King Faisal Specialist Hospital & Research Centre LLM Chat Data & Document Assistants

Prepared by: Bader Alshamrani

Prepared for: Digital Innovation Hub, KFSHRC

July 16, 2025

Contents

1 Executive Summary

This report gives a clear picture of two advanced Chat Assistant systems built using Large Language Models (LLMs) for King Faisal Specialist Hospital & Research Centre (KFSHRC). These innovative systems offer distinct but complementary functionalities: one is an expert assistant for analyzing structured blood bank transfusion records, and the other is a clever system (called RAG) for finding exact information within unstructured PDF documents. By letting staff ask questions about complicated data and vast document collections just by typing naturally, these assistants greatly improve how easy it is to access information. They also help make daily tasks smoother and empower people across different departments to make smarter choices. Each solution combines robust data processing, intelligent AI control, and a careful checking process, making sure all its answers are accurate and trustworthy.

2 Introduction

In a busy place like King Faisal Specialist Hospital & Research Centre (KFSHRC), healthcare workers are always dealing with a huge amount of information. This includes organized details, like accurate patient blood transfusion records, as well as lots of unorganized information found in medical research papers, clinic rules, and how-to guides, often in PDF files. Usually, getting useful information from organized databases needs special analytical skills or specific reports. And finding small details within many documents can take a lot of time and effort to do by hand. To address these big problems, KFSHRC is introducing two distinct LLM Chat Assistant systems. The main goal of these assistants is to make it easy for everyone to get both number-based data insights and information from documents, instantly. This means anyone who needs it can get it, no matter how tech-savvy they are. These separate but powerful capabilities are set to greatly improve how well things run, help with making medical decisions based on solid facts, and encourage a quicker, more data-focused way of working across the hospital.

3 System Architecture and Core Components

The King Faisal Specialist Hospital & Research Centre (KFSHRC) LLM Chat Assistant systems are engineered with sophisticated and modular architecture. This design allows for scalability, robustness, and the effective operation of distinct information processing capabilities.

3.1 User Interfaces (UIs)

The KFSHRC LLM Chat Assistant solutions provide separate, dedicated user interfaces for interacting with structured data and unstructured documents, ensuring a clear and tailored experience for each type of query. Both UIs are built using modern web frameworks, offering intuitive and highly interactive chat-based experiences.

User Interface for Structured Data Analysis

This interface features an **Interactive Chat Window** as the main way for users to ask questions about the structured blood bank data and receive responses, mimicking a natural conversation. A **Conversation History** is kept during each session, allowing the system to understand context for follow-up questions without needing to repeat everything. An **Expandable Dataset Overview** is provided, detailing available column names and their descriptions, guiding users on the scope of queryable information. **Example Questions** are prominently displayed to assist new users. For user convenience, there is also a **Clear Chat Functionality** to start a new conversation and dedicated **Evaluation Tabs** to check performance metrics.

User Interface for Document Queries (RAG)

This interface also utilizes an **Interactive Chat Window** for users to pose questions about PDF documents and receive answers. A **Conversation History** is maintained to support ongoing dialogue. A clear **Document Selection** mechanism allows users to specify which PDF they wish to interrogate, ensuring targeted information retrieval. **Example Questions** are available to guide users, alongside a **Clear Chat Functionality** for fresh starts and **Evaluation Tabs** for performance monitoring.

3.2 Data Management and Preparation

Getting useful information and analyzing it well really depends on having a strong system for managing data that fits the type of information.

- **Structured Data Management (Blood Bank Transfusion Records):** The system effectively manages a made-up dataset of blood bank transfusions, which it loads from a CSV file. This dataset contains important details like `ENCNTR_ID` (a unique number for each patient visit), `MRN` (medical record number), `AGE`, `GENDER`, `TRANSFUSED_VOL` (how much blood product was given), `PRODUCT_CAT` (the type of blood product), `CUR_ABO_CD` (the ABO blood type), `CUR_RH_CD` (the Rh blood type), `TRANSFUSION_DT` (the date of transfusion), and `MED_SERVICE`. As soon as the data is loaded, it goes through key preparation steps, like turning `TRANSFUSION_DT` into date and time information so it can be accurately analyzed over time. A built-in "memory" feature makes sure the dataset is loaded only once when the application starts. This greatly speeds up how the system works and how quickly it responds during every interaction with a user afterwards.
- **Unstructured Document Ingestion (PDF Knowledge Base):** This part focuses on getting a huge collection of unorganized complex PDF documents ready for smart searching. A separate process, usually a script that runs automatically, carefully reads and splits PDF documents into smaller, meaningful pieces of text. Importantly, it can also pull out images (like diagrams or charts) and link them to the right text parts. Each of these processed text pieces is then turned into a special numerical code, called an embedding, using a dedicated Sentence Transformer model (e.g., `all-MiniLM-L6-v2`). These embeddings are compact numerical codes that capture the true meaning of the

text. Finally, these embeddings, along with their linked details (including the source PDF filename, page number, the original text, and where to find any extracted `image_paths`), are stored in a special database for these codes (ChromaDB) that keeps them even after the system is turned off. This code database is built for very fast searches to find similar information, and it forms the core of the RAG system.

3.3 Intelligent Processing and Tool Orchestration (LLM Cores)

The core intelligence for each system comes from its dedicated Large Language Models (LLMs) provided by Fireworks AI.

LLM Core for Structured Data Analysis

This system uses a Qwen model, such as `Qwen3-30B-A3B`, specifically for general reasoning and orchestrating queries on structured data. At its heart is a carefully crafted "system prompt" that acts like the LLM's handbook. It tells the LLM to be an expert data analyst, explaining what tools it has (like `query_data` or `get_unique_values`), what they do, how the data is set up, specific rules for different types of data (like date formats or how to handle `MED_SERVICE`), and important rules for combining information (like knowing the difference between all events and unique patients). When a user types a question into the structured data UI, this LLM directly processes it and translates it into a precise tool call to extract insights from the blood bank data.

LLM Core for Document Queries (RAG)

This separate system uses a Vision-Language Model, specifically `Qwen2.5-VL-32B-Instruct`, which is designed for multimodal RAG tasks involving both text and images. Similar to the structured data LLM, it operates based on a "system prompt" that guides its role as a RAG assistant. When a user asks a question in the document query UI, this LLM orchestrates the RAG process by triggering the retrieval mechanism to find relevant information from the PDF knowledge base.

3.4 Specialized Data Analysis Tools (Structured Data)

The `query_data` function is a cornerstone for extracting insights from the blood bank's structured dataset. Its flexibility allows for a wide array of analytical questions:

- **Filtering Operations:** Users can apply sophisticated filters. For instance, `eq` (equals) for exact matches (e.g., `GENDER: {'eq': 'F'}`), `neq` (not equal), `gt` (greater than), `lt` (less than), `gte` (greater than or equal to), and `lte` (less than or equal to) are supported for numerical and date comparisons. Critically, for `MED_SERVICE` which often contains medical services names, the `contains` operator is exclusively used (e.g., `MED_SERVICE: {'contains': 'nephrology'}`), ensuring accurate matches. Dates are strictly handled in `'YYYY-MM-DD'` format, with date ranges always using inclusive `gte` and `lte` operators.

- **Diverse Aggregations:** The tool can perform various calculations. For questions about the total number of transfusion events or "how many patients received transfusions" (referring to events), `aggregations={'ENCNTR_ID': 'count'}` is the standard. Only when "unique patients" or "distinct individuals" are explicitly requested does the system use `aggregations={'MRN': 'nunique'}`. For inquiries about "total volume," "blood demand," or trends of specific `PRODUCT_CAT`s, `aggregations={'TRANSFUSED_VOL': 'sum'}` is applied. Average age is calculated using `aggregations={'AGE': 'mean'}`.
- **Categorical Grouping:** Results can be segmented by one or more categorical columns using `group_by` (e.g., `['PRODUCT_CAT', 'GENDER']`), providing detailed breakdowns.
- **Time-Series Analysis:** For trend questions, the `time_resample_period` parameter allows aggregation at daily (`'D'`), weekly (`'W'`), or monthly (`'M'`) intervals. For general "blood demand trend" questions, the system intelligently includes `group_by=['PRODUCT_CAT']` to show usage patterns across different product categories.

3.5 Retrieval Augmented Generation (RAG) Mechanism (Unstructured Documents)

The RAG mechanism is designed to provide highly accurate and sourced answers from the PDF knowledge base. When a user asks a question about a selected PDF:

- The `VectorStore.query` method is invoked. It takes the user's question, generates its embedding, and performs a semantic search within the ChromaDB vector store. This search is constrained to the currently selected PDF, ensuring targeted retrieval.
- The system retrieves the top k most semantically similar chunks of text and their associated metadata (including source PDF, page number, and any relevant `image_paths`). The number of retrieved chunks (k=15 by default) is configurable to optimize context provision.
- These retrieved chunks, encompassing both text and potentially visual information (which is encoded as Base64 images and sent to the multimodal LLM), form the context that is provided to the LLM. This direct, relevant context significantly reduces the likelihood of the LLM "hallucinating" information and ensures factual accuracy, a key advantage over traditional LLM approaches.
- The LLM then synthesizes an answer based only on this provided context, citing the precise source PDF and page number, allowing users to verify the information directly within the original document.

3.6 Response Post-processing

A dedicated `strip_model_thoughts` function is implemented to enhance the user experience by removing common LLM conversational fillers, internal thought processes (e.g., `<thought>...</thought>`), and direct tool code output that might appear in the raw response. This ensures that the final answer presented to the user is clean, direct, and professional.

4 Key Features and Capabilities

The King Faisal Specialist Hospital & Research Centre (KFSHRC) LLM Chat Assistant systems offer a comprehensive suite of functionalities, blending data analytics with document retrieval through their dedicated interfaces.

4.1 Capabilities for Structured Blood Bank Data Analysis

- **Intuitive Data Access:** Users can easily interact with the blood bank dataset through a natural language chat interface, eliminating the need for specialized query skills.
- **Context-Aware Interactions:** The system intelligently maintains conversation flow, understanding follow-up questions within the context of previous exchanges.
- **In-depth Data Exploration:** Allows users to easily identify and list unique values for categorical columns such as blood product types (`PRODUCT_CAT`) or medical services (`MED_SERVICE`), providing a clear understanding of the dataset's contents.
- **Granular Data Filtering and Analysis:** Enables precise filtering of transfusion records by diverse criteria, including patient demographics (age ranges, gender), specific blood characteristics (ABO/Rh types like "O negative"), and exact or range-bound transfusion dates (e.g., "all transfusions in 2021" or "in April 2021").
- **Diverse Quantitative Aggregations:** Supports the calculation of various metrics, such as the total number of transfusion events, the total volume of blood transfused (indicating demand), the count of unique patients, or the average age of patient cohorts.
- **Categorical Segmentation:** Provides the ability to group data by dimensions like `PRODUCT_CAT`, `GENDER`, or `MED_SERVICE`, allowing for segmented insights into transfusion patterns across different categories. **Time-Series Trend Analysis:** Crucially, the system can identify and report trends in blood demand or transfusion events over time, segmenting data by daily, weekly, or monthly periods. This includes complex queries like "Show me the monthly blood demand trend by product category in 2021," providing invaluable insights for inventory management and resource allocation.

4.2 Capabilities for Document Data

- **Targeted Question Answering from PDFs:** Users can obtain precise and concise answers to specific questions by directing their queries to particular PDF documents, reducing manual search efforts significantly.
- **Semantic Information Retrieval:** The underlying vector search ensures that the most semantically relevant sections of a document are retrieved, even if the user's query does not contain exact keywords from the document.
- **Multimodal Contextual Understanding:** The system's ability to process both text and images from retrieved document chunks means it can answer complex questions that might rely on visual information presented in diagrams, charts, or figures within the PDFs.

- **Verifiable Source Citation:** Every answer generated from PDF documents includes direct citations to the source PDF filename and the specific page number, fostering trust and enabling users to easily verify the information.
- **Scalable Knowledge Base:** The modular vector database architecture allows for seamless expansion of the system's knowledge base simply by ingesting new PDF documents, without requiring any retraining of the core LLM.

5 Model Evaluation and Quality Assurance

To ensure the highest standards of accuracy, relevance, and reliability, each system incorporates a rigorous evaluation framework, primarily leveraging the BERTScore metric for assessing the quality of natural language responses.

- **Conceptual Basis of BERTScore:** Unlike traditional lexical-overlap metrics, BERTScore assesses the semantic similarity between a generated text and a reference text. It achieves this by using contextual embeddings from pre-trained transformer models (like BERT), providing a more nuanced and accurate measure of how well the generated response captures the meaning of the expected answer. It produces Precision, Recall, and F1-scores, reflecting the quality of the generated response against the reference.
- **Evaluation Workflow for Structured Data Responses:** A comprehensive suite of test cases is prepared for structured blood bank data. Each test case consists of a natural language question and its corresponding factual "expected response" that would be generated if the underlying tool call was correct. During evaluation, the system simulates a full interaction: the LLM interprets the question, calls the appropriate data analysis tool (`query_data` or `get_unique_values`), receives the numerical or categorical output from the tool, and then calculates a natural language answer based on this output. BERTScore is then computed by comparing this generated natural language answer with the predefined expected natural language response. This process specifically evaluates the LLM's ability to accurately translate raw data results into coherent and correct human-readable text.
- **Evaluation Workflow for Document Responses:** A distinct set of test questions per PDF document is meticulously curated, with each question paired with a precise "expected response" directly sourced and verified from the content of that specific PDF. The evaluation process for RAG involves:

  1. The system takes a test question and the designated PDF.
  2. It then executes the RAG pipeline: the user's question is used to retrieve the most relevant text and image contexts from the specified PDF via the vector store.
  3. These retrieved contexts are provided to the multimodal LLM, which then generates an answer solely based on the provided information.
  4. BERTScore is calculated between this generated RAG answer and the pre-defined expected answer. This evaluation is critical for validating both the effectiveness of the

context retrieval mechanism and the LLM's ability to synthesize accurate information from potentially diverse, fragmented document chunks, including visual data.

- **Outcome:** The systematic application of BERTScore across both modalities provides quantitative insights into the models' performance, highlighting semantic accuracy and consistency. These metrics are invaluable for guiding iterative improvements in prompt engineering, optimizing embedding models, refining chunking strategies, and potentially fine-tuning the LLMs themselves, ensuring continuous enhancement of the assistants' reliability and precision.

6 Benefits and Impact for King Faisal Specialist Hospital & Research Centre

The King Faisal Specialist Hospital & Research Centre (KFSHRC) LLM Chat Data & Document Assistants are more than just tools; they are powerful assets that will bring big improvements across the hospital:

- **Faster, Smarter Decisions:** With quick, chat-based access to detailed data (like real-time blood trends) from one system, and specific information from documents (like clinic rules) from another, staff can make faster, better choices for things like managing resources, patient care, and daily plans.
- **Much Better Work Efficiency:** These systems let staff skip old, slow ways of finding data and searching documents. This saves a lot of time, allowing healthcare workers to spend more of their valuable time on patient care and important projects instead of searching for info.
- **Knowledge for Everyone:** The easy-to-use chat interfaces mean anyone at KFSHRC can get complex data insights and specialized knowledge, no matter their tech skills. This helps create a more informed team and encourages everyone to use the hospital's data.
- **Stronger Research and Patient Care:** Researchers can quickly find information from many studies using the RAG system. Doctors can quickly check official guidelines and best practices from documents, directly helping them provide care based on solid evidence and support research.
- **More Accurate and Trustworthy Answers:** The answers provided come directly from the data or documents. For document questions, knowing the exact PDF and page number makes the information more trusted and easy to check.
- **Easy to Grow and Ready for the Future:** Both systems are built in a flexible way, making it simple to add more capabilities. New datasets can be put into the data system, and new PDFs (like updated rules or research) can be added and searched in the document system, all without needing big changes or expensive re-training.

7 Conclusion

The King Faisal Specialist Hospital & Research Centre (KFSHRC) LLM Chat Data & Document Assistants represent a pivotal advancement in leveraging artificial intelligence to address the complex information management challenges within a leading healthcare institution. By providing two distinct yet powerful solutions for structured data analytics and cutting-edge Retrieval Augmented Generation for unstructured documents, the systems offer intuitive, efficient, and highly accurate means for KFSHRC personnel to interact with their entire spectrum of institutional knowledge. These comprehensive, dedicated solutions are destined to become indispensable assets, driving unprecedented levels of operational efficiency, fostering deeper data-driven insights, and ultimately contributing significantly to enhanced patient care and the ongoing digital transformation of KFSHRC.