Tiger Analytics

Building a GenAI based Information Retrieval System for a Leading Pharmaceutical Giant

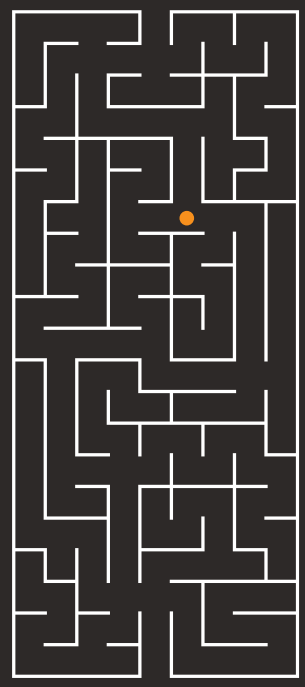
Developed a GenAI-powered retrieval QA system for leading pharmaceutical client, enabling primary market researchers to obtain generated answers and precise document references. The system also incorporates user feedback for continual improvement of conversational accuracy.

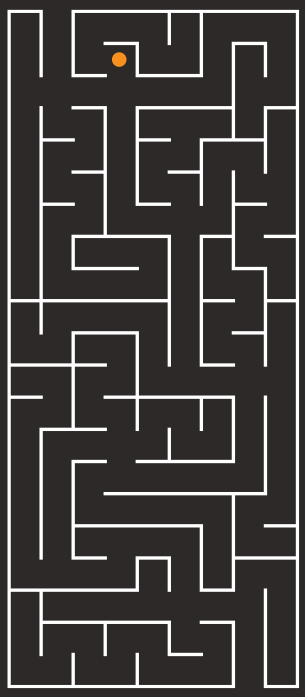
The Background

Our client is a leading global pharmaceutical company headquartered in the United States. The company focuses on researching, developing and commercializing innovative medicines, vaccines, and animal health products.

The market research division of the client engaged Tiger to

- To help on accurate data extraction from research documents..
- To build a mechanism to help researchers getting accurate documents as references based on questions asked.
- To analyze accuracy of tool and improve based on user feedback.





Key Challenges

Numerous challenges across Requirements, Implementation, and Operational phases

/ Requirements – Retrieving documents/information from the Pharma company's document repository was an arduous and time-consuming process based on given keywords. Requirement was to have a chat bot system to automate and generate answers from these market research documents

/ Implementation – System that has accurate retrieval module (for getting most similar references from questions) and leverage LLMs with prompt engineering to generate answers.

/ Operational – System that will extract, preprocess data from source and push into knowledge repository for refresh, RAG pipeline to generate answers and summary and pipeline to store conversations into DB and analyze on dashboard.

Our Solutions

/ The design of the overall solution is segmented into three stages.
- Data extraction, preprocessing and update knowledge repository (in vector store)
- RAG pipeline for retrieval and generation
- Storing topic wise conversation and enabling system as conversational chatbot.

/ Build a retrieval pipeline to get accurate documents and sub-content.
/ Leveraged LLM to generate answers based on given questions and retrieved references.
/ We also leveraged LLM to generate “n” follow-up questions based on the existing conversation.
/ In the pipeline, apart from RAG, we are also including memory where LLM can have enough context for top 3 earlier conversations and based on that it will rephrase the current question and will be passed to RAG pipeline.

Tech Stack

/ AWS SageMaker
/ Amazon S3
/ AWS Lambda
/ Amazon API Gateway
/ AWS Cloud Watch
/ Event Bridge Rule
/ Amazon SNS topics
/ Metric filters
/ DynamoDB
/ Amazon EKS

Functionality wise services we use in application:

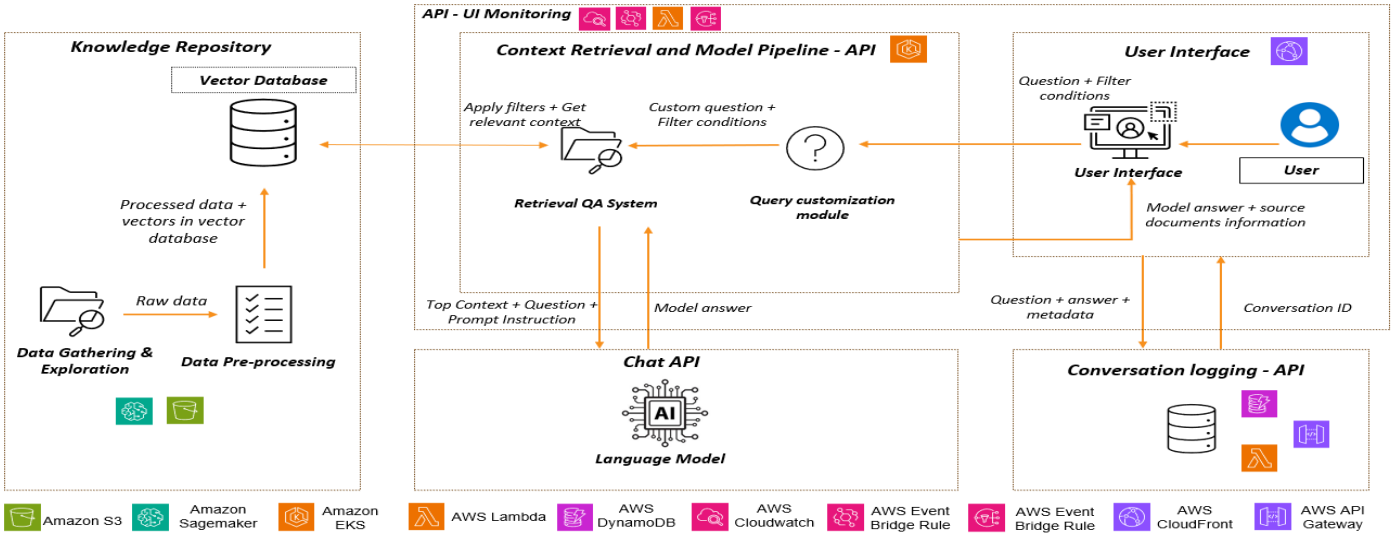
/ Data Storing and preprocessing - S3, Sagemaker
/ RAG pipeline (backend) - LLM, Vector Store, Lambda function, EKS
/ Storing topic wise conversation - Lambda function, API Gateway, DynamoDB
/ System monitoring - AWS Cloud Watch, Event Bridge Rule, SNS topics, Metric filters, Alarms
/ Open Source model deployment - Sagemaker
/ Application deployment (UI) - AWS Cloudfront, EKS

Product Features

/ GenAI based Chat Tool: Users can ask any question related to market research documents and apply filters for accurate references. Users will get a concise summary of the page wise document and answer generated from LLM.
/ Feedback mechanism: Users can give feedback to each conversation they did so far in the system.

/ Recommended questions - User will get top 3 follow-up questions after every conversation which is related to it.
/ Auto-apply filters - We use regex conditions to find which filters should be applied from question to narrow down documents and have accurate retrieval.

Solution Architecture



- The solution is to develop a conversational chat tool where users can ask questions about primary research documents and get answers.
- User query goes to API where it is rephrased based on certain keywords.
- Rephrased query goes to the vector store where the document repository is updated and gets accurate documents out of it.
- Query, and top N document references goes to LLM for generating answer.
- Answer is returned back to the UI for the user and is logged in the database.
- The system API is monitored by AWS Cloud Watch and event bridge rules.

Value Delivered

/ Built a framework that can scale to multiple business units, through specialized chatbots with minimal hallucinations.

/ With 75% acceptance from SMEs, the solution answers user queries on market research documents with a response time of 9 seconds.

/ Commercial Chatbot styled user experience to enable users input query, receive summarized answers, with an option to regenerate answers and feedback mechanism.

About Tiger Analytics

Tiger Analytics is a global leader in AI and analytics, helping Fortune 1000 companies solve their toughest challenges. We offer full-stack AI and analytics services & solutions to help businesses achieve real outcomes and value at scale. We are on a mission to push the boundaries of what AI and analytics can do to help enterprises navigate uncertainty and move forward decisively. Our purpose is to provide certainty to shape a better tomorrow.

Being a recipient of multiple industry awards and recognitions, we have 4000+ technologists and consultants, working from multiple cities in 5 continents.

www.tigeranalytics.com