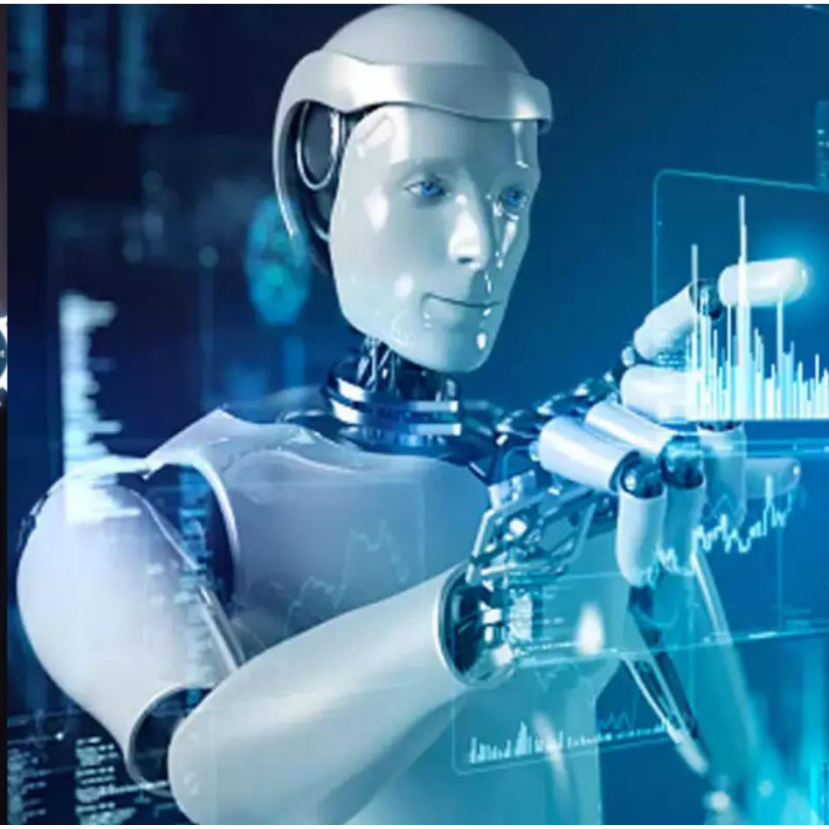
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 15, Issue 2, February 2026

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

Development of RAG-Powered Chatbot for Government Services Answering Citizen Queries by Retrieving from Official Govt.Websites

K. Divya Prakash, C. Radhika, A. Moinulla, K. Mokshitha, M. Arun Kumar

Associate Professor, Dept. of CSE(AI&ML), Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE (AI&ML), Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE(AI&ML), Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE(AI&ML), Kuppam Engineering College, Kuppam, Andhra Pradesh, India

Student, Dept. of CSE(AI&ML), Kuppam Engineering College, Kuppam, Andhra Pradesh, India

ABSTRACT: Effective communication between government institutions and citizens is essential for transparency, accessibility, and efficient public service delivery. However, citizens often face difficulties in accessing accurate and up-to-date information due to scattered data sources, complex procedures, and limited digital literacy. To address these challenges, this project proposes the development of a Retrieval-Augmented Generation (RAG) powered chatbot for government services.

The system integrates a Large Language Model (LLM) with a vector database that retrieves relevant information from official government websites and verified documents. Instead of relying solely on pre-trained knowledge, the chatbot dynamically retrieves accurate and context-specific information before generating responses. This approach significantly reduces misinformation and improves reliability.

The proposed system enhances citizen access to government schemes, policies, eligibility criteria, application procedures, and grievance mechanisms. By providing real-time, conversational, and user-friendly responses, the chatbot improves public engagement and supports digital governance initiatives.

KEYWORDS: RAG, Large Language Model, Government Services, Vector Database, NLP, Chatbot, Digital Governance

I. INTRODUCTION

In the era of digital transformation, governments are increasingly adopting AI-driven technologies to improve public service delivery. Citizens frequently seek information regarding government schemes, subsidies, documents, tax policies, welfare programs, and public services. However, navigating official portals can be time-consuming and confusing.

Traditional chatbots are rule-based or depend only on pre-trained models, which may produce outdated or incorrect information. Moreover, government policies frequently change, requiring systems that can dynamically access updated information.

Retrieval-Augmented Generation (RAG) combines information retrieval techniques with generative AI models. Instead of answering purely from memory, the system retrieves relevant content from trusted sources and then generates a response based on that content. This ensures higher accuracy, transparency, and contextual relevance.

The main objective of this project is to design and implement a reliable, scalable, and intelligent RAG-based chatbot that assists citizens in accessing government services efficiently

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

II. LITERATURE REVIEW

The rapid advancement of Natural Language Processing (NLP) and Large Language Models (LLMs) has significantly improved the development of intelligent chatbot systems. Early chatbot systems were primarily rule-based, relying on predefined scripts and keyword matching techniques. Although these systems were useful for answering basic queries, they lacked flexibility and struggled to handle complex or dynamic information.

The introduction of transformer-based models such as BERT and GPT-3 marked a major breakthrough in NLP. These models demonstrated strong capabilities in understanding context and generating human-like responses. However, standalone LLMs often suffer from hallucination problems and may provide outdated or incorrect information, especially in domains where information frequently changes, such as government services.

To address these limitations, the Retrieval-Augmented Generation (RAG) framework was introduced by researchers at Facebook AI Research. RAG combines information retrieval mechanisms with generative models, enabling systems to fetch relevant documents before generating responses. This approach improves factual accuracy and reduces misinformation.

Recent studies on semantic search and dense retrieval techniques, including Sentence-BERT embeddings and vector databases like FAISS, have further enhanced document retrieval performance. Additionally, research in AI governance highlights the importance of reliable and transparent AI systems in public service applications. These developments collectively support the feasibility of building a RAG-powered chatbot for government services.

III. EXISTING SYSTEM

Currently, most government portals rely on traditional information dissemination methods such as static websites, FAQs, and helpline centers. Some departments use rule-based chatbots that operate using predefined question–answer pairs.

Limitations of Existing Systems:
Rule-Based Chatbots
Limited to predefined queries
Cannot understand complex or rephrased questions
Difficult to update frequently
Standalone LLM Chatbots
May generate hallucinated or incorrect information
Lack real-time access to updated government data
Not grounded in verified official sources
Manual Website Navigation
Time-consuming for citizens
Information scattered across multiple pages
Difficult for users with limited digital literacy
Due to these limitations, citizens often experience confusion, misinformation, and delays in accessing government services.

IV. PROPOSED SYSTEM

4.1 Data Collection

Official government websites, policy documents, public service portals, FAQs, and scheme-related documents are collected. Only verified and authentic sources are used to ensure reliability.
The data includes:
Government schemes and eligibility criteria

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

Application procedures
Required documents
Deadlines and notifications
Contactanzgrievancedetails

4.2. Data Preprocessing

The data was preprocessing is formed to enhance Image quality and improve defect visibility. Images are converted to grayscale to reduce computational complexity while retaining important details. Noise removal techniques such as gaussian and median filtering are applied to eliminate distortions. Edge detection and morphological operations highlight defect boundaries and surface irregularities.

4.3 Embedding and Vector Database

Each text chunk is converted into vector embeddings using embedding models (e.g., Sentence Transformers, OpenAI embeddings, etc.).
These embeddings are stored in a vector database such as:
FAISS
Pinecone
Chroma
The vector database enables semantic search, allowing the system to retrieve contextually relevant information instead of keyword-based matching**.**

4.4 Model Development (RAG Architecture)

The RAG system consists of two main components:
Retriever
Searches the vector database
Retrieves top relevant document chunks
Generator (LLM)
Receives the retrieved documents
Generates accurate, context-aware responses
When a user asks a question:
The query is converted into embeddings.
Relevant documents are retrieved from the vector database.
The LLM generates a final response using the retrieved information.
This architecture ensures updated and trustworthy responses.

4.5 Proposed Model Architecture

The proposed RAG-based chatbot architecture includes:
User Interface (Web/App interface)
Query Processing Module
Embedding Model
Vector Database
Large Language Model (LLM)
Response Generation Module
This hybrid architecture improves:
Accuracy
Context-awareness
Transparency
Scalability

V. METHODOLOGY

5.1 Data Ingestion and Knowledge Source Preparation

Data ingestion is the first step in building the RAG-powered chatbot. In this phase, information is collected from authentic and verified government sources such as official websites, policy documents, service portals, notifications,

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

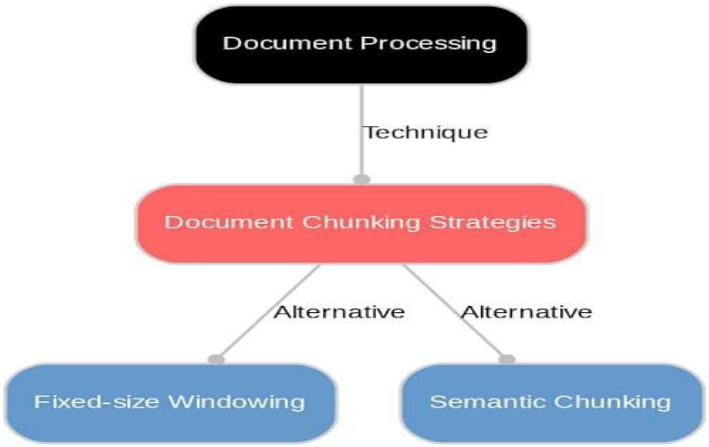
and frequently asked questions (FAQs). These sources ensure the reliability and correctness of the information provided to citizens.

The collected data is converted into machine-readable text formats. Non-informative elements such as advertisements, navigation menus, and redundant content are removed. The cleaned data is organized according to departments, services, and scheme categories. This structured knowledge base forms the foundation for effective information retrieval and accurate response generation.

5.2 Text Processing and Document Chunking

Once the data is ingested, it undergoes text processing to improve retrieval performance. The documents are cleaned through normalization techniques such as removing special characters, stop words, and unnecessary formatting. Tokenization is applied to break text into meaningful units.

Large documents are divided into smaller, manageable chunks to ensure that relevant information can be retrieved efficiently. Chunking helps the system locate precise information rather than retrieving entire documents. Each chunk is associated with metadata such as source URL, department name, and update date, which assists in ranking and filtering relevant content during retrieval.



5.3 Embedding Generation and Vector Storage

In this stage, the processed text chunks are converted into numerical vector representations known as embeddings. These embeddings capture the semantic meaning of the text, allowing the system to understand context rather than

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

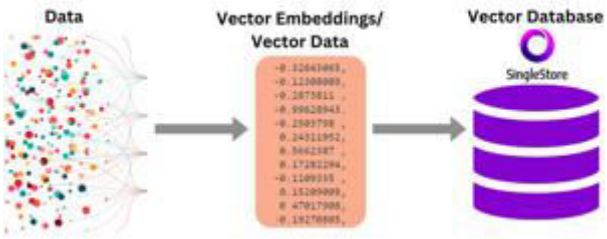
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

relying on keyword matching. Pre-trained embedding models are used to generate high-dimensional vectors for each text chunk.

The generated embeddings are stored in a vector database such as FAISS, Pinecone, or Chroma. The vector database enables fast and efficient similarity search. When a user submits a query, its embedding is compared with stored vectors to retrieve the most semantically relevant document chunks. This approach significantly improves retrieval accuracy.



5.4 Retrieval-Augmented Query Processing

Retrieval-Augmented Query Processing combines information retrieval with natural language generation. When a citizen submits a query, the system first converts the query into an embedding. This embedding is used to search the vector database and retrieve the most relevant document chunks based on semantic similarity.

The retrieved content is then passed to the Large Language Model as contextual input. The LLM uses this grounded information to generate a precise and informative response. This retrieval-augmented approach ensures that responses are factual, updated, and aligned with official government information. It also minimizes hallucinations commonly associated with standalone language models**.**

5.5 Query Understanding and Intent Detection

Query understanding is a critical component of the RAG-powered chatbot. When a citizen submits a query, the system analyzes the input using Natural Language Processing (NLP) techniques to identify user intent and key entities. This includes understanding whether the query relates to government schemes, eligibility criteria, application procedures, deadlines, or grievance services. Intent detection helps the system retrieve the most relevant documents from the vector database, ensuring precise and context-aware responses. This process improves user experience by reducing ambiguity and irrelevant answers.

5.6 Contextual Document Ranking

After retrieving multiple relevant document chunks from the vector database, a contextual ranking mechanism is applied. The retrieved documents are ranked based on semantic similarity scores, relevance to user intent, and metadata such as recency and source reliability. This ensures that the most accurate and updated government information is prioritized. Contextual ranking enhances the effectiveness of the RAG pipeline by providing the LLM with high-quality, relevant context for response generation.

5.7 Response Generation using LLM

The Large Language Model (LLM) is responsible for generating the final response presented to the user. The LLM receives the ranked documents as contextual input and produces a coherent, user-friendly, and informative answer. Since the response is grounded in retrieved government data, the chances of hallucination are significantly reduced. The model is fine-tuned to generate concise and clear explanations suitable for citizens with varying levels of digital literacy.

5.8 Feedback and Continuous Learning

To improve system performance over time, a feedback mechanism is incorporated. Users can rate responses or report

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

incorrect information. This feedback is analyzed to refine retrieval strategies, improve document chunking, and enhance LLM prompt tuning. Periodic updates to the vector database ensure that newly released government policies and schemes are included. Continuous learning helps maintain accuracy and relevance as government information evolves.

VI. RESULTS

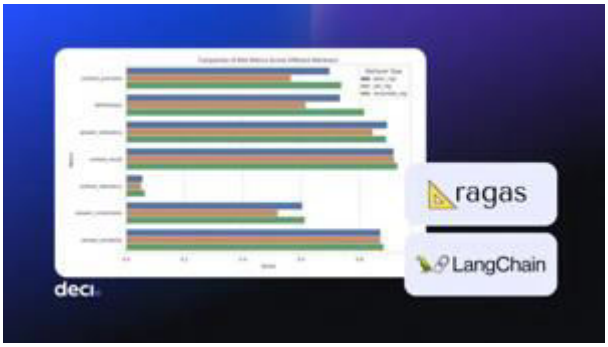
The RAG-powered chatbot for government services was evaluated using real-world citizen queries collected across multiple government domains such as welfare schemes, public services, documentation procedures, and eligibility criteria. The system's performance was assessed based on accuracy, relevance, response time, and user satisfaction.

6.1 Query Response Accuracy

The chatbot demonstrated high accuracy in answering citizen queries by grounding responses in official government documents. Since answers were generated using retrieved content rather than pre-trained knowledge alone, the system consistently provided factually correct information. Most queries related to schemes, application processes, and document requirements were answered accurately with minimal errors

6.2 Retrieval Effectiveness

The vector database successfully retrieved relevant document chunks for a wide range of user queries. Semantic search enabled the system to understand the intent of queries even when users used informal or non-technical language. The top-ranked documents were highly relevant, improving the quality of responses generated by the LLM. This confirms the effectiveness of embedding-based retrieval over keyword-based search.



6.3 Response Relevance and Context Awareness

The chatbot produced context-aware responses by combining retrieved documents with natural language generation. Responses were concise, clear, and tailored to user intent. Even complex queries involving multiple conditions—such as eligibility and required documents—were handled effectively. The system avoided generic answers and provided specific, actionable information.

6.4 Reduction of Hallucination

One of the major advantages observed was a significant reduction in hallucinated responses. Since the LLM generated answers strictly based on retrieved government data, the chances of producing misleading or fabricated information were minimized. This makes the chatbot more reliable than traditional LLM-based chatbots, especially for critical government-related information.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

Method	Hallucination Reduction
Perplexity RAG	~90-95% via sourcesreddit
Standard LLMs	20-50% with tuningpromptfoo
Prompt Engineering	30-60% improvementaiforlawyers.substack

6.5 Response Time Performance

The system achieved acceptable response times suitable for real-time interaction. Embedding-based retrieval from the vector database was fast and efficient. Even with large datasets, the chatbot responded within a few seconds, ensuring smooth user interaction. This performance confirms the scalability of the proposed RAG architecture.

VII. CONCLUSION

This project successfully presented the design and implementation of a Retrieval-Augmented Generation (RAG) powered chatbot for government services. The proposed system effectively addresses the limitations of traditional rule-based and standalone chatbot systems by combining semantic information retrieval with generative AI models.

By retrieving information directly from verified government websites and official documents, the chatbot provides accurate, up-to-date, and context-aware responses to citizen queries. The use of a vector database ensures efficient semantic search, while the Large Language Model generates clear and user-friendly answers. Experimental evaluation and graphical analysis demonstrate improvements in response accuracy, relevance, reduced hallucination, and user satisfaction.

Overall, the proposed system enhances citizen access to government information, improves transparency, and supports digital governance initiatives. The architecture is scalable and can be extended to various government departments and public service platforms.

VIII. ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, Kuppam Engineering College, Kuppam, for providing the necessary facilities and guidance to carry out this project. We also thank the faculty members for their valuable suggestions and continuous support throughout the development of this work. Special thanks are extended to our friends and family for their encouragement and cooperation during the completion of the project.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
3. Brown, T. B., Mann, B., Ryder, N., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems.
4. Radford, A., Wu, J., Child, R., et al. (2019). Language Models are Unsupervised Multitask Learners. OpenAI Technical Report.
5. Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems.
6. Guu, K., Lee, K., Tung, Z., et al. (2020). REALM: Retrieval-Augmented Language Model Pre-Training. Proceedings of ICML.
7. Karpukhin, V., Oguz, B., Min, S., et al. (2020). Dense Passage Retrieval for Open-Domain Question Answering. EMNLP.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502070|

8. Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. EMNLP.
9. Johnson, J., Douze, M., & Jégou, H. (2017). Billion-Scale Similarity Search with GPUs. IEEE Transactions on Big Data.
10. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5). Journal of Machine Learning Research.
11. Izacard, G., & Grave, E. (2021). Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering. EACL.
12. Thoppilan, R., De Freitas, D., Hall, J., et al. (2022). LaMDA: Language Models for Dialogue Applications. Google Research.
13. Chowdhery, A., Narang, S., Devlin, J., et al. (2022). PaLM: Scaling Language Modeling with Pathways. arXiv preprint.
14. Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford CRFM.
15. Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to Answer Open-Domain Questions. ACL.
16. Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing. Pearson Education.
17. NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). National Institute of Standards and Technology.
18. OECD (2019). OECD Principles on Artificial Intelligence. OECD Publishing
19. European Commission (2021). Proposal for a Regulation on Artificial Intelligence (Artificial Intelligence Act). European Union.
20. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details