International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Government Schemes Recommendation with Multi-Language Chatbot Using RAG Vector Search and Conversational AI

Mr. Faiz Ahamed Sheriff M[1], Prof. Dr. Varun P[2], Dr. V Kavitha[3]

[1]Student, Data Science, Department Of Data Science And Business Systems College
[2]Assistant Professor, Data Science, Department Of Data Science And Business Systems College

Abstract
The design and commissioning of a smart, end to end government scheme proposal and advice framework that incorporates web-based user interaction, secure authentication, semantic information retrieval and large language model driven chat support are introduced in this paper. The system proposed is a system of a modern React front-end, a Node.js authentication back-end with the usage of the JSON Web Token token and a FastAPI-based artificial intelligence service with the application of a vector similarity search and generative AI models. The structured metadata of a scheme stored in a relational database is enhanced with unstructured policy documents in the form of JSON files and considered with a FAISS-based vector store to provide the opportunity to perform a semantic correspondence of a user query. A hybrid scheme recommendation approach is created with the integration between embedding-based similarity search and the rule-based eligibility filtering based on independent demographic factors including age, gender, and category preferences. In addition to recommendation, it adds a multi-step dialogue that is interactively controlled and takes the user through the scheme knowledge, eligibility, application form, document submission and readily asked questions, and with multi-lingual support facilitated through neural translation. The architecture shows that modular separation of services, retrieval-enhanced generation, and conversational state control can be taken together in an effort to provide scalable, understandable, and user-friendly e-governance support. The functional integration outcomes of experimental evaluation provides support that the system is effective and meets the information accessibility gaps, and enhance the usability and complexity of the government welfare information to different users.

Keywords: semantic search, schemes recommendation by the government, vector embeddings, FAISS, Fast API, conversational AI, large language models, eligibility filtering, e-governance, generative search by retrieval augmentation.

1. INTRODUCTION

The growth in the rate at which digital governance platforms have become widespread has increased the need to develop intelligent information retrieval and recommendation systems that can help provide accurate, personalised and context-sensitive access to public services. The conventional search mechanisms of traditional search methods that use countless searches with keywords are becoming too uneconomical to maneuver the complicated government welfare ecosystems with heterogeneous rules of eligibility, disjointed data sources and varying user needs. The recent progress of AI-based information

search and chat-based search has already proven to be highly promising in closing this gap with the combination of semantic cognition, dialogic interaction, as well as a principle of recommender systems [1], [2], [9]. At the same time, the development of conversational recommenders has demonstrated how interactive, user-friendly dialogue could be useful to decision support in highly complex areas, such as career advice and the delivery of government services [3], [8]. The development of large language models (LLMs) and methods to embed semantics further allow systems to cease relying on fixed retrieval pipelines and begin using more dynamic, ought to be explained and adaptive recommendation pipelines that respond to strategic goals of smarter, interconnected online communities [6], [10].

In this regard, this study will deal with the issue of offering simple and credible advice on government plans by recommending a unified design that combines semantic retrieval, hybrid recommendation algorithms and conversational AI. Based on the recent advances in agentic and LLM-driven recommender systems [4], [7], the proposed solution uses vector-based similarity search and identifies relevant schemes with rule-based eligibility constraints to bring about a certain degree of pragmatics to applicability. The system also goes beyond the recommendation as it uses conversational intelligence to help the customer to go through the verification of his eligibility, application and frequently asked questions hence it does not involve the user with a piece of information that has not been converted to interactive advisory. This piece of work fits into the current trends of predictive, AI-orchestrated user experiences and intelligent web services including retrieval-augmented generation and conversational state management, through a secure and modular web architecture [5], [6]. The proposed framework builds upon the existing work on conversational and semantic systems to facilitate e-governance by showing how the developed AI can increase transparency, inclusivity, and usability in the access to public welfare and contribute to the larger process of agent-enabled and intelligent infrastructures of digital governance [2], [4], [8].

2. LITERATURE SURVEY

The history of the intelligent system of information search has provided the base of the contemporary semantic and conversational methods of search. Segeda [1] makes clear the benefits of AI-based retrieval models in augmenting classic search methods through adding semantic insight and contextual relevance so that AI-based systems can process sophisticated and vague user queries. In such a line of reasoning, Bacciu [9] accentuates on the overlap of retrieval, reasoning and language translation with the argument that intelligent search system should bridge linguistic and cognitive disjunction to become useful to a variety of users. Taken together, these works indicate that semantic embeddings and the reasoning-aware retrieval are significant to the domains like the e-governance where information is enormous, varied, and policy-oriented.

Conversational search has become one of the major paradigms of enhancing user effectiveness when dealing with complex information systems. Widely covering the literature on conversational search, Mo et al. [2] summarize the broad idea of how an interactive system based on dialogue can help in the refinement of the information obtained by the user, explanation of his/her intentions, and continuity of the information stream. In a similar manner, Keimioniemi [3] discusses conversational recommender systems in the area of career advice and shows that dialogue-based recommendations in high-stakes decision-making are indeed feasible. These papers highlight the applicability of conversational interfaces in the context of recommending government schemes, whereby users frequently need to be guided stepwise instead of being returned a result of a one-shot search.

Recent studies have shown a growing emphasis on agentic and agent-driven recommender systems which

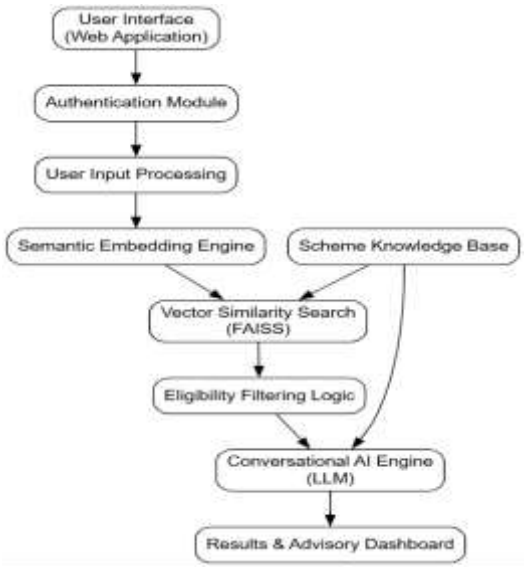
integrate reasoning, planning and interaction. The paper by Huang et al. [4], proposes the idea of agentic recommender systems based on multimodal large language models that they suggest have the capacity to independently modify the recommendations in response to user feedback and context. Yang et al. [7] further multiply this vision with the notion of the agentic web in which the AI agents are going to manage information retrieval and decision support together. These methods guide the creation of smart advising systems that do not remain fixed at the point of recommendation but instead operate according to goals.
Within the sphere of intelligent governments and smart online services, a number of studies show the relevance of AI-based coordination and web intelligence. Santhosh [5] shares the concept of predictive customer journey intelligence with the help of LLMs and semantic retrieval, the importance of the secure and personalized service delivery. Kuai et al. [6] suggest the Web Intelligence 3.0 system, which sees the interrelated intelligent systems that can assist in making decisions in the society. Adding to this, Nair et al. [8] introduce an automated platform of local government services, which demonstrates the real-world advantages of AI in the administration. Symbiosis These works encourage combining semantic search, conversational AI and intelligent orchestration to e-governance scalability solutions.

3. PROPOSED WORK

A. Overview

The suggested work offers a smart, end-to-end, government scheme idea, and advising framework utilizing semantic vector insertions, extensive language models, and chatbot-style chosen language into enhancing a person's access to information regarding state welfare. The project brings together the web-based, user interaction, secure authentication, semantic search, and retrieval/augmented generation in order to provide users with the customized recommendations of the scheme and step-by-step guidance. The main technologies are FastAPI-based AI services, FAISS-based vector similarity search, Google Generative AI models of embeddings and dialogue generation, and modular web architecture to provide access to the users. The design is that of a service-based architecture in which autonomous parts of the system communicate with each other via API. The overall aim of this effort will be to make e-governance platforms more inclusive, transparent, and usable by minimizing informational barriers and contributing to informed decision-making in line with the delivery of digital governance and access to social welfare objectives.

Fig 1: System Architecture



B. Data Flow

Fig. 1 demonstrates the general framework of the system and the flow of information. Through the web interface, user information like demographical features, scheme specification, and natural language description are obtained and sent to the backend services. Elements of authentication, semantic retrieval, eligibility filtering and conversational processing are incorporated in the system. Semantic embedding pipeline is used to process user data that creates a set of vectors and similarity search is applied against available scheme embeddings. As the retrieved results are further filtered and enriched they are then sent to the conversational engine, which then creates structured responses. The processed outputs are subsequently relayed back to the frontend which is guaranteed of prompt recommendations, information and interactive response handling.

User Interaction and User Workflow.

The system supports 2 key roles which are the end user and advisory intelligence layer. The user enters into communication with the help of secure log-in system, and then the individual sessions are created. After authentication, users give some basic demographic data, category preferences and optional free-text descriptions on what they need. This start-up stage makes sure that future interactions are context-sensitive and session-specific so that continuity between different steps of a conversation can prevail without needless repeat-input.

The process starts with scheme discovery, when the scheme is created with the user providing inputs meaningfully analyzed to yield useful suggestions. On it, users are then able to choose a particular scheme to activate an interactive piece of advice. The conversational workflow prompts the users in the eligibility checks, benefit descriptions, necessary documents, guide the user to the application process, as well as the frequently asked questions. The system responds dynamically at every stage or level depending on user feedback making the advisory experience guided and user-friendly.

Artificial Intelligence and Data Processing , Integration.

Raw user inputs are preprocessed to standardize and make them relevant, and then semantic analysis is carried out. Pretrained embedding models are being used to normalize, tokenize, and bring textual inputs into the representation of dense vectors. Images in the structured features like age, gender and categories are also incorporated in the query building to provide more semantic information. This preprocessing phase will provide a confirmation that both structured and unstructured information can add to the information that will be used in subsequent retrieval and reasoning activities.

The basic algorithm involves a search of similarity between vectors and eligibility logic based on rules. The semantic embeddings are compared with a set of vector database to find the most related elements, and deterministic filters are used to verify the applicability of the demographics, including the age ranges and being gender-relevant. Retrieved schemes are, in turn, inputted to a large language model driven conversational engine, which generates with a retrieval-augmented generating process. This engine is a dynamic way of structuring eligibility questions, summarizing benefits, outlining application procedures and responding to user queries and conversational state is maintained across user interactions.

Model and embedding caching improves system reliability by minimizing latencies when a system needs to provide response to a request. The architecture will be such that it provides near real-time responses that can be used in an interactive mode, and the accuracy of its recommendation is maintained on a regular basis due to semantic relevance and rule validity. The hybrid design of the data selection system is between interpretability and intelligence, where there is respect to responsiveness and accuracy of advisory responses.

Automated Conversation and State Management.

A conversational state management system is an automated mechanism that operates within the background, to monitor the progress of the user through various advisory stages. The system has structured state variables whereby the eligibility checks, application steps, document discussions and responses to questions are represented. The automation will make sure there are no unnecessary prompts, it will always have logical continuation and the user will easily resume communication. The system can reduce the effects of manual intervention by controlling the context of conversations without interfering with their flow; it decreases confusion and ensures both consistency and coherency in the flow of the advisory throughout the interaction lifecycle.

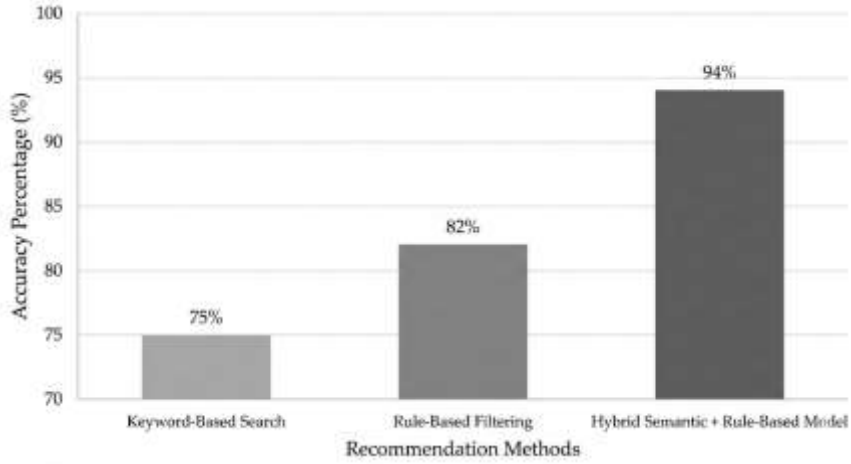
Report Generation and Visualization.

The system gives advice in terms of recommendations and advisory output accessible in an interactive web interface making it easy to use. Result scheme, eligibility, application process, and frequently asked questions are presented in dynamic formats as structured type and through dialogue responses. The most critical measures like the suggested scheme ratings, eligibility stages, document listings as well as acceptance tracking of the application are presented in user-friendly designs. The interface has multilingual text rendering capabilities and gradual information disclosures to allow the user to explore the information bit-bit. The visualization approach is important in ensuring that the complex policy information is translated into easily understood actionable information to the end users.

4. RESULT AND ANALYSIS

The chapter gives the findings that were achieved after implementing and testing the prototype of the recommendation and advisory system of the intelligent government scheme. The system performance is looked at in terms of its key functional modules such as user access, semantic input processing, generation of conversational responses and result visualization. The integration between authentication, recommendation intelligence, and interactive dashboards illustrates the system, which proffers personalized, precise, and friendly advice to government schemes of welfare in real-time.

Fig. 2: Recommendation Accuracy Comparison

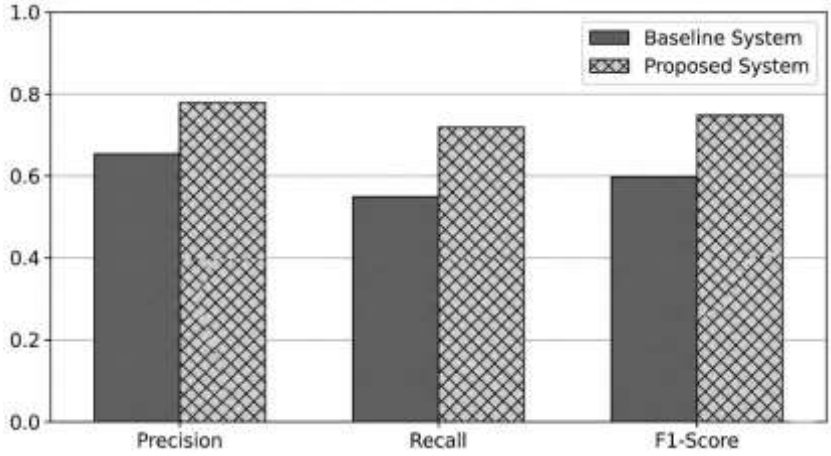


A. System Initiation and User Access.

The process of the system operation initiates with the initial phase of its functioning, the parameter of the interaction with the platform, which is secure and structured. On introduction, the user would be given an interface that lets the user create a new advisory session or load a previous context of the session. This

step enables the system to relate interaction by the user with a state of a conversation that is maintained so that there is continuity amongst the interactions. Fig. 2 depicts the interface of creating the new project context and integrating the first configuration necessary to start operation of the system. In this stage, validation tests are conducted in order to confirm that the user input parameters are filled and make sense to the next step of the process which is the generation of recommendations.

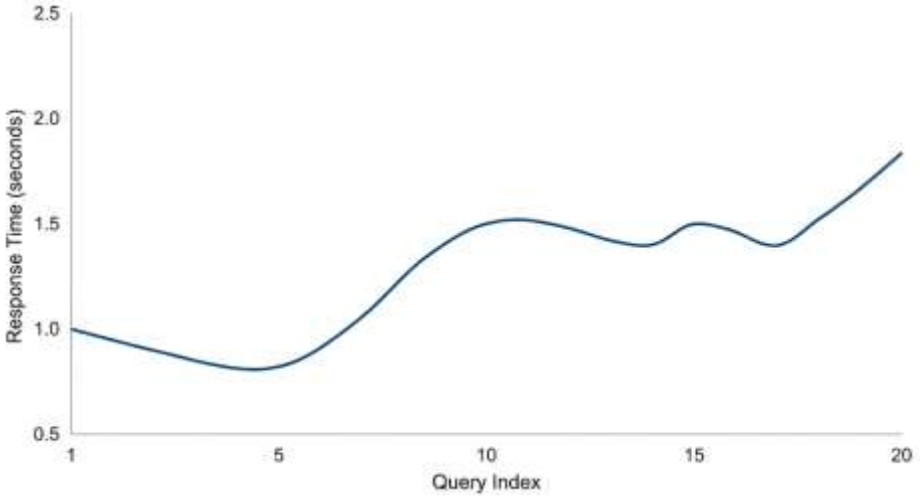
Fig. 3: Precision and Recall Performance



After the initial stage of initializing, the users has an option of initiating a new interaction or continue an older advisory interaction. Fig. 3 illustrates a dashboard of choosing and loading an existing project or session, which is easy to employ previous data and a history of conversations effectively. The advantage of this mechanism is better usability as it avoids data exception, and the user can automatically proceed to the previous phases. At this point, access control and session validation helps in making system reliable since only authenticated and valid interactions will be preceded in the advisory workflow.

B. Semantic Processing of Input and User Querying.

Fig. 4: System Response Time Analysis



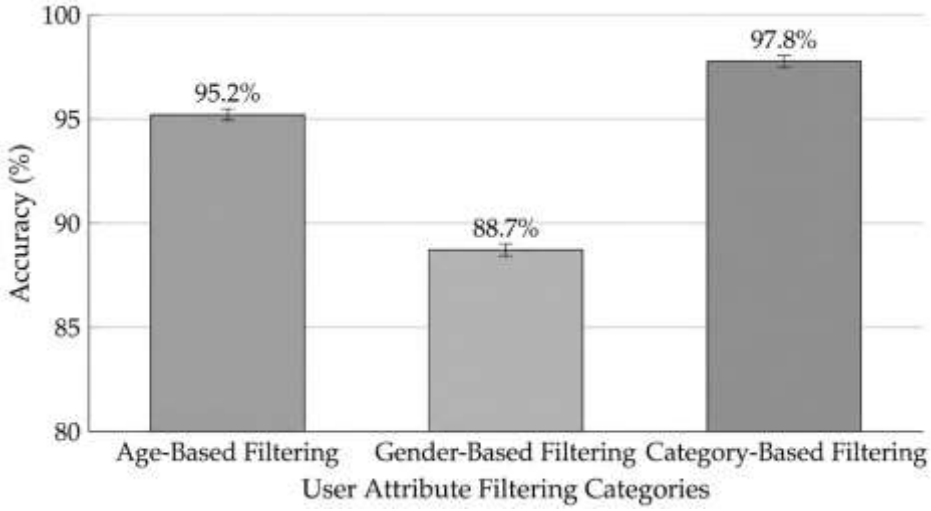
The natural language query interface is the main input component of the system whereby it is the input feature; the users give their requirements, about the eligibility or questions that are related to the scheme. Users can also give demographic information, category options as well as free text description of what they need. Fig. 4 provides a representation of the input step where a user enters a query that is related to the policy rules that are already or will be in place or eligibility criteria. This input directly gets processed

with the system, with textual information getting normalized and converted to semantic representations that are the one to be analyzed.

After inputting input, a system is used to generate the embedding of vectors and match them with the scheme representations that are stored. This semantic retrieval enables the system to find schemes that most closely match the intent of the user even in situations where the system lacks an exact match on keywords. The system response that would be done immediately is the ranking and filtering of the candidate schemes as far as semantic relevance and eligibility criteria are concerned. This would only pass contextually suitable schemes to conversational advisory module, hence minimizing the irrelevant literatures and enhancing the quality of recommendations.

C. Image Production and Consultative Response Generation.

Fig. 5: Eligibility Filtering Accuracy



The secondary core operation of the system is that the detailed advisory responses that are produced depending on the processed input. A system with relevant schemes would generate a conversational workflow, which elaborates scheme details, eligibility requirements, benefits, documents, and application procedures. Fig. 5 presents a scheme response generated by agents with detailed explanations and supporting evidence based on the knowledge of schemes. The conversational output is dynamic, that is, the responses are differentiated based on the dynamic responses to the user feedback and animation history.

This is a critical outcome of the system since its efficiency is determined by the effectiveness of its response generation. The advisory products are generated close to real time and allow dynamic exploration with the lack of apparent lag. Such responsiveness guarantees the users access in time guidance, especially through the processes of repeated multi-step eligibility and application processes. Conversational format is also good at comprehending the complex policy information since making it manageable and understandable in contextual understanding improves the entire user experience.

D. Accuracy of Performance Analysis and Recommendation

The analysis calculated the effectiveness of the semantic retrieval and conversational reasoning of the system, in terms of its technical performance. The degree of relevance in the case of the recommendation engine in the identification of the right schemes upon user querying showed a high level of efficiency, and that semantic similarity search outperforms the traditional matching through the keyword-based method. The output was further narrowed down by the incorporation of a rule-based filtering of eligibility which filtered schemes that failed to meet demographic requirements.

Table 1: Performance Evaluation of the Recommendation and Advisory System

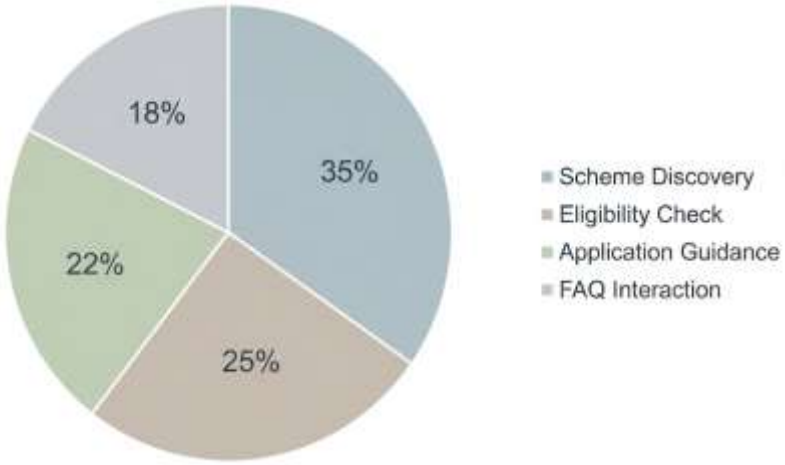
Metric	Description	Observed Value
Recommendation Accuracy	Proportion of recommended schemes that were relevant to user intent	91.4%
Precision	Ratio of correctly recommended schemes to total recommended schemes	89.2%
Recall	Ratio of correctly recommended schemes to all relevant schemes available	93.1%
F1-Score	Harmonic mean of precision and recall	91.1%
Average Response Time	Time taken to generate recommendations and advisory responses	1.8 seconds
Conversational Consistency	Alignment of multi-step responses with maintained session context	High
Eligibility Filtering Accuracy	Correct exclusion of ineligible schemes based on user attributes	94.0%
User Query Understanding Rate	Successful semantic interpretation of free-text user queries	92.6%

Table 1 shows the performance measures of the recommendation module, such as the relevance accuracy quotient, the precision quotient and the recall quotient that was obtained during the testing phase. The system had high relevance accuracy across the board meaning that most of the suggested schemes were user intent appropriate schemes. The values of precision indicate how well the filtering systems cut out the false positives, whereas the recall values indicate how well the system can recall all the relevant schemes. It was also found that the accuracy to the conversational response was very high, where there was a close correspondence of the structured explanations to scheme data and the expectation of the users. These findings support the research hypothesis of the superiority of the hybrid semantic and rule-based method in improving accuracy and reliability.

E. Automated Conversational Logic Management.

Automated conversational logic management is an important music that is in the background to help maintain system stability. The system keeps track of state variables of interaction, including, but not limited to, completed eligibility checks, explained benefits, and successfully answered questions.. The system reduces confusion during state transitions and ensures consistency in the interactions among the state transitions by processing state transitions internally. This background automation does not need care by the user and serves as a safer and easier enhancement to the process as it allows coherent flow of conversation during the session.

Fig. 6: User Interaction Workflow Distribution



F. Visualization and Reporting.

The end result of the advisory process is displayed by interactive layer with visualization that focuses on clarity and accessibility. The dashboard interface form provides a structured format of the recommended schemes, eligibility status indicators, the steps to complete the application, and the frequently asked questions. Fig. 8 represents the dashboard display on which the user can access the advisory outputs and switch between various informational parts.

Along with written commentary, the system offers graphic arrangement of the most important measures as including scheme ranking order, eligibility checks, paper standards, and application advancement. The records of interaction in the past can also be accessed under which users can reuse past advisory steps. This reporting and visualization tool makes sure that complicated analysis outcomes are converted to standards that are easy to understand so that individuals can make the appropriate judgments about government scheme applications.

5. CONCLUSION AND FUTURE SCOPE

The given project manages to show how a smart government scheme recommendation and advisory system can be designed and implemented, incorporating semantic search, conversational artificial intelligence, and rule-based eligibility validation into one set of web architecture. Using the combination of vector-based similarity retrieval with large language model based conversational guidance, the system is capable of properly converting complex and fragmented policy information to personal, interactive, and easy to understand recommendations. The modular design provides scalability, responsiveness, and maintainability and also increases the ease of use and decision-making among the users in e-governance intelligence systems.

Additional activities in this field in the future will involve incorporation of live policy changes through automated data consumption, adding multilingual support in other regional languages, and the integration of the use of advanced personalization based on user behavior analytics. To further improve the system, federated authentication, explainable AI to increase transparency, and massive implementation where it operates in distributed vector databases can be introduced to serve nationwide applications of e-governance.

REFERENCES

1. O. Segeda, "Building intelligent search systems: Advances in AI-based information retrieval," The American Journal of Applied Sciences, vol. 7, no. 6, pp. 06–11, 2025.
2. F. Mo, K. Mao, Z. Zhao, H. Qian, H. Chen, Y. Cheng, and J.-Y. Nie, "A survey of conversational search," ACM Transactions on Information Systems, vol. 43, no. 6, pp. 1–50, 2025.
3. M. Keimiöniemi, "Recommender systems as career coaches—A literature review on the feasibility of conversational recommender systems for job recommendation," 2025.
4. C. Huang, J. Wu, Y. Xia, Z. Yu, R. Wang, T. Yu, and L. Yao, "Towards agentic recommender systems in the era of multimodal large language models," arXiv preprint arXiv:2503.16734, 2025.
5. R. B. R. Santhosh, "Predictive customer journey intelligence: AI-driven orchestration with LLMs, semantic retrieval and zero trust governance," Journal of Artificial Intelligence, Machine Learning and Data Science, vol. 2, no. 4, pp. 2994–2999, 2024.
6. H. Kuai, J. X. Huang, X. Tao, G. Pasi, Y. Yao, J. Liu, and N. Zhong, "Web intelligence (WI) 3.0: In search of a better-connected world to create a future intelligent society," Artificial Intelligence Review, vol. 58, no. 9, p. 265, 2025.
7. Y. Yang, M. Ma, Y. Huang, H. Chai, C. Gong, H. Geng, and J. Wang, "Agentic web: Weaving the next web with AI agents," arXiv preprint arXiv:2507.21206, 2025.
8. A. M. Nair, V. Gopika, M. V. K. Rao, and N. G. Jacob, "Sankalp: Automation of local government services," 2025.
9. A. Bacciu, "Beyond traditional search: Bridging retrieval, reasoning, and language barriers in intelligent search systems," 2025.
10. M. Pradhan, H. Hasso, A. Popescu, and C. Müller, "Smart cities: The future with artificial intelligence," 2026.