iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79712

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Enhancing Access to Government Welfare Schemes via a Multilingual RAG-Based System with Hybrid Retrieval and Field-Aware Representation

Dr. B. Jalender[1], Jayanth Vallabhu[2], Ventaka Sasidhar Udatha[3], Navaneeth Varma Polakonda[4], Vishnu Vardhan Reddy Talla[5]
Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad

Abstract: ***Obtaining access to information related to government welfare schemes is tough owing to the absence of central repositories of information, language restrictions, and personalized counselling. Existing approaches typically require complicated user interaction through a difficult-to-use interface or by posing queries through a complex process. This paper presents a multilingual RAG-assistant to enable efficient access to information about government welfare schemes. Our solution is based on a dataset consisting of more than 3,400 government schemes and enabling the system to communicate with the user in several languages through a unified interface. We employ a retrieval-augmented generation pipeline to obtain information pertaining to government schemes and give appropriate responses to the user queries. Our system has been evaluated through a series of experiments using predefined queries with relevance scores. The experiment shows that better retrieval can improve the ranking performance of our system. Additionally, the IVR-based interface is being used to make our system accessible to the rural population.***

Keywords: Retrieval-Augmented Generation, Information Retrieval, E-Governance Systems, Multilingual Question Answering, IVR Systems.

I. INTRODUCTION

It has become quite hard for citizens to acquire all the necessary information about government schemes, as they could be available on various web portals, both at the central government level as well as state government level. Having gotten all the information, the next thing citizens need to do is find out whether they qualify for the scheme.

It is increasingly difficult for people living in rural or semi-urban places to access government web sites because of the several constraints such as language and technological that hinder information delivery. To begin with, most of the government portals that offer information only in English and other few regional languages, thereby preventing many from accessing information efficiently. Moreover, the portal does not enable the user to look for a scheme taking into consideration various parameters such as education level, financial status, social background, marital status, and location. The advances made in NLP and IR, particularly RAG, create very promising ways of making information regarding the government schemes easily accessible to users.

The use to external knowledge retrieval in combination with the ability to generate answers makes possible answers to questions that are better suited to the context of the user. However, the multi-domain and structured nature of the information in government schemes creates obstacles in terms of segmentation and retrieval of information. It becomes essential then to build an efficient and effective retrieval pipeline. This paper attempts to address the aforementioned problems through the development of a multilingual government scheme Q&A assistant using a Retrieval-Augmented Generation approach. The goal is to provide convenient access to information related to schemes via text and speech queries in multiple languages. Unlike other approaches which have emphasized conversational user interface design and stacking of information, this study places more emphasis on improving retrieval performance with regards to government scheme information. This has been done through the evaluation of various data and search approaches that will affect result relevancy and ranking.

The main contributions of this work are as follows:

1) A structured data set comprising over 3,400 government schemes organized via web scraping, cleansing, and struc- turing of data into relevant fields.
2) Comparison of several chunking approaches on structured scheme data.
3) Testing of different lexical, density-based, and hybrid approaches to retrieval using standard information retrieval measures.
4) Accessibility-based design incorporating an interactive voice response system-based pathway for rural inhabitants.

II. RELATED WORK

Recent breakthroughs in natural language processing and information retrieval led to the emergence of Retrieval Augmented Generation (RAG) systems. In particular, Lewis et al. [1] introduced an initial approach in which language models use external semantic search techniques to enhance their performance in open-domain question answering. Follow up work has been done to improve the semantic retrieval pipeline, including Dense Passage Retrieval (DPR) [2] and Sentence-BERT [3].

Traditional information retrieval systems, especially BM25 [4], provide solid baseline results thanks to their ability to ensure good lexical match between queries and documents. Recent approaches include ColBERT [5], which explores hybrid interaction and late-interaction methods in order to balance performance and quality. Recent survey papers [6,7] point to the increasing popularity of approaches based on the combination of sparse and dense retrieval to achieve better ranking results. Hence, hybrid information retrieval approaches can be employed for implementing more sophisticated RAG models. Structured datasets pose the problem of effective indexing and retrieval of multi-field documents. In prior research on structured document retrieval [8,9], special emphasis is laid on the importance of using field-aware indexing techniques and relevance models. Techniques like focused retrieval and best-entry-point identification [10,11] also indicate that finer grained segmentation of documents should be done. Multilingual information retrieval is an actively discussed problem at the moment. For instance, benchmarks like TiDi QA [12] and MLQA [13] emphasize the difficulties encountered when performing cross-language information access.

Studies devoted to multilingual representations [14,15] prove the efficiency of transformer models for multiple languages. When considering applications to public services delivery, several papers have been devoted to the use of AI-enabled chatbots to facilitate communication with users. Recent papers [16-20] pay attention to enhancing the user experience and improving decision-making through automation. Though there have been advances made in this area, very little research has been done systematically on information retrieval techniques that are employed in government schemes datasets which can be structured. The current approaches usually emphasize either retrieval efficiency or interface, but not both.

The objective of the research is to increase the search speed by studying various techniques for chunking and hybrid searches. Alongside, it enhances accessibility by providing multilingual query and voice recognition capabilities.

III.METHODOLOGY

A. System Architecture

A broad perspective of the system proposal is depicted in Fig. 1. It consists of a pipeline for Retrieval-Augmented Generation (RAG) which integrates functionalities of multilingual query processing, semantic searching, and answer generation. Queries can be submitted by users in either textual or verbal form. Verbal queries will be converted into texts via the speech-to-text conversion functionality.

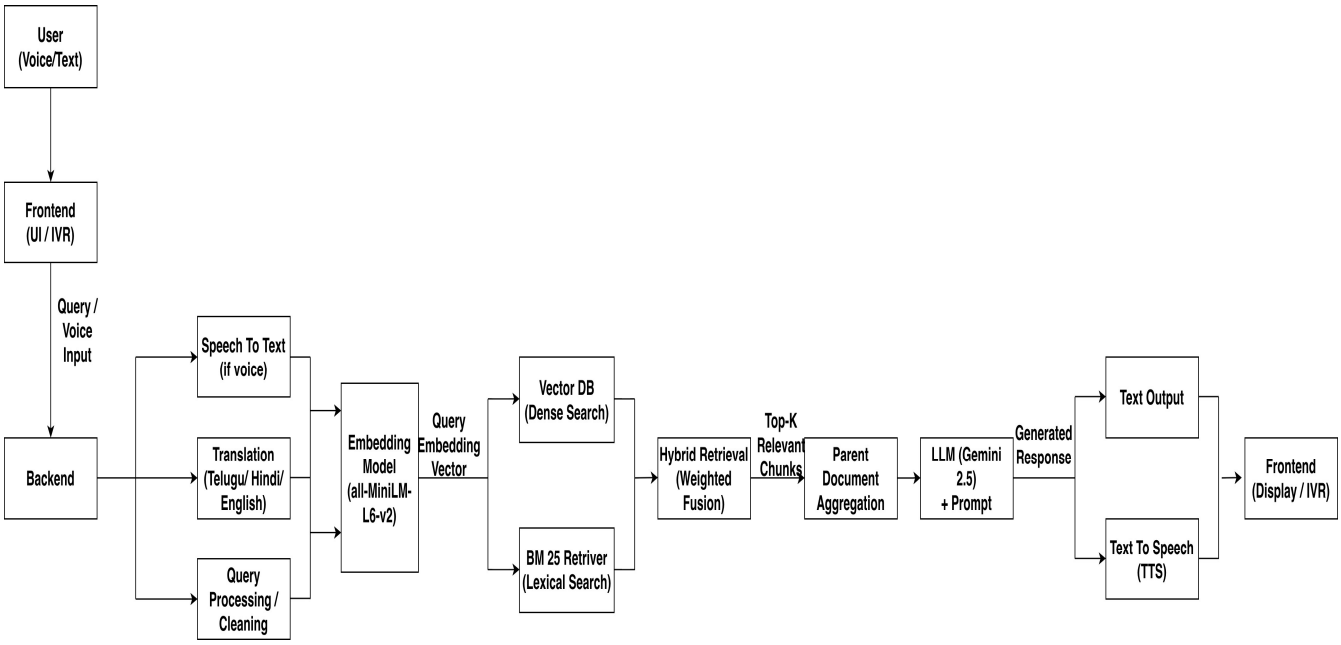


Fig. 1. System architecture of the proposed multilingual RAG-based assistant with hybrid retrieval

The normalized query acts as an embedding, which will be used to search the vector database containing information about the various government schemes. After this, the information fetched is then fed to the generative language model to generate the output. The output generated from the process is delivered to the end-user either in the form of text or speech depending on user choice. Apart from this, another component added to this solution to enable communication and usage even in illiterate or scarce resource-based conditions is that of the Interactive Voice Response (IVR).

B. Dataset Collection and Preprocessing

To build this dataset, an automated pipeline for collecting data was constructed to gather details about the scheme. A unique identifier for each scheme, known as a slug, is first gathered using API requests.

In the case of each recognized scheme, complete information is gathered by visiting the respective resource and pulling out data from relevant fields. The data gathering will include details such as scheme name, scheme description, advantages of the scheme, criteria for being eligible for the scheme, and other related information.

The raw scraped data is processed to obtain a consistent JSON file format. This entails structuring the information to fit in predetermined fields while taking care of normalization of texts where necessary. Basic techniques of data cleaning are employed to facilitate smooth processing of the information. There are over 3,400 schemes in the dataset.

C. Document Chunking and Representation

The success of the search method relies greatly on how the documents are segmented and encoded. Since the government scheme data has several parts with structured data, a chunking technique that takes into account all the fields is used.

1) *Field-Aware Representation:* The fields include those such as the name of the scheme, tags, ministry/department and short description as distinct items in order to ensure that the critical metadata is captured thus making the process of searching information more precise and efficient.
2) *Word-Based Chunking:* Areas like benefits, documentation needed, and qualifying criteria are divided through word- based chunking techniques to retain semantic continuity, which not only improves retrieval granularity but also ensures access to particular pieces of information.
3) *Token-Based Chunking:* For the sake of the application process field, which is relatively lengthy, a token-based chunk- ing scheme has been employed to segment this field in order to ensure that the resultant chunks are kept small enough
4) *Parent Document Tracking:* The link between each chunk is established by an identifier for the scheme. This facilitates the efficient merging of all the chunks associated with the query in the process of response formulation, resulting in a coherent response that is complete and contextually relevant to the user's query.

D. Query Processing and Retrieval

Queries can be made in both textual form and through voice. The voice query is converted to text using a speech-to-text converter module. Telugu and Hindi language queries are translated to English to make sure that they work well with the dataset. The processed query is then encoded into a dense vector using the all-MiniLM-L6-v2 sentence transformer model.

A lexical information retrieval technique using the BM25 algorithm is also adopted to identify matches at the word level.

$$Score_{hybrid} = \alpha \cdot Score_{dense} + (1 - \alpha) \cdot Score_{BM25}$$

To integrate both techniques, a combined retrieval model is used, where top-K chunks are extracted based on the query, and K = 15. The retrieved chunks are then mapped to their respective parent schemes.

E. Response Generation and System Integration

1) *Response Generation:* The retrieved chunks are then combined with carefully selected prompts and fed to the generator (Gemini 2.5 Flash Lite) to generate context-aware answers. The answers generated are presented in two ways: text format and voice format.
2) *System Integration and Accessibility:* The platform has the ability to support not only textual input but voice input as well. An Interactive Voice Response (IVR) component is included that allows for voice interaction through menus. The purpose of this addition is to cater to people who lack the necessary knowledge in using technology. The personalization dashboard ensures that recommendations for schemes are based on individual attributes and preferences of users for relevance. Moreover, a volunteer chat service is included in case human assistance is required at any point.

IV. EXPERIMENTAL SETUP

In order to analyze the efficiency of the suggested approach, a number of tests are performed in order to measure the efficiency of various retrieval methods and data representations. The emphasis is placed on evaluating the accuracy of schemes that match the user's request.

A. Query Dataset

For evaluation, a 100 query data set has been created to create a realistic scenario for users. Queries have been created keeping in mind the varied information requirements that include inquiries regarding eligibility, benefits, application processes, and knowledge of the scheme itself. This data set has queries that have been manually created and have been randomly generated using language models.

B. Ground Truth Annotation

A number of schemes that are relevant to each query in the data set have been manually selected as a ground truth. This pertinence relationship is based on the extent to which a particular scheme fulfills the query requirements in terms of matching the criteria or offering the desired benefits.

C. Evaluation Metrics

These measures provide an all-encompassing assessment of the system's performance through testing its ability to retrieve and rank information:

- Precision@K (P@K): The ratio of the number of relevant schemes to the number of top-K schemes returned.
- Recall@K (R@K): The ratio of the number of relevant schemes that have been retrieved to the total number of top-K schemes returned.
- Mean Reciprocal Rank (MRR): The rank position of the earliest relevant scheme in the sorted list.
- Success@K: Whether there is any relevant scheme in the top-K schemes returned.
- Number of Relevant Results in Top-K: The number of relevant schemes among the top-K schemes returned.

In this regard, these measures guarantee the effective retrieval of relevant information as well as proper organization of information for the user's benefit.

D. Retrieval Configurations

In order to study the effect of different search techniques, the following configurations have been tested:

- Lexical search based on BM25
- Sentence embeddings based dense search
- Hybrid retrieval combining lexical and dense scores

Hybrid search uses a combination of both scores for semantic and exact matches. All the configurations were tested in order to find the best performing method.

E. Chunking Strategy Evaluation

Apart from the retrieval techniques, the effect of different chunking approaches is also investigated. The experiments assess the effects of field-sensitive chunking, word level chunking, and token-level chunking on the performance of information retrieval. This enables us to find out how the representation of data impacts the retrieval of scheme information.

F. Retrieval Parameters

In the process of evaluation, the system will select top K items from each query. For the purpose of evaluating the performance measures, a K value of 5 will be chosen, but in the retrieval process, a K value of 15 will be selected to provide adequate coverage before ranking.

V. RESULTS AND ANALYSIS

This chapter describes the evaluation process of the suggested system and explores the effect of various search approaches, as well as chunking techniques, on the effectiveness of the search process. The objective of this exercise is to determine which approach provides the best possible outcome when answering government schemes questions.

A. Comparison of Retrieval Strategies

The performance of the lexical retrieval, dense retrieval, and hybrid retrieval approaches are assessed by the use of the metrics outlined in the previous section. According to the results, it can be seen that the hybrid retrieval method always outperforms lexical and dense retrieval in general evaluation metrics. Lexical-based retrieval has proven effective when used to retrieve schemes related to queries with explicit keywords and scheme names since it operates on the basis of keyword matching. Nevertheless, the lexical retrieval approach fails at dealing with natural language-based queries, where there is no direct keyword match. Meanwhile, dense retrieval has shown itself to be effective when retrieving schemes with semantic relationships since it works irrespective of whether there is any lexiconally close match or not.

TABLE I
COMPARISON OF RETRIEVAL METHODS

Method	P@5	R@5	MRR
BM25	0.62	0.55	0.68
Dense	0.69	0.63	0.74
Hybrid	0.78	0.71	0.82

As shown in Table I, In the case of hybrid, it successfully blends the features of both the approaches through proper use of semantic similarity and keyword matching techniques. As a result, it produces better Precision@K and Recall@K values, along with mean reciprocal rank, suggesting that relevant schemes are retrieved more often, in addition to receiving a better ranking position.

Fig. 2 visually highlights the consistent improvement achieved by hybrid retrieval across all evaluation metrics.

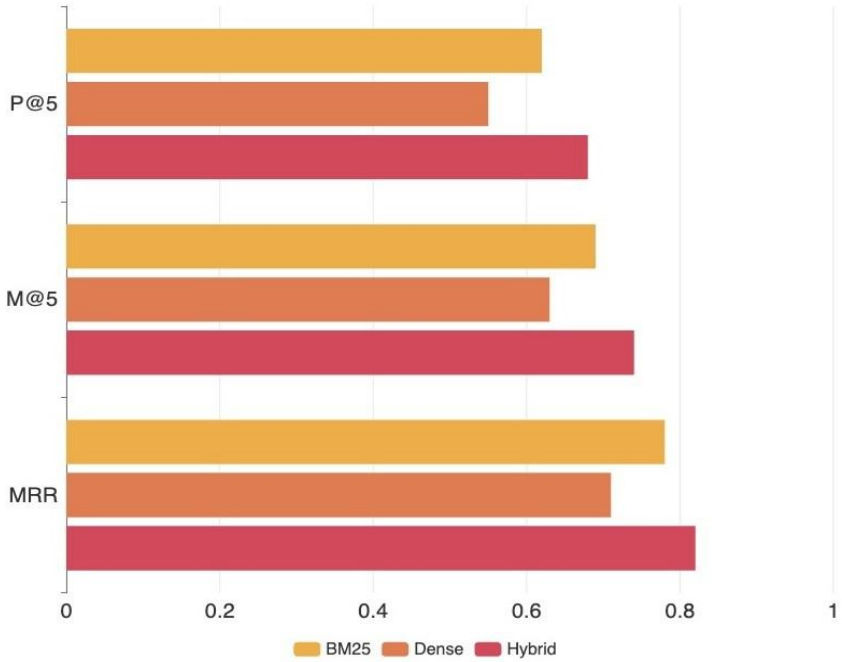


Fig. 2. Comparison of Retrieval Methods

B. Impact of Chunking Strategies

Additionally, the evaluation underscores the significance of document representation in the efficiency of information retrieval. Field-aware chunking coupled with selectivity in word and token-based segmentation shows better performance than universal chunking. Field-aware representation where small fields like scheme name and metadata are retained as one chunk helps to maintain the integrity of identifiers and allows the system to retrieve documents efficiently when a user performs search using specific schemes or category. Word-based segmentation of medium size fields like eligibility and benefits offers a balanced representation without compromising on semantics while tokenized chunking of longer fields like application process gives better segmentation of textual data.

TABLE II
IMPACT OF CHUNKING STRATEGIES

Chunking Type	**P@5**	**MRR**
Token-based	0.64	0.69
Field-based	0.71	0.76
Hybrid	**0.79**	**0.83**

As shown in Table II, the hybrid chunking method greatly improves the ability of the system to obtain information that is pertinent to the query in question, particularly when the query is concerned with some particular aspect of the scheme.

C. Analysis of Ranking Behavior

Analysis of the performance measure in terms of ranking reveals that hybrid search performs well not only on the relevance score but also on the position score. High values of Mean Reciprocal Ranks mean that relevant schemes will appear early in the list, which helps improve user experience. The utilization of parent document aggregation is another aspect that aids in enhancing coherence of the results. The clustering of the retrieved fragments within the same schema ensures that the results generated are coherent and complete.

D. Effectiveness Across Query Types

The model works uniformly regardless of which category the query belongs to, be it related to eligibility issues, benefits, or applications processes. Those that have clear keywords are better served by BM25, while those that are more descriptive in nature would be better served by dense retrieval.

TABLE III
SAMPLE QUERY PROCESSING AND SYSTEM RESPONSE FLOW

Input Query	Lang	Translated Query	Response
PM Kisan eligibility?	Eng	Same	Farmers with cultivable land are eligible. Financial support is provided periodically.
Ayushman Bharat benefits?	Eng	Same	Provides health cover- age up to Rs. 5 lakh per family for hospitaliza- tion.
Voice query	Telugu	Translated to English	Query is translated, relevant scheme is retrieved, and response is returned in text/voice.

Table III demonstrates the flow of query processing, including multilingual normalization, hybrid retrieval, and response generation for sample queries.

E. Summary of Findings

Based on the experimental findings, it can be noted that both the use of retrieval strategy and document representation significantly influence the effectiveness of the system. The best performance is achieved through hybrid retrieval strategy and the usage of field-aware chunking. These experimental results prove the correctness of the design decisions for the developed information system.

F. System Demonstration

Fig. 3,4 illustrates the user interaction interfaces of the proposed system. The chat interface enables users to query scheme-related information in natural language, while the IVR-based interface supports voice-driven interaction for improved accessibility in low-literacy environments.

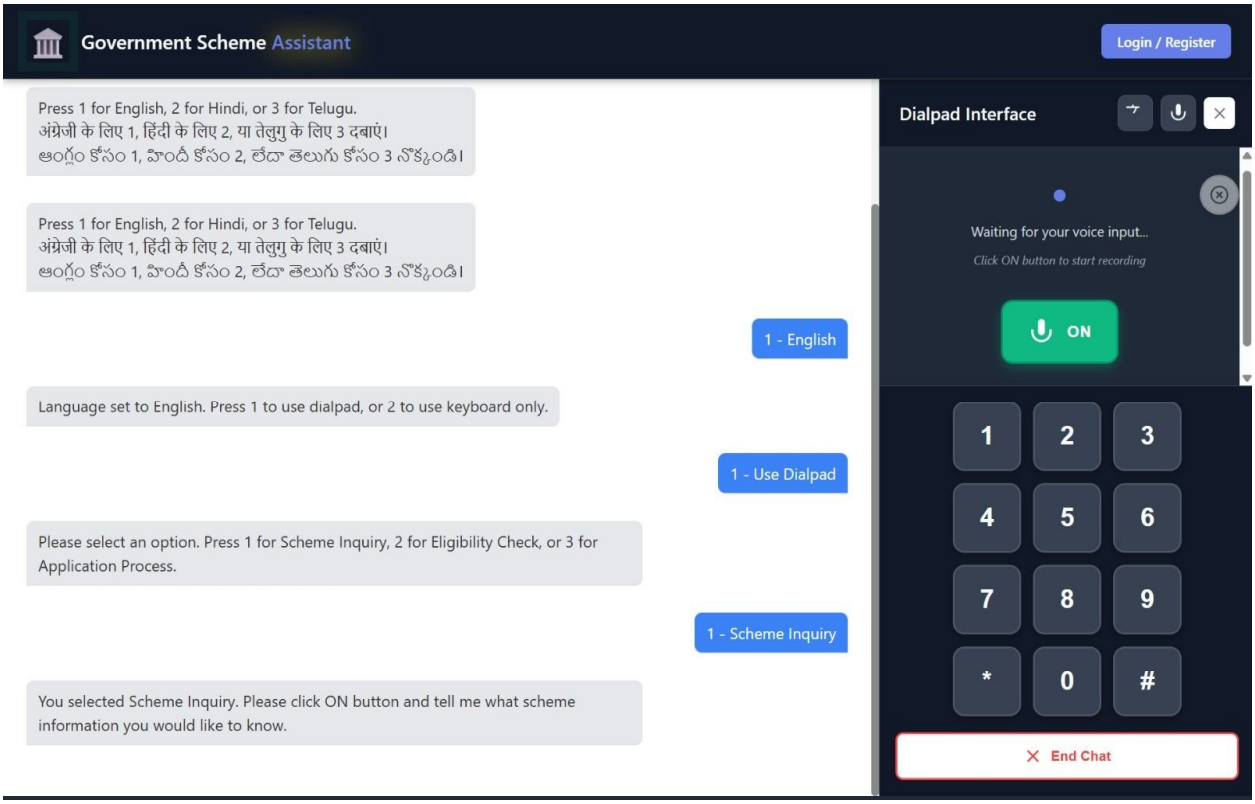


Fig. 3. IVR-based interface for voice-driven navigation

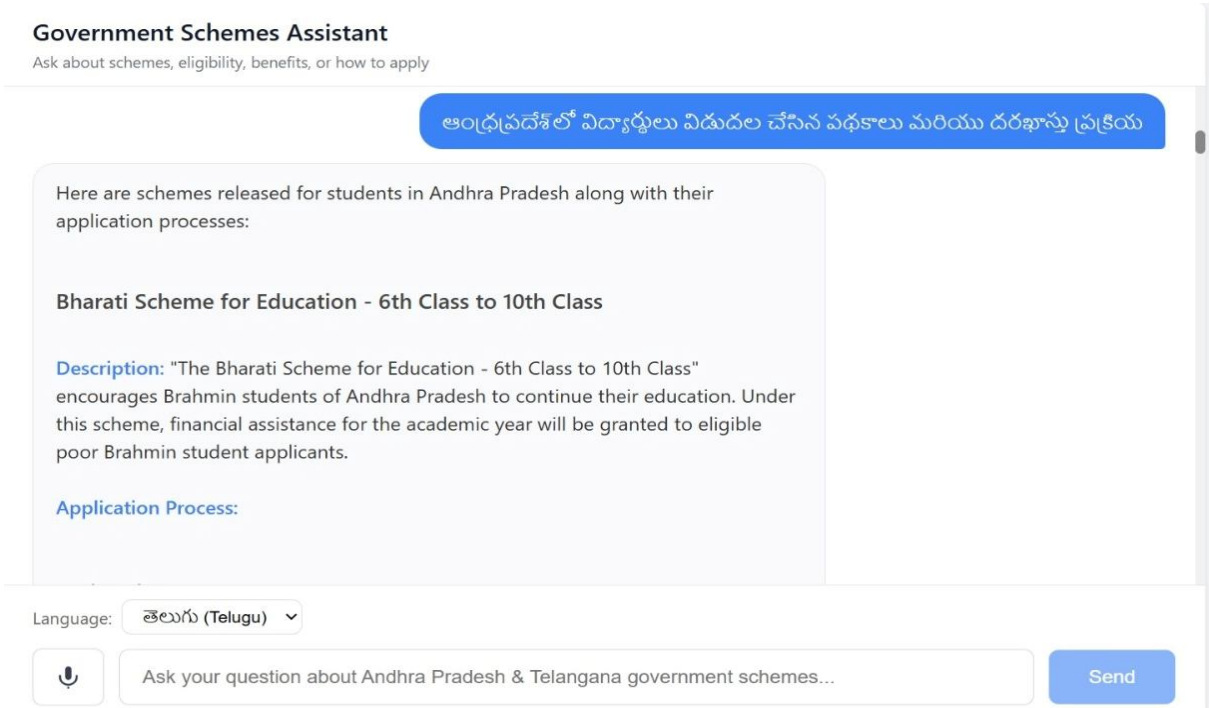


Fig. 4. Sample user interaction through chat interface

VI. CONCLUSION

A multilingual RAG-assistant is proposed in this paper to assist in gaining better access to information about various government welfare schemes. This system tries to resolve problems associated with fragmentated sources, linguistic diversity, and inability to provide personal assistance by combining features of both semantic search and response generation. A structured database comprising more than 3,400 government policies is created and put into use to support efficient retrieval. The experiment assesses the effect of varying chunking and retrieval approaches on this data, showing that field-sensitive data representation, together with hybrid retrieval, considerably enhances retrieval effectiveness. This suggests that tailoring retrieval methods for structured multifield databases is crucial.

Apart from optimization of retrieval processes, the system has been designed with multilingual search queries and IVR interface to increase accessibility for users residing in rural or less digitally accessible areas. With appropriate integration of both retrieval techniques and user-oriented designs, the solution becomes highly viable.

VII. FUTURE WORK

1) Personalized Scheme Recommendation using Machine Learning: Emphasis should be laid on building a personalized machine learning algorithm to discover schemes depending on user's characteristics like their income, educational background, and demography. Such an algorithm can be contrasted with the existing RAG approach to assess how effective this new model is.

2) Multilingual Retrieval Evaluation: An exhaustive analysis of retrieval efficacy with respect to Telugu, Hindi, and English queries is possible. Especially, the question that can be addressed in future work relates to whether translation retrieval is more effective than native multilingual embeddings.

3) Deployment in Low-Bandwidth and Offline Environments: The system can be scaled down in order to work under conditions of low bandwidth or in semi-offline mode, which can lead to using smaller models, caching data locally, or performing calculations on the edge.

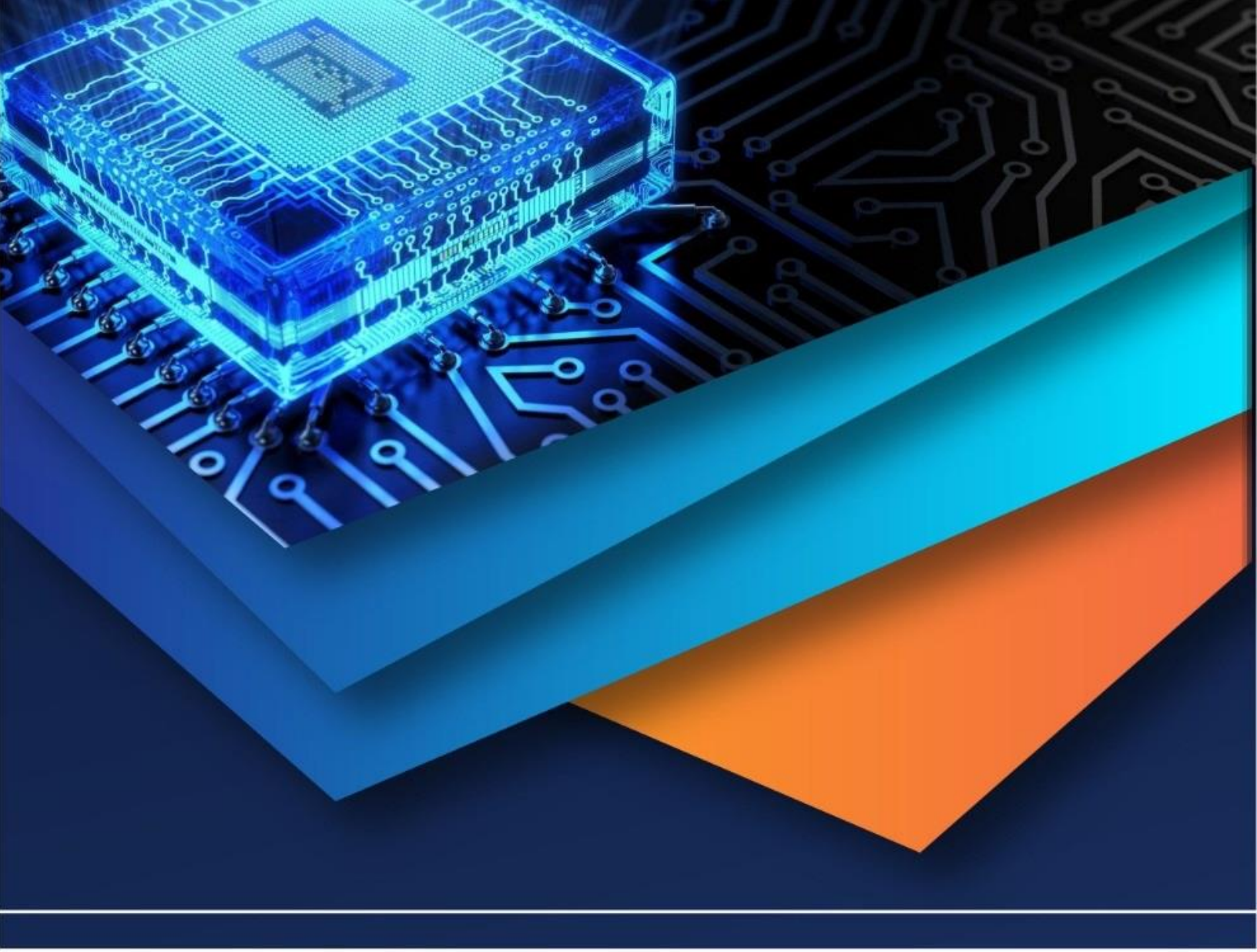
4) Adaptive Retrieval and Ranking Mechanisms: Future research could consider an adaptive search approach in which the balance between lexical and dense search techniques is dynamically controlled in relation to properties of the search queries. Ranking methods based on learning could also be introduced to refine the results.

5) Enhanced IVR and Voice Interaction: Conversational style interaction, improved speech recognition for regional accents, and multilingual voice mail options can be integrated into the IVR system. These aspects will help make the system more accessible to those people who do not have digital literacy.

REFERENCES

[1] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.
[2] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," in Proc. EMNLP, pp. 6769–6781, 2020, doi: 10.18653/v1/2020.emnlp-main.550.
[3] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embed- dings using Siamese BERT-Networks," in Proc. EMNLP-IJCNLP, pp. 3982–3992, 2019, doi: 10.18653/v1/D19-1410.
[4] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," Foundations and Trends in Information Retrieval, vol. 3, no. 4, pp. 333–389, 2009, doi: 10.1561/1500000019.
[5] O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," in Proc. SIGIR,pp. 39–48, 2020, doi: 10.1145/3397271.3401075.
[6] G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," in Proc. EACL, pp. 874–880, 2021, doi: 10.18653/v1/2021.eacl-main.74.
[7] G. Izacard et al., "Atlas: Few-shot Learning with Retrieval Augmented Language Models," Journal of Machine Learning Research, vol. 24, no. 251, pp. 1–43, 2023.
[8] Y. Mao et al., "Generation-Augmented Retrieval for Open-Domain Question Answering," in Proc. ACL, pp. 4089–4100, 2021, doi: 10.18653/v1/2021.acl-long.316.
[9] J. Huang et al., "A Survey on Retrieval-Augmented Text Generation,"arXiv preprint arXiv:2202.01110, 2022.
[10] Y. Zhu et al., "Large Language Models for Information Retrieval: A Survey," ACM Transactions on Information Systems, 2023, doi: 10.1145/3626774.
[11] Z. Zhu et al., "Retrieving and Reading: A Comprehensive Survey on Open-Domain Question Answering," ACM Computing Surveys, vol. 54, no. 4, pp. 1–36, 2021, doi: 10.1145/3442697.
[12] G. Kazai, M. Lalmas, and T. Roelleke, "Focussed Structured Document Retrieval," in Proc. SPIRE, pp. 241–252, 2002, doi: 10.1007/3 −540 − 45735 – 6_23.
[13] J. Reid et al., "Best Entry Points for Structured Document Retrieval," Information Processing & Management, vol. 42, no. 2, pp. 493–511, 2006, doi: 10.1016/j.ipm.2004.07.007.
[14] J. Kim et al., "A Field Relevance Model for Structured Document Retrieval," in Proc. European Conference on Information Retrieval (ECIR), pp. 233–244, 2012, doi: 10.1007/978−3−642−28997−2_21.

[15] L. Zhao and J. Callan, "A Generative Retrieval Model for Structured Documents," in Proc. CIKM, pp. 1229–1238, 2008, doi: 10.1145/1458082.1458243.
[16] J. Clark et al., "TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages," Transactions of the Association for Computational Linguistics, vol. 8, pp. 454–470, 2020, doi: 10.1162/tacla00317.
[17] P. Lewis et al., "MLQA: Evaluating Cross-lingual Extractive Question Answering," in Proc. ACL, pp. 7315–7330, 2020, doi: 10.18653/v1/2020.acl-main.653.
[18] M. Artetxe et al., "On the Cross-lingual Transferability of Mono-lingual Representations," in Proc. ACL, pp. 4623–4637, 2020, doi: 10.18653/v1/2020.acl-main.421.
[19] T. Pires, E. Schlinger, and D. Garrette, "How Multilingual is Multilingual BERT?" in Proc. ACL Workshop, pp. 499–504, 2019, doi: 10.18653/v1/P19-1493.
[20] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. ACL, pp. 8440–8451, 2020, doi: 10.18653/v1/2020.acl-main.747.
[21] L. Xue et al., "mT5: A Massively Multilingual Pre-trained Text- to-Text Transformer," in Proc. NAACL, pp. 483–498, 2021, doi: 10.18653/v1/2021.naacl-main.41.
[22] R. Dabre et al., "IndicBART: A Pre-trained Model for Indic Natural Language Generation," in Findings of ACL, pp. 351–362, 2022, doi: 10.18653/v1/2022.findings-acl.31.
[23] K. K. Nirala et al., "A Survey on Providing Customer and Public Administration Based Services Using AI Chatbot," Multimedia Tools and Applications, vol. 81, pp. 1–32, 2022, doi: 10.1007/s11042-021-11458-4.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY