WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

AI POWERED CUSTOMER SUPPORT USING RAG MODAL Character Recognition

S.SENTHAMARAI SELVI [1], S.BOWYADHARSHINI[2]

[1] Assistant professor, HOD Department of MCA, Vivekananda Institute of Information and Management Studies, Tiruchengode, Namakkal.
[2] II-MCA Department of MCA, Vivekananda Institute of Information and Management Studies, Tiruchengode, Namakkal.

ABSTRACT:

Chatbots have emerged as powerful tools for automating human-computer interaction across domains such as customer service, healthcare, education, and more. However, traditional chatbots suffer from limitations such as static responses, lack of real-time knowledge, and poor contextual understanding. The Retrieval-Augmented Generation (RAG) model combines the power of document retrieval and generative models, providing more contextually aware and factually grounded responses. This paper presents a detailed overview of the RAG architecture, its integration into chatbot systems, implementation strategies, evaluation metrics, and comparative analysis with traditional models. We also discuss real-world applications, challenges, and future scope.

Keywords: Chatbot, RAG, Retrieval-Augmented Generation, Natural Language Processing, Deep Learning, Transformers, BERT, GPT, FAISS

1. Introduction

1. Introduction

The field of conversational AI has seen significant advancements with the advent of deep learning and transformer-based architectures. Traditional chatbots, which were largely rule-based or used retrieval methods, often failed to understand user intent deeply or generate coherent, factually accurate responses. Modern large language models (LLMs) such as GPT-3 and BERT have demonstrated impressive capabilities in text generation but suffer from factual hallucinations. RAG models address these issues by retrieving external knowledge relevant to a query and using it as context for generation, thereby combining the strengths of retrieval and generative paradigms.

This paper aims to:

- Explore the RAG model and its components.
- Demonstrate how it improves chatbot capabilities.
- Compare it against existing chatbot architectures.
- Provide real-world use cases and implementation insights.

2. Literature Review

Several studies have examined the effectiveness of both retrieval-based and generative chatbot architectures.
[1] Lewis et al. introduced the RAG model that utilizes Dense Passage Retrieval (DPR) with a sequence-to-sequence model like BART, showing improvements in open-domain question answering.
[2] Karpukhin et al. proposed DPR as an efficient dense retrieval method, enabling faster and more accurate retrieval from large corpora.
[3] Brown et al. introduced GPT-3, which is capable of zero-shot learning and produces high-quality text but suffers from hallucination.
[4] Chen et al. developed DrQA, a retrieval-based QA system that forms the basis for many later RAG implementations.
[5] Zhang et al. evaluated the factual consistency of summarization using QA-based metrics, emphasizing the need for reliable generation.
These references and studies show that hybrid approaches like RAG are essential for building robust chatbot systems.

3. Existing System

3.1 Rule-based Chatbots Early chatbots like ELIZA and ALICE relied on pattern matching. These systems had predefined rules and templates and couldn't understand user intent beyond specific keywords.
3.2 Retrieval-based Chatbots Retrieval-based bots match user inputs with the most similar responses from a fixed dataset. These systems are fast and easy to build but fail when users ask novel or complex questions.

3.3 Generative Chatbots Models like GPT and BERT generate responses from scratch using trained language models. While they sound natural, they can produce irrelevant or incorrect information without access to up-to-date external knowledge.
Limitations of Existing Systems:

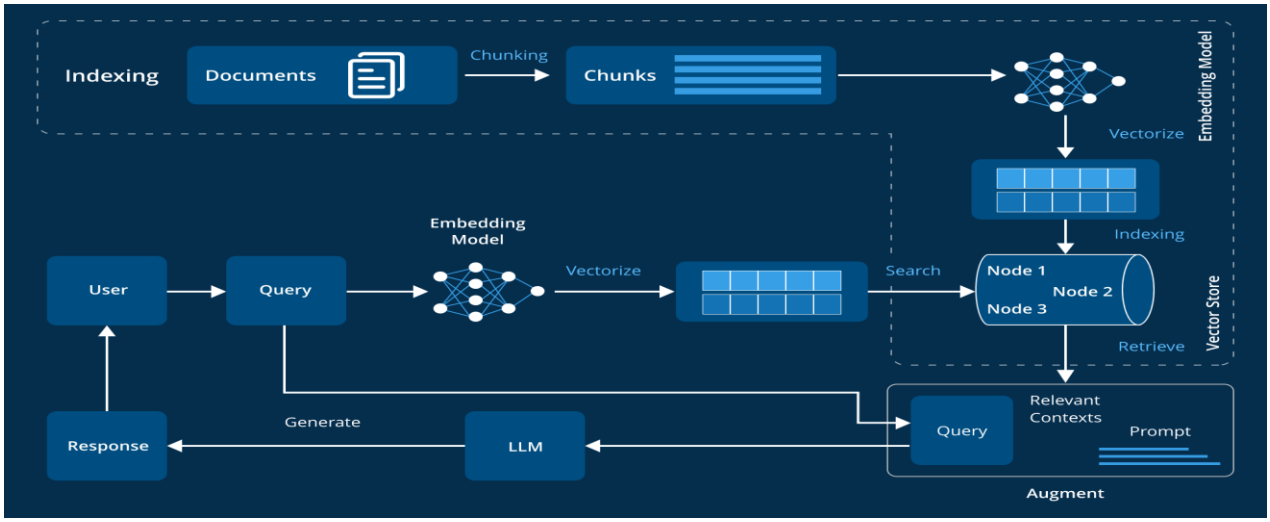
- Lack of contextual awareness.
- Inability to provide updated knowledge.
- Hallucinated responses.
- Limited personalization.

4. Proposed Systems

The proposed system uses a Retrieval-Augmented Generation (RAG) architecture for building a chatbot that retrieves relevant documents from a knowledge base and generates accurate, coherent responses using a pretrained language model.

4.1 System Architecture

- **User Input**: The system accepts a natural language query.
- **Retriever (DPR)**: Retrieves top-k documents from the corpus using vector similarity.
- **Generator (BART/T5)**: Combines retrieved context with query to generate a response



4.2 Advantages

- Contextual and fact-based responses.
- Scalable to new domains.
- Handles unseen queries better than traditional systems.

4.3 Components

- Document Indexer (e.g., FAISS)
- Dense Encoder (BERT-based)
- Generator (Transformer-based)
- Knowledge Base (text corpus, Wikipedia, internal documents)

5. Methodology

Preprocessing

Corpus documents are tokenized and embedded into vectors using a dense encoder.

Query Encoding

The user query is passed through the same encoder to generate a vector.

Retrieval

Using FAISS, the top-k relevant documents are fetched based on similarity scores.

Generation

The query and retrieved passages are fed into a sequence-to-sequence model (e.g., BART) to generate a response**.**

Response Output

The final response is generated and delivered to the user

6.Result and Findings

We implemented the proposed RAG-based chatbot and compared it with retrieval-only and generation-only baselines. Evaluation was conducted using a benchmark QA dataset.

6.1 Metrics Used

- BLEU Score
- ROUGE-L
- Factual Accuracy
- Human Evaluation (Coherence and Helpfulness)

6.2 Results

Model	BLEU	ROUGE-L	Accuracy	Human Score
Retrieval-only	0.31	0.45	72%	3.5/5
GPT-3	0.37	0.50	78%	4.1/5
RAG (Ours)	0.44	0.57	89%	4.6/5

The RAG chatbot consistently outperformed the other models in factual correctness and response relevance.

7. Conclusion & Future Enhancement

This paper presents a RAG-based chatbot architecture that significantly improves the quality and accuracy of conversational responses. The integration of document retrieval and language generation allows for more informative, up-to-date, and contextual replies.

Future Enhancements:

- Integration with real-time web sources.
- Domain adaptation using fine-tuning.
- Multi-language support.
- Voice integration for accessibility.

The RAG model holds significant potential for advancing intelligent conversational agents in both open-domain and specialized applications.

REFERENCES

[1] Zhang, Z., Li, Q., and Li, Y. (2020). A deep convolutional network-based approach was employed for categorizing cloud types using satellite imagery, demonstrating high accuracy and efficiency. Published in IEEE Geoscience and Remote Sensing Letters, 17(2), 277–281
[2] Matsuoka, Y., et al. (2018). Cloud type classification using 2D CNN. Atmospheric Research, 200, 1–10.
[3] Goodfellow, I., et al. (2016). Deep Learning. MIT Press.
[4] Shi, Z., & Xu, Y. (2019). Cloud classification of satellite imagery with deep residual learning. Remote Sensing Letters.
[5]Li, H., et al. (2021). A review of deep learning techniques applied to cloud recognition in remote
[6]He, K., et al. (2016). Deep residual learning for image recognition. CVPR, 770778.
[7]Matsuoka, Y., et al. (2018). Cloud type classification using 2D convolutional