Medical Data Extraction using LLM

Problem : Traditionally, medical records in India and many other countries are handwritten, resulting in unstructured free-flowing text. This format makes it difficult to analyse and utilise valuable medical data for research, improved care, and public health initiatives.

Prior Limitation: Extracting insights from these unstructured records was previously cumbersome and inefficient.

New Opportunity: The emergence of large language models (LLMs) like generative AI presents a solution. These models can potentially:

- **Extract key information:** LLMs can process the unstructured text in medical records and identify crucial details.
- **Convert to a usable format:** The extracted information can be converted into a structured format, like tables or databases, making it easier to analyse and utilise.
- **Unlock data potential:** By structuring medical data, LLMs can unlock its potential for improving healthcare in various ways:
    - **Research:** Structured data can be used for medical research, leading to better treatments and preventative measures.
    - **Personalised care:** Doctors can leverage the structured data to provide more personalised and informed patient care.
    - **Public health initiatives:** Analysing structured data can help identify trends and patterns, informing public health strategies to combat diseases and improve overall health outcomes.

Overall, LLMs offer a promising solution to bridge the gap between unstructured medical records and unlock their valuable insights for a more data-driven and improved healthcare system.

Data - **https://huggingface.co/datasets/AGBonnet/augmented-clinical-notes**

The Data consists of following columns

Field	Description	Source
idx	Unique identifier, index in the original NoteChat-ChatGPT dataset	NoteChat
note	Clinical note used by NoteChat (possibly truncated)	NoteChat
full_note	Full clinical note	PMC-Patients
conversation	Patient-doctor dialogue	NoteChat
summary	Patient information summary (JSON)	ours

Use the **conversation** columns and extract the following information:

```
fields = """{
 "summary": [
     {{
     "chief_complaints": ,
     "symptoms": ,
     "medical_examinations": ,
     "patient medical history": ,
     "surgeries": ,
     "allergies to medicines": ,
     }}
 ],
 "patient information": [
     {{
```

```
      "age": ,
      "sex": ,
      "ethnicity": ,
      "weight": ,
      "height": ,
      "family medical history": ,
      "recent travels": ,
      "socio economic context": ,
      "occupation":
    }}
  ],
   "symptoms": [
    {{
      "name_of_symptom": "",
      "intensity_of_symptom": "",
      "location": "",
      "time": "",
      "temporalisation": "",
      "behaviours_affecting_the_symptom": "",
      "details": ""
    }}
  ],
  "medical_examinations": [
    {{
      "name": "",
      "result": "",
      "details": ""
    }}
```

```
],
diagnosis_tests": [
  {{
    "test": ,
    "severity": ,
    "result": ,
    "condition": ,
    "time": ,
    "details":
  }}
],
"surgeries": [
  {{
    "reason": ,
    "Type": ,
    "time": ,
    "outcome": ,
    "details":
  }}
],
"patient medical history": [
  {{
   "physiological context": ,
   "psychological context": ,
   "vaccination history": ,
   "allergies": ,
   "exercise frequency": ,
   "nutrition": ,
```

```
        "sexual history": ,
        "alcohol consumption": ,
        "drug usage": ,
        "smoking status":
      }}
    ],
    "treatments": [
      {{
        "name": ,
        "related condition": ,
        "dosage": ,
        "time": ,
        "frequency": ,
        "duration": ,
        "reason for taking": ,
        "reaction to treatment": ,
        "details":
      }}
    ],
    "discharge": [
      {{
        "reason":,
        "referral": ,
        "follow up":,
        "discharge summary":
      }}
    ]
}"""
```

1. Data Processing and Extraction:

- Utilise LLMs to process the "conversation" text columns in the medical records.
- Fine-tune prompts to guide the LLM in accurately extracting specific fields of interest from the text, such as: (replace with your desired fields)
    - Patient Name
    - Diagnosis
    - Medication Prescribed
    - Symptoms Described
    - Treatment Plan
    - Other data available

2. Data Storage and Management:

- Store the extracted information in a structured format, like a PostgreSQL database, for efficient analysis and retrieval.

3. User Interface for Accessibility:

- Develop a user interface (UI) using Gradio, a web framework, to facilitate data extraction.
- Users can input the conversation text as input, and the UI will display the extracted fields as the output.

4. Evaluation and Accuracy Measurement:

- To gauge the effectiveness of the LLM in extracting information, calculate metrics like BLEU and ROUGE scores.
- These metrics compare the extracted data with the "ground truth" - a column summary present in the dataset, indicating the expected information for each field.

Overall, the expected outcome is a system that automatically extracts crucial medical information from unstructured conversations, stores it in a structured format, provides a user-friendly interface for access, and ensures accuracy through evaluation metrics.