



International Journal of Research in Academic World



Received: 05/February/2025

IJRAW: 2025; 4(SP3):27-32

Accepted: 19/March/2025

GEN-AI Vault: Framework for Deploying Gen-Ai Applications

^{*1}Dr. Krishnaveni Sakkarapani, ²Vishnuvardhini M and ³Kiruthika R

^{*1}Assistant Professor, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

²PG Student, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

³Research Scholar, Department of Data Analytics (PG), PSGR Krishnammal College for Women, Coimbatore, Tamil Nadu, India.

Abstract

The Gen-AI Vault is a modular, scalable, self-hosted framework for implementing AI-powered applications in the automotive sector. It employs Role-Based Access Control for effective user permission management and secure authentication via JSON Web Token. A document summarizer, a data visualizer, and a chat assistant are the three main AI applications that make up the framework. It is the perfect option for contemporary AI-driven apps because it is built with a React frontend and FastAPI backend, which boost performance, lower latency, and improve scalability.

Keywords: Gen-AI, Role-Based Access Control, JWT, Chat Assistant, Data Visualizer, Document Summarizer.

1. Introduction

As the world becomes increasingly dependent on AI-powered solutions, businesses need scalable and secure architectures to deploy intelligent applications applicable to their domain needs. The automotive industry is very data-intensive, and therefore AI solutions should aid knowledge extraction, data visualization, and document summarization. Gen-AI Vault addresses this need by providing a secure, modular, and self-hosted platform to deploy LLM-based applications with an effective and secure user management system.

Security is one of the critical issues in applications that are AI-based, particularly when it comes to industry-specific data. Gen-AI Vault protects users with JWT, providing a stateless and lightweight way of verifying user credentials. In addition, the use of RBAC is necessary to allow users to access an application and capability based on assigned roles, reducing the risk of unauthorized access as well as data leakage.

The platform offers three primary AI-driven applications that seek to increase productivity in the automobile industry. The Chat Assistant allows users to talk to an AI model that is trained on vehicle data and cancel out irrelevant queries. The Data Visualizer gives users the option to develop interactive dashboards from uploaded data sets, allowing decision-making through data. The Document Summarizer, which uses Mistral and GraphRAG, extracts essential points from lengthy documents, minimizing time and effort in data acquisition.

For optimal performance, the project was initially developed using Flask but later shifted to FastAPI due to its improved asynchronous capabilities and lower response times. The capability of FastAPI to handle simultaneous requests

***Corresponding Author:** Dr. Krishnaveni Sakkarapani

seamlessly and its seamless integration with WebSockets made it the optimal choice for real-time AI applications, offering smooth and responsive frontend-backend interactions.

2. Literature Review

Coming in industries different from its own with applications that range from taking charge of smart chatting to teaching interactive dashboards in data visualization is the specialty of Generative AI (GEN-AI). GEN-AI VAULT is an end-to-end solution that encompasses everything from Mistral-based chatbot applications, a retrieval-augmented generation platform, interactive dashboards for CSV and Excel data visualization, and a new document summarization engine based on Mistral and GraphRAG. These applications automate workflows, support decision-making, and allow swift and sensible adoption of AI across industries. As organizations in the broadest sense leap onto the AI-tools bandwagon, it becomes a pressing disadvantage for every sector to lose views of GEN-AI apps regarding innovation, productivity, and competitiveness to massive data.

Implementation of Mistral coupled with RAG in a chatbot exploits expectations of human behaviour to smartly deliberate conversation management solutions. Real-time data access is offered by Mistral's architecture, enabling context-aware interactions with rapid response. Such a hybrid combination is exceedingly beneficial in user cases that require instantaneous, precise, and personalized responses so as to maximize user experience within the domain involved (William H Overholt, 2018). Data visualization applications

for generative AI convert the input of raw data into interactive dashboards, allowing the end-user to see some unforeseen trends and patterns. Research reveals AI visualization platforms enhance analytical competency and usability of workflow, hence better decision (Dhoni P, 2023) [14]. This truly proves to be a precious asset where actionable intelligence is essential to the strategic decision-making process.

Document summarization through Mistral and GraphRAG is primarily developed to deal with this concern of coping with huge volumes of text information by compressing it to the least relevant minimum information. This enhances the usability of information in the fields of medicine and education. Summarization techniques based upon AI largely bolster clinical decision support systems by providing timely and succinct information to patients and medical personnel (Nia MF *et al.*, 2025) [15]. In education, it enhances integrity and motivation among students (Namoun A *et al.*, 2024) [16]. Thus, as AI moves, the processes of AI-based summarization are gaining traction, putting more and more people in a position to understand and act upon this information.

GEN-AI is inventively marking the paradigm shift in industries by bettering communication, analysis of data, and information management. Along with the enhancement of algorithms, Mistral and GraphRAG fosters ease of use in the healthcare, marketing, and education industries. Certain studies reveal that in diagnostics, marketing, and user experience, there is an ever-increasing usage of GEN-AI (Grewal D *et al.*, 2024) [17]. While the benefits arising from these converging innovations are enormous, the complexities of AI implementation cannot be compromised if organizations want to reap its full benefits in the constantly changing digital landscape.

3. Methodology

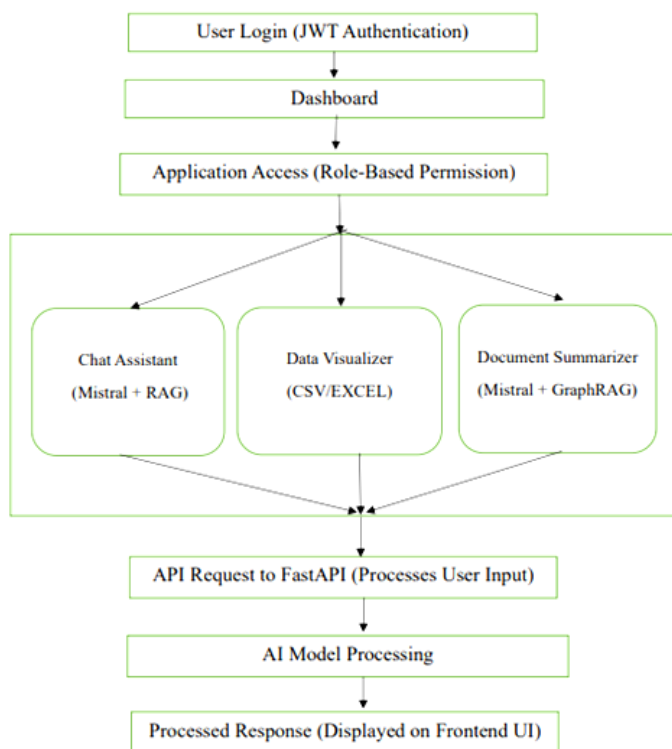


Fig 1: Process Flow Diagram

The figure 1 indicates the process flow and is descriptive of the task as done in its sequential flow in this project. The process shown is a complete representation of how the

different elements interact: user authentication to answer generation, based on the inputs from the users and processed through AI techniques.

3.1. System Workflow

Frontend (React + Axios): The necessary frontend development is done through React to have an interactive and responsive interface for users. The benefit includes secure authentication, feature choice, file upload as well as visualizing results. The front-end uses Axios for API calls to the backend for smooth communication.

Backend (FastAPI + AI Models + FAISS): The backend developed in FastAPI is responsible for accepting user requests, authentication, and communication with AI models. JWT authentication is used for secure access and role-based authorization functionality based on user roles limitation. The AI models like Mistral + RAG are applied for chat support and Mistral + GraphRAG for document summarization with use of FAISS for efficient retrieval of relevant data. This is a straight workflow of the system showing how user input flows right from authentication to response generation by AI.

i). User Login (JWT Authentication)

User authentication starts the system. Only an end-user authenticated will be able to access this computing platform. The credentials are filled by users in the Login Page, from which the system authenticates them by means of JWT (JSON Web Token). This enables stateless authentication because it does not require the server to maintain session information, all of which is carried in the token.

Once authentication is successful, it issues a JWT token, which is sent to the browser of the user. This token is then expected for all the subsequent API calls made by the user, thus eliminating the need for repetitive logins. Re-authentication avoids unauthorized entry since it has a token expiration and a refresh mechanism attached to the whole authentication system.

Once found a token invalid or expired, will log out the user and reroute to a login page. This process of secure authentication ensures that all user activity is kept safe from unauthorized access attempts.

ii). Dashboard

After having logged in successfully, the users enter the Dashboard, where the most important platform for navigating the platform features resides. Designed for interactivity and intuitive user experience allows users to access AI-powered tools, upload files, and use AI-powered applications.

The dashboard automates the view according to user roles, hence showing an AI application only to which the specific user has privilege. For example, an admin can use all AI applications while another user will be restricted to basic AI applications such as the Chat Assistant and Data Visualizer. Role-based access control ensures that important functions remain under the purview of permitted users only.

The user interface has indicators on the status of various AI processing tasks, such as document summarization or data visualization. The frontend is smart and responsive, giving users a lag-free experience while interacting with different applications.

iii). Application Access (Role-Based Permission)

RBAC system is used for permission management per user. The backend assigns roles to the user after logging in, therefore targeting which AI applications are open for his or

her use. This function thus denies all unauthorized users access to restricted AI applications.

For example:

- A default User can use only the assigned AI applications as per their profile. A team leader can use Document Summarization and Data Visualizer.
- An administrator has access to all AI applications like Document Summarization, Chat Assistant, and Data Visualizer.

The RBAC makes sure that in the frontend, unauthorized AI applications do not show up on user's dashboards; in the backend, it authenticates user's roles and denies unauthorized access. This security scheme offers more protection to data and obedience to access regulations stated in the organization.

iv). AI Apps Interaction

- Chat Assistant (Mistral + RAG):** The interface of the chat assistant offers basically an online forum to users to put their questions. From this point on, the user's input goes to WebSocket, which transports the command to the backend, where FastAPI processes this, transfers it to the Mistral + RAG AI model that attempts to obtain context-based answers via FAISS vector database and then sends back the response that takes note of the previous conversation for its embedded memory and ultimately sends the processed output to Frontend to be displayed on the interface chat window. The work of WebSocket here mainly promotes real-time communication within this seamless chat experience.
- Data Visualizer (CSV/Excel Dashboards):** The uploads of structured data in CSV/Excel format can be done by users from the UI. Data transformation techniques after coming into the backend will clean the data and structure it. Thereafter dashboards are thrown up using visualization libraries like Plotly and put up for the user to view insights graphically. This is of great interest for data analysts and business folks, allowing them to explore trends with minimal technical involvement.
- Document Summarizer (Mistral + GraphRAG):** Users upload documents via the PDF/TXT document upload feature to fastapi, which passes ahead the analysis and extraction of key insights to the backend Mistral + GraphRAG AI model. The AI model understands the text context and summarizes its important points into a shorter text. The user will view this on the UI for fast understanding without actually reading the entire document. The status and tracking of document processing will get updated on real time through WebSockets in the system.

v). API Call to FastAPI (Processes User's Input Data)

Once the AI application is selected from the front-end interface, an API call to the FastAPI back end is made. This is just a process handler that checks the request and authenticates the user before sending the request to the actual AI model.

The request carries structured data, including:

- JWT token-authenticating the user.
- The particular AI application selected-chat assistance, dataviz, or summary of documents.
- Input data, which could be a text query, uploaded file, and the like.

Besides, these features-fast-and-efficient-with-asynchronous

request processing allowing multiple simultaneous servicing of users, thus eliminating any delays from the system processing API requests to FastAPI.

vi). AI Model Processing

A request that is well validated will route by the backend on to the appropriate AI model for processing according to the desired feature:

Chat Assistant (Mistral + RAG): A human query is forwarded to Mistral-an optimal AI model for conversational responding. The system is thus enhanced with the Retrieval-Augmented Generation (RAG) to ensure more accurate responses from a contextually digested answer. This is where the FAISS (Facebook AI Similarity Search) Vector Database comes to play. FAISS holds previous interactions and embedding knowledge that allows the model to complement similar past responses in enhancing the current. This guarantees the correctness of the fact-based and context-based responses of FAISS, thus enhancing the productivity of the AI assistant.

Data Visualizer (CSV/Excel Dashboards): For the data visualisation part, the backend shall clean and shape the uploaded CSV or Excel files and, by using visualisation libraries like Plotly, shall produce the charts, graphs, and statistical insights. Since this functionality does not need historical context or retrieval-based AI, it won't use FAISS, either.

Document Summarizer (Mistral + GraphRAG):

The document summarization feature concerning PDF or TXT files is processed by the Mistral model with GraphRAG. FAISS is clever because it is handy for retrieving segments from past documents together with all similar contents stored within its vector database-quickly able to compare the current document with the embeddings kept in storage and derive key insights. This entirely different approach to summary generation assigns a better contextual understanding, thus accomplishing high accuracy and brevity in summarization. Designed in the spirit of doing high efficiency and accuracy optimization for immediate users with specific and meaningful outputs, the AI models are made; meanwhile, FAISS addition improves the live nature of response retrieval and thus makes the whole interaction smarter.

vii). Displaying The Response On Frontend UI

Once the AI model processes the request, it then sends the response to the frontend through WebSocket for real-time update for Chat Assistant and Document Summarizer. The frontend dynamically renders the response by progressively updating the UI. Here, the conversation mode applies.

4. Process Visualization

The Gen-AI Vault framework securely integrates AI-powered applications into authentication and role-centric access control. The below results demonstrate what the system can do.

- **Login Page:** The login organization uses JWT for secure user identity, with no chance for unauthorized access.
- **Dashboard:** The dashboard is a central place that dynamically displays AI applications based on user roles, with one-click access to use and control.
- **Chat Assistant:** This AI assistant powered by Mistral model and RAG allows users to interact in a domain-specific way and capture insights from the uploaded documents.
- **Data Visualizer:** The system automatically creates

dashboards using the uploaded CSV or Excel worksheets, making it easier for users to interrogate and analyze datasets.

- **Document Summarizer:** With the aid of GraphRAG, the

summarization tool leverages key insights derived from the uploaded documents to present them in a digestible manner, thereby increasing efficiency and productivity.

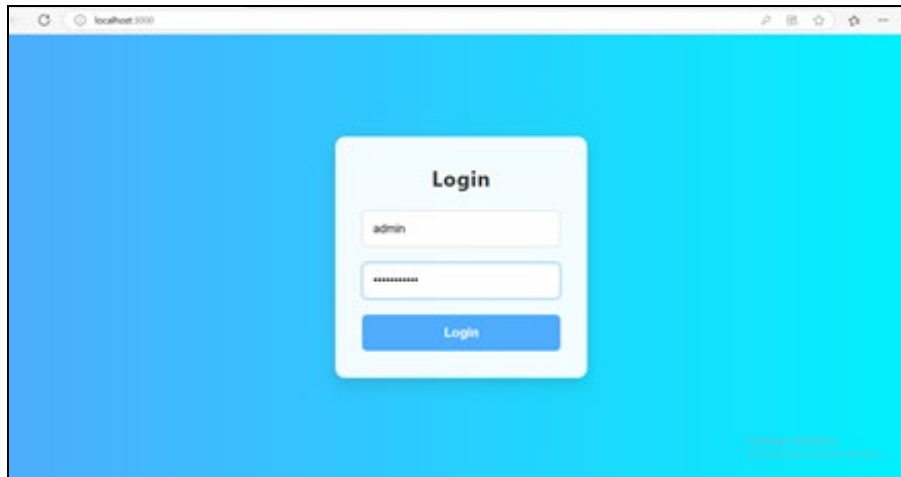


Fig 2: Log-in Page

The above figure 2 Log-in Page acts as the entry point for users who enter a username and password to log in to the system. The system verifies the credentials using JWT authentication for secure access. Upon success, users are

authorized based upon their role and permissions. Based on the authentication, the user will be redirected to the Dashboard for other interactions.

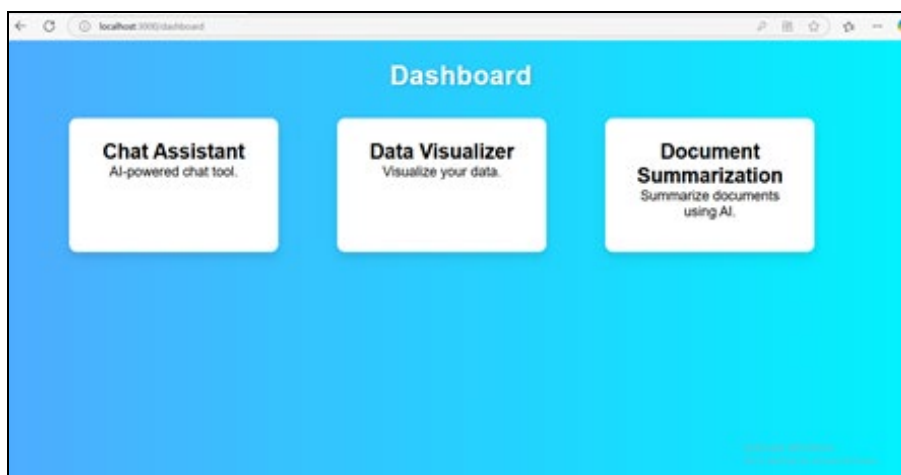


Fig 3: Dashboard

The following figure 3 Dashboard shows the main interface through which users utilize the various AI-powered features based on their roles and permissions. It provides a very user-friendly UI for navigation. Users can select Chat Assistant,

Data Visualizer, or Document Summarizer. The dashboard dynamically updates the available applications and enhances the user experience by integrating Role-Based Access Control.

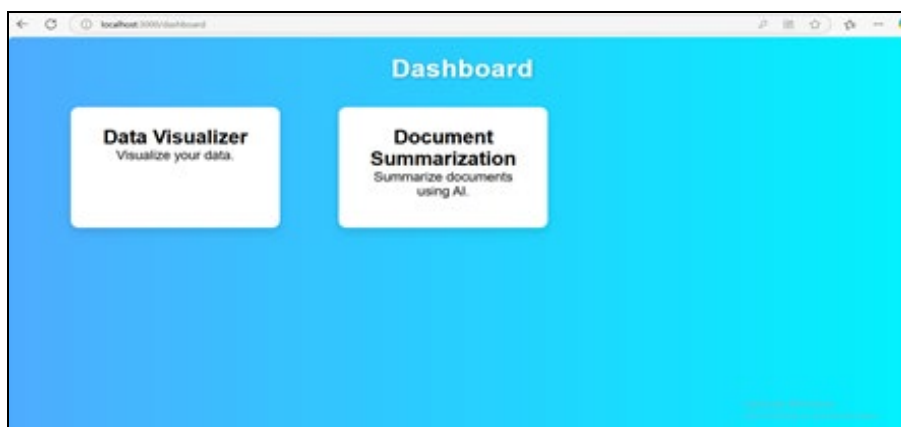


Fig 4: Accessible Application for "Team Leader" User

Dashboard view for the Team Leader role user shown in Figure 4. As per the Access Control mechanism, only applications and features allowed for the Team Leader to access are rendered visible. The system implements the role-

based access control (RBAC) policy that allows a user to have access and control only over those applications the user has permission to interact with so as to protect sensitive information and comply with the exigencies of law.

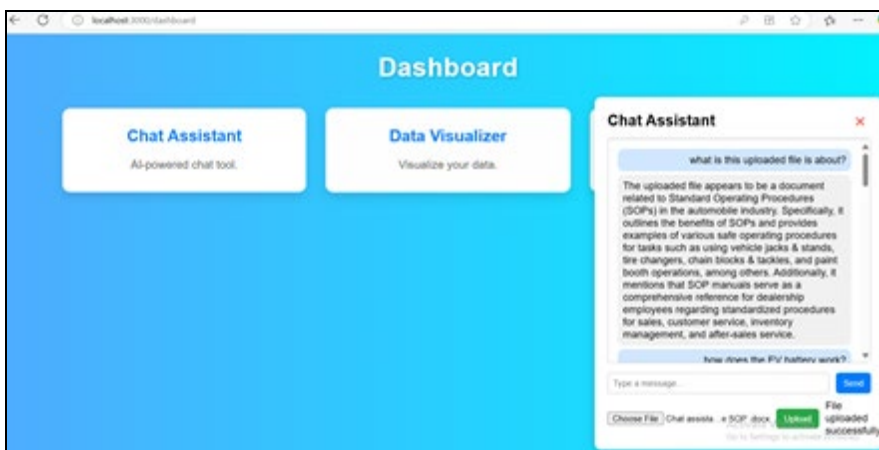


Fig 5: Output of Chat Assistant when the user asks a query about an uploaded file

This Figure 5 shows the output of the Chat Assistant when the user asks a query related to an uploaded file. The system processes the query using Mistral + RAG, pulling relevant information up from the document stored in the FAISS Vector Database. The AI generates

meaningful insights and formulates an accurate answer based upon the content of the file. The answer is then displayed in the chat dialog box so users can interactively engage with the document.

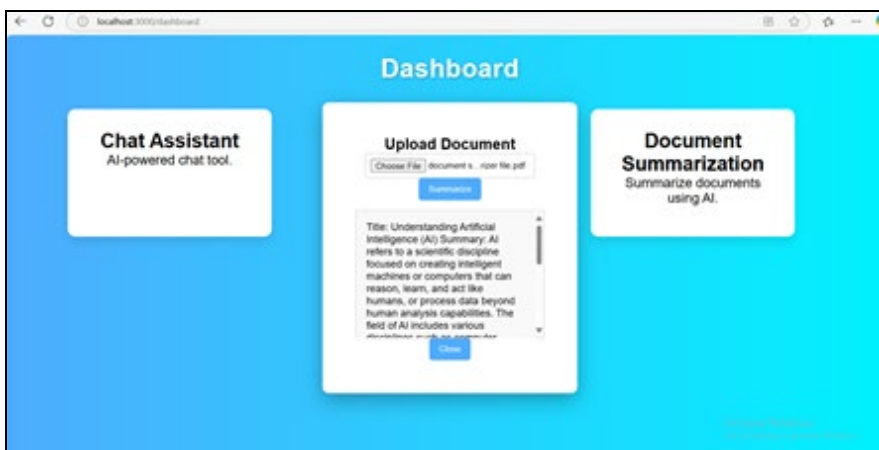


Fig 6: Output of Document summarizer

This Figure 6 is the final output summarized in the Document Summarizer application. After processing the uploaded file, Mistral + GraphRAG extracts key insights and creates a terse summary. The summarized text is displayed on the UI, and

therefore, users can make quick sense of the key information from long documents. Such features offer significant time savings by presenting well-structured and relevant insights to the users.

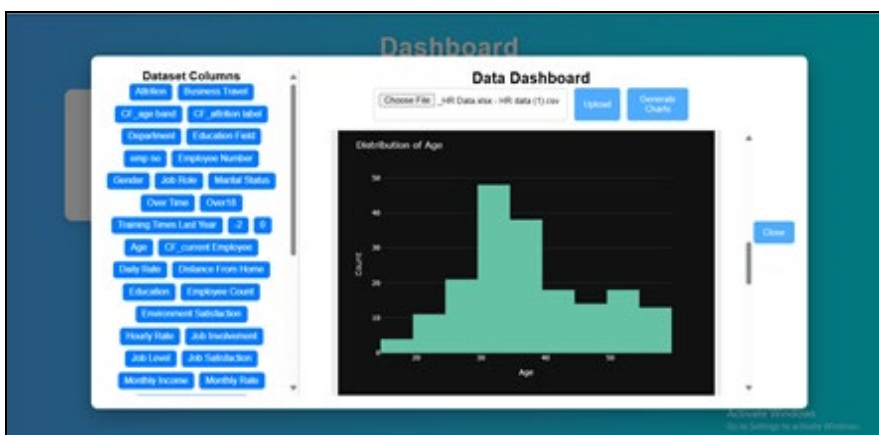


Fig 7: Chart generated on distribution of Age

This Figure 7 consists of a histogram depicting the age distribution in the sample set. The histogram presents frequency counts, that is, how many persons of a certain age range exist in the sample population. By fixing the age data into bins, the histogram is useful to identify patterns such as what age range gets its highest frequency or whether the dataset trends toward normality. All in all, this visualization serves as a means for performing demographic analysis and raising age-related insights for decision-making.

Conclusion

The use of AI technology has now become much more efficient, safer, and more usable. AI tools such as Chat Assistant and Document Summarizer have made the use of technology superior. This has been largely made possible by using seamless and responsive user experience with FastAPI for authentication and through WebSockets for real-time interactions. This also helps in fetching all the past interactions and delivers context-aware responses for the system. By such an approach, performance optimization is done, and the cost will be considerably low when compared to the traditional database techniques to make the system lightweight and scalable.

Security and access mechanisms are central to keeping a system's integrity. The framework comprises efficient JWT authentication and role-based access, and an API Gateway handling all incoming requests with request validation to create a solid framework keeping unauthorized access at bay while ensuring smooth communication across the frontend and backend. Processing of an AI model using frameworks like Mistral and GraphRAG would also add similar quality to the responses generated and a more intelligent and adaptive assistant. All the processed responses, then, will be reflected on the frontend UI so that real-time responses are derived towards accurate and relevant information for the user.

This overall paper is an indication of the fact that an AI solution could possess the ability to dramatically enhance user interaction, security, and efficiency. FastAPI, FAISS, and the WebSockets put together with advanced AI models have hence built a scalable framework for real-time processed AI applications. The future will see improvements in things like multi-vector database support and sophisticated natural language understanding. There will also be development towards tight partnerships with third-party application through ongoing work as AI develops into the near future.

References

- Rahul C. Basole, Timothy Major, Generative AI for Visualization: Opportunities and Challenges, *IEEE Computer Graphics and Applications*. 2024; 44(2):55-64.
- Ayodeji Ismail Moshood, Zoe Jeffrey, An In-Depth Approach to Strengthening Security in Open-Access Libraries Utilizing JSON Web Tokens (JWT), *International Journal of Recent Technology and Engineering (IJRTE)*. 2025; 13(5):14-19.
- Yilmaz Uygun, Victor Momodu, Local large language models to simplify requirement engineering documents in the automotive industry, *Production & Manufacturing Research*. 2024; 12(1):1-34
- Diash Firdaus, Idi Sumardi, Yuni Kulsum, Integrating Retrieval-Augmented Generation with Large Language Model Mistral 7b for Indonesian Medical Herb, *JISKA (Jurnal Informatika Sunan Kalijaga)*. 2024; 9(3):230-243
- Tiberiu Boros, Radu Chivoreanu, Stefan Daniel Dumitrescu, Fine-Tuning and Retrieval Augmented Generation for Question Answering Using Affordable Large Language Models, *The Third Ukrainian Natural Language Processing Workshop (UNLP)*, 75-82
- Alexander Tobias Neumann, Yue Yin, Sulayman Sowe, An LLM-Driven Chatbot in Higher Education for Databases and Information Systems, *Ieee Transactions On Education*. 2025; 68(1):103-116.
- Neumann T. *et al.*, Chatbots as a tool to scale mentoring processes: Individually supporting self-study in higher education, *Front. Artif. Intell.* 2021; 4:64-71.
- Md. Shahidul Salim A, Sk Imran Hossain A, Tanim Jalal A, Dhiman Kumer Bose, LLM based QA chatbot builder: A generative AI-based chatbot builder for question answering, *Software X* 29 102029|December 2024
- Yunfan Gao, Yun Xiongb, Xinyu Gao, Retrieval-Augmented Generation for Large Language Models: A Survey, *arXiv preprint arXiv:2312.10997v5* |Mar 2024
- Darren Edge, Ha Trinh, Newman Cheng, From Local to Global: A GraphRAG Approach to Query-Focused Summarization, *arXiv preprint arXiv:2404.16130v2*|Feb 2024
- Ayush Thakur, Raghav Gupta, Introducing Super RAGs in Mistral 8x7B-v1, *arXiv preprint arXiv:2404.08940v1* |Apr 2024
- Sonal Prabhune, Donald J. Berndt, Deploying Large Language Models With Retrieval Augmented Generation, *arXiv preprint arXiv:2411.11895v1*|Nov 2024
- Jens Kohl, Luisa Gloger, Rui Costa, Otto Kruse, Generative Ai Toolkit-A Framework For Increasing The Quality of Llm-Based Applications Over Their Whole Life Cycle, *arXiv preprint arXiv:2412.14215v1*|Dec 2024
- Pan Dhoni, Exploring the Synergy between Generative AI, Data and Analytics in the Modern Age| August 2023
- Masoumeh Farhadi Nia, Mohsen Ahmadi, Elyas Irankhah, Transforming dental diagnostics with artificial intelligence: advanced integration of ChatGPT and large language models for patient care, *Frontiers in Dental Medicine*, 2025, 5.
- Abdallah Namoun, Isa Ali Ibrahim, Ehaz Mustafa, Ahmed Alrehaili, Ali Tufail, Junaid Shuja, Kashif Bilal, *et al.*, Generative artificial intelligence in education: an umbrella review of applications and challenges, *Research Square*, 2024.
- Dhruv Grewal, Cinthia B. Saturnino, Thomas H. Davenport, Abhijit Guha, How generative AI Is shaping the future of marketing, *Journal of the Academy of Marketing Science*, 2024.
- Lifeng Shang, Zhengdong Lu, Hang Li, Neural Responding Machine for Short-Text Conversation, In: *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing*. 2015; 1:1577-86.
- Shibi, Krithick, R. Kingsy Grace, M. Sri Geetha., Abstractive Summarizer using Bi-LSTM. In *2022 International Conference on Edge Computing and Applications (ICECAA)*, 2022, 1605-1609.
- Sangita Pokhrel, Swathi Ganesan, Tasnim Akther, Lakmali Karunarathne, Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit, *Journal of Information Technology and Digital World*. 2024; 6(1):70-86.