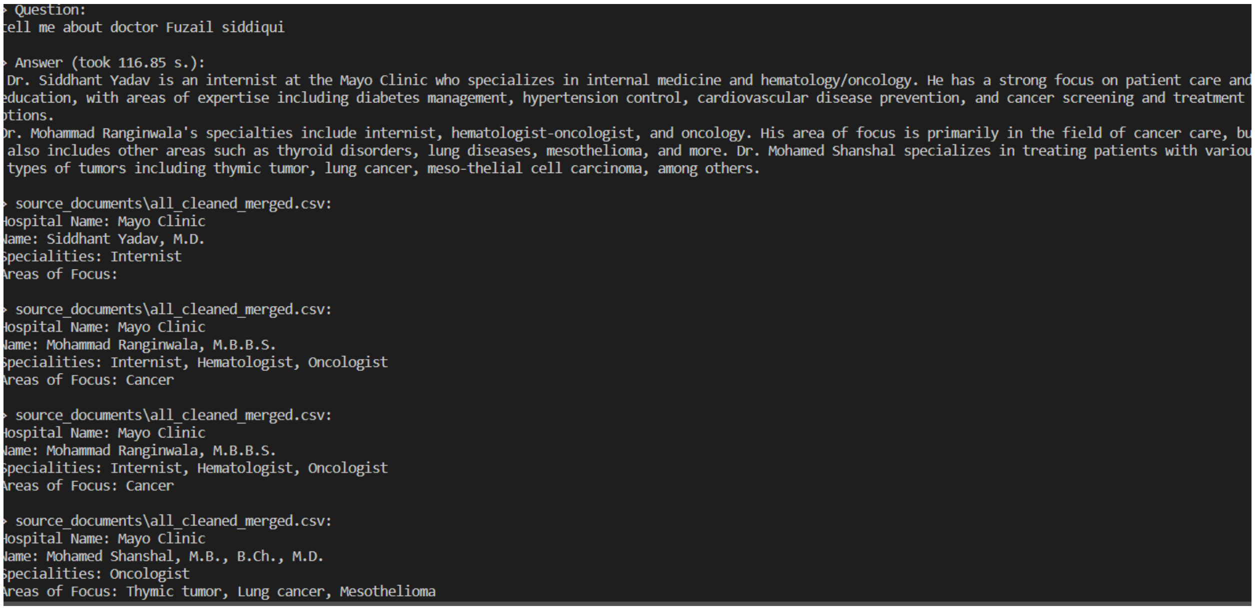
. I tried some prompts to make the model understand the data better.

Example prompt

“In the dataset, the 'Hospital Name' column contains the names of different hospitals, with each row representing a specific hospital. The 'Name' column contains the names of doctors, while the subsequent columns contain hospital-related information. I request assistance based solely on the available data. If the data doesn't contain specific information, please respond with 'No information”

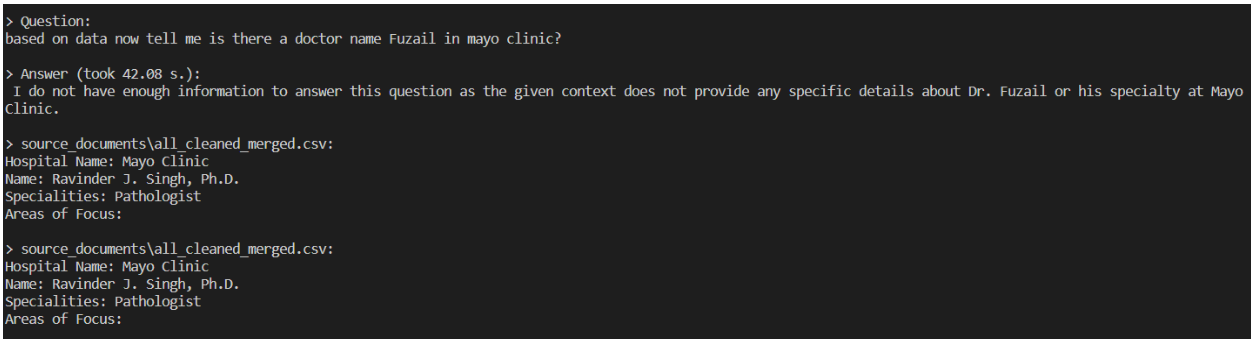
Another Example

Without prompt engineering. Hallucinating.



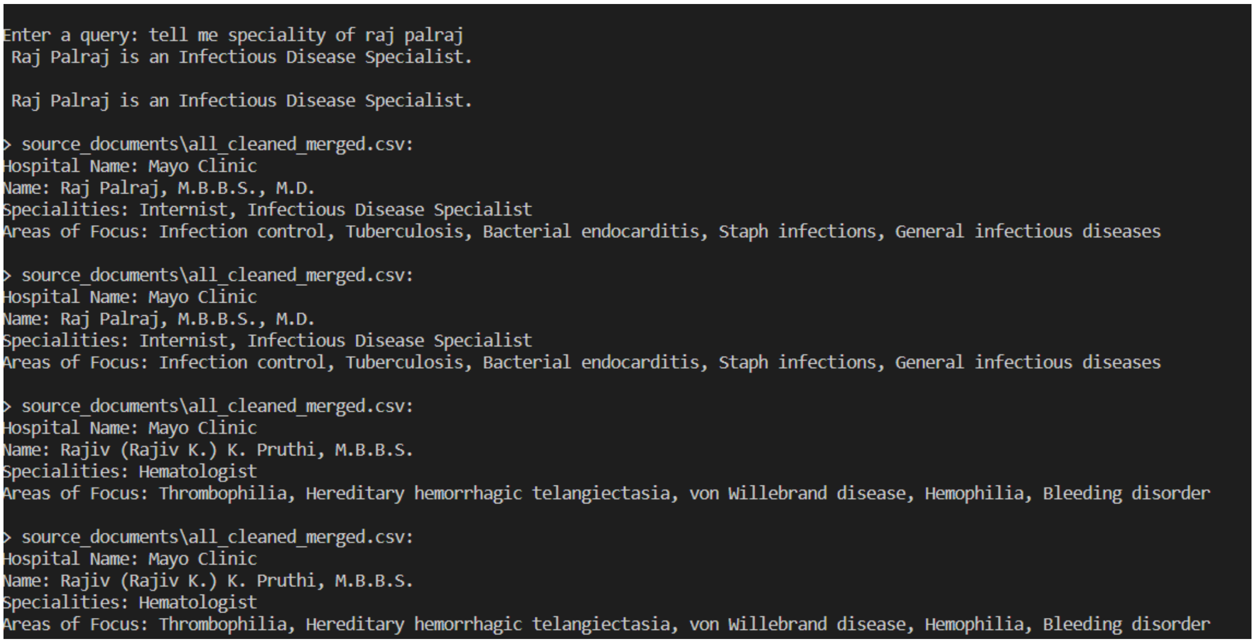
```
> Question:
tell me about doctor Fuzail siddiqui

> Answer (took 116.85 s.):
 Dr. Siddhant Yadav is an internist at the Mayo Clinic who specializes in internal medicine and hematology/oncology. He has a strong focus on patient care and
education, with areas of expertise including diabetes management, hypertension control, cardiovascular disease prevention, and cancer screening and treatment
options.
Dr. Mohammad Ranginwala's specialties include internist, hematologist-oncologist, and oncology. His area of focus is primarily in the field of cancer care, bu
 also includes other areas such as thyroid disorders, lung diseases, mesothelioma, and more. Dr. Mohamed Shanshal specializes in treating patients with variou
 types of tumors including thymic tumor, lung cancer, meso-thelial cell carcinoma, among others.

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Siddhant Yadav, M.D.
Specialities: Internist
Areas of Focus:

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Mohammad Ranginwala, M.B.B.S.
Specialities: Internist, Hematologist, Oncologist
Areas of Focus: Cancer

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Mohammad Ranginwala, M.B.B.S.
Specialities: Internist, Hematologist, Oncologist
Areas of Focus: Cancer

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Mohamed Shanshal, M.B., B.Ch., M.D.
Specialities: Oncologist
Areas of Focus: Thymic tumor, Lung cancer, Mesothelioma
```

With prompt



```
> Question:
based on data now tell me is there a doctor name Fuzail in mayo clinic?

> Answer (took 42.08 s.):
 I do not have enough information to answer this question as the given context does not provide any specific details about Dr. Fuzail or his specialty at Mayo
Clinic.

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Ravinder J. Singh, Ph.D.
Specialities: Pathologist
Areas of Focus:

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Ravinder J. Singh, Ph.D.
Specialities: Pathologist
Areas of Focus:
```

Declined according to prompt.



```
Enter a query: tell me speciality of raj palraj
 Raj Palraj is an Infectious Disease Specialist.

 Raj Palraj is an Infectious Disease Specialist.

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Raj Palraj, M.B.B.S., M.D.
Specialities: Internist, Infectious Disease Specialist
Areas of Focus: Infection control, Tuberculosis, Bacterial endocarditis, Staph infections, General infectious diseases

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Raj Palraj, M.B.B.S., M.D.
Specialities: Internist, Infectious Disease Specialist
Areas of Focus: Infection control, Tuberculosis, Bacterial endocarditis, Staph infections, General infectious diseases

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Rajiv (Rajiv K.) K. Pruthi, M.B.B.S.
Specialities: Hematologist
Areas of Focus: Thrombophilia, Hereditary hemorrhagic telangiectasia, von Willebrand disease, Hemophilia, Bleeding disorder

> source_documents\all_cleaned_merged.csv:
Hospital Name: Mayo Clinic
Name: Rajiv (Rajiv K.) K. Pruthi, M.B.B.S.
Specialities: Hematologist
Areas of Focus: Thrombophilia, Hereditary hemorrhagic telangiectasia, von Willebrand disease, Hemophilia, Bleeding disorder
```