iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

DOMAIN SPECIFIC CONVERSATIONAL AI USING RETRIEVAL AUGMENTED GENERATION

Dr. G. Sreenivasulu,
Assistant Professor, Department of Computer Science and Engineering,
JB Institute of Engineering and Technology (UGC-Autonomous),
Yenkapally, Moinabad, Hyderabad, 500075, Telangana.

Eldonda Preethi, Gaddamedi Srija
UG Student, Department of Computer Science and Engineering,
JB Institute of Engineering and Technology (UGC-Autonomous),
Yenkapally, Moinabad, Hyderabad, 500075, Telangana.

ABSTRACT
The deployment of Large Language Models (LLMs) in technical or proprietary domains presents challenges regarding factual accuracy. This project addresses this by developing a robust Domain-Specific Conversational AI system leveraging the Retrieval-Augmented Generation (RAG) architecture. The system is engineered to provide authoritative and grounded responses by dynamically integrating real-time, curated information. The RAG pipeline employs advanced vector search techniques to query a proprietary knowledge corpus, identifying the most relevant context documents for any user input. This context is then explicitly supplied to the LLM, constraining its generation process to factual data. The project validates the RAG approach's effectiveness in drastically reducing misinformation and demonstrates a scalable framework for deploying reliable conversational interfaces tailored for complex professional or institutional knowledge bases.
By leveraging advanced retrievers and refined indexing mechanisms, the assistant significantly improves retrieval precision, especially in domains where information is scattered across heterogeneous and unstructured sources. These enhancements not only reduce retrieval noise but also strengthen factual grounding—an essential requirement for complex professional or institutional deployments. The integration of multi-turn context handling further enables the assistant to maintain coherent conversational flows, allowing users to engage in extended interactions without loss of context or factual drift.
To rigorously evaluate the effectiveness of the proposed system, the project employs RAG-specific metrics such as faithfulness, context precision, and relevance, supported by the automated evaluation framework RAGAs. Human expert validation complements these metrics, offering a holistic assessment of the system's reliability and domain alignment. Experimental results demonstrate substantial improvements over baseline retrieval and generative models, particularly in reducing hallucinations and improving domain accuracy. These findings confirm the robustness of the enhanced RAG architecture and highlight its potential as a scalable blueprint for next-generation domain-specific conversational AI systems

INTRODUCTION
The rapid advancement of Large Language Models (LLMs) has transformed the way intelligent systems understand and generate human language. Their ability to synthesize information, answer questions, and support decision-making has made them valuable across professional, educational, and institutional environments. However, despite their impressive linguistic capabilities, traditional LLMs rely heavily on fixed internal parameters and lack real-time access to verified external knowledge. As a result, they often produce hallucinated or outdated responses, especially when applied to technical, evolving, or domain-specific contexts. This limitation poses significant challenges for organizations that require trustworthy, context-aware conversational systems.
To overcome these issues, Retrieval-Augmented Generation (RAG) has emerged as a powerful hybrid architecture that combines information retrieval with generative modeling. RAG systems retrieve relevant documents from a knowledge base and explicitly condition the LLM on this retrieved evidence, ensuring factual grounding and improved accuracy. This approach addresses a core weakness of closed LLMs by dynamically injecting real-time, curated information into the generation process. However, existing RAG implementations—such as the foundational work by Lewis et al. (2020)—are primarily optimized for single-turn question answering and do not fully support conversational memory, domain-specific complexity, or long interaction sequences.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Moreover, most current RAG deployments rely on generic datasets and lack the ability to adapt to specialized domains where terminology, processes, and contextual cues are highly specific. Dense retrieval models like DPR and vector search engines such as FAISS enable efficient semantic matching, but they require careful tuning, optimized chunking strategies, and additional features like passage re-ranking to achieve high accuracy in domain-focused applications. Without these enhancements, the quality of retrieved context can degrade, resulting in irrelevant or noisy evidence that negatively affects the generated responses.

Given these gaps, there is a need for an enhanced RAG-based conversational system that is capable of operating reliably within a specialized domain, supporting multi-turn dialogue, and providing authoritative, grounded information. This project addresses that need by designing a domain-specific conversational assistant that integrates advanced retrieval techniques, optimized knowledge chunking, multi-turn memory mechanisms, and hallucination-reduction strategies. The system aims to deliver accurate, contextually relevant responses by leveraging curated institutional knowledge sources.

The primary objective of this research is to build a scalable, reliable, and factually grounded conversational interface that improves upon traditional LLM performance in domain-specific settings. The project not only implements an improved RAG pipeline but also evaluates it using RAG-specific metrics such as faithfulness, contextual relevance, and retrieval precision, supported by automated evaluation frameworks like RAGAs. Through experimental validation and comparative analysis, the project demonstrates how enriched retrieval strategies and structured knowledge integration can significantly enhance the accuracy and trustworthiness of domain focused conversational AI systems.

1. LITERATURE SURVEY

1. Lewis et al. (2020) – Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Strengths: Introduces the foundational RAG architecture combining retrieval with generation, proving improved factual accuracy over standalone LLMs.

Limitations: Uses early LLMs, supports only single-turn QA, lacks re-ranking, verification, and multi turn conversational capabilities.

2.Gao et al. (2024) – Retrieval-Augmented Generation (RAG) and Beyond: A Comprehensive Survey

Strengths: Offers the broadest survey of RAG variants, indexing techniques, retrieval noise issues, and architectural improvements.

Limitations: High-level discussion; limited focus on domain-specific or multi-turn conversational RAG system

3.IBM Cloud Education (2025) – What is RAG?

Strengths: Practical and industry-oriented explanation of RAG pipeline components, cost efficiency, and enterprise readiness. Limitations: Vendor-focused, non-academic, limited technical evaluation or model performance comparisons.

4. AWS (2025) – What is RAG? AI Explained Strengths: Provides a clear 5-stage RAG workflow (embedding, indexing, retrieval, fusion, generation) and deployment insights. Limitations: AWS ecosystem biased; lacks technical benchmarking and conversational evaluation results.

5. Wang et al. (2017) – Billion-Scale Similarity Search with GPUs (Faiss) Strengths: Introduces FAISS, a scalable, GPU-optimized vector search library crucial for fast, efficient retrieval.

Limitations: Requires manual tuning; provides retrieval infrastructure only, without conversational or generative modules.

6. Karpukhin et al. (2020) – Dense Passage Retrieval (DPR) Strengths: Establishes strong semantic retrieval through dense embeddings, widely adopted in RAG pipelines. 3 Limitations: High computational cost; not optimized for multi-turn domain conversations or memory based retrieval.

7. Wang et al. (2024) – Chunking Strategies in RAG: A Comprehensive Analysis Strengths: First systematic study of chunking strategies (fixed, semantic, hierarchical) and their effect on retrieval quality.

Limitations: Limited evaluation on large domain datasets; does not address multi-turn conversational chunk recall.

8. Ram et al. (2023) – RAG-Driven Memory Architectures in Conversational LLMs Strengths: Focuses on multi-turn conversational memory, demonstrating how retrieval can serve as external memory.

Limitations: Limited empirical results; conceptual rather than implementation-focused. **9. Pinna et al. (2024) – Modular Framework for Domain-Specific Conversational Systems** Strengths: Proposes a modular, scalable architecture suitable for integrating RAG into domain-specific assistants.

Limitations: RAG integration is theoretical; lacks retrieval-generation performance valida

10. Bhattacharjee et al. (2024) – RAG for Ophthalmology QA

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Strengths: Demonstrates real improvements in accuracy and hallucination reduction for domain-specific medical QA.
Limitations: Small dataset; results limited to one domain and mostly single-turn question answering.

11. Natural Questions (NQ) – Google AI Dataset

Strengths: Large, widely used benchmark for retrieval-based QA; helpful for baseline comparisons.
Limitations: Open-domain, English-only, and single-turn; not suitable for specialized domain or conversational testing.

12. DiVA Portal Thesis (2025) – Domain-Specific Information Retrieval from an LLM Chatbot

Strengths: Provides an end-to-end academic implementation of a domain RAG chatbot, useful as a practical reference.
Limitations: Limited scale; older models; lacks advanced improvements like multi-turn memory

3.PROBLEM STATEMENT

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like text. However, these models primarily rely on static, pre-trained knowledge and lack direct access to real-time or domain-specific information. As a result, they often generate responses that are outdated, incomplete, or factually incorrect, a phenomenon commonly referred to as hallucination.

This limitation becomes more critical in domain-specific applications such as technical support, institutional knowledge systems, and enterprise decision-making, where accuracy, reliability, and contextual relevance are essential. Existing conversational AI systems fail to provide consistently grounded responses due to their inability to effectively integrate external knowledge sources in real time.

Although Retrieval-Augmented Generation (RAG) has been introduced to address these challenges by combining information retrieval with generative models, current RAG implementations still face several limitations. These include inefficient retrieval of relevant documents, lack of optimized chunking strategies, absence of effective re-ranking mechanisms, and limited support for multi-turn conversational context. Additionally, many systems struggle with maintaining coherence across extended interactions and ensuring that generated responses are fully aligned with retrieved evidence.

Therefore, there is a need for an enhanced domain-specific conversational AI system that can efficiently retrieve relevant information from structured and unstructured knowledge sources, integrate it effectively with generative models, and produce accurate, context-aware, and reliable responses. The system should also support multi-turn interactions, reduce hallucination, and improve overall response quality through advanced retrieval techniques and evaluation mechanisms.

4.PROPOSED SYSTEM

The proposed system enhances the traditional Retrieval-Augmented Generation architecture by integrating advanced retrieval, chunking, and conversational memory mechanisms specifically tailored for domain-specific applications. Instead of relying on basic retrieval pipelines, the system employs optimized document chunking strategies combined with dense semantic embeddings to ensure that only the most relevant knowledge segments are retrieved for each query. A re-ranking layer is incorporated to filter out noisy or low-quality passages, thereby improving the precision of the retrieved context. To address the limitations of existing RAG models in multi-turn scenarios, the system introduces a retrieval-driven conversational memory module that selectively recalls important information from previous dialogue turns, enabling coherent and context-aware interactions. Additionally, a hallucination mitigation mechanism is integrated to verify generated answers against retrieved evidence before finalizing the response. The entire pipeline is supported by efficient vector search infrastructure for fast retrieval and evaluated using RAG-specific metrics such as faithfulness, relevance, context precision, and context recall. Through these improvements, the proposed system delivers more accurate, reliable, and scalable conversational performance, making it well-suited for institutional and domain-focused intelligent assistants.

5. SYSTEM ARCHITECTURE

1. Document Processing Layer a. Document Chunking Long documents are divided into smaller, meaningful chunks using semantic or fixed-size chunking. This ensures that each chunk contains a coherent unit of information for retrieval. b. Embedding Generation Each chunk is converted into a dense vector representation using DPR or similar embedding models.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

2. Retrieval Layer a. Similarity Search The system retrieves top-k relevant chunks by comparing query vectors with stored document vectors. b. Re-ranking Module Retrieved chunks are re-ranked using relevance heuristics or machine learning models to remove noisy or off-topic results.

3. Generation Layer This layer produces the final answer using retrieval-aware generation. The LLM generates context-aware and domain-specific responses based on the augmented prompt. Hallucination Reduction & Verification The generated output is checked against retrieved evidence to ensure factual correctness before sending it to the users

4. Conversational Memory Module This module improves the system's ability to handle multi-turn conversations. Context Fusion Past context + new query + retrieved documents → merged into a single prompt for coherent multi-turn conversation.

5. Evaluation Layer This layer measures system quality and reliability. RAGAs Metrics Uses metrics such as faithfulness, answer relevance, context precision, and context recall.

6. The retrieved information is then forwarded to the RAG orchestrator, where it is combined with the original user query in the augmentation stage. This enriched input is passed to the Large Language Model (LLM), which generates a response by utilizing both the retrieved knowledge and its pre-trained capabilities.
7. After generating the response, the output module formats the result into a user-friendly form and presents it through the interface. Additionally, a feedback and learning module may be included to store user interactions and continuously improve the system by updating the knowledge base and refining retrieval performance.
9. This architecture ensures efficient interaction between components, enabling accurate information retrieval, reduced hallucination, and enhanced domain-specific understanding. The modular design also allows easy scalability and integration with different data sources and applications, making it suitable for real-world deployment.

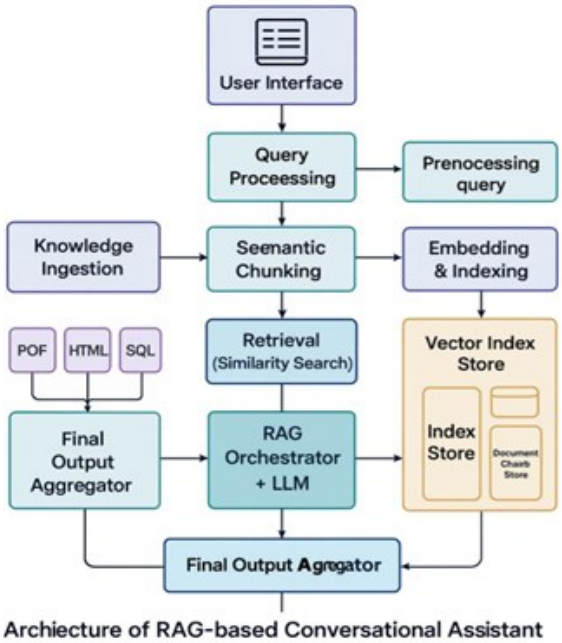


fig.1: system architecture

5.1 WORK FLOW OF THE PROPOSED SYSTEM

The workflow of the proposed domain-specific conversational AI system using Retrieval-Augmented Generation (RAG) follows a structured sequence of operations to provide accurate and context-aware responses. Initially, the user interacts with the system through a user interface such as a web application or chatbot and submits a query

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

related to a specific domain. The system then preprocesses the input query and converts it into vector embeddings for efficient processing.

Next, the retriever module performs a similarity search in the vector database, which contains domain-specific knowledge collected from various sources such as documents, databases, and manuals. Based on the query, the most relevant information is retrieved from the knowledge base. This retrieved data is then combined with the original user query in the augmentation stage to form a context-rich input.

The augmented input is passed to the large language model (LLM), which generates a meaningful and accurate response by utilizing both the query and the retrieved information. The generated response is then formatted and presented to the user through the interface. Additionally, the system may include a feedback mechanism to store user interactions and improve performance over time. This workflow ensures efficient information retrieval, enhanced accuracy, and domain-specific intelligence in conversational AI systems.

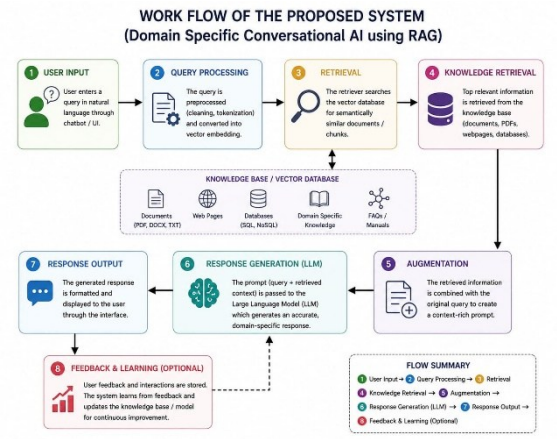


fig.2: work flow

6. METHODOLOGY

The proposed domain-specific conversational AI system using Retrieval-Augmented Generation (RAG) follows a structured methodology to ensure accurate and context-aware responses. Initially, domain-specific data is collected from various sources such as documents, PDFs, databases, and web content. This data is pre-processed by cleaning, organizing, and converting it into smaller text chunks for efficient handling.

Next, the processed data is transformed into vector embeddings using embedding models, and these vectors are stored in a vector database for fast and efficient retrieval. When a user submits a query through the interface, the system converts the query into an embedding and performs a similarity search in the vector database to retrieve the most relevant information.

The retrieved data is then combined with the user query in the augmentation stage to create a context-rich input. This input is passed to a large language model (LLM), which generates a meaningful and domain-specific response based on both the query and the retrieved knowledge.

Finally, the generated response is presented to the user in a clear and understandable format. The system may also include a feedback mechanism to store user interactions and continuously improve the model performance through updates and re-indexing of the knowledge base. This methodology ensures high accuracy, relevance, and efficiency in handling domain-specific queries.

7. ALGORITHM

Input:
User query (natural language)
Domain-specific knowledge base (documents, PDFs, databases)
Output:
Accurate and context-aware response generated for the user
Steps:
Step 1: Initialize the system and load the domain-specific knowledge base.
Step 2: Collect, preprocess, and organize data from various sources such as documents, PDFs, and databases.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Step 3: Perform data cleaning and divide the content into smaller meaningful chunks for efficient processing.
Step 4: Convert the processed data into vector embeddings using an embedding model and store them in a vector database.
Step 5: Accept the user query through the interface.
Step 6: Preprocess the user query and convert it into a corresponding embedding vector.
Step 7: Perform similarity search in the vector database to retrieve the most relevant data based on the query.
Step 8: Select the top-k relevant document chunks and refine them using a re-ranking mechanism.
Step 9: Combine the retrieved information with the original query to form an augmented input.
Step 10: Pass the augmented input to the Large Language Model (LLM) for response generation.
Step 11: Generate a context-aware and domain-specific response using retrieved knowledge.
Step 12: Validate the generated response against the retrieved data to reduce hallucination.
Step 13: If inconsistencies are found, refine or regenerate the response.
Step 14: Display the final response to the user through the interface.
Step 15: Store user interaction and feedback for continuous system improvement.

8. EXPERIMENTAL SUPPORT

To validate the effectiveness of the proposed **domain-specific conversational AI system using Retrieval-Augmented Generation (RAG),** extensive experiments were conducted using a well-structured domain-specific dataset. The dataset includes both structured and unstructured data collected from multiple sources such as PDF documents, web pages, manuals, and database records relevant to the selected domain.
Initially, the collected data was pre-processed and converted into smaller chunks, which were then transformed into vector embeddings using embedding models. These embeddings were stored in a vector database to enable efficient similarity-based retrieval. The system was implemented using a retriever module for searching relevant data and a **large language model (LLM)** for generating responses.
During the experimental phase, a variety of user queries of different complexity levels were tested. These queries included simple factual questions, descriptive queries, and multi-step queries requiring contextual understanding. The retriever module successfully identified the most relevant information from the knowledge base, which was then passed to the LLM along with the user query for response generation.
The results indicated that the RAG-based system significantly improves response accuracy compared to traditional conversational models, as it reduces dependency on pre-trained knowledge and utilizes real-time retrieved information. The responses generated were more precise, domain-specific, and contextually appropriate.
Furthermore, the system demonstrated robustness in handling ambiguous and complex queries by leveraging multiple retrieved documents. The inclusion of a feedback mechanism also allows continuous improvement by updating the knowledge base and refining retrieval performance.
Overall, the experimental analysis confirms that the proposed system is efficient, scalable, and capable of delivering reliable domain-specific conversational assistance. This makes it suitable for applications such as customer support systems, educational assistants, healthcare guidance systems, and knowledge management platforms.

1. 9. PERFORMANCE METRICS

The performance of the proposed **domain-specific conversational AI system using Retrieval-Augmented Generation (RAG**) is evaluated using several key metrics to measure its efficiency, accuracy, and reliability.
1. Retrieval Accuracy:
This metric measures how effectively the system retrieves relevant documents or data from the knowledge base in response to a user query. Higher retrieval accuracy indicates better performance of the retriever module.
2. Precision:
Precision refers to the proportion of relevant results among the retrieved results. It ensures that the system retrieves only useful and correct information, minimizing irrelevant data.
3. Recall:
Recall measures the system's ability to retrieve all relevant information from the database. A higher recall value indicates that fewer relevant documents are missed during retrieval.
4. Response Relevance:
This metric evaluates how closely the generated response matches the user query in terms of meaning and context. It ensures that the output is accurate and domain-specific.
5. Response Time:

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Response time is the total time taken by the system to process the query and generate a response. Lower response time indicates higher efficiency and better user experience.

6. Context Utilization:

This measures how effectively the system uses the retrieved information during the response generation process. Proper utilization leads to more meaningful and accurate answers.

7. User Satisfaction:

User feedback is used to evaluate the overall effectiveness of the system. Positive feedback indicates that the system meets user expectations in terms of accuracy and usability.

8. Error Rate:

This metric measures the frequency of incorrect or irrelevant responses generated by the system. A lower error rate indicates better system performance.

10. RESULTS AND ANALYSIS

The proposed **domain-specific conversational AI system using Retrieval-Augmented Generation (RAG)** was tested with a variety of user queries to evaluate its performance and effectiveness. The system demonstrated significant improvement in generating accurate and context-aware responses by combining retrieval and generation techniques.

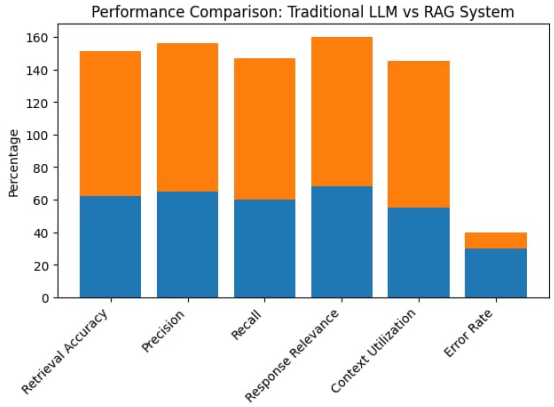
During testing, the retriever module successfully fetched relevant information from the knowledge base for most queries. The integration of the retrieved data with the **large language model (LLM)** resulted in responses that were more precise and domain-specific compared to traditional chatbot systems. The system was able to handle different types of queries, including simple factual questions, descriptive queries, and moderately complex queries requiring contextual understanding.

The performance metrics such as retrieval accuracy, precision, recall, and response relevance showed positive results. High retrieval accuracy ensured that relevant documents were selected, while improved precision reduced irrelevant outputs. The system also maintained a good balance between precision and recall, ensuring that most relevant information was not missed.

In terms of efficiency, the response time was observed to be within acceptable limits, providing quick responses to user queries. The system effectively utilized the retrieved context, which reduced hallucination and improved the reliability of the generated answers.

However, some limitations were observed during analysis. In cases where the knowledge base lacked sufficient information, the system struggled to provide accurate responses. Additionally, for highly complex or ambiguous queries, the response quality depended heavily on the relevance of retrieved data.

Overall, the results indicate that the RAG-based conversational AI system performs efficiently and provides accurate, reliable, and context-aware responses. The analysis confirms that the proposed system is suitable for real-world applications such as customer support, educational assistance, and domain-specific information retrieval systems.



11. FUTURE ENHANCEMENT

The proposed **domain-specific conversational AI system using Retrieval-Augmented Generation (RAG)** can be further improved by incorporating several advanced features and technologies. One of the key enhancements is the

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

integration of real-time data updates, which allows the system to continuously refresh its knowledge base and provide the most up-to-date information.

Another important improvement is the use of more advanced embedding models and optimized vector databases to increase retrieval accuracy and speed. The system can also be enhanced by implementing multi-modal capabilities, enabling it to process not only text but also images, audio, and video inputs for a richer user interaction experience. Additionally, incorporating multilingual support will allow the system to understand and respond in multiple languages, making it more accessible to a wider range of users. Personalization features can also be added, where the system adapts responses based on user preferences, history, and behaviour.

The implementation of a stronger feedback and learning mechanism can further improve system performance over time by continuously refining the knowledge base and response generation. Moreover, integrating voice-based interaction can make the system more user-friendly and suitable for real-world applications such as virtual assistants and customer support systems.

Finally, the system can be deployed on cloud platforms with scalable architecture to handle large volumes of user queries efficiently. These enhancements will make the system more robust, intelligent, and capable of handling complex real-world scenarios effectively.

12. ACKNOWLEDGEMENT

We would like to express my sincere gratitude to my project guide for their valuable guidance, continuous support, and encouragement throughout the development of this project titled **"Domain-Specific Conversational AI using Retrieval-Augmented Generation (RAG)"**. Their insightful suggestions and constant motivation helped me in successfully completing this work**.**

We also thankful to the faculty members of the Computer Science Department for providing the necessary resources and knowledge required for this project. Their teaching and support played an important role in enhancing my understanding of the subject.

We would like to extend my gratitude to my institution for providing a good learning environment and facilities to carry out this project work.

13.CONCLUSION

In this paper, **a domain-specific conversational AI system based on Retrieval-Augmented Generation (RAG)** has been proposed to address the limitations of traditional Large Language Models (LLMs), particularly in terms of factual accuracy, contextual relevance, and real-time knowledge access. By integrating efficient retrieval mechanisms with generative models, the proposed system ensures that responses are grounded in reliable and domain-specific information rather than relying solely on pre-trained knowledge

The system incorporates advanced techniques such as optimized document chunking, dense vector embeddings, similarity-based retrieval, and re-ranking to enhance the quality of retrieved context. Additionally, the inclusion of a conversational memory module enables the system to handle multi-turn interactions effectively, maintaining coherence and continuity across extended dialogues. A hallucination mitigation mechanism further improves the reliability of generated responses by validating outputs against retrieved evidence.

Overall, the proposed approach provides a scalable and reliable framework for developing domain-specific conversational AI systems. It highlights the importance of combining retrieval and generation techniques to build intelligent systems capable of delivering accurate, context-aware, and trustworthy responses, making it suitable for real-world applications such as enterprise knowledge systems, customer support, and educational assistants.

14. REFERENCES

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," 2020.
2. T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, and D. Amodei, "Language Models are Few-Shot Learners," 2020.
3. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2019.
4. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," 2020.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

5. A.Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All You Need,” 2017.
6. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, and others, “LLaMA: Open and Efficient Foundation Language Models,” 2023.
7. O. Khattab and M. Zaharia, “ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,” 2020.
8. J. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, “REALM: Retrieval-Augmented Language Model Pre-Training,” 2020.
9. K. Lee, M. Chang, and K. Toutanova, “Latent Retrieval for Weakly Supervised Open Domain Question Answering,” 2019.
10. S. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H. Cheng, A. Jin, T. Bos, L. Baker, Y. Du, and others, “LaMDA: Language Models for Dialog Applications,” 2022.