Project Support Session - Medical Assistant

Agenda

- Project Problem Statement and Dataset Overview
- Open Q&A
- Project Evaluation Rubric
- Project Submission Guidelines
- FAQs

Medical Assistant - Business Context

The healthcare industry is rapidly evolving, and professionals face increasing challenges in managing vast volumes of medical data while delivering accurate and timely diagnosis. Quick access to comprehensive, reliable, and up-to-date medical knowledge is critical for improving patient outcomes and ensuring informed decision-making in a fast-paced environment.

Healthcare professionals often encounter information overload, struggling to sift through extensive research and data to create accurate diagnoses and treatment plans. This challenge is amplified by the need for efficiency, particularly in emergencies, where time-sensitive decisions are vital. Furthermore, access to trusted, current medical information from renowned manuals and research papers is essential for maintaining high standards of care.

To address these challenges, healthcare centers can focus on integrating systems that streamline access to medical knowledge, provide tools to support quick decision-making and enhance efficiency. Leveraging centralized knowledge platforms and ensuring healthcare providers have continuous access to reliable resources can significantly improve patient care and operational effectiveness.

Medical Assistant - Objective

As an AI specialist, your task is to develop a RAG-based AI solution using renowned medical manuals to address healthcare challenges. The objective is to understand information overload, apply AI techniques to streamline decision-making, analyze its impact on diagnostics and patient outcomes, evaluate its potential to standardize care practices, and create a functional prototype demonstrating its feasibility and effectiveness.

Medical Assistant - Questions to Answer

1) What is the protocol for managing sepsis in a critical care unit?

2) What are the common symptoms of appendicitis, and can it be cured by medicine? If not, what surgical procedure should be followed to treat it?

3) What are the effective treatments or solutions for addressing sudden patchy hair loss, commonly seen as localized bald spots on the scalp, and what could be the possible causes behind it?

4) What treatments are recommended for a person who has sustained a physical injury to brain tissue, resulting in temporary or permanent impairment of brain function?

5) What are the necessary precautions and treatment steps for a person who has fractured their leg during a hiking trip, and what should be considered for their care and recovery?

Medical Assistant - Data Dictionary

The Merck Manuals are medical references published by the American pharmaceutical company Merck & Co., that cover a wide range of medical topics, including disorders, tests, diagnoses, and drugs. The manuals have been published since 1899 when Merck & Co. was still a subsidiary of the German company Merck.

The manual is a PDF with over 4,000 pages divided into 23 sections.

Medical Assistant - Q&A

Medical Assistant - Evaluation Rubric

Section	Points
Question Answering using LLM	8
Question Answering using LLM with Prompt Engineering	11
Data Preparation for RAG	8
Question Answering using RAG	12
Output Evaluation	9
Actionable Insights & Recommendations	4
Presentation/Notebook - Overall Quality	8

Medical Assistant - Evaluation Rubric

Section 1: Question Answering using LLM

What to Cover?	Guiding Questions and Considerations
Load the large language model from Hugging Face	Load any appropriate LLM from Hugging Face
Create a function to define the model parameters and response	Define a function that accepts a text query along with other parameters needed for the LLM
Apply the response generation function to get answers to the questions provided in the problem statement	Use the previously defined function on all the given questions to generate the responses

Medical Assistant - Evaluation Rubric

Section 1: Question Answering using LLM

What to Cover?	Guiding Questions and Considerations
Provide comments/observations for the answers received	How do the responses look like? Are they too generic?

Medical Assistant - Evaluation Rubric

Section 2: Question Answering using LLM with Prompt Engineering

What to Cover?	Guiding Questions and Considerations
Apply prompt engineering and LLM parameter tuning (at least 5 combinations) and get answers to the questions provided in the problem statement	What factors to consider when engineering a better prompt for the desired output? What LLM parameters can be tuned for creativity vs specificity?
Provide comments/observations for the answers received	How does the response change compared to a vanilla LLM?

Medical Assistant - Evaluation Rubric

Section 3: Data Preparation for RAG

What to Cover?	Guiding Questions and Considerations
Load the data file provided	Use the appropriate PDF loader to load the PDF data file.
Split the data using a text splitter with necessary attributes	What are the considerations for choosing chunk size and chunk overlap?
Load the embedding model	How to choose an embedding model?

Medical Assistant - Evaluation Rubric

Section 3: Data Preparation for RAG

What to Cover?	Guiding Questions and Considerations
Load the vector database	The vector database enable faster retrieval of relevant document(s)
Define the retriever with appropriate search method and k value	Small values of k (e.g., 3–5) for precise, focused responses and faster retrieval Large values of k (e.g., 10–20) if the context is broad or the query requires more supporting information

Medical Assistant - Evaluation Rubric

Section 4: Question Answering using RAG

What to Cover?	Guiding Questions and Considerations
Get answers to the questions provided in the problem statement	Get the responses for all the questions provided
Fine-tune the chunking, retriever, and LLM parameters (at least 5 combinations) to check different results	Try different combinations of RAG and LLM parameters
Provide comments/observations for the answers received	How does the response change on using RAG?

Medical Assistant - Evaluation Rubric

Section 5: Output Evaluation

What to Cover?	Guiding Questions and Considerations
Define the evaluation prompt for groundedness	Ensure that the evaluation criteria, metric and instructions are present
Define the evaluation prompt for relevance	Ensure that the evaluation criteria, metric and instructions are present
Evaluate all the responses for the questions provided in the problem statement	Does the LLM being used as a Judge provide a score and relevant rationale for the score?

Medical Assistant - Evaluation Rubric

Section 5: Actionable Insights and Recommendations

What to Cover?	Guiding Questions and Considerations
Conclude with the key takeaways for the business	What are the key things that we learned from the data that's useful for the business? Logical next steps are a key inclusion

Medical Assistant - Evaluation Rubric

Section 6: Presentation/Notebook - Overall Quality

Low-code Considerations	Full-code Considerations
Clear structure, flow, and visual appeal - everything sits well in a story	Clear structure and flow
Crispness - Focus on key points	Well-commented and executed code
All sections of the rubric included, even if in an appendix	All sections of the rubric included, even if in an appendix

Medical Assistant - Low-Code Version

For learners who aspire to be in managerial roles in the future, focusing on solution review, interpretation, recommendations, and communication with business stakeholders

Download the dataset and the *Learner Notebook - Low Code* (this is a template notebook)

Fill in the blanks in the notebook to complete and execute the code to solve the questions and perform all the tasks as per the grading rubric

Once the notebook is completely executed and necessary outputs obtained, a business presentation (using Microsoft PowerPoint, Google Slides, etc.) has to be created

Medical Assistant - Low-Code Version

The presentation should contain observations, insights, and recommendations for the business problem

The presentation template provided can be referred to as a sample

Once the presentation is complete, convert the presentation to .pdf format

The presentation should be submitted as a PDF file (.pdf) and NOT as a .pptx file

Please make sure that all the sections mentioned in the grading rubric have been covered in the submission

Medical Assistant - Full-Code Version

For learners who aspire to be in hands-on coding roles in the future, focusing on building solution codes from scratch

Download the dataset and the *Learner Notebook - Full Code* (this is a template notebook containing high-level steps to perform and insight-based questions)

Write necessary code to solve the questions and perform all the tasks as per the grading rubric

Clearly write down observations, insights, and recommendations for the business problem based on the analysis performed

Once the notebook is complete, download it as a **.ipynb** file and convert it to a **.html** file

Medical Assistant - Full-Code Version

The notebook should be submitted as an HTML file (.html) and NOT as a notebook file (.ipynb)

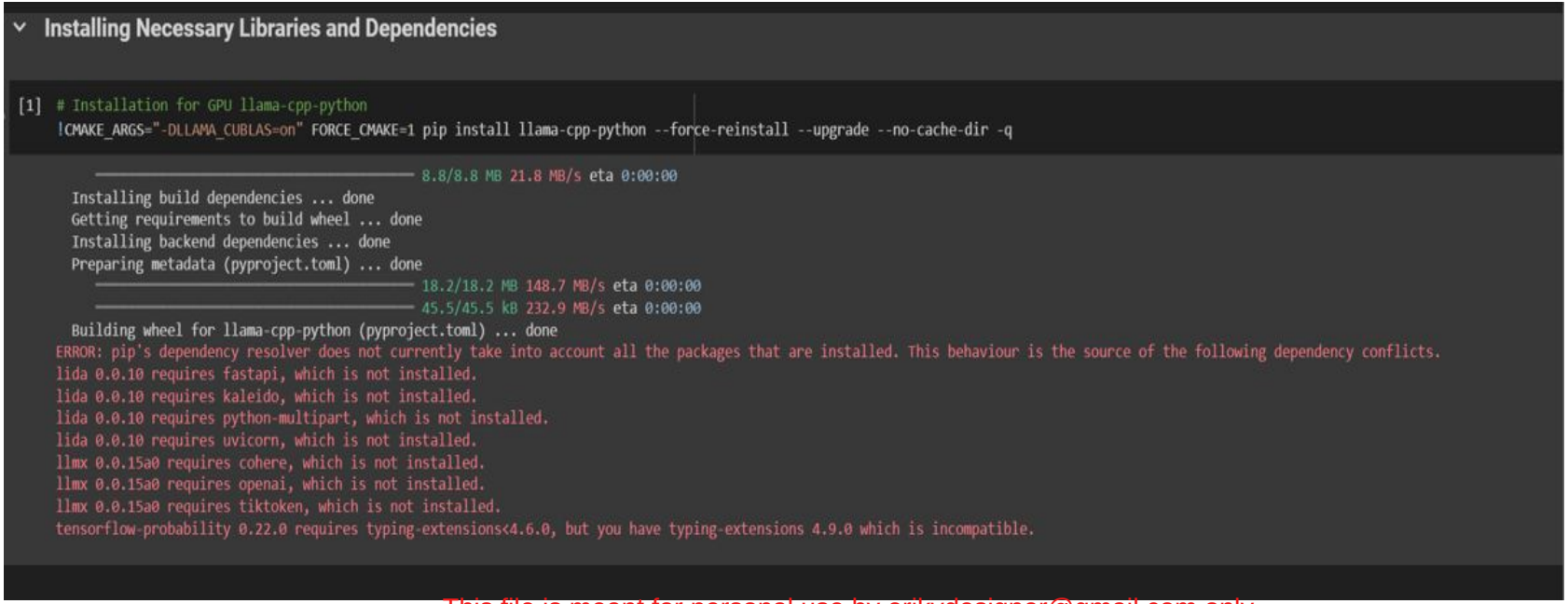
The conversion can be done via one of the following ways:

Jupyter Notebook: ***Download as HTML*** from the ***File*** menu
Google Colab: Use free online tools

Please make sure that all the sections mentioned in the grading rubric have been covered in the submission

Medical Assistant - FAQ

I am getting the below error. How do I resolve it?



Medical Assistant - FAQ

Kindly ensure that the runtime is connected to a T4 GPU, as DLLAMA_CUBLAS is set to 'on'

Also, this is not an error but a dependency conflict issue. It can be safely ignored, as the packages have been successfully installed.

Kindly restart the session after the installation command is executed to ensure that the environment is updated correctly.

Medical Assistant - FAQ

The vector database is taking longer than expected to be created

Chunk size and chunk overlap control the time and space complexity of the vector database

Lowering the chunk size will help reduce the memory needed to store these chunks

Lowering the chunk_overlap parameter will reduce the number of generated chunks.

Great Learning

Power Ahead!