Assignment: Develop a simple LLM-powered pipeline to extract information from medical docs.

At Co:Helm, we are building a Prior Authorization tool to enhance the productivity of Utilization Review nurses. We are leveraging LLMs (large language models) to extract critical information from medical docs before integrating custom models and techniques.

Your task is to build a small part of our tool that handles medical record ingestion and querying. As well as the performance of your overall pipeline, you will be assessed on the quality of your code, ability to critically reflect on your solution and show deeper awareness of how the technology works. We will also test your code on a hold-out test set, so give thought as to how best to make your code generalisable.

There are two parts to this assignment:
Submission - a take-home task to be completed on your own and uploaded to GitHub (3hrs)
Technical Interview - an in-person or virtual session which may include:

- Presentation to walk through and talk about your submission (30mins)
- Q&A (30mins)
- Peer Programing exercise on your submission (30mins)

Submission:

Write an application or script that satisfies the following requirements:

1. Ingest a sample medical PDF provided by email. (Note: this is based on a hospital fax which has been converted using OCR. Therefore, artifacts may exist.)
2. Extract the following information:
    a. Patient’s chief complaint
    b. Treatment plan the doctor is suggesting
    c. A list of allergies the patient has
    d. A list of medications the patient is taking, with any known side-effects
3. Answer the following questions using the PDF:
    a. Does the patient have a family history of colon cancer in their first-degree relatives?
    b. Has the patient experienced minimal bright red blood per rectum?
    c. Has the patient had significant loss of blood?
    d. Does the patient have a history of skin problems?
    e. Has the patient used hydrocortisone cream for the haemorrhoids that they are currently experiencing?
    f. Were any high risk traits found on the patient’s genetic test?
    g. Has the patient had a colonoscopy in the last 5 years?
    h. Has the patient had any recent foreign travel?
    i. How long has the patient been known to healthcare services?
4. For each answer above, the application must:
    a. Justify its reasoning, ie provide conclusive evidence for each answer
    b. Provide a confidence score where 10/10 is very confident of its answer
5. Finally, your application should suggest whether the treatment plan is appropriate. The metric and clinical accuracy does not need to be accurate and can simply be a comparison of how many questions in (3) are answered as Yes vs No.
6. The final output should be a JSON object with an appropriate structure to represent the information retrieved in the above steps. The expectation is the response will be used to show information on a front-end application.

Bonus points are awarded for submissions that:

- demonstrate wider knowledge by introducing at least one significant optimisation beyond prompt stuffing
- generalize well across other sample medical records and queries
- provide an executable or dockerfile for the script
- demonstrate an understanding of applying code to build Product solutions through a thoroughly designed output structure

We encourage you to use large language models to complete this task. Submissions without a Readme or instructions on how to run the project will be rejected.

If you would like to keep your GitHub repository private, please invite zaarheed and chris-lovejoy as a collaborator.