AI Club Fall 2024: Database Specific Chatbot Group Project Schedule

Week	Goals	Notes	AI Club Schedule
Week 1 (9/23/24)	- Get familiar with Python - Start researching the project - Gather list of potential databases - Finalize initial presentation - Take a look at this article to get an idea on the project		Initial Presentations (9/27/24)
Week 2 (9/30/24)	- Experiment with coding Python on Google Colab - Take a look at different web scraping tools and libraries - *requests* for HTTP requests - *BeautifulSoup* for parsing HTML - *pdfminer.six* for parsing PDF files		Industry Presenters: LPL Financial (10/4/24)
Week 3 (10/7/24)	- Gather data - Experiment with data scraping the chosen database - If not possible try a different database - **Decide if we will use a structured dataset for fine tuning or stick with the pdf text for RAG**	Data mined text requires sufficient structuring for fine tuning. A prepared dataset is better. Look into RAG for unstructured data	Programming Challenge (10/11/24)
Week 4 (10/14/24)	- Research LangChain for RAG Question-Answering on a PDF - Ex. RAG tutorial on Game Manuals - Finalize loading pdf into text, splitting text into chunks, and embedding chunks - Find a Text Embedding model to convert text chunks to vector embeddings	We have reinterpreted the project as a Medical Domain Specific RAG bot for pdf files. We met this week to look at different vector DBs / LLMs.	Workshop: Chaining Models (10/18/24)
Week 5	- Find a free open-source vector database to store	We have finalized	Midterm

(10/21/24)	embedding - Chroma, Weaviate, Milvus - Implement vectorization method - Test vector db similarity retrieval with an example question	our selected vector DB as Chroma which can be loaded easily with LangChain	Presentations (10/25/24)
Week 6 (10/28/24)	- Find an effective and free LLM to load along with a method for loading it - Ollama lets us locally run Gemma LLM but answer generation is lengthy - Optimize response generation time while maintaining code accessibility, minimize API usage	Not so sure about Ollama but does allow anyone to run the code w/o API	No Meeting (11/1/24)
Week 7 (11/4/24)	- Focus on RAG chain definition - Find how to reduce response generation time - Continue Week 6 objectives	Alternative vector retrieval method found to reduce runtime by 25%	ANCS Collab (11/8/24)
Week 8 (11/11/24)	- Meet this week to discuss progress - Test if the LLM retains memory from previous conversation - Gemma model did pass tests	Found the correct way to install Ollama for colab	Workshop: TBD (11/15/24)
Week 9 (11/18/24)	- Test Mistral model for response time generation, accuracy, and memory retention - Meet this week to start work on UI via Gradio	Response time reduced to ~4s, based on data, and did retain memory	Work on Projects (11/22/24)
Week 10 (11/25/24)	- Final evaluations on chatbot - Prepare prompts for final presentation - Finalize final presentation slides		No Meeting (11/29/24)
Week 11 (12/2/24)	- Meet this week to practice final presentation and chatbot demo		Final Presentations (12/6/24)