International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

ENTERPRISE RAG-BASED INTELLIGENT CUSTOMER SUPPORT ASSISTANT

[1] Siddamma CM, [2] K Pravalika, [3] A Shiva Raj, [4] K Ankitha
[1]AssistantProfessor, [234]Students
Department of IOT
Siddhartha Institute of Technology & Sciences, Narapally
siddammacm@siddhartha.org.in , 23tq1a6931@siddhartha.co.in, 23tq1a6928@siddhartha.co.in, 23tq1a6941@siddhartha.co.in

Abstract

Customer support is an essential part of modern enterprises because customers expect quick, accurate, and personalized responses to their questions. Organizations commonly maintain large collections of product manuals, frequently asked questions, service policies, troubleshooting guides, and internal support documents. Finding the correct information from these resources manually can take considerable time and may result in inconsistent responses.

The proposed **Enterprise RAG-Based Intelligent Customer Support Assistant** is an AI-powered platform designed to provide accurate and context-aware responses to customer queries. The system uses **Retrieval-Augmented Generation (RAG)** to retrieve relevant information from authorized enterprise knowledge sources before generating an answer. This approach allows the assistant to provide responses based on current organizational information rather than relying only on the language model's general knowledge.

The system processes enterprise documents through document extraction, cleaning, chunking, embedding generation, and vector indexing. When a customer submits a question, the query is converted into a semantic representation and compared with the indexed knowledge base. Relevant documents or content sections are retrieved and supplied as context to the Generative AI model, which produces a natural-language response.

The platform can also provide source references, confidence indicators, conversation history, escalation mechanisms, and feedback collection. When the system cannot confidently answer a question, it can recommend transferring the conversation to a human support representative. This helps reduce unsupported responses while ensuring that complex customer issues can receive appropriate human attention.

I. Introduction

Customer support plays a major role in maintaining customer satisfaction and building long-term relationships with users. Enterprises receive large numbers of questions related to products, services, account procedures, troubleshooting, policies, and general assistance. Handling these queries manually can require significant time and human resources, particularly when the same questions are repeatedly asked by different customers.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

Traditional customer-support systems commonly use FAQs, static knowledge bases, keyword searches, and predefined chatbot responses. These approaches can be effective for simple and frequently repeated questions, but they may struggle when customers use different wording or ask questions that require information from multiple documents. Support representatives may also need to search through several internal resources before providing an accurate answer.

Generative AI provides an opportunity to improve customer-support systems by allowing users to interact using natural language. Large language models can understand questions and generate fluent responses, but relying solely on a general-purpose model can result in inaccurate or unsupported information. Enterprise customer-support applications therefore require mechanisms that connect AI-generated responses to trusted organizational knowledge.

Retrieval-Augmented Generation (RAG) addresses this challenge by combining information retrieval with Generative AI. Instead of generating an answer only from the model's learned knowledge, the system first retrieves relevant information from an approved enterprise knowledge base. The retrieved content is then provided to the language model as context, enabling the system to generate responses that are more closely grounded in enterprise-specific information.

II. Literature Survey

Traditional customer-support systems have relied on FAQs, knowledge bases, ticketing systems, and rule-based chatbots. These systems provide structured access to frequently requested information but generally depend on predefined responses and keyword matching. Their effectiveness can decrease when customers use unexpected terminology or ask questions requiring contextual understanding.

Natural-language processing has improved the ability of customer-support applications to understand user queries. NLP techniques can identify user intent, extract important entities, and classify customer requests. These capabilities allow support systems to route queries to appropriate services or provide relevant predefined responses. However, intent-based systems may have limitations when answering open-ended questions requiring information from large document collections.

Large language models have introduced more flexible conversational capabilities into customer-support applications. They can understand natural-language questions, summarize information, maintain conversational context, and generate human-readable responses. However, language models may produce unsupported or inaccurate information when they lack access to current or organization-specific knowledge.

Retrieval-Augmented Generation has emerged as an approach for combining the strengths of information retrieval and language generation. In a RAG architecture, a user's query is processed and relevant documents are retrieved from a knowledge repository. These documents are then provided as context to a language model, which generates a response based on the retrieved information. This approach can improve knowledge grounding and support enterprise-specific applications.

Vector databases and embedding models have become important components of modern semantic-retrieval systems. Documents can be converted into numerical representations that capture semantic relationships between pieces of content. When a customer submits a question, semantic similarity can be used to identify relevant knowledge even when the exact words in the query do not appear in the source document.

Research and practical implementations of AI customer-support systems demonstrate the benefits of conversational interfaces, semantic retrieval, RAG, automated responses, and human escalation. However, enterprise deployments still need to address issues such as access control, knowledge freshness, response accuracy, source attribution, privacy, and monitoring. The proposed system combines these concepts into a centralized RAG-based customer-support platform.

III. System Analysis

System analysis for the proposed Enterprise RAG-Based Intelligent Customer Support Assistant focuses on the requirements needed to provide reliable AI-assisted customer service. The system should provide secure access for customers, support agents, and administrators according to their roles. It should allow administrators to connect approved enterprise knowledge sources such as FAQs, product manuals, policies, and troubleshooting documents. The document-processing module should extract, clean, divide, and organize enterprise content into meaningful sections. The processed content should be converted into embeddings and stored in a vector-based retrieval system. When a customer submits a question, the query-processing module should understand the question and generate a semantic search representation. The retrieval engine should identify relevant enterprise content based on semantic similarity and metadata. Retrieved information should be passed to the Generative AI model as contextual input for response generation. The system should provide source references where available and identify low-confidence responses. A conversation-management component should maintain relevant session context while respecting privacy and retention policies. Complex or unsupported queries should be escalated to human support agents. Analytics should monitor query volumes, response quality, retrieval performance, and escalation patterns.

Existing System

Existing customer-support environments commonly use call centers, email support, FAQs, static knowledge bases, ticketing systems, and rule-based chatbots. Customers may need to search through FAQs or wait for support representatives to answer their questions. Keyword-based search can return many results without understanding the actual meaning of the customer's request. Rule-based chatbots generally provide predefined responses and may struggle with questions outside their programmed scenarios. Support representatives may need to manually search multiple enterprise documents before answering complex questions. Response quality can vary

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

depending on the experience and knowledge of individual agents. High query volumes can increase customer waiting times and support costs. Traditional systems may also have limited semantic understanding of customer questions. Historical conversations may not be effectively used for improving knowledge retrieval. Generating detailed reports and analyzing support trends can require separate tools. These limitations can reduce scalability and make it difficult to provide consistent, context-aware customer assistance.

Disadvantages of Existing System

- Heavy dependence on manual customer-support agents.
- Long response times during high query volumes.
- Keyword search may return irrelevant results.
- Rule-based chatbots have limited conversational flexibility.
- Limited semantic understanding of customer queries.
- Limited automated source-based answer generation.
- Difficult to provide support continuously at scale.
- Limited analysis of customer-query and escalation patterns.

Proposed System

The proposed Enterprise RAG-Based Intelligent Customer Support Assistant provides an intelligent conversational platform for answering customer queries using authorized enterprise knowledge. The system allows administrators to connect approved sources such as product documentation, FAQs, policies, troubleshooting guides, and service information. The document-ingestion module extracts, cleans, chunks, and prepares enterprise content for retrieval. Each content segment is converted into a semantic embedding and stored in a vector database along with relevant metadata. When a customer submits a question, the query-processing module analyzes the request and performs semantic retrieval against the enterprise knowledge base. Retrieved content is filtered according to applicable access and security policies before being provided to the Generative AI model. The RAG engine generates a contextual response based on the retrieved enterprise information. Source references can be displayed to improve transparency and allow users or support agents to verify the answer. The system can evaluate retrieval and response confidence and identify questions that require human intervention. Complex or unsupported requests can be transferred to authorized support representatives with relevant conversation context. Customer feedback can be collected to evaluate response usefulness and improve the system. Administrators can monitor knowledge sources, usage, retrieval performance, and escalation patterns through analytics. Overall, the proposed system provides a structured workflow for enterprise knowledge retrieval, AI response generation, customer assistance, and human escalation.

Advantages of Proposed System

- Provides AI-powered customer support.
- Uses RAG to ground responses in enterprise knowledge.
- Supports natural-language customer queries.

- Provides semantic rather than keyword-only retrieval.
- Generates contextual and concise responses.
- Can provide references to source documents.
- Supports centralized enterprise knowledge management.
- Provides customer-query and support analytics.
- Can be updated as enterprise knowledge changes.
- Helps reduce customer waiting time.
- Scales to large volumes of customer queries.

IV. Methodology

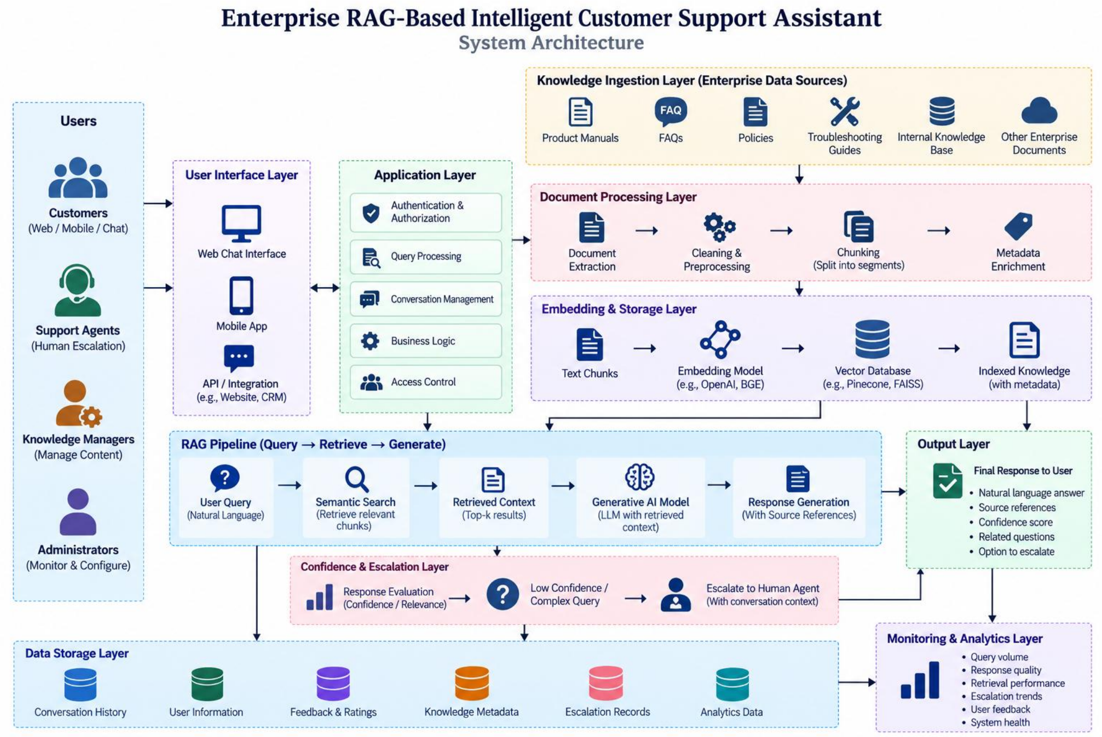
The methodology begins with authenticating users and defining appropriate customer-support and administrative access levels. Approved enterprise knowledge sources such as FAQs, product manuals, policies, and troubleshooting documents are connected to the platform. The document-ingestion module extracts and cleans information from these sources. Documents are divided into meaningful chunks and enriched with metadata such as document title, category, and update information. Embedding models convert the processed content into semantic vectors that are stored in a vector database. When a customer submits a question, the query-processing module analyzes the request and creates a semantic representation. The retrieval engine searches the vector database and identifies the most relevant enterprise knowledge segments. Retrieved content is filtered according to applicable access and security policies. The RAG component passes the selected content to the Generative AI model to generate a grounded response. The system can attach relevant source references to help verify the generated answer. Confidence and retrieval-quality checks can identify questions that should be escalated to human support. Customer feedback and conversation outcomes are stored according to organizational policies for quality analysis. Administrators can monitor performance, update knowledge sources, review analytics, and improve the retrieval and response pipeline.

System Architecture

The system architecture consists of several interconnected layers designed to provide intelligent and secure enterprise customer support. The User Layer contains customers, support agents, knowledge managers, and administrators with role-based access. The Customer Interface Layer provides the conversational chatbot, query submission, response display, source references, feedback, and escalation options. The Application Layer manages authentication, user sessions, query processing, conversation management, business logic, and access control. The Knowledge Ingestion Layer connects to approved enterprise sources such as FAQs, product manuals, policies, troubleshooting documents, and support knowledge bases. The Document Processing Layer extracts, cleans, chunks, and enriches enterprise content with appropriate metadata. The Embedding Layer converts document chunks and user queries into semantic vector representations. The Vector Database Layer stores searchable document embeddings and associated metadata for efficient semantic retrieval. The Retrieval Layer searches the knowledge base and selects relevant information based on semantic similarity and applicable filtering rules. The RAG Layer combines retrieved enterprise content with the customer query and sends the

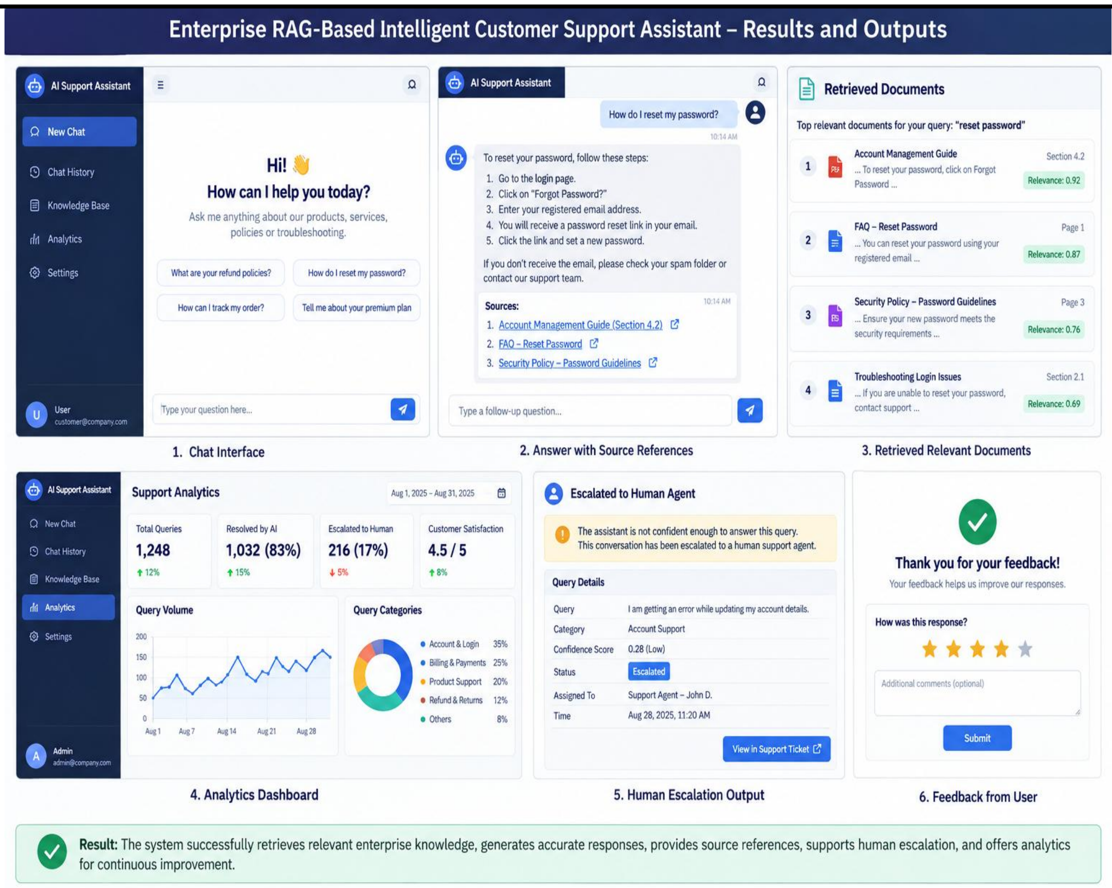
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

contextual information to the Generative AI model. The Response Generation Layer produces a natural-language answer based on the retrieved context and can provide source references. The Confidence and Escalation Layer evaluates response quality and routes uncertain or complex queries to authorized human support agents. The Database Layer stores customer-support records, conversation information, feedback, knowledge metadata, escalation records, and audit information according to organizational policies. The Analytics Layer provides query statistics, response-quality metrics, retrieval performance, customer feedback, and escalation trends. The Administration Layer manages users, permissions, knowledge sources, system configurations, and model settings. Security controls, access management, encryption, audit logging, and appropriate data-retention policies protect customer and enterprise information. The overall architecture follows the workflow **Ingest → Process → Embed → Retrieve → Generate → Verify → Respond/Escalate → Analyze**, providing an intelligent and scalable customer-support solution.



V. Result and Output

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



VI. Conclusion

The Enterprise RAG-Based Intelligent Customer Support Assistant provides an effective solution for improving enterprise customer support through the integration of Retrieval-Augmented Generation (RAG), semantic search, document processing, and Generative AI. The system retrieves relevant information from authorized enterprise knowledge sources and uses it to generate contextual and reliable responses to customer queries.

The proposed system reduces the dependency on manual document searching and helps customers receive faster answers to frequently asked questions, product-related issues, policies, and troubleshooting requests. By providing relevant source references along with generated responses, the system improves transparency and makes the retrieved information easier to verify.

The system also supports confidence-based response evaluation and human escalation for complex or low-confidence queries. This ensures that questions requiring expert assistance can be transferred to human support agents with the relevant conversation context. Analytics and feedback mechanisms further help organizations evaluate response quality and continuously improve the knowledge base.

Overall, the proposed RAG-based assistant can improve response efficiency, knowledge accessibility, customer experience, and support-team productivity while maintaining controlled access to enterprise information. The system can be further enhanced with multilingual support, voice interaction, advanced retrieval techniques, real-time knowledge updates, enterprise application integration, and improved security and evaluation mechanisms.

References

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.
[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.
[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm
[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.
[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, "Real-Time Object Detection in Drone Surveillance Using YOLOv5," in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.
[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, "Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks," in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.
[7] R. D. Kumar, V. N. S.Manaswini, "Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology," in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.
[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, "An advanced movie recommender using collaborative filtering and sentiment analysis," International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.
[9] Ravi Kumar Banoth, Ramana Murthy B V, "Automatic crop recommendation system using LightGBM and decision tree machine learning models," Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.
[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, "Smart agriculture through IoT and machine learning for analyzing carbon footprints," in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.
[11] Ravi Kumar Banoth, B. V. Ramana Murthy, "Soil image classification using transfer learning approach: MobileNetV2 with CNN," SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.