ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

Enhancing IVF Patient Care Using an AI-Powered Chatbot with Retrieval-Augmented Generation (RAG)

Rajitha Maduri[1]; Venkata Siva Gatta[2]; Naveen Kumar[3]; Bharani Kumar Deparu[4]; Sreeja Deparu[5]; Bhargavi Depuru[6]; Mukesh Marwade[7]; Gayathri K[8]

[1,2,3]Research Associates, [4,5,6]Directors, [7]Project Mentor, [8]Team Lead
[1,2,3,4,5,6,7,8] AiSPRY, Hyderabad, India

Publication Date: 2026/04/20

Abstract: Infertility impacts a significant number of couples globally, and patients undergoing In-Vitro Fertilization (IVF) frequently require continuous access to accurate and reliable medical guidance. However, availability of healthcare professionals is often limited beyond clinical hours. This study presents an AI-driven IVF Patient Support Chatbot designed to assist patients by delivering timely and context-aware responses throughout their treatment journey.

The proposed system combines retrieval-based information access with advanced language generation capabilities to improve response reliability. It utilizes Groq's LLaMA 3.3 70B model for generating human-like responses, while semantic understanding is achieved through Sentence Transformers (all-MiniLM-L6-v2). FAISS is used as a vector database to efficiently store and retrieve IVF-related knowledge. This integrated approach ensures that responses are both relevant and grounded in domain-specific information.

Keywords: *IVF Chatbot, Retrieval-Augmented Generation (RAG), LLaMA 3.3 70B, Sentence Transformers, FAISS, Healthcare AI, Patient Support Systems, NLP in Healthcare.*

How to Cite: Rajitha Maduri; Venkata Siva Gatta; Naveen Kumar; Bharani Kumar Deparu; Sreeja Deparu; Bhargavi Depuru; Mukesh Marwade; Gayathri K (2026) Enhancing IVF Patient Care Using an AI-Powered Chatbot with Retrieval-Augmented Generation (RAG). *International Journal of Innovative Science and Research Technology*, 11(4), 1068-1077. https://doi.org/10.38124/ijisrt/26apr670

I. INTRODUCTION

The adoption of artificial intelligence in healthcare has accelerated in recent years, particularly in enhancing patient support systems and improving operational workflows. IVF treatment involves multiple stages that require continuous communication between patients and healthcare providers, including guidance on procedures, medications, scheduling, and emotional support.

Despite this need, IVF clinics often face difficulties in managing a high volume of patient queries. Patients frequently seek immediate clarification regarding treatment steps and medication usage. Traditional communication methods such as phone calls and emails can result in delayed responses and increased administrative burden.

To address these challenges, this research proposes an AI-Powered IVF Patient Support Chatbot that leverages Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs) to provide accurate, context-aware responses. The system integrates Groq's LLaMA 3.3 70B as the LLM, Sentence Transformers (all-MiniLM-L6-v2) [2] for embedding generation, and FAISS as the vector database for efficient semantic retrieval of IVF-related knowledge.

The development of the chatbot follows the CRISP-ML(Q) methodology, which provides a structured approach for building machine learning systems with quality assurance.

➢ *CRISP-ML(Q) Methodology:*

- *Business Understanding:*

In this work, the problem is approached from an operational perspective rather than a theoretical one. IVF clinics receive a high number of repeated queries, which slows down response time. The objective here is to reduce that dependency by introducing a system that can handle routine questions instantly while maintaining accuracy. The chatbot aims to improve patient satisfaction, reduce staff workload, and enhance operational efficiency in IVF clinics.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

- *Data Understanding:*

Instead of treating data as static documents, the information is analyzed based on how frequently patients interact with it. Priority is given to treatment steps, medication instructions, and commonly asked questions to ensure relevance during response generation.

Exploratory analysis is performed to understand the structure of the documents, identify relevant content, and determine how the information can be effectively used to support chatbot responses.

- *Data Preparation:*

The collected content is not used in its raw form. It is simplified, segmented, and organized into smaller meaningful units so that the system can quickly identify and retrieve only the required portions during interaction.

Each text chunk is converted into numerical vector representations using **Sentence Transformers (all-MiniLM-L6-v2) [2]** embeddings. These embeddings capture the semantic meaning of the text and are stored in **FAISS**, which serves as the vector database for efficient similarity search.

- *Model Building:*

The implementation focuses on combining retrieval with generation in a practical way. Rather than generating answers directly, the system first identifies supporting information and then constructs responses using that context.

When a user submits a query, the system converts the query into an embedding vector and retrieves the most relevant document segments from the vector database. These retrieved contexts are then passed to the Groq's LLaMA 3.3 70B Large Language Model, which generates a coherent and context-aware response.

- *Model Evaluation:*

Evaluation is carried out by observing how the system behaves in real scenarios, including response clarity, relevance, and speed, rather than relying only on theoretical metrics.

Experimental results indicate that the chatbot achieves more than 95% response accuracy while maintaining response times of less than 30 seconds.

Model Deployment: The system is deployed in a way that supports continuous interaction, ensuring users can access responses without delay while maintaining consistent performance under multiple queries.

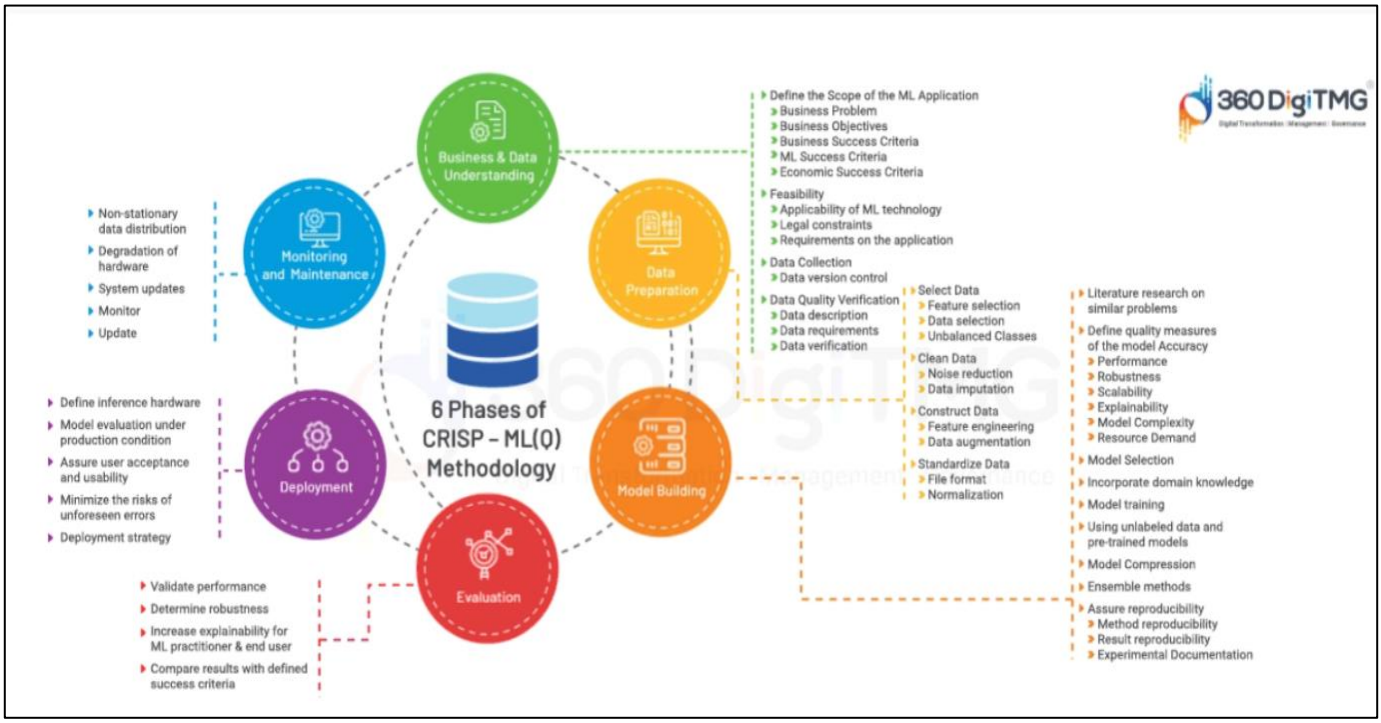

Fig 1 CRISP-ML(Q) Methodology

II. BACKGROUND AND MOTIVATION

IVF clinics experience high volumes of repetitive patient queries regarding treatment procedures, medication instructions, and appointment management. Manual handling of these queries leads to delayed responses and operational inefficiencies.

AI-based conversational systems can automate routine patient interactions, reduce staff workload, and improve patient satisfaction. The integration of RAG frameworks allows chatbots to retrieve relevant domain knowledge before generating responses, making them highly suitable for healthcare applications where accuracy is critical.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

III. SYSTEM ARCHITECTURE

The proposed AI-Powered IVF Patient Support Chatbot is designed using a modular architecture that integrates natural language processing, vector-based semantic retrieval, and large language models to provide accurate responses to patient queries. The system architecture consists of several components including the knowledge base, document ingestion pipeline, embedding generation module, vector database, LLM inference engine, and user interface.

The architecture enables efficient retrieval of IVF-related knowledge and ensures that responses are generated based on relevant medical information. Figure 2 illustrates the overall architecture of the chatbot system.

➢ *Knowledge Base*

The knowledge base forms the foundation of the chatbot system and contains IVF-related information collected from medical guidelines, treatment documentation, and patient support materials.

These documents provide essential information regarding:

- IVF procedures and treatment stages
- Medication instructions and dosage guidance
- Appointment scheduling and clinical processes
- General fertility treatment information

The collected documents are processed and stored in a structured format to enable efficient retrieval during chatbot interactions.

➢ *Natural Language Processing (NLP)*

The chatbot utilizes Large Language Models (LLMs) for understanding patient queries and generating meaningful responses.

In this system, the Groq's LLaMA 3.3 70B model is used as the primary LLM for conversational response generation. The model processes patient queries and combines them with retrieved contextual information to generate accurate and coherent responses.

The use of LLMs enables the chatbot to handle natural language queries and provide conversational responses that improve patient engagement.

➢ *Embedding Generation*

Text is converted into a representation that reflects meaning rather than exact wording, allowing the system to match queries based on intent instead of keywords.

To enable semantic search, the system converts text documents into numerical vector representations using Sentence Transformers (all-MiniLM-L6-v2) [2].

Embeddings capture the semantic meaning of text and allow the system to identify relevant document segments based on similarity with user queries. Both the knowledge documents and patient queries are converted into embeddings to enable efficient information retrieval.

➢ *Vector Database*

The generated embeddings are stored in FAISS, which functions as the vector database for the chatbot.

FAISS enables fast similarity search by comparing query embeddings with stored document embeddings. When a patient submits a query, the system retrieves the most relevant IVF knowledge documents using vector similarity search.

This retrieval process ensures that the chatbot responses are grounded in domain-specific knowledge.

➢ *Retrieval-Augmented Generation (RAG) Pipeline*

The chatbot uses a Retrieval-Augmented Generation (RAG) framework to improve response accuracy. The RAG pipeline consists of the following steps:

- Patient query input
- Query embedding generation
- Similarity search in FAISS
- Retrieval of relevant document chunks
- Context injection into LLM prompt
- Response generation using Groq's LLaMA 3.3 70B

This approach ensures that the LLM generates responses based on verified IVF knowledge rather than relying solely on its internal knowledge.

➢ *User Interface*

The chatbot is accessible through a web-based interface that allows patients and clinic staff to interact with the system easily.

Key interface features include:

- Patient chat interface
- Query input system
- Conversation history
- Admin monitoring tools

The interface provides a user-friendly environment for submitting queries and receiving responses in real time

.

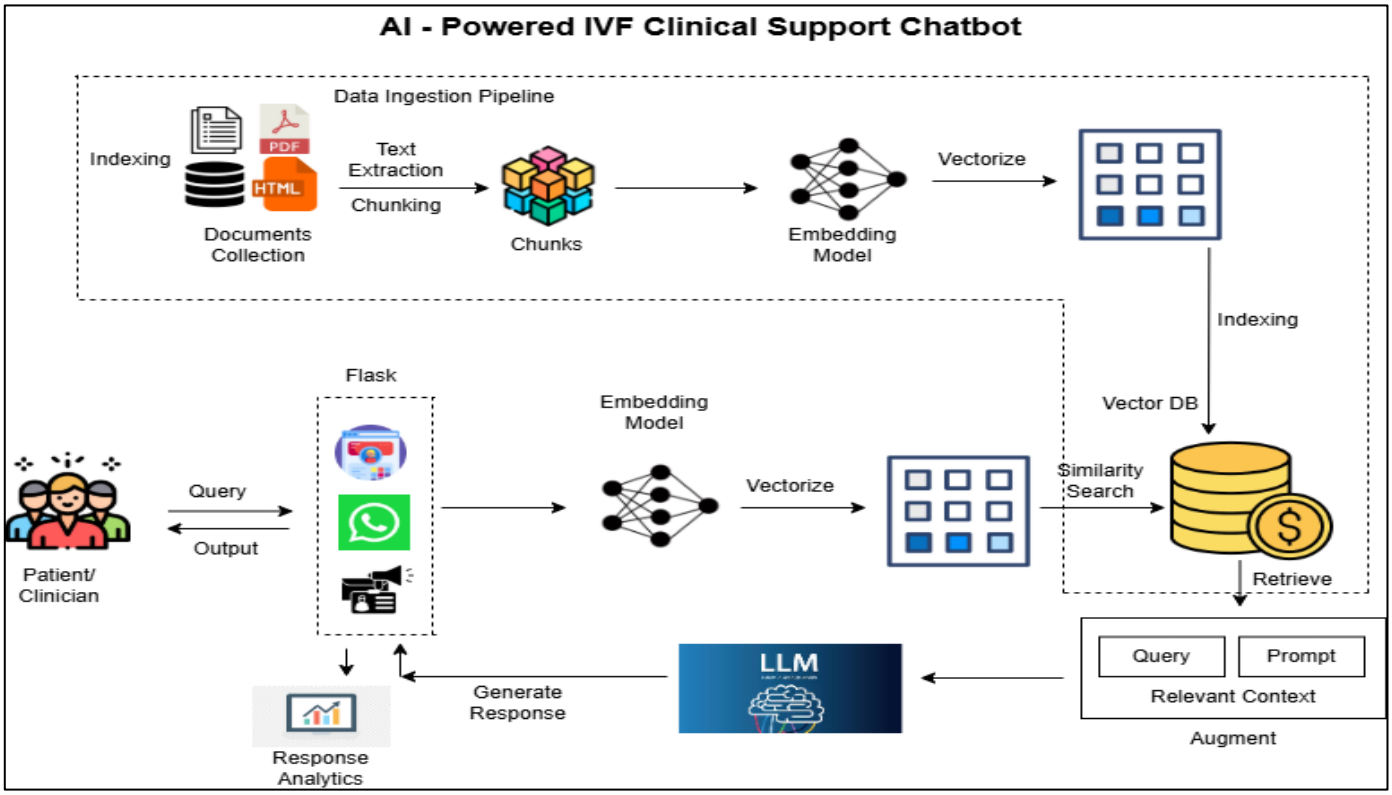

Fig 2 Architecture of AI-Powered IVF Patient Support Chatbot

IV. METHODOLOGY

The development of the AI-Powered IVF Patient Support Chatbot follows a structured methodology to ensure accuracy, scalability, and reliability. The system integrates Retrieval-Augmented Generation (RAG) [1] with Large Language Models to provide context-aware responses to patient queries. This section describes the key stages involved in building the chatbot system.

➢ *Data Collection*

The data used in this system is selected based on its usefulness in real patient interactions rather than just availability. Only relevant IVF-related content is included to improve response quality.

The first stage involves collecting domain-specific knowledge related to IVF treatment and patient care. The knowledge base includes medical guidelines, IVF treatment documentation, fertility clinic resources, and patient support materials.

These documents provide essential information regarding:

- IVF treatment procedures
- Medication instructions
- Treatment timelines
- Patient care guidelines
- Frequently asked IVF questions

The collected data serves as the primary knowledge source for training and supporting the chatbot responses.

➢ *Data pre-processing*

Instead of applying generic preprocessing, the focus is on making the data easier for the system to interpret and retrieve. This includes simplifying language, organizing content, and removing unnecessary complexity.

After data collection, the documents undergo several preprocessing steps to prepare them for semantic retrieval.

The preprocessing steps includes:

- Text cleaning and normalization
- Named entity recognition
- Synonym mapping
- Intent labelling
- Handling out-of-scope queries
- Data augmentation
- Noise handling
- Response formatting
- Safety and filtering layer

Large documents are divided into smaller segments using text splitting techniques.

These smaller chunks help improve retrieval efficiency and maintain contextual relevance during response generation

➢ *Embedding Generation*

Text is converted into a representation that reflects meaning rather than exact wording, allowing the system to match queries based on intent instead of keywords.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

To enable semantic search, the processed text chunks are converted into numerical vector representations known as embeddings.

The chatbot uses Sentence Transformers (all-MiniLM-L6-v2) [2] to generate embeddings for both documents and user queries. These embeddings capture the semantic meaning of the text and allow the system to retrieve the most relevant information based on similarity search.

The embeddings generated from the IVF knowledge documents are stored in a vector database for efficient retrieval.

➢ *Vector Database Indexing*

The processed information is stored in a structured way that allows quick comparison and retrieval, ensuring that only the most relevant data is used for each query.

The generated embeddings are stored and indexed in **FAISS**, which acts as the vector database for the chatbot system.

FAISS enables fast similarity search by comparing query embeddings with stored document embeddings. When a user submits a query, the system searches the vector database to retrieve the most relevant IVF-related information.

This retrieval mechanism ensures that responses are grounded in domain-specific medical knowledge.

➢ *Model Integration*

The language model is used as a response generator guided by retrieved information, ensuring outputs remain aligned with actual data instead of assumptions.

The chatbot integrates a Large Language Model (LLM) for generating conversational responses. In this system, Groq's LLaMA 3.3 70B is used as the core language model.

The model receives the user query along with the retrieved context from the vector database.

Based on this information, the LLM generates a response that is both context-aware and medically relevant.

This integration ensures that the chatbot can handle natural language queries and provide meaningful answers to patients.

➢ *Response Generation using RAG*

The response process follows a guided approach where retrieved information is combined with the user query before generating the final answer, improving both accuracy and clarity.

The chatbot follows a Retrieval-Augmented Generation (RAG) workflow to generate accurate responses.

The process includes the following steps:

- Patient submits a query through the chatbot interface
- Query is converted into an embedding vector
- Similarity search is performed in FAISS
- Relevant IVF knowledge documents are retrieved
- Retrieved context is combined with the user query
- The prompt is sent to the Groq's LLaMA 3.3 70B model
- The model generates a context-aware response

➢ *Model Evaluation*

The chatbot system is evaluated based on its ability to provide accurate and timely responses to patient queries.

Key evaluation metrics include:

- Response Accuracy: Measures correctness of chatbot responses
- Response Time: Time taken to generate answers
- User Satisfaction: Improvement in patient experience
- Operational Efficiency: Reduction in staff workload

Table 1 Model and Embeddings Evaluation

Model Name	Link	Use	Advantages	Disadvantages	Findings
Mistral-7B-Instruct-v0.2	mistralai/Mistral-7B-Instruct-v0.3 · Hugging Face	Generation	Very good Quality.	Heavy for local machine.	This model is very good model compared to all the models i have tried and this model is heavy but the responses is great and response very quality but when i run it in my local machine it is giving me answers in 1 min avg but when i am accessing it from Huggingface API Key it is working very fastly. the response time is min 5sec to max 30 sec. that is great for a ivf model.
microsoft/Phi-3-mini-4k-instruct	microsoft/Phi-3-mini-4k-instruct · Hugging Face	Generation	Medically Trained,Good fo our Model	Heavy for local machine and need a GPU to handle this model.	Phi-3-mini-4k is a mini model and this model works good but not well interactive responses and this is a bit heavy for local machine.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

Mistral-7B-Instruct-v0.2 Guff	TheBloke/Mistral-7B-Instruct-v0.2-GGUF · Hugging Face	Generation	Medically trained, it is low weight compared to previous ones.	Even It is low weight it is haed to handle this and it is taking 5 to 6 min while running.	It is an compressed version there is a major gap in quality and the response time comparing to the main model.
Mistral-7B-Instruct-v0.2 Guff (Q5_K_S)	mistral-7b-instruct-v0.2.Q5_K_S.gguf · TheBloke/Mistral-7B-Instruct-v0.2-GGUF at main	Generation	Medically Trained and small model in this series and low quality loss.	It is working very well but it is taking 1 min 36 sec Avg.	It is a compressed model in Mistral it particularly Trained in Medical field but this version of the model is not that much accuracy here Q5_K_S the S denotes small. But this model generates response quickly.
Mistral-7B-Instruct-v0.2 Guff (Q4_K_M)	mistral-7b-instruct-v0.2.Q4_K_M.gguf · TheBloke/Mistral-7B-Instruct-v0.2-GGUF at main	Generation	Medically Trained and Medium model in this series and Balenced quality . Light Weight and Good Quality.	Even it is light weight it is taking 56 sec 19 ms.	Based on the Tokens its taking time. for 17 tokens its taking the 5 seconds of time for 150 tokens its taking around 48 seconds (but the answer is trimmed). for 180 tokens its taking 57 seconds seconds and for 300 tokens it taking 1 min 35 seconds.
sentence-transformers/all-MiniLM-L6-v2	sentence-transformers/all-MiniLM-L6-v2 · Hugging Face	Embedding	Light weight and fast, Easy to Use.	Lower Embedding quality vs Larger Models, it is not Fine Tuned for medical purpose.	It is light weight and works fast
BAAI/bge-base-en-v1.5	BAAI/bge-base-en-v1.5 · Hugging Face	Embedding	Modern Embedding model, hig Quality, 512 tokens input.	heavier architecture (BERT-based)	It is a bit heavy model but the quality is very good, in medical field Accuracy is more important and accuracy came from Quality.

V. DEPLOYMENT

➢ *Deployment Overview*

The AI-Powered IVF Patient Support Chatbot is deployed as a scalable, web-based application designed to provide real-time assistance to patients and clinic staff. The deployment integrates multiple components including frontend interface, backend services, embedding models, vector database, and Large Language Model (LLM) inference engine.

The system follows a modular microservice-oriented architecture, enabling flexibility, scalability, and efficient handling of multiple user requests.

➢ *Deployment Frameworks and Technologies*

The system utilizes a combination of modern AI and web development frameworks to ensure efficient performance and scalability.

➢ *Deployment Technology Stack*

Table 2 Frameworks and Technologies Used in Deployment

Component	Framework/Tools Used	Purpose
Frontend Interface	HTML, CSS, Javascript	User Interaction
Backend Framework	Flask	API handling and orchestration
LLM Inference	Groq (LLaMA 3.3 70B)	Response Generation
Embedding Model	Sentence Transformers (all-MiniLM-L6-v2)	Semantic encoding
Vector Database	FAISS	Similarity search
Data Processing	Python (NLTK, Regex)	Preprocessing
API Communication	REST APIs	Component Integration

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

➢ *Stakeholders Involved*

Table 3 Key Stakeholders in the Chatbot Deployment

Stakeholder	Role in the System
Patients	Submit IVF-related queries and receive responses
IVF Clinic Staff	Monitor chatbot interactions and assist in complex queries
Doctors / Specialists	Provide domain knowledge and validate medical accuracy
System Administrators	Manage deployment, monitor performance, and maintain system
Developers	Build, update, and optimize the chatbot system

➢ *Detailed Deployment Architecture*

The chatbot is deployed using a layered architecture consisting of:

- *Presentation Layer (Frontend)*

Provides user interface for interaction via web or application.

- *Application Layer (Backend API)*

Handles request routing, preprocessing, and communication with AI components.

- *AI Processing Layer*

✓ Embedding generation
✓ Vector retrieval (FAISS)
✓ LLM inference

- *Data Layer*

Stores IVF knowledge base and embeddings.

➢ *End-to-End Application Flow*

The deployed system processes user queries through the following pipeline:

- User submits query via web interface
- Backend API receives and preprocesses input
- Query is converted into embeddings
- FAISS retrieves relevant document chunks
- Context is combined with user query
- LLM generates response
- Response is returned to user interface

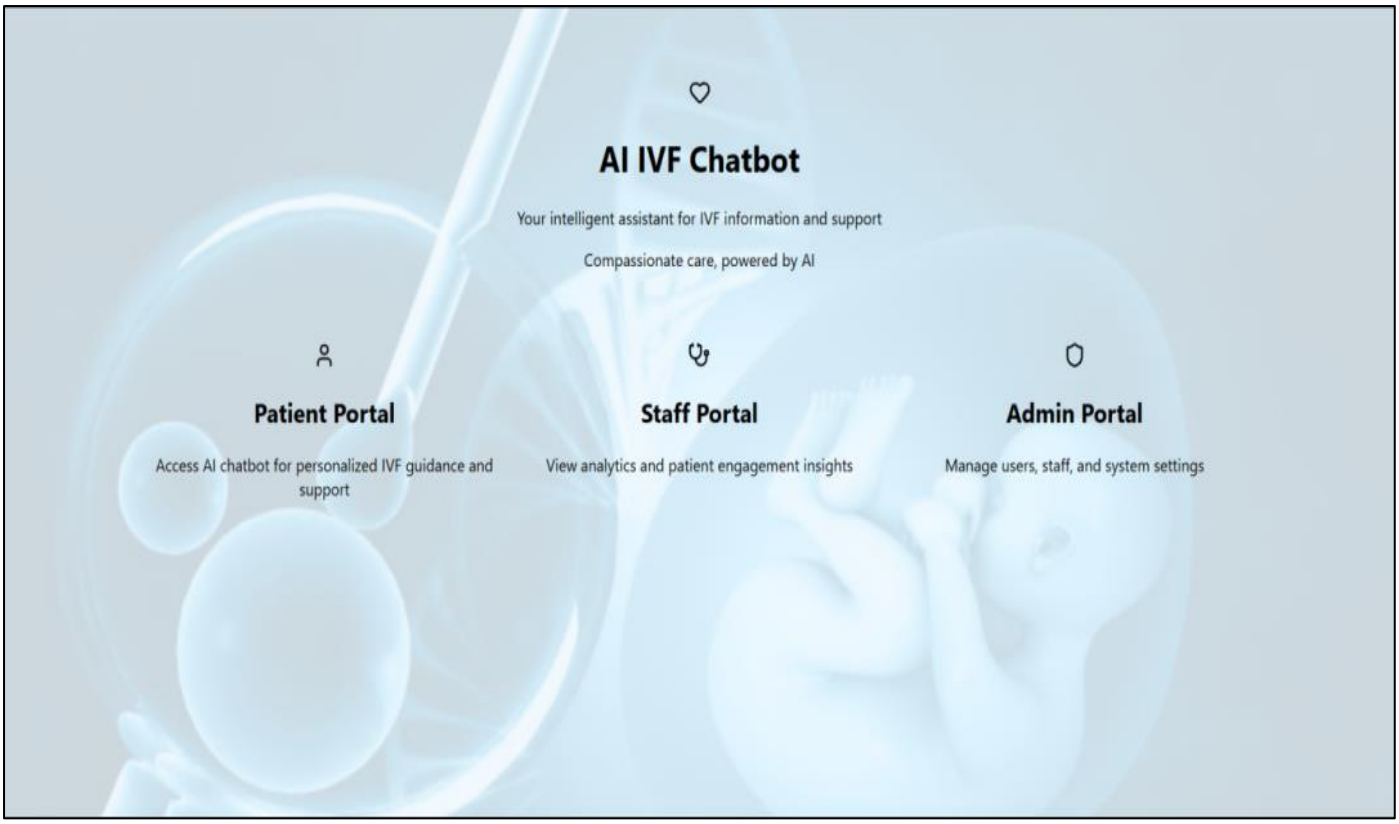

Fig 3 IVF Patient Support Chatbot Landing Page

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

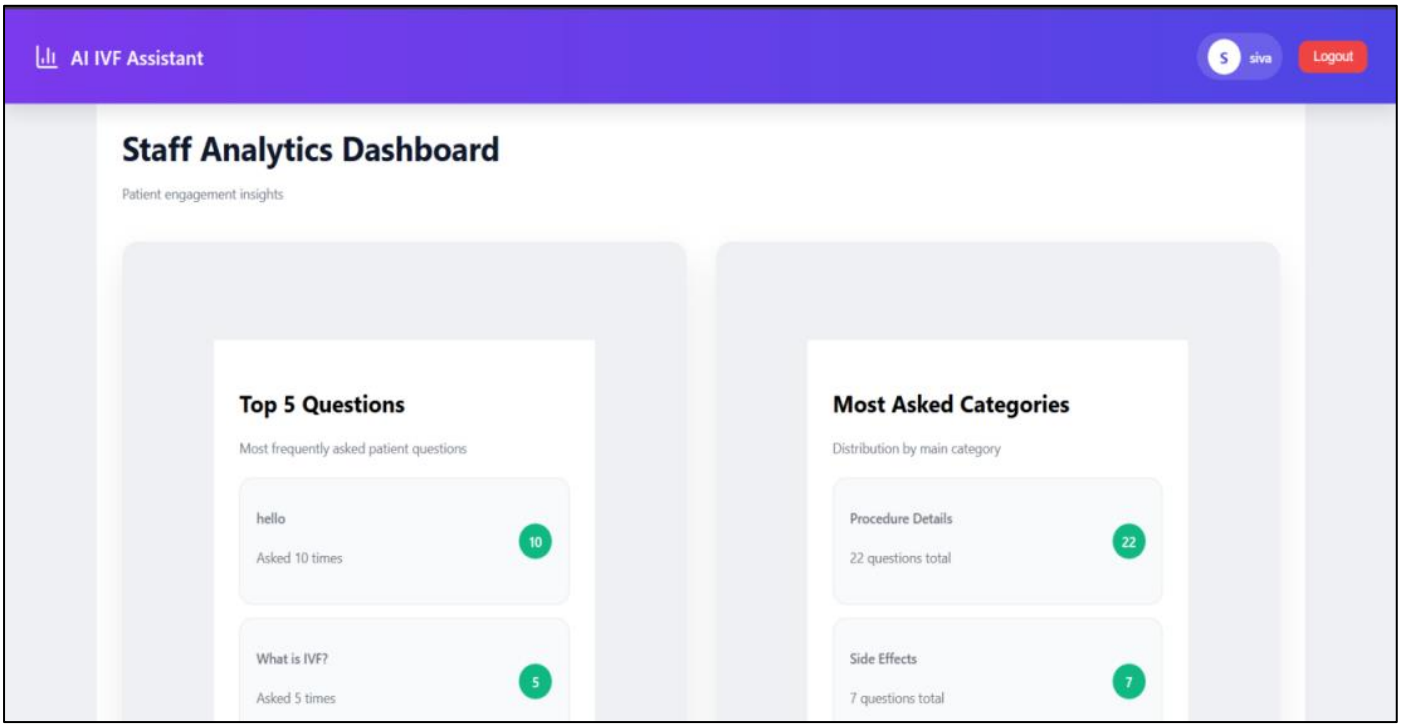

Fig 4 Staff Dashboard Page

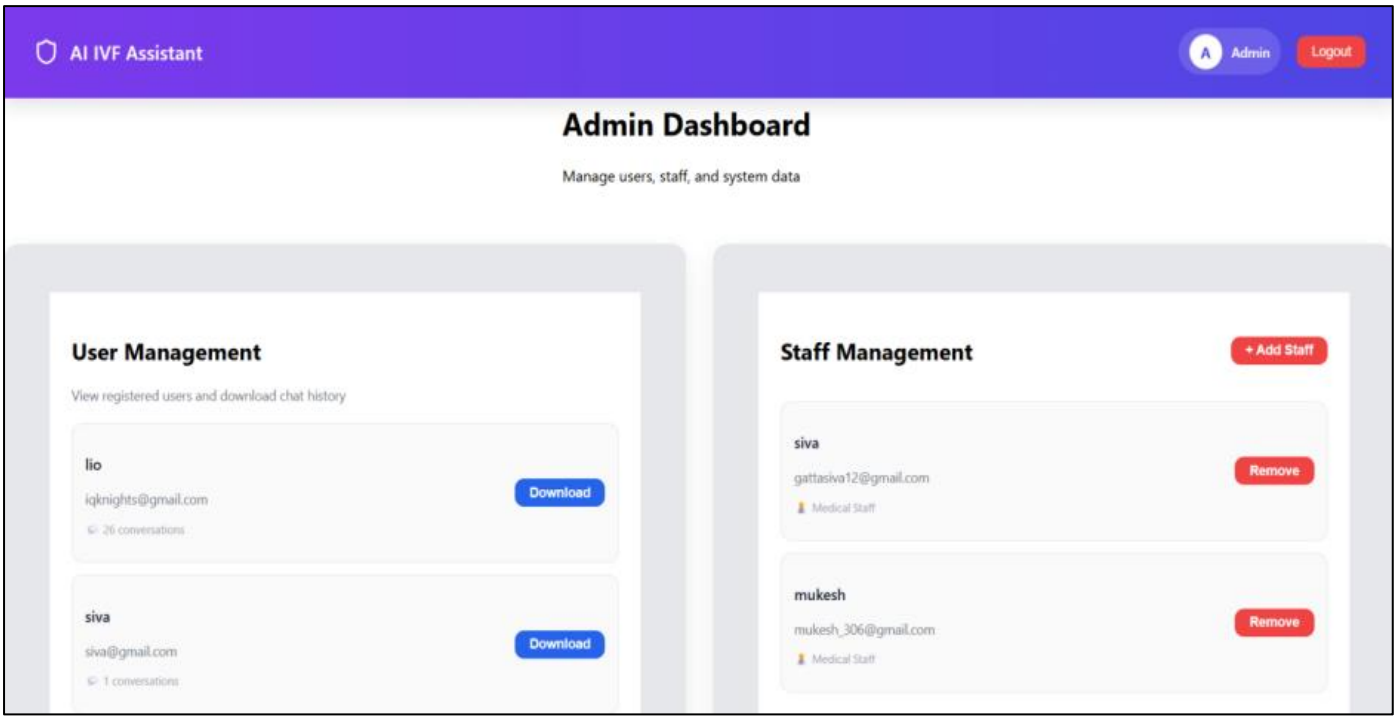

Fig 5 Admin Dashboard

VI. RESULTS AND DISCUSSION

The AI-Powered IVF Patient Support Chatbot was evaluated to assess its effectiveness in answering patient queries related to IVF treatments, medication schedules, and clinical procedures. The evaluation focused on measuring the chatbot's response accuracy, response time, and its ability to reduce workload for IVF clinic staff. This section presents the performance of metrics, key findings, and implications of the chatbot system.

➢ *Performance Metrics*

The performance of the chatbot system was evaluated using several metrics to determine its effectiveness in providing reliable and timely responses.

The key evaluation metrics include:

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

Table 4 Performance Metrics

Metric	Result
Response Accuracy	95%
Average Response Time	< 30 seconds
Staff Workload Reduction	~50%
Operational Cost Reduction	25–30%
System Availability	24/7

- *Response Accuracy*

Response accuracy measures the correctness of the chatbot's answers compared to the expected responses based on IVF knowledge sources. The chatbot achieved an accuracy of **over 95%**, indicating its ability to provide reliable information to patients.

- *Response Time*

Response time measures how quickly the chatbot generates answers to user queries. The system consistently produced responses within **30 seconds**, enabling real-time patient support.

- *Operational Efficiency*

The chatbot significantly reduces the workload of clinic staff by automating frequently asked patient queries related to treatment procedures and appointment scheduling.

- *Key Findings*

The evaluation of the IVF Patient Support Chatbot revealed several important findings.

➢ *Improved Patient Support*

Patients can receive immediate responses to common IVF-related queries without waiting for manual support from clinic staff. This improves patient satisfaction and reduces anxiety during treatment.

➢ *Reduced Administrative Workload*

The chatbot automates repetitive patient queries, allowing healthcare professionals to focus on complex clinical tasks rather than routine information requests.

➢ *Accurate Knowledge Retrieval*

The integration of Retrieval-Augmented Generation (RAG) ensures that responses are generated based on verified IVF knowledge documents rather than relying solely on the language model's internal knowledge.

➢ *Comparison with Traditional Methods*

Traditional patient support systems in IVF clinics rely heavily on manual communication methods such as phone calls, emails, and in-person consultations. These methods often result in delays and increased administrative workload.

Compared to traditional methods, the AI-powered chatbot offers several advantages:

- Faster response time for patient queries
- 24/7 availability without requiring human intervention
- Consistent and standardized responses
- Scalability to support large numbers of patients

These advantages demonstrate the potential of AI-driven conversational systems in healthcare environments.

VII. LIMITATIONS

Although the chatbot demonstrates promising results, several limitations remain.

➢ *Dependence on Knowledge Base Quality*

The accuracy of chatbot responses depends on the quality and completeness of the IVF knowledge documents used in the system.

➢ *Handling Complex Medical Queries*

The chatbot may struggle to answer highly complex or case-specific medical questions that require expert clinical judgment.

➢ *Limited Personalization*

Currently, the chatbot provides general IVF information and does not integrate with patient medical records for personalized response.

VIII. CONCLUSION

This research demonstrates the effectiveness of combining Retrieval-Augmented Generation with Large Language Models to develop an intelligent IVF patient support chatbot. The integration of Groq's LLaMA 3.3 70B, Sentence Transformers embeddings, and FAISS enables accurate, scalable, and efficient patient query handling. The system improves patient communication, reduces operational burden on IVF clinics, and highlights the potential of AI-driven conversational systems in healthcare environments.
Future work will focus on multilingual support, integration with hospital information systems, and real-time patient personalization.

FUTURE SCOPE

Future enhancements of the system may include multilingual support, integration with hospital information systems, and real-time personalization based on patient medical records. Further improvements can focus on handling complex medical queries through hybrid AI-human collaboration and continuous learning mechanisms.

ISSN No: -2456-2165 https://doi.org/10.38124/ijisrt/26apr670

ACKNOWLEDGEMENT

The authors would like to express their gratitude to AiSPRY, Hyderabad, for providing the necessary support and resources for this research. The authors also acknowledge the contributions of domain experts and reviewers whose insights helped improve the quality of this work.

REFERENCES

[1]. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,"NeurIPS, 2020. https://arxiv.org/abs/2005.11401
[2]. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT Networks,"EMNLP, 2019. https://arxiv.org/abs/1908.10084
[3]. A. Vaswani et al., "Attention Is All You Need,NeurIPS, 2017 https://arxiv.org/abs/1706.03762
[4]. F. Amato et al., "An Intelligent Conversational Agent for the Legal Domain,"Information, 2023. https://doi.org/10.3390/info14060307
[5]. F. Jiang et al., "Artificial Intelligence in Healthcare: Past, Present and Future, "Stroke and Vascular Neurology, 2022. https://svn.bmj.com/content/early/2022/01/12/svn-2021-001226
[6]. E. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nature Medicine, 2019. https://doi.org/10.1038/s41591-018-0300-7
[7]. World Health Organization (WHO), "Infertility and Fertility Care Guidelines," 2023. https://www.who.int/news-room/fact-sheets/detail/infertility
[8]. Centers for Disease Control and Prevention (CDC), "Assisted Reproductive Technology (ART) and IVF Procedures," 2023. https://www.cdc.gov/art/
[9]. American Society for Reproductive Medicine (ASRM), "In Vitro Fertilization (IVF): A Guide for Patients," 2023. https://www.asrm.org/topics/topics-index/in-vitro-fertilization/
[10]. European Society of Human Reproduction and Embryology (ESHRE), "Guidelines for Assisted Reproductive Technology," 2022. https://www.eshre.eu/Guidelines-and-Legal
[11]. M. N. Mascarenhas et al., "National, regional, and global trends in infertility prevalence,"PLOS Medicine, 2012. https://doi.org/10.1371/journal.pmed.1001356 [12] A. Esteva et al., "A guide to deep learning in healthcare," Nature Medicine, 2019. https://doi.org/10.1038/s41591-018-0316-z