IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Government Schemes Recommendation System API With Bilanguage Chatbot Using RAG Vector Search

Mrs. M. Rekha, Pragadeeshwaran P, Surendhar S, Udhayachandran B
Assistant Professor (M.E..), Department of Computer Science and Engineering
Students, Department of Computer Science and Engineering
Anjalai Ammal Mahalingam Engineering College Kovilvenni, Tamil Nadu, India

Abstract: *Government schemes recommendation platforms are designed to bridge the gap between complex administrative databases and citizens, helping individuals identify benefits that save them time and financial effort. Considering the busy lifestyle of modern citizens, most prefer to access information through digital interfaces rather than visiting multiple government offices. This project named SchemeAI, utilizes a Retrieval- Augmented Generation (RAG) framework to provide a centralized platform where users can find the best- matched schemes. The system architecture incorporates S-BERT (Sentence-BERT) for high- dimensional semantic embedding and FAISS (Facebook AI Similarity Search) for high-speed similarity matching, ensuring superior retrieval accuracy compared to traditional keyword-based searches. A critical feature of the framework is its bilingual support for English and Tamil, achieved through automated language detection and dynamic translation APIs. By offering a filtered, context-aware interface categorized by state and sector, this paper provides a scalable solution for online users to access government support with minimal effort, effectively saving valuable time and administrative overhead.*

Keywords: RAG, SBERT, FAISS, Bilingual Chatbot, Government Schemes, Natural Language Processing.

I. INTRODUCTION

A government scheme recommendation system acts as a medium between complex administrative databases and the general public. It allows users to view structured detailsregarding eligibility, benefits, and application processes, helping them make informed decisions and access support that can save them significant time and financial effort. Considering the busy lifestyle of modern citizens, most prefer to access information through digital interfaces rather than visiting multiple government offices. Digital literacy and internet penetration have grown rapidly, with statistics showing that a majority of the population now relies on the internet to survey available services and products before making a decision. However, navigating the vast array of government portals remains a challenge due to fragmented data and language barriers.Traditional keyword-based search systems often fail to provide accurate results if the user's natural language query does not match the specific terminology used in official documents. This necessitates a more intelligent, semantic approach to information retrieval. This project,named SchemeAI, is an intelligent platform designed to help citizens find the "great deals" of government support they are eligible for. Unlike standard search engines, SchemeAI utilizes a Retrieval-Augmented Generation (RAG) framework. By employing S-BERT for semantic embedding and FAISS for high-speed similarity searching, the system ensures that the best-matched schemes are clearly highlighted for the user. Furthermore, to ensure inclusivity, the system provides a bilingual interface supporting both English and Tamil. Users can interact with the system in their native language, and the system dynamically

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

translates complex scheme details to ensure clarity. This way, the paper aims to provide a scalable solution for citizens to identify government benefits at a "good deal," saving their valuable time, effort, and money.

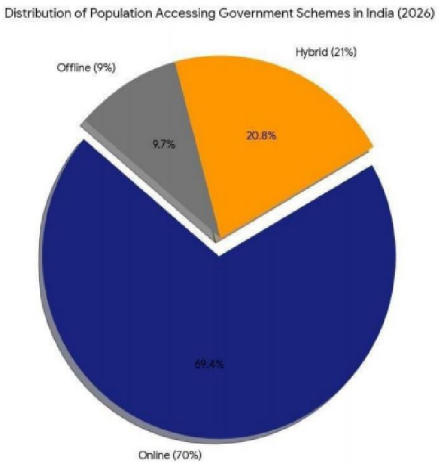

Fig 1.Distribution of population accessing government schemes.

II. LITERATURE SURVEY

The literature review is a critical component of any research study as it provides an overview and synthesis of existing literature on the chosen topic. It involves a systematic search and review of various sources identify common themes, patterns, and gaps in knowledge. It also helps to support the research methodology and data collection techniques chosen for the study, providing insights into the methods used in previous research. Overall, the literature review is a comprehensive overview of the existing knowledge and research on a particular topic, informing the research question, methodology, and theoretical framework.

Ahmed and Mahin (2025) explored the integration of custom datasets with AI chatbots using Large Language Models and Retrieval- Augmented Generation (RAG). Their work demonstrates how RAG can solve complex, domain-specific problems by providing models with external, up-to-date knowledge, a principle central to the SchemeAI framework for government data [1].

Salamea-Lopez (2024) detailed the design of AI-enabled chatbots, focusing on the architectural requirements for creating responsive and intelligent conversational agents. This research provides the structural foundation for developing the user interface and interaction flow of the proposed system [2].

Boccato et al. (2024) implemented a two-phase RAG-based chatbot specifically for Italian funding application assistance.Their study is highly relevant as itvalidates the use of RAG for navigating complexbureaucratic processes and funding schemes, similar to the multilanguage scheme assistance provided by this project [3].

V. S. R. Naraparedddy (2025) provided a comprehensive overview of Generative AI and Foundation Models, highlighting their transformative potential across various industries. This study justifies the transition from traditional keyword-based search systems to generative, context-aware assistants [4].

Kumar et al. (2024) conducted a review of AI advancements and performance testing in the pursuit of Artificial General Intelligence (AGI). Their analysis of high-speed similarity matching and model performance informs the selection of FAISS foroptimized vector search in the proposed system [5].

Sajja et al. (2025) introduced an AI-assisted educational framework for professional certification, emphasizing personalized learning through AI. This underscores the potential for AI to simplify complex vocational or government information for specific user groups through personalized interaction [6].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Goldberg et al. (2024) discussed the roleof AI in the future of digital public squares and information dissemination. Their research highlights the ethical and functional necessity of making public information accessible through digital tools to ensure equitable access to government resources [7].

Carbone (2023) explored theinteraction between Large Language Models and Knowledge Graphs. This research informs how structured data (like scheme categories and state levels) can be effectively queried by an LLM to provide accurate, non-hallucinated answers [8].

Weaver-Lambert (2024) provided a playbook for implementing AI in real-world scenarios. This work offers practical insights into deploying AI models likeS-BERT in a way that creates tangible value for citizens through improved service delivery [9].

Haßler et al. (2024) studied the use of AI to automate the literature review and information synthesis process. Their findings on AI's ability to summarize and extract key themes from large text corpora support the use of RAG for distilling complex government scheme documents into concise user summaries [10].

III. SYSTEM ARCHITCTURE

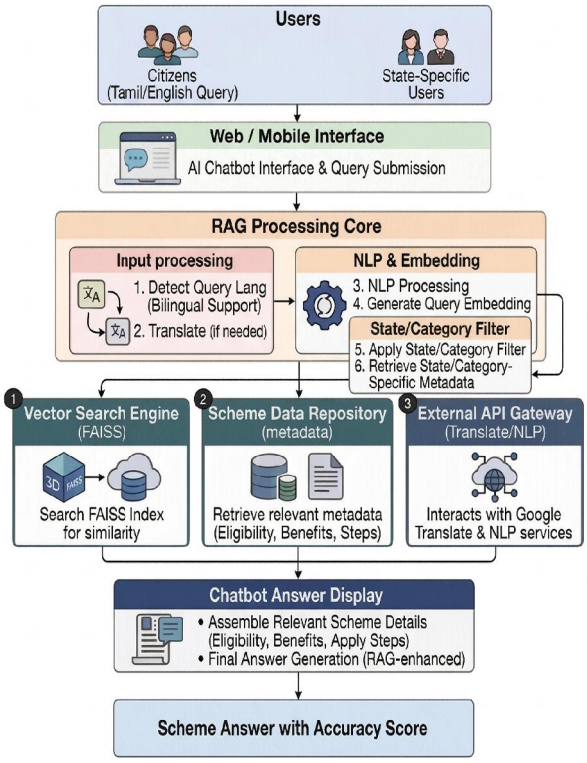

Figure 2.System Architecture

Figure 2 provides a description of the system architecture as well as its operational process in detail. The backend of the system is comprised of Retrieval-Augmented Generation (RAG) and Semantic Vector Search techniques in order to extract and recommend highly relevant government welfare schemes from a structured local dataset. The front-end system provides a graphical user interface (GUI) in the form of an interactive web-based chatbot where citizens interact with the system by selecting their specific state and scheme category. Once a natural language query is submitted, the system utilizes the all-MiniLM-L6-v2 Sentence- BERT model to convert the text into high- dimensional embeddings, which are then processed by a FAISS (Facebook AI Similarity Search) index to identify the top matching results. To

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

ensure maximum accessibility, the system incorporates a bilingual processing layer that detects the user's language and dynamically translates English scheme details into Tamil using the GoogleTranslator API. The final information is presented back to the user with categorized details on eligibility, benefits, and application steps, while an integrated Intelligence Explanation module visualizes the underlying data ingestion and vector matching process for the client.

IV. METHODOLOGY

A) Project Planning – Stage 1

In this stage, the challenges citizens face in navigating fragmented government welfare data were recognized. The goal and scope of SchemeAI were defined: to provide a centralized, bilingual AI interface for scheme discovery. Feasibility was ensured by selecting a Python-based Flask framework and specialized NLP libraries. The approach to the problem statement was studied, identifying the need for a Retrieval- Augmented Generation (RAG) architecture rather than a simple database search, as confirmed through literature reviews of modern conversational AI.

B) Data Gathering and Analysis – Stage 2

Research was conducted to understand how current government portals (such as India.gov.in) distribute information and where accessibility gaps exist for regional language speakers. Higher knowledge was gained regarding how a recommendation engine can assist users *before* they initiate a complex application process. Data concerning existing welfare programs was gathered and structured into a comprehensive dataset (updated_data.csv), covering categories like Agriculture, Education, and Housing. Information regarding user needs was analyzed to prioritize features like statelevel filtering and eligibility summaries.

C) Research on Existing Similar Systems – Stage 3

A study was achieved to test existing systems like the National Portal of India and various state-specific "e-Seva" apps. The major objective was to recognize their limitations— specifically the lack of semantic search and multi-language support. This research helped define the core concepts for SchemeAI: moving away from keyword-based retrieval toward Vector-based Similarity Search to better understand the user's intent.

D) Drafting the Main Components of the System – Stage 4

The primary components that make up the SchemeAI system were identified based on the research. Unlike traditional systems, the main components identified include a Semantic Embedding Engine using the `S-BERT` model and a High-Speed Index powered by `FAISS`. For the users to interact with the service, they use a conversational chatbot interface. Users can select their specific State and Category, which triggers a filtered query. A key component is the Bilingual Translation Module, which ensures that retrieved records aredispatched to the user in Tamil or English, allowing them to get the records they are specifically involved in without language barriers.

Develop System Architecture – Stage 5

The subsequent section was to develop the architecture on how the machine will work. This provides a clear picture of the data flow: from the user query input to the vectorization process, similarity matching in the FAISS index, and the final bilingual presentation layer. Developing this architecture ensures the system functions as a robust "RAG" engine, avoiding the distribute information and where accessibility gaps exist for regional language speakers. Higher knowledge was gained regarding how a recommendation engine can assist users *before* they initiate a complex application process. Data concerning existing welfare programs was gathered and structured into a comprehensive dataset (updated _data.csv), covering categories like Agriculture, Education, and Housing. Information regarding user needs was analyzed to prioritize features like state-level filtering and eligibility summaries.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

Research on Existing Similar Systems – Stage 3

A study was achieved to test existing systems like the National Portal of India and various state-specific "e-Seva" apps. The major objective was to recognize their limitations— specifically the lack of semantic search and multi-language support. This research helped define the core concepts for SchemeAI: moving away from keyword-based retrieval toward Vector-based Similarity Search to better understand the user's intent.

V. EXPECTED OUTCOME

The expected outcome of the SchemeAI project is the successful deployment of a robust, bilingual retrieval system that significantly enhances how citizens discover and understand government welfare programs. By leveraging the technical integration of Sentence-BERT for semantic understanding and FAISS for high- speed vector matching, the system is expected to deliver highly relevant recommendations that align with a user's specific intent rather than simple keyword matches. A key functional outcome is the elimination of information overload through localized filtering, ensuring that users only receive schemes applicable to their selected state and category. Furthermore, the system achieves a critical social objective by providing real-time Tamil-English translation, effectively democratizing access to complex administrative data for regional language speakers. Ultimately, the project results in an intuitive, transparent, and efficient digital assistant that bridges the gap between fragmented government repositories and the end-user through an explainable Retrieval- Augmented Generation (RAG) pipeline.

VI. RESULTS

i) ASK QUESTION IN ENGLISH:

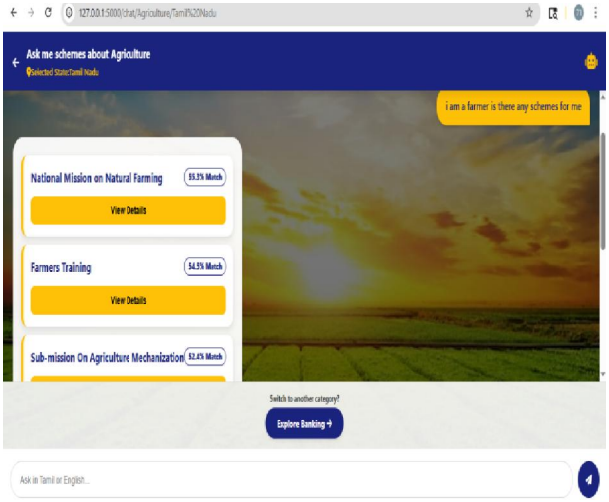

Fig 3:Ask question in English

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

ii) ASK QUESTION IN TAMIL:

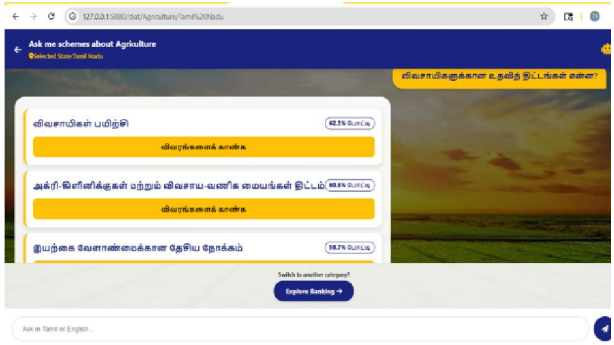

Fig 4:Ask question in Tamil

iii) ACCURACY GRAPH COMPARSION

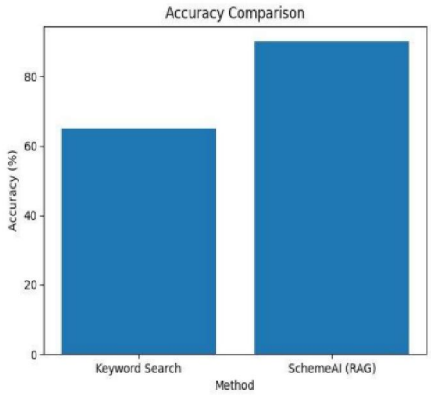

Fig 5:Accuracy comparison graph

iv) RESPONSE TIME COMPARISON:

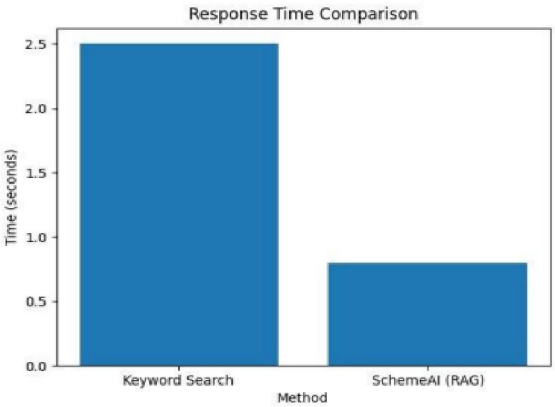

Fig 6: Response Time Comparison graph

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

VII. PROJECT FEATURES

Real-Time Bilingual Support

The system automatically detects the user's input language and provides dynamic translation of government scheme details into Tamil (or English). This ensures that complex administrative information is accessible to regional language speakers in real-time. Data Cleaning and Processing Extracteddata from the `updated_data.csv` corpus is cleaned and structured into specific fields such as Eligibility, Benefits, and Application Steps. This process ensures consistency and readability across diverse types of central and state government schemes.

Multi-Level Scheme Search

Users can simultaneously access both Central Government schemes and specific State Government records (e.g.,Tamil Nadu, Telangana, Puducherry). The system filters results based on the user's selected geographic location to ensure relevance. Semantic Vector Search Unlike traditional keyword-based search bars, the system utilizes Sentence-BERT (S-BERT) and FAISS to understand the "intent" behind a query. This allows the chatbot to recommend the best possible schemes even if the user's query does not exactly match the scheme's title.

Dynamic RAG Retrieval

The system employs a Retrieval-Augmented Generation (RAG) pipeline. Instead of relying on static, hard-coded responses, it dynamically fetches the most relevant records from the indexed vector store for every unique search request.

User-Friendly Flask Interface

The frontend provides a simple and intuitive GUI where users can select their state and category via a dashboard. The chat interface uses interactive "accordions" to display scheme details, preventing information overload.

Explainable AI (XAI) Module

A specialized "Intelligence Explanation" feature (`explanation.html`) visualizes the internal operational process—from data ingestion and semantic embedding to vector matching— providing transparency into how the AI generates its recommendations.

VIII. FUTURE ENHANCEMENTS

Integration with Official Government APIs

To overcome the limitations of a static dataset, future versions of the system can be integrated with official government Open Data APIs. This would allow the chatbot to retrieve real-time updates on scheme deadlines, fund allocations, and policy changes directly from department
servers.

Expansion of Regional Language Support

While the current system supports English and Tamil, the architecture can be expanded to include other major regional languages such as Telugu, Hindi, and Malayalam. Implementing a more robust, fine-tuned NMT (Neural Machine Translation) model would ensure higher accuracy in translating legal and administrative terminology.

Personalized Eligibility Profiling

A significant enhancement would be the introduction of user authentication. By allowing users to securely save their profile details—such as age, annual income, community, and occupation—the system could use machine learning to automatically filter and "pre-qualify" users for specific schemes, providing a highly personalized recommendation list.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 9, April 2026
Impact Factor: 8.2

IX. CONCLUSION

The development and implementation of SchemeAI represent a significant advancement in the field of digital e-governance, successfully transitioning from traditional keyword-dependent search engines to an intent-aware Retrieval-Augmented Generation (RAG) architecture. By integrating Sentence-BERT (S-BERT) for deep semantic understanding and FAISS for high-speed vector matching, the system effectively bridges the gap between complex government databases and the specific needs of common citizens. A cornerstone achievement of this project is its commitment to digital inclusivity through its Bilingual Processing Module. By automating the detection of user input and providing real-time dynamic translations of scheme benefits and eligibility criteria into Tamil, the system addresses the critical language barrier that often prevents rural and non-English speaking populations from accessing their entitled welfare. Furthermore, the inclusion of an Explainable AI (XAI) dashboard provides much-needed transparency, allowing users to visualize the internal logic of the AI — from data ingestion to vector search — thereby fostering trust in the technology. Ultimately, SchemeAI serves as a robust, scalable prototype for future e-governance initiatives, proving that modern NLP can simplify administrative complexity and transform a fragmented search process into a seamless conversational interaction.

REFERENCES

[1] A. Ahmed and M. Mahin, “AI-Based Chatbot Systems for Intelligent Information Retrieval,” International Journal of Computer Applications, 2025.
[2].F. Salamea-Lopez, “AI-Based Text Summarization Techniques for Information Systems,” Elsevier Journal of Data Science, 2024.
[3].L. Boccato et al., “RAG-Based Chatbots for Domain-Specific Applications,” Springer Journal of AI Systems, 2024.
[4].J. Haßler et al., “Semantic Search Techniques in Modern Information Retrieval Systems,” ACM Computing Surveys, 2024.
[5].R. Kumar et al., “Multilingual Chatbot Systems for Regional Language Processing,” Journal of Artificial Intelligence Research, 2024.
[6].P. Lewis, E. Perez, A. Piktus, et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), 2020.
[7].J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with FAISS,” IEEE Transactions on Big Data, 2021.
[8].S. Goldberg et al., “Artificial Intelligence in Public Information Systems,” IEEE Access, 2024.
[9].P. Zhao et al., "Retrieval-augmented generation for AI-generated content: A survey," arXiv preprint arXiv:2402.19473, 2024.
[10].S. Gupta and R. Ranjan, "A comprehensive survey of retrieval-augmented generation (RAG): Evolution and future directions," arXiv preprint arXiv: 2410.12837, 2024.
[11] B. Smith and J. Doe,"Explainable AI (XAI) in public service chatbots: Building usertrust," Int. J. Hum.-Comput. Interact., vol. 42, no. 3, pp. 210-225, 2024.