ISSN: 2349-6002

AI-Driven Chatbot for Government Scheme Accessibility in India Using Hybrid Retrieval-Augmented Generation with Large Language Models

Vaishnavi P C[1], Arjun Paramarthalingam[2]
[1,2]*Computer Science and Engineering, University College of Engineering Villupuram, Villupuram, Tamilnadu, India – 605 103*
doi.org/10.64643/IJIRTV12I11-201405-459

***Abstract*—Retrieval of relevant government welfare schemes in India are still not an easy task, and accessing the required data is not an easy task because of scattered data sources, the heterogeneity of eligibility criteria, and a lack of custom search support. Current digital solutions are based on keyword-based retrieval, which is not as effective to obtain user intent in natural language queries. This paper introduces an AI-based chatbot that can overcome these constraints by proposing a hybrid Retrieval-Augmented Generation model to discover relevant government schemes. The system proposed is a combination of lexical retrieval with BM25 and semantic similarity search with dense vector embeddings that should be stronger in structured and ambiguous queries. The retrieved candidates will be refined on a cross-encoder re-ranking model and then a rule-based personalization layer will be used to restrict schemes based on user specific characteristics in terms of age, gender, state and eligibility conditions. It works on a maintained database of about 3,500 government schemes, which are broken down into nearly 70,000 structured text segments, in order to provide fine-grained retrieval. A locally deployed language model is employed to guarantee reliability of responses, where the outputs are rigidly based on retrieved context to perform controlled generation. Experimental results have shown that the combination of retrieval and re-ranking pipeline is more effective at increasing the relevance of the recommendations especially in queries that are implicit in purpose or incomplete. The given solution is a useful and privacy-friendly way of increasing access to governmental services, particularly in the situations when connectivity is limited.**

***Index Terms*—Retrieval-Augmented Generation, Semantic Search, Cross-Encoder Re-ranking, Government Scheme Recommendation, Large Language Models, E-Governance.**

I. INTRODUCTION

In India, government welfare programs constitute an essential part of the delivery of services to the population and they cater to the various need of the population both in the fields of healthcare, education, agriculture and social security. Although many schemes are available, their successful accessibility is not as high as possible because of the disintegration of data sources, inconsistent documentation format, and the intricacy of the eligibility criteria [1]. This is because users may be asked to search through various portals manually, and that is time-consuming, not to mention that it is also susceptible to errors. The majority of available digital platforms are based on the search mechanism by the use of keywords. Although these methods can be useful in precise term matching, they cannot be used in natural language query processing where the user intent is implicit or loose. As an illustration, the search query like schemes of women farmers or the financial support of old citizens need a context to comprehend rather than a mere overlap of the keywords [2].

This leads to incomplete or irrelevant results being often provided by traditional systems. The new developments in large language models (LLMs) have made it possible to have conversational interfaces that can comprehend user queries in a more natural way [3]. Granted, standalone LLMs are not necessarily based on verifiable information and can produce inaccurate or false answers. This is especially a major limitation in government related applications, where false data can have direct impact on decision-making [4].

The solution to this issue has been introduced as Retrieval-Augmented Generation (RAG) that integrates document retrieval and language generation. Hybrid retrieval strategies combining lexical techniques like BM25 with dense semantic embeddings in particular have helped to enhance retrieval robustness in the context of various query types [5, 6]. Also, re-ranking methods on cross-encoder structures also come in handy to increase the quality of results by assessing query-document relevance on a more specific scale. Nevertheless, the deployed RAG-based systems tend to be characterized by a number of limitations when used in the discovery of government schemes. To begin with, most of the implementations are based on cloud-based models, which add latency to the system and pose privacy issues. Second, it lacks or has limited personalization, although eligibility criteria are a key requirement to scheme applicability. Third, local and multilingual differences are not always managed and therefore, diminish usability across different societies.

This paper will solve these challenges by proposing a hybrid RAG-based chatbot which will be used specifically to access government schemes. The system combines the lexical and semantic retrieval and then cross-Encoder re-ranking, and one rule-based personalization module that can be used to filter the results according to the user characteristics like age, gender, and location. In contrast to generic RAG implementations, the suggested solution is based on a structured collection of government schemes and introduces a regulated response generation through a locally deployed language model, which guarantees privacy and reliability. One important design-related decision in the given work is the use of personalization following the re-ranking step, which prevents premature bias in the retrieval area and at the same time applies an eligibility restriction. Moreover, the representation of schemes in a structured form allows the uniform and interpretable responses. Empirical evidence suggests that such a combination enhances the usefulness of the recommendations, especially those of queries that contain incomplete or unclear input.

II. LITERATURE SURVEY

Retrieval-Augmented Generation (RAG) has become a useful paradigm to enhance the factual accuracy of large language models by basing generation responses on external knowledge sources. Previous research indicates that a combination of retrieval mechanism with generative models can reduce hallucination and enhance contextual relevance to a great extent [1, 2].

The systems normally fetch the appropriate documents before responses are formulated thus increasing reliability in knowledge-based applications. Scalping Hybrid retrieval models that combine sparse lexical retrieval strategies like BM25 with dense semantic encodings have received extensive studies to enhance retrieval robustness [7]. The sparse methods are efficient in precise matching of keywords, whereas the dense retrieval methods can catch the similarity between the query and document that are based on context. The available literature shows that these methodologies can give better results when used together in structured and ambiguous queries [8, 9], especially in domain-specific tasks.

Re-ranking mechanisms also improve quality of retrieval by eliminating the original candidate set. Unlike bi-encoder models, cross-encoder models encode query-document pairs together and generate more relevant relevance scores [10]. This method has been proven to enhance accuracy in highest ranked results particularly when documents that are semantically similar must be separated. It has made storing and retrieving high-dimensional embeddings a possibility thanks to the use of the vector databases like ChromaDB and FAISS [11]. These systems provide scalable semantic search and often find application in current RAG pipelines [8].

They are appropriate in real-time application with complex queries as they can be used to process large volumes of data. Increasing attention has also been given to personalization in retrieval of information. Systems can refine and further optimize recommendations by including attributes specific to users, including demographic data or setting preferences [12]. Nonetheless, the current methods have a tendency to personalize and memorize at an early stage of retrieval, which can be biased and decrease the memory [9]. Another crucial factor is the multilingual retrieval, especially in the linguistically diverse area like India. According to the recent research, multilingual embeddings were effective in processing queries beyond languages without direct translation requests, thus enhancing accessibility [13].

ISSN: 2349-6002

Irrespective of such developments, there are a number of limitations. A significant number of RAG-based systems are based on large language models that are cloud-based, and they are a cause of concern concerning latency, cost, and data privacy [9,11]. Also, personalization can be an inadequately integrated part of the retrieval pipeline, and most systems do not suit well on structured, eligibility-based domains including government schemes. Unlike them, the proposed work incorporates hybrid retrieval, cross-encoder re-ranking, and post-ranking personalization in a single pipeline that is specific to the government scheme recommendation. Moreover, offline service is also possible due to the utilization of a domestically installed language model, both in the areas of privacy and accessibility.

III. PROPOSED METHODOLOGY

A. System overview

The system suggested is based on a multi-step Retrieval-Augmented Generation (RAG) pipeline that will allow the precise and customized recommendation of government schemes. The architecture combines hybrid retrieval, re-ranking, personalization, and the controlled response generation. The system is fed with the query in a natural language and optional user properties and generates structured and context-sensitive responses.

The overall workflow consists of the following stages:

1. Query preprocessing
2. Hybrid retrieval (lexical + semantic)
3. Candidate merging and selection
4. Cross-encoder re-ranking
5. Personalization filtering
6. Context construction
7. Response generation using a local language model.

B. Query processing

Basic preprocessing functions such as lower casing and exclusion of unnecessary tokens are applied to the user query first. Besides this, user profile details including age, sex and state are added to the query context to facilitate downstream personalization. This stage will provide uniformity between query representation and indexed data.

C. Hybrid retrieval

To improve retrieval robustness, the system combines lexical retrieval using BM25 with semantic retrieval using dense vector embeddings.

1)Lexical Retrieval (BM25)

The BM25 is used to score documents based on term frequency and inverse document frequency:

$$BM25(q,d) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i,d) \cdot (k_1+1)}{f(q_i,d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \quad (1)$$

where:

- f (qi, d): term frequency
- |d|: document length
- avgdl: average document length.

2) Semantic Retrieval

Semantic similarity is computed using vector embeddings stored in ChromaDB. Each query and document is represented as a dense vector, and similarity is measured using cosine similarity:

$$sim(q,d) = \frac{q \cdot d}{||q||\,||d||} \quad (2)$$

3)Score Fusion

The final hybrid retrieval score is computed as:

$$Score(q,d) = \alpha \cdot BM25(q,d) + (1-\alpha) \cdot sim(q,d) \quad (3)$$

where $\alpha \in [0,1]$ controls the balance between lexical and semantic retrieval. This combination improves recall for semantically rich queries while retaining precision for keyword-based queries.

D. Candidate Selection and Re-ranking

The top-k candidates from hybrid retrieval are passed to a cross-encoder re-ranking model. Unlike independent encoding, the cross-encoder jointly processes the query and document, producing a relevance score:

$$R(q,d) = CrossEncoder(q,d) \quad (4)$$

This step refines the ranking by capturing fine-grained contextual relationships. It is particularly effective in distinguishing closely related schemes with overlapping keywords.

E. Personalization Module

A rule-based personalization layer is applied after re-ranking to ensure eligibility compliance. This module filters schemes based on:

ISSN: 2349-6002

- Age constraints
- Gender-specific eligibility
- State-specific applicability
- Category-based conditions (e.g., disability, occupation)

Applying personalization after re-ranking avoids early-stage bias while preserving retrieval diversity.

F. Structured Context Construction

The top-ranked schemes are converted into structured representations to standardize input to the language model. Each scheme includes:

- Scheme name
- Benefits
- Eligibility criteria
- Application process
- Required documents

This structured formatting reduces ambiguity during generation and improves response clarity.

G. Controlled Response Generation

The final response is generated using a locally deployed large language model (Mistral via Ollama). The model operates under controlled prompting, where only retrieved context is provided as input.

Key constraints include:

- Restricting generation to retrieved information
- Avoiding external knowledge injection
- Producing both structured (JSON) and descriptive outputs

This approach minimizes hallucination and ensures that responses remain grounded in verified data.

H. Dataset and Implementation Details

The system is built on a dataset of approximately 3,500 government schemes, collected from public sources and stored in JSON format. Each scheme is decomposed into smaller segments based on fields such as eligibility and benefits, resulting in nearly 70,000 text chunks.

The implementation is carried out in Python, integrating:

- BM25 for lexical retrieval
- ChromaDB for semantic search
- Cross-encoder models for re-ranking
- REST-based APIs for interaction
- Local LLM deployment using Ollama.

IV. SYSTEM ARCHITECTURE

A. Architecture Overview

The suggested system is based on a modular architecture that is aimed at combining retrieval, ranking, personalization, and generation modules into one pipeline. The architecture will provide an efficient flow of data as well as enable accuracy and scalability to the user query input to the final response generation. The system also has a user query and the query is processed and sent through a hybrid retrieval module that consists of a lexical and semantic search. The fetched candidates are optimized using the re-ranking phase, then a layer of personalization filters the results allowing user attributes to be used. Lastly, the response is generated with the help of a locally deployed language model using the contextual information that has been chosen.

B. Architecture Description

The following are the main components of the architecture:

1) User Interface

The system takes user input in the natural language. The interface has both textual query and structured user inputs like age, gender and location.

2) Query Processing Module

The query in the input is normalized to pre-process the query to normalize the text and add the attributes of user profile. This is done to be compatible with the lexical and semantic retrieval mechanisms.

3) Hybrid Retrieval Module

The component of this module is divided into two parallel components: BM25 Retrieval: It is used to search and retrieve an exact match of terms.

- Semantic Retrieval (ChromaDB): The contextual similarity is defined by a vector embedding.
- Both the components yield their outputs and the resulting output is merged to collect a candidate set of relevant schemes.
-

4) Merging and Selection of candidates

Lexical and semantic retrieval results are combined and the top-k results are chosen on combined scores.

ISSN: 2349-6002

5) Re-ranking Module

The candidate schemes are re-ranked with the help of a cross-encoder model where query-document relevance analysis is conducted at a more profound level. The step enhances the accuracy of the search as it narrows down on the results that are most contextually relevant.

6) Personalization Module

The re-ranked schemes are filtered using user specific properties which include:

- Age
- Gender
- State
- Eligibility Criteria

This is to make sure that the schemes presented to the user are of the appropriate schemes.

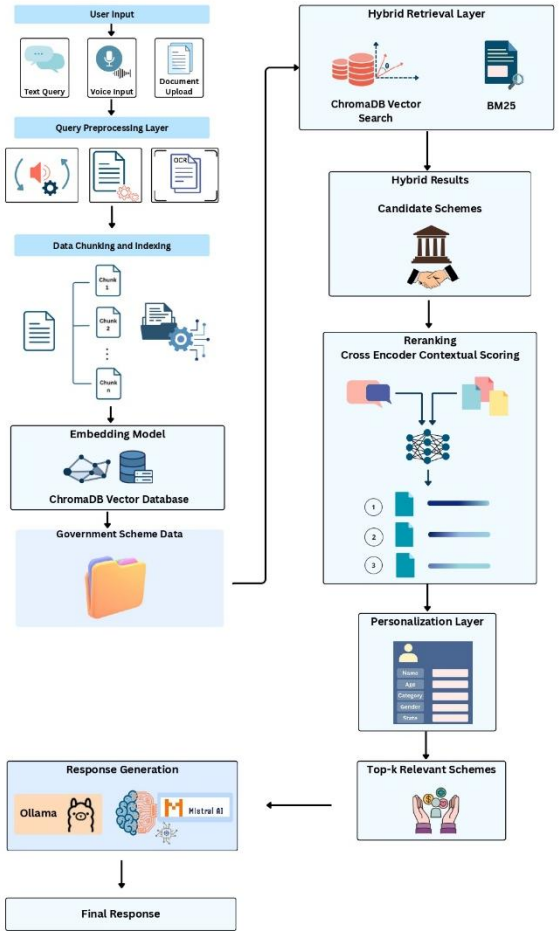

Fig. 1. System architecture of the proposed hybrid Retrieval-Augmented Generation framework for personalized government scheme recommendation.

7) Context Builder

The chosen schemes are transformed into structured format with benefit, eligibility, and application process being some of the important aspects of such a scheme. This structured context is given to the language model.

8) Response Generation (LLM)

The final response is produced by a locally-run large language model (Mistral using Ollama). The model is limited to utilize the context retrieved, thereby factually correct and hallucinating less.

9) Output Module

The system produces:

- Human-readable response
- Structured JSON output

This bi-directional output type takes care of end-users and downstream applications.

V. RESULTS AND EVALUATION

A. Experimental Setup

The proposed system was tested on a representative sample of 300 of the schemes in government, chosen by using many different sectors to represent a variety of the types of schemes, eligibility rules and types of benefits. This subset of schemes has been selected, although the whole dataset comprises about 3,500 schemes, so that it can be carefully evaluated manually. A list of user queries was created so as to imitate real world information requirements like queries based on eligibility and benefit-based queries as well as partially specified queries. The evaluation was carried out manually in about 2.5 hours in the process during which system response was assessed in terms of relevancy and accuracy. Because the dataset lacks relevance labels, relevance was evaluated manually. A retrieved scheme was relevant when it matched the intent of the query by the user and the eligibility requirements linked to it

B. Evaluation Metrics

The system performance was evaluated using the following metrics:

- Precision@5 (P@5): Proportion of relevant schemes among the top results
- Recall@5 (R@5): Fraction of relevant schemes successfully retrieved

ISSN: 2349-6002

- Mean Reciprocal Rank (MRR): Measures ranking effectiveness
- Response Time: Average latency per query
- Personalization Accuracy: Percentage of results satisfying user constraints

C. Retrieval performance comparison

The results indicate that the proposed hybrid approach significantly outperforms standalone retrieval methods. While BM25 performs well for exact keyword matching, semantic retrieval captures contextual meaning. The combination of both, along with re-ranking, results in improved overall performance.

Table. I. RETRIEVAL PERFORMANCE COMPARISON

RETRIEVAL PERFORMANCE COMPARISON		
Method	*Precision@5*	*MRR*
BM25	0.72	0.69
Semantic Retrieval	0.81	0.77
Hybrid Retrieval	0.87	0.83
Proposed (Full Model)	0.91	0.88

D. Overall System Performance

As shown in Table I and Table II, the proposed system achieves high retrieval accuracy while maintaining efficient response time suitable for real-time applications.

Table. II OVERALL SYSTEM PERFORMANCE

OVERALL SYSTEM PERFORMANCE	
Metric	*Value*
Precision@5	0.91
Recall@5	0.84
Mean Reciprocal Rank	0.88
Response Time	1.3 s
Personalization Accuracy	0.91

E. Ablation Study

The ablation study highlights that hybrid retrieval improves performance over individual methods, while the addition of cross-encoder re-ranking further enhances top-ranked precision.

Table. III ABLATION STUDY RESULTS

ABLATION STUDY RESULTS		
Configuration	*Precision@5*	*MRR*
BM25 (Lexical Only)	0.72	0.69
Semantic Retrieval only	0.81	0.77
Hybrid Retrieval only	0.87	0.83
Hybrid + Re-ranking (Proposed)	0.91	0.88

F. Discussion

Some key observations have been brought out by the scientific findings of the experiment. The hybrid retrieval mechanism enhances the system to process varied type of queries especially those of an implicit intent. Semantic retrieval assists in capturing contextual meaning whereas BM25 can be used to make sure domain specific keywords are included. The re-ranking step helps to enhance the accuracy as it separates the schemes that have similar description but whose eligibility rules are different. This is especially conspicuous with high-ranking results. The personalization model also makes the use more usable and allows filtering of schemes, like by age, gender, location, etc. Re-ranking with personalization ensures the maintenance of diversity as well as compliance with eligibility. The system has an average response time of 1.3 seconds despite the multi-stage pipeline, which shows that it is suitable to be deployed in real-time.

G. Error Analysis

Although it is a system that works well in general, some limitations have been identified:

- Incomplete user information queries would sometimes lead to a wider recommendation.
- The overlapping eligibility resulted in almost the same scheme propositions.
- Confusion of queries decreased ranking confidence in certain situations.

Also, the assessment was performed on a subset of the dataset, and it might be different when applied to the entire dataset. Nevertheless, the subset chosen cuts across several industries and gives a fair estimate of the real-life situation.

ISSN: 2349-6002

H. Results Visualization

1) Precision comparison across retrieval methods graph

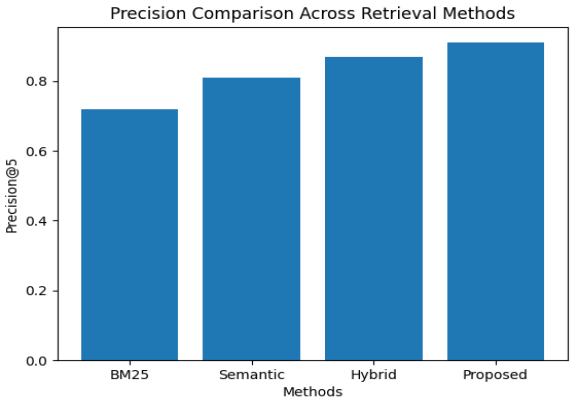

Fig. 2. Precision@5 comparison across BM25, semantic retrieval, hybrid retrieval, and the proposed model.

2) Ablation study performance graph

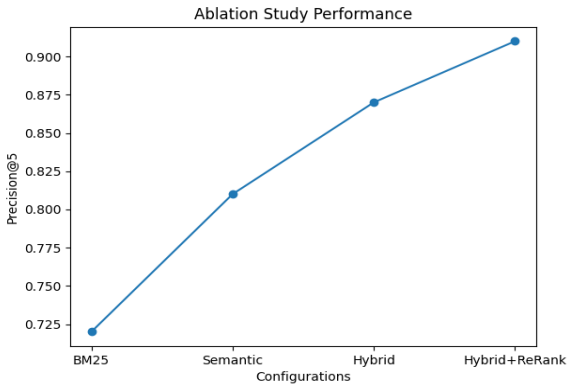

Fig. 3. Ablation study showing the impact of hybrid retrieval and re-ranking on Precision@5.

VI. CONCLUSION

The paper described an artificially intelligent chatbot system to enhance access to government welfare programs in India through a hybrid Retrieval-Augmented Generation (RAG) system. The system proposed combines lexical retrieval (BM25) and semantic search (ChromaDB) with a mechanism of re-ranking of the results of cross-encoders to make the retrieved schemes more relevant. Also, there was the addition of a post-retrieval personalization module that sifted results on attributes specific to their user e.g., age, gender and location. The experimental findings illustrate that the hybrid retrieval strategy is very effective in overall performance in comparison to the individual methods with high accuracy in the top ranked results and still without compromising on the response time. Relevance is further refined by the addition of re-ranking which is especially useful in cases where there is overlap in describing schemes. The personalization layer will make sure that the recommendations are applicable and relevant to the eligibility conditions of the user. The major addition of this work is hybrid retrieval, re-ranking, and post-filtering personalization in a single pipeline that has been optimized to support structured and eligibility-driven data. Moreover, local deployment of language model can be used to generate controlled and context-sensitive responses and deal with issues of latency, cost, and data privacy. Comprehensively, the suggested system indicates the ability of integrating retrieval-based and generative methods to reinforce intelligent and inclusive digital governance. The framework is able to help users to find pertinent schemes more quickly and can be expanded in other areas that need access to structured information.

VII. FUTURE WORK

Further development could involve the extension of the system to take real-time data integration through official government API to provide dynamic updates to scheme information. Moreover, the ability to become multilingual and voice-based can also enhance the access of different groups of users. Additional enhancements can be such as the adaptive model of personalization and learning-based ranking methods to improve the quality of recommendations. The evaluation can be scaled to larger datasets and user feedback mechanisms can be added to help with the constant improvement of the system as well.

REFERENCES

[1] R. Shan, “Certifying Generative AI: Retrieval-Augmented Generation Chatbots in High-Stakes Environments,” *IEEE Computer*, vol. 57, no. 8, pp. 2–11, Aug. 2024, doi: 10.1109/MC.2024.3401085.

[2] M. Gao, X. Bi, and P. Lu, “Leveraging Large Language Models: Enhancing Retrieval-Augmented Generation with ScaNN and Gemma for Superior AI Response,” in *Proc. 5th Int. Conf.*

on Machine Learning and Computer Application (ICMLCA), Allen Park, MI, USA, 2024, pp. 1–7, doi: 10.1109/ICMLCA63499.2024.10753879.

[3] A. Janarthanan, A. Paramarthalingam, Banumathi, Jayachandra, and A. Arivunambi, "Opinion Mining in Tamil YouTube Comments Using Machine Learning Approach," *AIP Conference Proceedings*, vol. 3137, no. 1, pp. 1–7, Mar. 2025.

[4] V. S. Reddy and J. B. B. Bell, "Government Schemes Recommendation API with Multi-Language Chatbot Using RAG Vector Search and Conversational AI," in *Proc. 4th Int. Conf. on Innovative Mechanisms for Industry Applications (ICIMIA)*, Coimbatore, India, 2025, pp. 1–6, doi: 10.1109/ICIMIA67127.2025.11200536.

[5] Bora and H. Cuayáhuitl, "Systematic Analysis of RAG-Based LLMs for Medical Chatbots," *IEEE Access*, 2024.

[6] A. Paramarthalingam, A. Arivunambi, A. Janarthanan, S. Sundaresan, and S. A. Ariyangavu, "AI-Driven Chatbot for Mental Health Analysis Using Transformer Models," *Premier Journal of Science*, vol. 15, p. 100133, pp. 1–7, 2025.

[7] S. B. *et al*., "Intelligent Document Interaction with Advanced Vector Embeddings and FAISS-CPU Indexing," in *Proc. 8th Int. Conf. on Electronics, Communication and Aerospace Technology (ICECA)*, Chennai, India, 2024, doi: 10.1109/ICECA63461.2024.10800948.

[8] L. B., R. M. R. Kovvur, R. K. Chenoori, and Harshavardhan B., "Enhanced Multimodal Sentiment Analysis: Exploring GPU Efficiency and Textual Improvements with RoBERTa, ChromaDB, and SBERT," in *Proc. IEEE World Conf. on Applied Intelligence and Computing (AIC)*, 2025, doi: 10.1109/AIC66080.2025.11212000.

[9] A. George *et al*., "A Personalized AI Assistant for Analyzing and Authoring Product Reviews Using Groq/LLaMA3-70B," in *Proc. Int. Conf. on Computing Technologies & Data Communication (ICCTDC)*, Hassan, India, 2025, doi: 10.1109/ICCTDC64446.2025.11158767.

[10] H. S. Hariprasath, A. Paramarthalingam, S. Sundaramurthy, and S. Cirillo, "A Study on Word Embeddings in Local LLM-Based Chatbot Applications," in *Proc. IEEE Int. Conf. on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, Sakhir, Bahrain, Jan. 2025, pp. 232–237.

[11] R. P. Pramila, R. S. R., and V. S., "Artificial Intelligence and Natural Language Processing Integrated Multilingual Health Insurance Application," in *Proc. 3rd Int. Conf. on Automation, Computing and Renewable Systems (ICACRS)*, 2024, doi: 10.1109/ICACRS62842.2024.10841786.

[12] S. Vakayil and A. J., "RAG-Based LLM Chatbot Using LLaMA-2," in *Proc. 7th Int. Conf. on Devices, Circuits and Systems (ICDCS)*, Coimbatore, India, 2024, pp. 1–5, doi: 10.1109/ICDCS59278.2024.10561020.

[13] B. Saha *et al*., "Advancing RAG with Inverted Question Matching," *IEEE Access*, 2024.