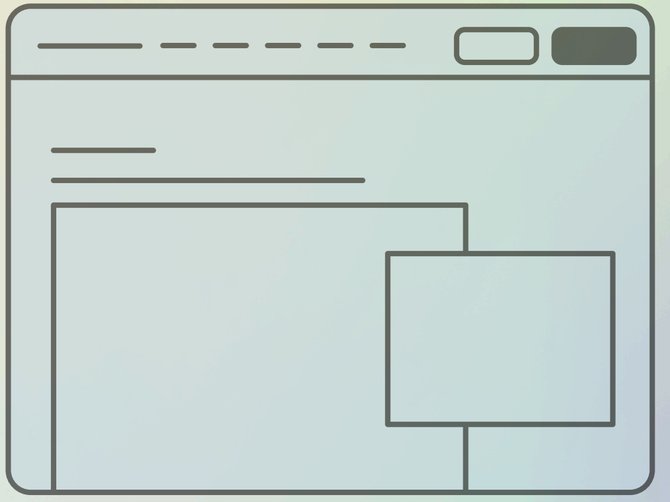
13 JULY 2025

LLM-Powered Q&A System for PDF and Excel Documents



BADER ALSHAMRANI

Problem Understanding

Tackling Information Overload in Scientific PDF Documents

1

Context

Hospitals and researchers have tons of important information saved in files like **PDFs** (documents) and **Excel sheets**

2

Problem

Finding answers by hand is slow, especially from complicated tables, pictures, or big sheets

3

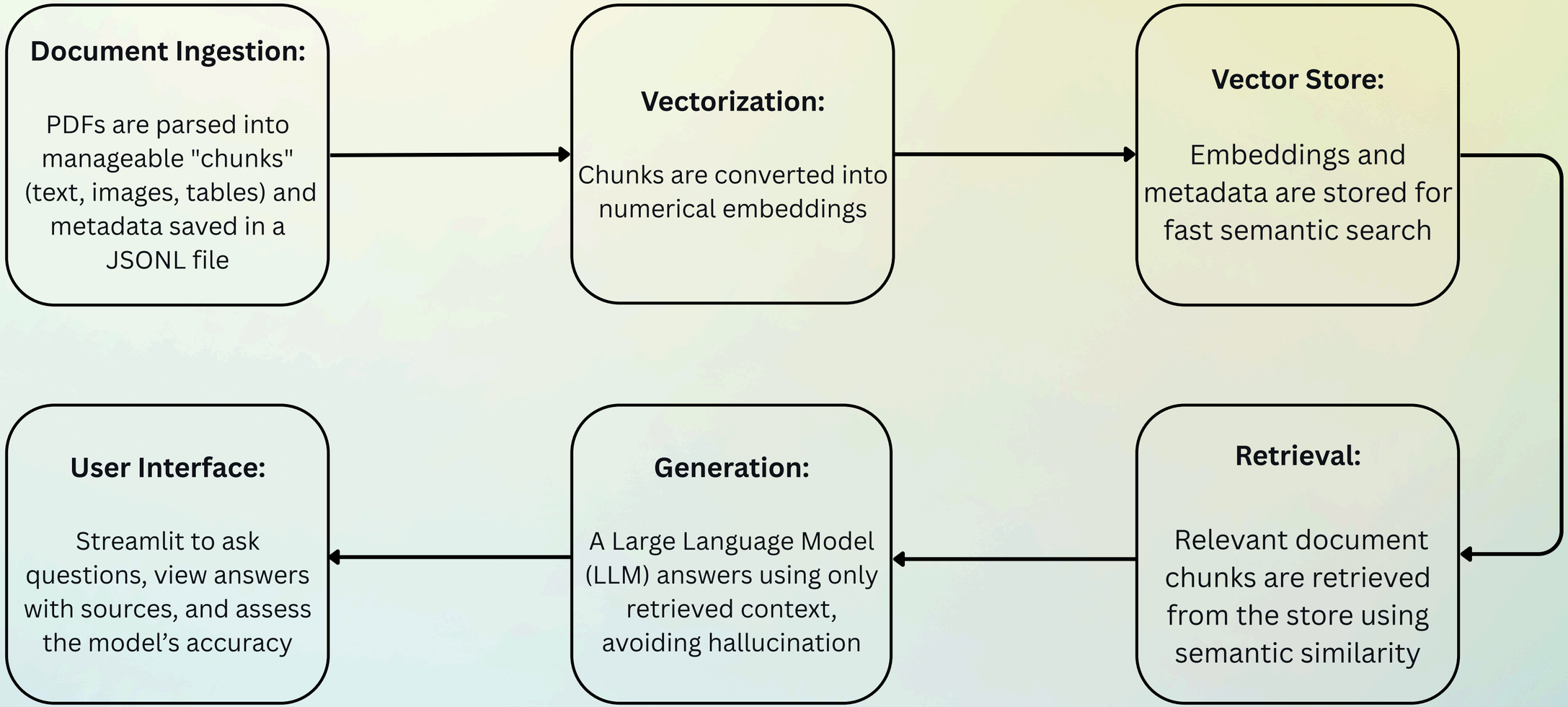
Goal

A chatbot that gives easy answers by finding facts from documents and sheets

LLM-Powered Document Assistant for Research Papers (PDF)

Solution Design & Approach – A RAG Pipeline

Retrieval-Augmented Generation (RAG) System



Model/tools used and reasoning

Tool/Model	Specifics	Role/Reasoning
FastPDF Custom Parser	Parses PDFs into structured chunks stored as JSON records	Extracts detailed text, tables, and images for precise retrieval
SentenceTransformer (MiniLM)	Sentence embedding model (all-MiniLM-L6-v2)	Converts text chunks into vectors to improve search accuracy
ChromaDB	Local vector database for embeddings	Stores and indexes embeddings for fast semantic search
Qwen2.5-VL-32B-Instruct	Multimodal LLM (text + image + tables + diagrams)	Generates rich answers using both text and image context
Streamlit	Python web app framework	Creates an easy-to-use chat interface for interaction

System Evaluation Approach

Automated RAG Evaluation (BERTScore)

Uses predefined questions with ground-truth answers per PDF for evaluation.

System generates responses, then semantic similarity is measured vs. ground truth.

Reports average Precision, Recall, and F1 scores to assess model performance.

Test Case

Detailed Results:

Question	Generated Response	Expected Response	BERTScore Precision	↓ BERTScore Recall	BERTScore F1
What was the estimated total cost of incomplete appointments	The estimated total cost of incomplete ap	The Veterans Health Administration (VHA) estimated the total co	0.8242	0.9451	0.8805
What modeling technique was adopted for the predictive mode	The modeling technique adopted for the p	Logistic regression was adopted as the modeling technique beca	0.8266	0.9104	0.8665
Which three variables were consistently related to a patient's nc	The three variables consistently related to	The three variables consistently related to a patient's no-show p	0.8309	0.9071	0.8673
According to the pilot study results for successfully contacted pa	According to the pilot study results for suc	After the intervention, the no-show rate in the pilot group was re	0.8457	0.9009	0.8724
What was the primary objective of the study as stated in the abs	The primary objective of the study, as stat	The primary objective was to develop and test a predictive mode	0.8353	0.8984	0.8657

Limitations

Fixed & Static Test Questions Per File

Evaluation questions are fixed and per PDF

Single PDF Query Limit

Chat currently only supports querying one selected PDF at a time

API Key Dependency

Manual setup needed for external LLM access

Future Improvements

Multi-Document Querying

Enable users to ask questions across multiple selected PDFs simultaneously

In-App PDF Upload

Enable direct PDF upload, including drag-and-drop functionality, within the UI

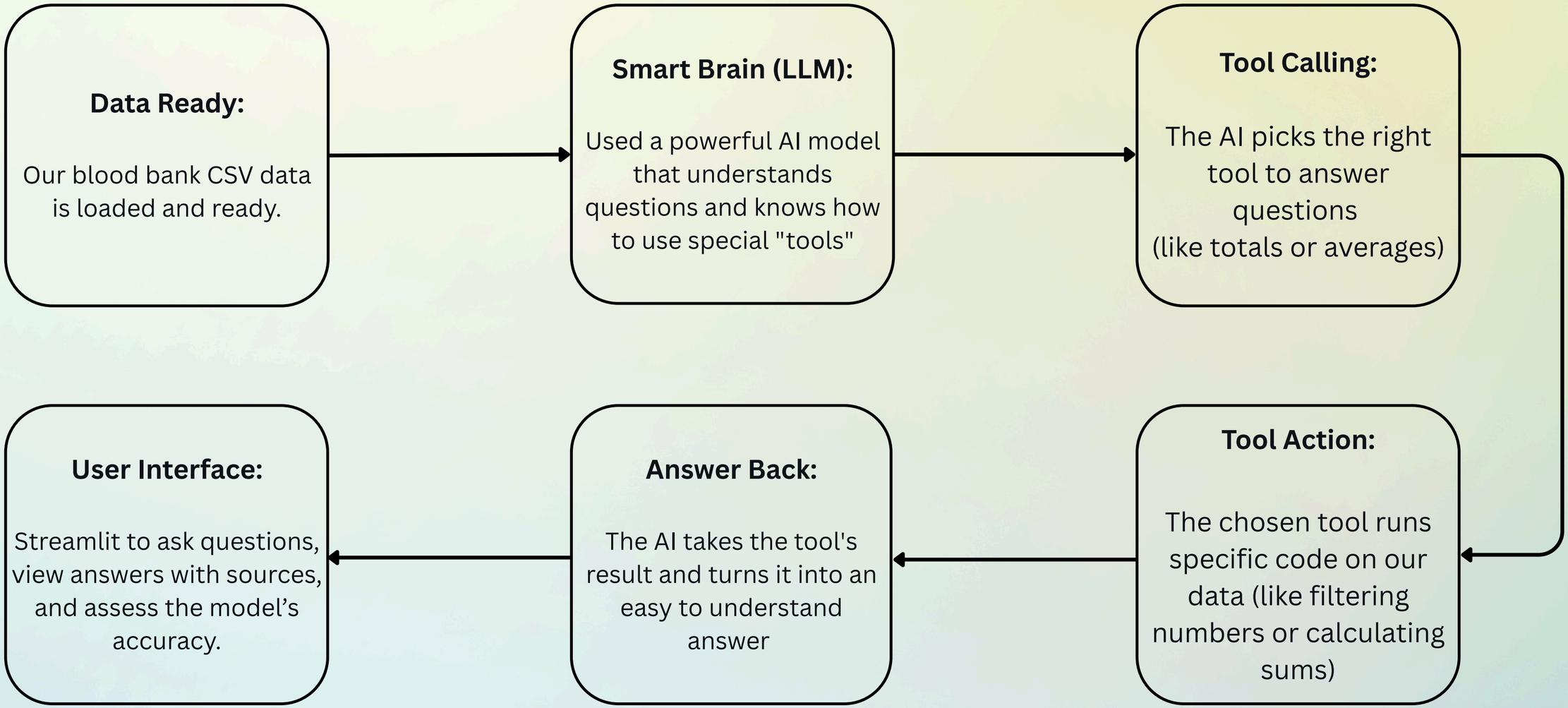
Use LangChain

To simplifies the development, production, and deployment

LLM-Powered Blood Bank Chat Assistant (Excel Dataset)

Solution Design & Approach – A RAG Pipeline

A Chatbot for Data Analysis



Model/tools used and reasoning

Tool/Model	Specifics	Role/Reasoning
Pandas (df)	Python data analysis library	Loads and manages the CSV data for all operations.
Custom Data Tools	query_data() get_unique_values()	Python functions that perform data filtering, aggregation, and retrieve unique values.
Qwen3-30B-A3B	LLM	The **"brain"** that understands questions and calls tools
Streamlit	Python web app framework	Creates an easy-to-use chat interface for interaction.

System Evaluation Approach

Automated RAG Evaluation (BERTScore)

Uses predefined questions with ground-truth answers per PDF for evaluation.

System generates responses, then semantic similarity is measured vs. ground truth.

Reports average Precision, Recall, and F1 scores to assess model performance.

Test Case

Detailed BERTScore Results:

Question	Generated Response	Expected Response	BERTScore Precision	↓ BERTScore Recall	BERTScore F1
What is the average transfused volume for patients with O blood type?	The average transfused volume for patie	The average transfused volume for patients wi	0.9729	0.9913	0.982
How many patients have AB-negative blood type?	There are **2,381 unique patients** with	There are 2,492 patients with AB-negative bloo	0.845	0.969	0.9028
How many transfusions occurred in 2021?	The total number of transfusion events ir	There were 19,941 transfusions recorded in the	0.9117	0.9499	0.9304
What is the average age of male patients?	The average age of male patients is 53.79	The average age of male patients who received	0.9849	0.9495	0.9669
How many transfusions were done in KFHI-Adult Cardiac Surgery?	1351 transfusions were done in KFHI-Adu	There were 1,351 transfusions performed in th	0.958	0.9475	0.9527
What is the total volume of Bone Marrow transfusions?	The total volume of Bone Marrow transfu	The total volume of Bone Marrow transfusions	0.8866	0.9369	0.911
How many female patients received Plasma Thawed?	The number of female patients who recei	A total of 1,970 female patients received Plasm	0.9024	0.9207	0.9114
How many patients older than 50 received any product?	The number of patients older than 50 wh	A total of 10,789 patients older than 50 receive	0.8865	0.9069	0.8966
What is the earliest transfusion date recorded?	Okay, the user is asking for the earliest tr	The earliest transfusion date recorded in the d	0.7631	0.8678	0.8121
What is the maximum transfused volume recorded?	Okay, the user is asking for the maximum	The highest single transfused volume recorded	0.7545	0.8655	0.8062

Limitations

Prompt Dependency

Relies entirely on explicit data schema and system prompt instructions

No External Knowledge

Operates purely on the provided dataset, lacking broader medical context

No Visualization

Results are textual, missing visual insights

Future Improvements

Data Visualization

Empower the LLM to generate relevant charts and graphs directly

Predictive Modeling

Build ML models to predict future blood needs, helping manage supplies early and avoid shortages

Interactive "What-If" Scenarios

If O-negative supply drops by half, which services would be most affected?

Thank You
Any Questions?