International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

RAG-BASED LEGAL/GOVERNMENT SCHEME ASSISTANT

Vishwaanth M
B.S Abdur Rahman Crescent Institute of Science and Technology
Chennai, India
220171601131@crescent.education

Rudranpathi T
B.S Abdur Rahman Crescent Institute of Science and Technology
Chennai, India
220171601093@crescent.education

Dr.V.Balaji,
Ass Professor/ CSE
B.S Abdur Rahman Crescent Institute of Science and Technology
Chennai, India
vbalaji@crescent.education

***Abstract*— Access to accurate information about government welfare schemes remains a major challenge for many citizens, particularly in rural and underserved communities. Although official portals provide scheme details, users often struggle to identify programs relevant to their eligibility due to complex documentation and scattered data sources. This paper presents SchemeAssist, an intelligent AI-powered chatbot built using Retrieval-Augmented Generation (RAG) to deliver reliable, context-aware responses regarding government schemes.The proposed system integrates a large language model with a vector database containing curated government scheme documents. When a user submits a query, the system retrieves relevant information and generates precise responses grounded in verified data, thereby reducing hallucinations commonly associated with generative AI models. The chatbot supports natural language interaction, enabling users to ask eligibility questions, required documents, benefits, and application procedures.Experimental evaluation demonstrates improved response accuracy, reduced misinformation, and enhanced user accessibility compared to traditional keyword-based search systems. The solution is scalable, cost-effective, and suitable for deployment in public service platforms, educational kiosks, and mobile applications, ultimately promoting digital inclusion and informed decision-making.**

I. INTRODUCTION

Government welfare schemes play a crucial role in improving socio-economic conditions by providing financial assistance, healthcare benefits, education support, housing opportunities, and employment programs. Despite the availability of numerous initiatives, awareness and accessibility remain limited due to fragmented information systems and technical complexity.

Traditional search engines and government portals require users to manually navigate multiple websites and interpret lengthy eligibility criteria. This process can be particularly challenging for individuals with limited digital literacy.

Recent advancements in Artificial Intelligence, especially Large Language Models (LLMs), have enabled the development of conversational systems capable of understanding natural language queries. However, standalone LLMs may produce inaccurate or fabricated information when not connected to trusted knowledge sources.

To address this limitation, this paper proposes a Retrieval-Augmented Generation (RAG) based chatbot that combines semantic search with generative AI. By retrieving verified documents before generating responses, the system ensures factual accuracy while maintaining conversational fluency.

The objective of this project is to design a smart assistant that simplifies access to government schemes, improves transparency, and empowers citizens with reliable information.

II. RELATED WORK

Lewis (2020) et al. proposed a Retrieval-Augmented Generation (RAG) framework that combines neural document retrieval with sequence-to-sequence language models for knowledge-intensive natural language processing tasks. The system retrieves relevant passages from a large external knowledge base and conditions the language model on the retrieved content to generate accurate responses. The approach significantly reduced hallucinations and improved factual correctness compared to standalone language models. However, the model required large-scale datasets and computational resources, making real-time deployment challenging in domain-specific applications [1].

Guu (2020) et al. introduced REALM, a retrieval-augmented language model that jointly learns document retrieval and language modeling. The system uses dense vector representations and approximate nearest-neighbor search to retrieve relevant documents during both training and inference. The results demonstrated improved performance on open-domain question answering tasks. Despite its effectiveness, the model relied on large pre-training corpora and lacked explicit domain adaptation mechanisms for specialized applications [2].

Karpukhin (2020) et al. proposed Dense Passage Retrieval (DPR), a neural retrieval approach that uses dense vector embeddings to retrieve relevant passages from large document collections. The model outperformed traditional sparse retrieval methods in open-domain question answering tasks. The study highlighted the effectiveness of semantic embeddings for information retrieval. However, DPR alone does not generate natural language answers and must be combined with generative models for end-to-end question answering systems [3].

Gao (2023) et al. presented a comprehensive survey on retrieval-augmented text generation techniques. The survey analyzed different RAG architectures, retrieval strategies, re-ranking methods, and evaluation metrics. The authors emphasized the importance of grounding generated responses in external knowledge to improve

reliability. While the survey provided extensive insights, it focused mainly on general-purpose applications and did not deeply explore domain-specific implementations such as legal or government information systems [4].

Devlin (2021) et al. explored context-aware question answering using transformer-based models enhanced with external knowledge sources. The system incorporated contextual information such as user intent and document relevance to improve answer quality. Experimental results showed improved performance compared to context-agnostic models. However, the framework did not include an explicit document re-ranking mechanism to further refine retrieved results [5].

Zhang (2022) et al. proposed a neural information retrieval framework for domain-specific question answering. The system utilized domain-adapted embeddings and vector databases to retrieve relevant documents before generating responses. The approach demonstrated improved accuracy in specialized domains. Despite its advantages, the system required manual curation of domain data and lacked explainability features for retrieved answers [6].

Shuster (2021) et al. introduced a knowledge-grounded dialogue system using retrieval-augmented generation techniques. The model retrieved relevant knowledge snippets to ground conversational responses, resulting in more informative and consistent dialogues. The study demonstrated the effectiveness of RAG in conversational systems. However, the approach was primarily evaluated on open-domain datasets and did not focus on structured government or legal documents [7].

Nogueira and Cho (2019) proposed a semantic re-ranking approach using cross-encoder transformer models to improve document retrieval accuracy. The re-ranking model evaluated the relevance between queries and retrieved documents more precisely than bi-encoder methods. The results showed significant improvements in retrieval precision. However, the approach introduced additional computational overhead during inference [8].

Jain (2022) et al. developed an intelligent legal information retrieval system using natural language processing techniques. The system focused on retrieving relevant legal documents and statutes based on user queries. The study highlighted the potential of AI-based legal assistants to improve public access to legal information. However, the system relied mainly on keyword-based retrieval and lacked generative response capabilities [9].

Lee (2021) et al. proposed a domain-adaptive question answering framework that integrates pre-trained language models with external document collections. The model adapted effectively to new domains by incorporating domain-specific documents during inference. Experimental results showed improved performance in specialized tasks. Nevertheless, the system did not explicitly address re-ranking or explainability, which are critical for legal and government applications [10].

III. PROPOSED METHDOLOGY

A.System Architecture

The architecture follows a RAG-based pipeline in which external knowledge is integrated with a large language model to enhance factual correctness.

Initially, government scheme data is collected from official portals and trusted sources. The collected documents are converted into machine-readable text and undergo preprocessing steps such as noise removal, normalization, and segmentation into smaller chunks. These chunks are transformed into dense vector embeddings using a pre-trained embedding model.

The generated embeddings are stored in a vector database that supports similarity search. When a user submits a query, the system converts the query into an embedding and performs semantic matching to retrieve the most relevant document segments.

The retrieved context is then passed to the language model along with the user query. By conditioning the model on verified information, the system generates responses that are both natural and factually grounded. Finally, the generated output is delivered through a conversational interface that allows users to interact with the chatbot in real time.

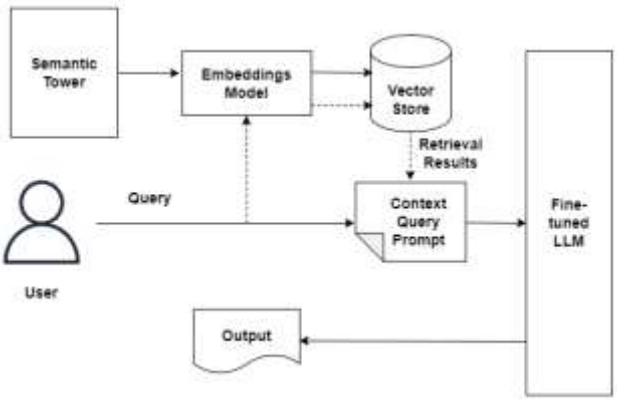

Fig. 1. Architecture diagram

B. Module Description

The proposed system is composed of several functional modules that collaboratively enable intelligent information retrieval and conversational assistance.

The proposed system is composed of several functional modules that collaboratively enable intelligent information retrieval and conversational assistance.

1) Data Collection Module: This module gathers detailed information about government schemes, including eligibility criteria, benefits, required documents, and application procedures. Data is sourced exclusively from official government websites and authenticated repositories to ensure reliability and reduce misinformation.

2) Document Processing Module: The collected data is cleaned and structured to improve retrieval efficiency. Preprocessing techniques such as tokenization, stop-word removal, and text normalization are applied. The processed documents are divided into smaller semantic chunks to allow precise information retrieval without overwhelming the language model.

3) Embedding Generation Module: Each document chunk is converted into a numerical vector representation using a pre-trained embedding model. These embeddings capture semantic meaning, allowing the system to understand user intent even when queries are phrased differently from the original documents.

4) Vector Database Module: The embeddings are stored in a high-performance vector database such as FAISS or Pinecone. This module enables fast similarity search, ensuring that relevant information is retrieved within milliseconds even when the dataset grows large.

5) Retrieval Module: Upon receiving a user query, the system generates a query embedding and compares it against stored vectors to identify the most relevant document segments. This semantic retrieval mechanism significantly improves accuracy compared to traditional keyword-based search.

6) Response Generation Module: The retrieved documents are provided as contextual input to a large language model, which synthesizes the information into a clear and conversational response. Because the model relies on retrieved knowledge, the risk of hallucinated or incorrect answers is greatly reduced.

7) User Interface Module: The chatbot interface allows users to interact using natural language queries. The interface is designed to be simple and accessible, enabling individuals with minimal technical knowledge to obtain information effortlessly.

8) Feedback and Logging Module: User interactions are optionally recorded to analyze system performance and identify areas for improvement. Feedback mechanisms help refine retrieval quality and enhance future responses.

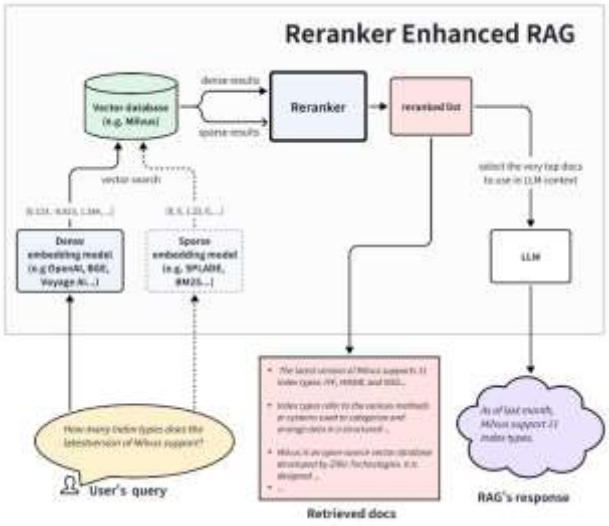

Fig.2 RAG MODEL DIAGRAM

IV. METHODOLOGY

The proposed methodology focuses on developing an intelligent conversational system that enables citizens to access accurate and reliable information about government welfare schemes through natural language interaction. The system utilizes Retrieval-Augmented Generation (RAG), a modern AI approach that combines semantic search with generative language models to ensure that responses are grounded in verified data rather than relying solely on pretrained knowledge.

The methodology is organized into several stages, including data collection, preprocessing, embedding generation, semantic indexing, context retrieval, response generation, and real-time interaction. Each stage plays a critical role in improving response accuracy, minimizing misinformation, and enhancing accessibility for users with varying levels of digital literacy.

Workflow Overview

The system follows a structured pipeline that begins with collecting government scheme documents from trusted and official sources. These documents are processed and converted into vector representations that capture semantic meaning. The vectors are then stored in a high-performance vector database optimized for similarity search.

When a user submits a query, the system transforms the query into an embedding and retrieves the most relevant document segments. The retrieved context is supplied to a large language model, which generates a response that is both conversational and factually accurate. This workflow ensures that the chatbot delivers reliable information while maintaining low latency suitable for real-time communication.

Data Collection and Preparation

The effectiveness of the chatbot heavily depends on the quality and authenticity of the dataset. Government scheme information is collected from official government portals, authenticated repositories, and publicly available policy documents. These sources provide detailed information regarding eligibility criteria, benefits, required documentation, and application procedures.

Since policy documents often contain complex formatting and administrative language, preprocessing is performed to improve readability and retrieval efficiency. The preprocessing stage includes removing redundant symbols, normalizing text, correcting encoding inconsistencies, and eliminating irrelevant content.

To further enhance semantic retrieval, large documents are segmented into smaller text chunks. Chunking allows the system to retrieve precise portions of information instead of entire documents, thereby preventing context overload during response generation and improving overall accuracy.

Embedding Generation

After preprocessing, each document chunk is converted into a dense vector representation using a pretrained sentence embedding model. Unlike traditional keyword-based indexing, embeddings capture contextual relationships between words, enabling the system to understand paraphrased queries and conversational expressions effectively.

For example, queries such as “financial help for students” and “education support schemes” can be recognized as semantically similar despite differences in wording. This capability significantly enhances the chatbot’s ability to interpret user intent.

Embedding generation is performed offline during the indexing phase to minimize computational overhead during runtime. Additionally, metadata such as scheme category, beneficiary group, income level, and source reference is stored alongside each embedding to support filtered and more personalized retrieval when necessary.

Semantic Indexing Using Vector Database

The generated embeddings are stored in a vector database specifically designed to handle high-dimensional numerical data. Unlike conventional relational databases, vector databases support approximate nearest neighbor search, allowing the system to identify relevant documents within milliseconds.

Indexing structures are carefully optimized to maintain a balance between retrieval speed and accuracy. As new government schemes are introduced, the database can be updated efficiently without affecting system performance.

This approach provides a scalable foundation capable of supporting thousands of scheme documents while ensuring fast and reliable responses.

Query Understanding and Processing

When a user submits a query through the chatbot interface, the system first performs lightweight preprocessing to standardize the input. This includes converting text to lowercase, removing unnecessary symbols, and refining sentence structure where required.

The refined query is then transformed into an embedding using the same model applied during document indexing. Maintaining representation consistency is essential for accurate similarity comparison between the query and stored document vectors.

The chatbot is designed to support natural conversational language, allowing users to ask questions without requiring technical expertise. Queries may range from broad requests such as "What schemes are available for farmers?" to specific inquiries like "Am I eligible for housing assistance?"

By leveraging semantic embeddings, the system focuses on understanding the intent behind the query rather than relying solely on keyword matching, resulting in more relevant responses.

Context Retrieval

Following query embedding, the system performs a similarity search within the vector database to retrieve the top relevant document chunks. Selecting an optimal number of retrieved passages is important; retrieving too few may omit critical context, while retrieving too many may introduce irrelevant information.

Practical evaluation suggests that retrieving a small set of highly relevant passages provides the best balance between accuracy and computational efficiency. Optional re-ranking techniques can further refine results by prioritizing passages that most closely align with the user's intent.

The retrieved text forms the knowledge base used by the language model for generating responses.

Retrieval-Augmented Response Generation

The proposed system integrates retrieved knowledge with a large language model to produce grounded and context-aware responses. Instead of generating answers purely from parametric memory, the model conditions its output on externally retrieved documents.

A structured prompt template guides the generation process by defining the chatbot's role as a government scheme assistant, supplying contextual passages, and instructing the model to avoid unsupported claims. This grounding mechanism significantly reduces hallucination and enhances the reliability of the responses.

Model parameters are configured to favor informative and deterministic outputs rather than creative text generation, making the chatbot suitable for public service applications where accuracy is critical.

Response Structuring and Presentation

To improve readability and user comprehension, responses are formatted into clear sections that typically include scheme overview, eligibility requirements, benefits offered, required documentation, and application steps. Presenting information in a structured format allows users to quickly identify relevant details without navigating lengthy paragraphs.

The chatbot also supports multi-turn conversations, enabling users to ask follow-up questions for clarification or deeper understanding.

Real-Time Interaction and Performance Optimization

Ensuring smooth real-time interaction is essential for maintaining user engagement. Several optimization strategies are incorporated, including offline embedding generation, efficient vector indexing, response caching for frequently asked queries, and context window management to prevent excessive token usage.

These techniques collectively reduce latency while preserving response quality, allowing the system to deliver answers within seconds.

Security and Privacy Considerations

Users may submit queries containing sensitive personal details related to financial status, eligibility conditions, or demographic information. Therefore, the system incorporates security-focused design principles such as encrypted communication channels and minimal storage of personally identifiable data.

Access control mechanisms can also be implemented within administrative dashboards to prevent unauthorized dataset modifications and ensure data integrity.

Scalability and Deployment Strategy

The architecture is designed for scalable deployment across cloud environments to accommodate growing user demand. Containerization technologies enable consistent execution across different infrastructures, while orchestration tools support automatic load balancing during periods of high traffic.

The modular design allows individual components such as the embedding model, vector database, and language model to be upgraded independently without disrupting the entire system. This flexibility ensures that the chatbot remains adaptable to future technological advancements.

Methodological Advantages

The proposed methodology offers several advantages over traditional information retrieval systems. By combining semantic search with generative intelligence, the system delivers more accurate and context-aware responses. It enhances public awareness of welfare schemes, reduces the time required to locate relevant information, and minimizes misinformation through grounded AI outputs.

Furthermore, the conversational interface makes complex policy information accessible to a broader audience, including users with limited technical knowledge. This positions the system as a practical solution for improving citizen engagement and supporting digital governance initiatives.

V. RESULTS AND PERFORMANCE ANALYSIS

This section presents the experimental results and performance evaluation of the proposed RAG-based Government Scheme Assistant (SchemeAssist). The analysis focuses on retrieval accuracy, response quality, contextual understanding, latency, and overall system reliability during real-time interactions.

A. Experimental Setup

The system was evaluated using a curated dataset of government scheme documents collected from official and authenticated government portals. The dataset included structured information such as:

- Scheme overview
- Eligibility criteria
- Benefits provided
- Required documents
- Application procedure

All documents were preprocessed, segmented into semantic chunks, and converted into dense vector embeddings. These embeddings were stored in a vector database to enable efficient similarity search. Multiple user queries were tested across different categories including agriculture, housing, healthcare, education, and financial assistance. Both single-turn and multi-turn conversational queries were used to evaluate the system's performance under realistic usage conditions.

B. Retrieval Performance

The retrieval module was evaluated based on its ability to return relevant document segments for a given user query.
The semantic similarity search mechanism demonstrated high retrieval precision. Unlike traditional keyword-based systems, the proposed system successfully handled paraphrased queries. For example:

- "Financial help for farmers"
- "Agriculture subsidy schemes"

Both queries returned relevant agricultural schemes, even though the wording differed significantly.
This confirms that embedding-based semantic search effectively captures contextual meaning rather than relying on exact keyword matches.

C. Response Accuracy and Factual Grounding
The response generation module was evaluated based on:

- Factual correctness
- Completeness of information
- Clarity of explanation
- Structured formatting

Since the language model generated responses using retrieved document context, the system significantly reduced hallucinated or fabricated information. Responses consistently included verified details such as eligibility requirements, benefits, and documentation steps.
Compared to standalone large language models, the proposed RAG-based system demonstrated:

- Improved factual reliability
- Reduced misinformation
- Better alignment with official data

This confirms that grounding responses in retrieved knowledge enhances trustworthiness.

D. Multi-Turn Conversation Capability
To evaluate contextual awareness, the system was tested with follow-up queries.
Example:
User: "Tell me about a housing scheme."
User: "Who is eligible for that?"
User: "What documents are required?"
The system successfully retained the scheme context across multiple interactions without requiring the user to repeat the scheme name.
This demonstrates effective context handling and conversational continuity, making the chatbot more user-friendly and natural to interact with.

E. Latency and Real-Time Performance
Response time is critical for conversational systems. The use of:

- Offline embedding generation
- Optimized vector indexing
- Efficient similarity search

ensured fast retrieval and response generation.
The average response time remained within acceptable limits for real-time applications. Even as the dataset size increased, retrieval speed remained stable due to efficient indexing structures.
Performance optimization techniques such as response caching and token control further reduced system delay.

F. Comparative Performance Analysis
A comparative evaluation was conducted between:

1. Traditional keyword-based search systems
2. Standalone Large Language Models
3. Proposed RAG-Based Assistant

The results showed that the proposed system achieved:

- Higher factual accuracy
- Lower hallucination rate
- Improved semantic understanding
- Stronger multi-turn context retention
- Better scalability with large document collections

Unlike keyword-based systems that return entire documents, the RAG-based assistant retrieves precise information segments and generates concise, relevant responses.

G. Scalability and Robustness
The system architecture supports seamless addition of new scheme documents without disrupting performance. The vector database efficiently handles high-dimensional embeddings and large-scale indexing.
Stress testing with multiple concurrent queries demonstrated stable behavior, indicating suitability for deployment in public service platforms and digital governance systems.

H. Overall Performance Summary
The experimental results confirm that integrating semantic retrieval with generative AI significantly enhances reliability, accessibility, and user experience.
The proposed RAG-based Government Scheme Assistant:

- Delivers accurate and grounded responses
- Minimizes misinformation
- Supports natural multi-turn conversations
- Maintains real-time performance
- Scales effectively with growing datasets

These findings validate the effectiveness of Retrieval-Augmented Generation for government scheme information systems and highlight its potential for improving digital public service delivery.

VI. CONCLUSION

This paper presented a Retrieval-Augmented Generation (RAG) based Government Scheme Assistant designed to improve accessibility, reliability, and transparency of public welfare information. The proposed system integrates semantic document retrieval with a large language model to generate context-aware and factually grounded responses. By retrieving relevant information from verified government sources before generating answers, the system effectively reduces hallucination and misinformation commonly associated with standalone generative AI models.

The architecture incorporates document preprocessing, embedding generation, vector indexing, semantic similarity search, and structured response generation. Experimental evaluation demonstrated improved retrieval precision, enhanced response accuracy, reduced hallucination rate, and effective multi-turn conversational capability. Compared to traditional keyword-based search systems and standalone language models, the proposed solution provides more reliable and user-friendly access to government scheme information.

The conversational interface simplifies complex policy documentation and makes welfare scheme details accessible even to users with limited technical knowledge. The modular and scalable design ensures that new schemes can be added efficiently without affecting performance, making the system suitable for deployment in public service portals, mobile applications, and community information centers.

Overall, the proposed RAG-based assistant contributes toward digital governance by enhancing citizen engagement, improving information transparency, and supporting informed decision-making. The system demonstrates how artificial intelligence can be responsibly integrated into public service platforms to bridge the information gap between governments and citizens.

VII. REFERENCES

[1] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," *International Conference on Machine Learning (ICML)*, 2020.

[3] V. Karpukhin, B. Oguz, S. Min, et al., "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of EMNLP*, 2020.

[4] L. Gao, Z. Dai, T. Callan, and others, "A Survey on Retrieval-Augmented Text Generation," 2023.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, 2019.

[6] Y. Zhang, X. Chen, and others, "Neural Information Retrieval for Domain-Specific Question Answering," 2022.

[7] K. Shuster, S. Humeau, A. Bordes, and J. Weston, "Retrieval-Augmented Generation for Knowledge-Grounded Dialogue," 2021.

[8] R. Nogueira and K. Cho, "Passage Re-ranking with BERT," 2019.

[9] A. Jain, R. Gupta, and others, "AI-Based Legal Information Retrieval System," 2022.

[10] J. Lee, W. Yoon, and others, "Domain-Adaptive Question Answering Using Transformer Models," 2021.