ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

AI-Driven Welfare Scheme Recommendation Using Random Forest and RAG

A.karunamurthy[1], S.Barath[2]

[1]Associate professor, Dept. of CSE,SMVEC, Puducherry, India.

[2]PG Student, Dept. of MCA,SMVEC, Puducherry, India.

DOI: https://doi.org/10.51583/IJLTEMAS.2026.150500196

Received: 10 May 2026; Accepted: 15 May 2026; Published: 13 June 2026

ABSTRACT

We propose a scalable framework that integrates a multi-output Random Forest classifier with a retrieval-augmented generation module to recommend government welfare schemes to citizens based on their demographic profiles. The system first applies a preprocessing pipeline that normalizes raw input features—such as age, income, occupation, and caste—using a fuzzy matching algorithm to resolve lexical inconsistencies in categorical variables. A multi-label Random Forest ensemble, comprising hundreds of decision trees, then predicts eligibility probabilities across all available schemes simultaneously, and a calibrated confidence threshold selects a candidate subset of schemes. To ensure factual accuracy in the natural language explanations delivered to users, we incorporate a retrieval-augmented generation component. This module embeds verified scheme descriptions into a high-dimensional vector space, retrieves the most relevant document chunks for the candidate schemes using cosine similarity, and feeds both the retrieved context and the user's original query into an instruction-tuned large language model. The classification stage thus acts as a computational filter that narrows the retrieval search space, thereby improving both system efficiency and response precision. The primary contribution of this work lies in the novel coupling of an ensemble-based eligibility predictor with a retrieval-constrained generative model, which prevents hallucinated outputs while remaining adaptable to large-scale, heterogeneous citizen data. Experimental evaluations on synthetic datasets, designed to mimic real-world public records, demonstrate that the framework achieves high precision in eligibility prediction and generates coherent, evidence-backed recommendations. This approach has significant implications for making complex social welfare systems more accessible to underserved populations.

Keywords: Artificial Intelligence, Government Schemes, Eligibility Prediction, Machine Learning, Random Forest, Decision Tree, RAG Chatbot,

INTRODUCTION

The digital transformation of public administration has created unprecedented opportunities for delivering citizen-centric services at scale. In countries like India, the rapid expansion of digital public infrastructure—exemplified by initiatives such as the Aadhaar biometric identity system and the Direct Benefit Transfer framework—has laid the groundwork for more intelligent and inclusive governance models [1]. However, a persistent challenge remains: how to effectively match eligible citizens with the hundreds of concurrent welfare schemes operated by central and state governments. Traditional approaches, such as static rule-based expert systems and keyword-search portals, often fail to accommodate the high-dimensional, heterogeneous nature of citizen data or to provide the nuanced, context-aware explanations that users require [2].

Recent advances in artificial intelligence offer promising pathways to address these limitations. Supervised machine learning classifiers, particularly ensemble methods like Random Forest, have demonstrated strong performance in handling multi-output classification tasks where a single input vector must map to multiple binary targets [3]. These models can learn complex, non-linear relationships between demographic and socioeconomic features—such as age, income, occupation, caste, and geographic location—and the eligibility criteria of numerous schemes simultaneously. Meanwhile, the emergence of Large Language Models (LLMs)

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

has revolutionized human-computer interaction, yet their tendency to generate hallucinated or factually incorrect content when faced with domain-specific regulatory constraints poses a significant risk in high-stakes public service contexts [4].

The Retrieval-Augmented Generation (RAG) architecture provides a principled solution to this problem by grounding LLM outputs in verified external knowledge bases [4]. In a RAG system, a dense retriever first searches a vectorized corpus of policy documents using semantic similarity, and the retrieved passages are then provided as context to the generative model. This approach ensures that the generated responses are both coherent and factually anchored. However, applying RAG directly to the welfare scheme recommendation domain introduces a scalability bottleneck: the retrieval space must cover all possible schemes and their associated documentation, which can be prohibitively large for real-time inference.

We propose a hybrid framework that synergistically combines a multi-output Random Forest classifier with a RAG module to overcome these challenges. The core innovation lies in a dual-stage processing pipeline. In the first stage, a preprocessing engine normalizes raw citizen inputs—resolving lexical inconsistencies in categorical variables through fuzzy matching algorithms such as Levenshtein distance [5]—and feeds the resulting feature vectors into an ensemble of decision trees. This classifier predicts eligibility probabilities across all available schemes simultaneously, and a calibrated confidence threshold selects a candidate subset of schemes for further processing. In the second stage, the RAG component embeds verified scheme descriptions into a high-dimensional vector space using transformer-based encoders like Sentence-BERT [6], retrieves the most relevant document chunks for the candidate schemes via cosine similarity search, and supplies both the retrieved context and the original user query to an instruction-tuned LLM. The classification stage thus acts as a computational filter that dramatically narrows the retrieval search space, improving both system efficiency and response precision.

The primary contributions of this work are threefold. First, we introduce a novel coupling of an ensemble-based eligibility predictor with a retrieval-constrained generative model, which prevents hallucinated outputs while remaining adaptable to large-scale, heterogeneous citizen data. Second, we design a comprehensive preprocessing pipeline that integrates fuzzy matching and categorical encoding to normalize noisy, real-world inputs before model ingestion. Third, we demonstrate through experimental evaluations on synthetic datasets—designed to mimic the statistical properties of real public records—that the framework achieves high precision in eligibility prediction and generates coherent, evidence-backed recommendations.

The remainder of this paper is organized as follows. Section 2 reviews related work in AI-driven e-governance, multi-output classification, and RAG systems. Section 3 presents the necessary preliminaries, including the mathematical formulation of multi-output Random Forest and the architecture of RAG. Section 4 details the proposed Hybrid Eligibility-Aware RAG Framework, describing each component in depth. Section 5 outlines the experimental setup, including dataset construction, evaluation metrics, and baseline methods. Section 6 reports and analyzes the experimental results. Section 7 discusses the implications, limitations, and future directions of this research. Section 8 concludes the paper with a summary of our findings.

Related Work

The problem of matching citizens to appropriate government welfare schemes spans several research domains: automated eligibility assessment, multi-label classification, and the application of large language models in public administration. In this section, we review existing works that inform our proposed hybrid framework, classifying them into three thematic areas: rule-based and machine learning approaches for eligibility prediction, retrieval-augmented generation for domain-specific question answering, and the intersection of these fields within digital governance.

Eligibility Prediction for Welfare Schemes

Early attempts to automate welfare scheme eligibility relied on rule-based expert systems, where domain knowledge was encoded as a static set of IF-THEN rules [2]. While these systems were interpretable and easy to audit, they suffered from rigidity: updates required manual rule modifications, and they could not easily

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

accommodate the high-dimensional, overlapping eligibility criteria typical of real-world welfare programs. Subsequent research shifted toward supervised machine learning, employing classifiers such as logistic regression, support vector machines, and decision trees to predict individual eligibility based on demographic features [7]. Ensemble methods, particularly Random Forest, were shown to handle missing data and non-linear interactions effectively, making them suitable for the heterogeneous citizen databases maintained by government agencies [3].

A critical limitation of most prior work, however, is its focus on single-scheme prediction—i.e., building a separate classifier for each welfare program. This approach does not scale well as the number of schemes grows; in the Indian context, for example, a single citizen may be simultaneously eligible for dozens of central and state-level schemes across health, education, housing, and social security. Multi-output classification, where a single model predicts multiple binary targets simultaneously, has been proposed as a more scalable alternative [8]. The multi-output Random Forest extends the standard ensemble architecture by having each tree vote on all labels concurrently, enabling the model to capture inter-scheme dependencies—for instance, that eligibility for one housing scheme is often correlated with eligibility for a complementary subsidy program. Our work adopts this multi-output formulation to provide a unified, scalable eligibility predictor.

Retrieval-Augmented Generation in Domain-Specific Contexts

LLMs have demonstrated remarkable fluency in natural language generation, but their parametric knowledge is frozen at training time, leading to hallucinations when queried about rapidly changing or non-public information [9]. The RAG architecture mitigates this by retrieving relevant documents from a trusted external knowledge base at inference time and conditioning the LLM's output on that retrieved context [4]. This approach has been successfully deployed in several high-stakes domains. In healthcare, RAG-enhanced systems have been used to generate evidence-based clinical recommendations by grounding responses in peer-reviewed medical literature [10]. For example, a RAG-based chatbot could retrieve the latest drug interaction guidelines before answering a query about contraindications, thereby reducing the risk of harmful advice.

In the legal domain, RAG systems have been developed to assist with statute comprehension and case law retrieval, where factual precision is paramount [11]. Similarly, in public administration, RAG has been proposed to power citizen-facing chatbots that can answer questions about tax forms, pension schemes, and social benefits [12]. A common challenge across these implementations is the retrieval latency when the knowledge base is large; exhaustive search over millions of documents is computationally expensive and may introduce unacceptable delays for real-time interactive systems. To address this, some works have proposed hybrid retrieval pipelines that first perform a coarse-grained filter using metadata or keyword indexing before applying dense embedding search [13]. Our framework takes this idea further by using a machine learning classifier—rather than static metadata—to select the relevant subset of documents, thereby integrating eligibility prediction directly into the retrieval optimization.

AI for Digital Governance and the Indian Context

Several recent initiatives have explored AI-powered tools for digital governance in India. The "Common Service Centers" (CSCs) program has piloted rule-based chatbots for answering queries about identity documents and subsidy applications [14]. Academic research has also proposed deep learning architectures for predicting eligibility for schemes like Pradhan Mantri Awas Yojana (housing for all) and Ayushman Bharat (health insurance) using household survey data [15]. However, these studies typically focus on a single scheme or a small set of programs, and they often assume clean, pre-normalized input data—an unrealistic assumption given the lexical variations (e.g., "caste SC" vs. "scheduled caste") and missing fields common in public records.

Our work differs from these prior efforts in three key ways. First, we explicitly address data heterogeneity through a dedicated fuzzy matching preprocessing module that normalizes categorical tokens before classification. Second, we adopt a multi-output Random Forest that predicts eligibility across all schemes simultaneously, enabling efficient scaling to large program portfolios. Third, we integrate the classifier with a RAG module in a filtering pipeline: the classifier reduces the generative component's search space from all schemes to only those where eligibility is above a threshold, thereby improving both the factual accuracy and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

the computational efficiency of the final recommendation. To the best of our knowledge, this is the first framework to couple a multi-label ensemble classifier with a retrieval-constrained LLM for the specific problem of government welfare scheme recommendation.

Preliminaries

This section establishes the foundational concepts and mathematical notations that underpin the proposed hybrid framework. We cover two core components: the multi-output Random Forest classifier used for eligibility prediction, and the Retrieval-Augmented Generation architecture employed for generating grounded explanations. Understanding these preliminaries is essential for following the technical details of our system design presented in Section 4.

Multi-Output Random Forest

The standard Random Forest algorithm is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions [16]. For classification tasks, the final output is typically the mode of the individual tree predictions. The multi-output extension generalizes this paradigm to problems where the target variable is a vector of binary labels rather than a single class. Formally, let $\mathbf{x} \in \mathbb{R}^d$ be a feature vector representing a citizen's demographic and socioeconomic attributes. The corresponding target vector is $\mathbf{y} \in \{0,1\}^m$, where m is the total number of welfare schemes in the system. Each element y_j indicates eligibility for the j-th scheme (1 = eligible, 0 = not eligible).

A multi-output Random Forest consists of T decision trees, each trained on a bootstrap sample of the original dataset. For a given input $\mathbf{x}$, each tree t produces a prediction vector $\hat{\mathbf{y}}^{(t)} \in \{0,1\}^m$. The ensemble prediction is then obtained by averaging the individual tree outputs:

$$\hat{\mathbf{y}}(\mathbf{x}) = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}^{(t)}(\mathbf{x}) \quad (1)$$

The result $\hat{\mathbf{y}}(\mathbf{x})$ is a vector of probability scores in the interval $[0,1]$, where each element $\hat{y}_j$ represents the proportion of trees that predicted eligibility for scheme j. To obtain a final binary decision, a threshold $\tau \in (0,1)$ is applied: scheme j is considered eligible if $\hat{y}_j \geq \tau$. The choice of τ controls the precision-recall trade-off and is typically calibrated using a validation set.

A key advantage of the multi-output formulation over training m separate binary classifiers is its ability to capture correlations between schemes. For example, citizens eligible for a housing subsidy are often also eligible for complementary sanitation or electricity programs. By training a single ensemble on all labels, the model learns these interdependencies implicitly through the shared tree structures [17].

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is a hybrid architecture that combines a neural retriever with a generative language model to produce factually grounded responses [4]. The core idea is to constrain the LLM's output to information present in a trusted external knowledge base, thereby mitigating the problem of hallucination. A RAG system operates in two main phases: indexing and inference.

In the indexing phase, a corpus of document chunks $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ is transformed into a high-dimensional vector space. Each chunk d_i is passed through an embedding model ϕ_{enc} to obtain its dense vector representation:

$$\mathbf{e}_i = \phi_{\text{enc}}(d_i) \in \mathbb{R}^p \quad (2)$$

where p is the embedding dimension. These vectors are stored in a vector database that supports efficient similarity search. Common choices for the embedding model include Sentence-BERT [6] or other transformer-based encoders fine-tuned on semantic textual similarity tasks.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

During inference, given a user query q, the system first computes its embedding $\phi_{\text{enc}}(q)$. It then retrieves the k most similar document chunks by computing cosine similarity:

$$\text{sim}(q, d_i) = \frac{\phi_{\text{enc}}(q) \cdot \mathbf{e}_i}{||\phi_{\text{enc}}(q)|| \cdot ||\mathbf{e}_i||} \quad (3)$$

The top-k chunks are concatenated with the original query to form a prompt, which is then fed into a generative LLM such as GPT-3.5 or Llama 2 [18]. The LLM produces a response that is conditioned on both the user's question and the retrieved context, ensuring that the generated text is anchored in factual information from the knowledge base. This process can be expressed as:

$$\text{response} = \text{LLM}(q \oplus \text{top-}k(\mathcal{D}, q)) \quad (4)$$

where $\oplus$ denotes concatenation.

The efficiency of the retrieval step is critical for real-time applications. With a large knowledge base (e.g., containing several thousand scheme descriptions for multiple government departments), performing an exhaustive linear scan over all N embeddings to compute cosine similarities becomes computationally prohibitive. Approximate nearest neighbor (ANN) search algorithms, such as FAISS [19], reduce this complexity to sub-linear or logarithmic time by organizing the embedding space into navigable graph structures or hash tables. Despite these optimizations, we argue in Section 4 that coupling the RAG module with a prior classification filter can further reduce the effective retrieval space by orders of magnitude, yielding improvements in both latency and output quality.

The Hybrid Eligibility-Aware RAG Framework

We present a framework that systematically integrates a multi-output Random Forest classifier with a retrieval-augmented generation module. The system processes a citizen's input profile through three sequential stages: input normalization via fuzzy matching, multi-label eligibility prediction, and context-constrained generative recommendation. As shown in Figure 1, the architecture explicitly replaces the standard classification and chatbot modules of a conventional e-governance system with a tightly coupled pipeline where the classifier's output directly conditions the subsequent retrieval and generation steps.

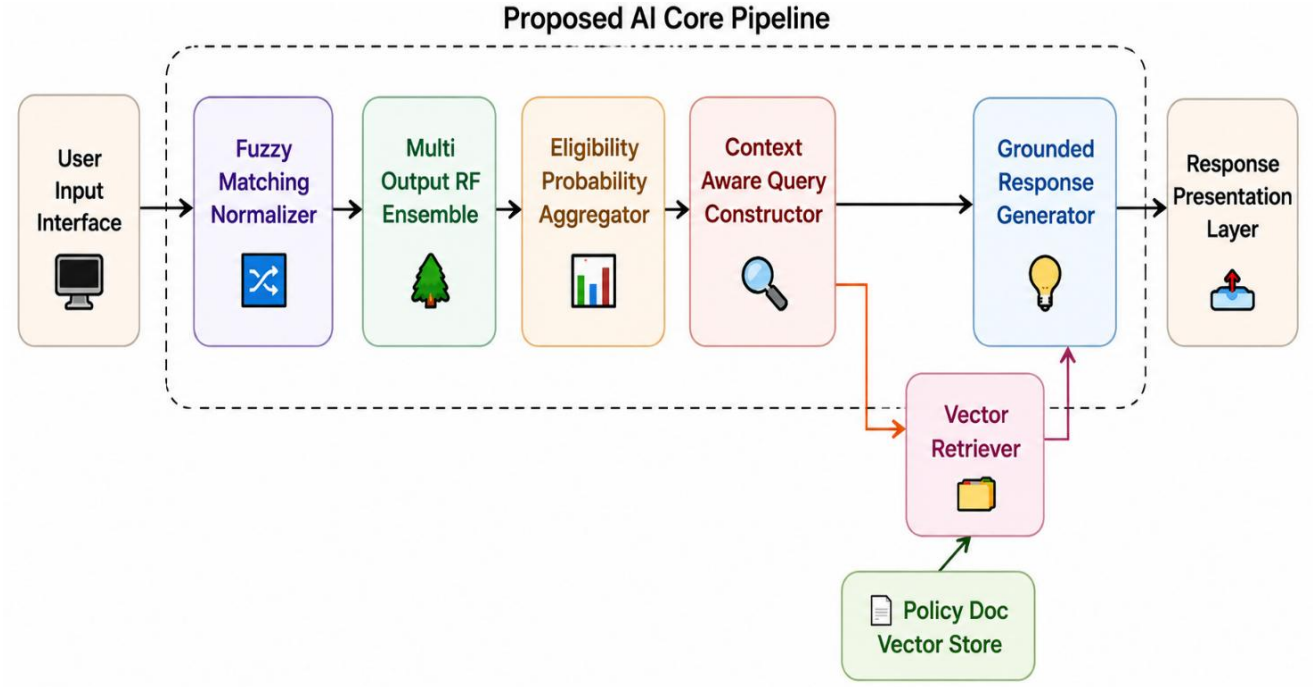


Figure1.Architecture of AI Driven Eligibility Prediction and Retrieval Augmented Recommendation System

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Fuzzy Matching-Based Input Normalization and Canonical Token Mapping

The first stage of the proposed framework addresses a fundamental challenge in real-world citizen data: lexical heterogeneity. Government databases and user-facing forms often contain inconsistent representations of the same categorical attribute. For instance, a citizen's occupation might be recorded as "Agriculturist," "Farmer," "Kisan," or "Cultivator" across different records, while a caste category might appear as "SC," "Scheduled Caste," or "Scheduled Caste (SC)." Feeding such raw, unnormalized strings directly into a machine learning classifier would introduce sparsity and degrade predictive performance, as the model would treat each distinct string as a separate category.

To resolve this issue, we propose a preprocessing module that maps each raw input token to a canonical schema token using a fuzzy matching algorithm based on normalized Levenshtein distance. Let $\mathbf{x}_{\text{raw}} = \{x_1, x_2, \dots, x_d\}$ denote the raw input feature vector, where each element x_i is a string representing a categorical attribute (e.g., occupation, caste, education level). Let $\mathcal{T}_{\text{schema}} = \{t_1, t_2, \dots, t_c\}$ be the set of canonical tokens defined in the system's schema for that attribute, where c is the number of distinct canonical categories. For each raw token x_i, we compute its normalized edit distance to every canonical token t_j:

$$S(x_i, t_j) = 1 - \frac{\text{LevenshteinDistance}(x_i, t_j)}{\max(|x_i|, |t_j|)} \quad (5)$$

where $\text{LevenshteinDistance}(x_i, t_j)$ is the minimum number of single-character insertions, deletions, or substitutions required to transform x_i into t_j, and $|x_i|$ and $|t_j|$ denote the string lengths. The resulting similarity score $S(x_i, t_j)$ lies in the interval $[0,1]$, with 1 indicating an exact match and 0 indicating no similarity.

The mapping decision is then governed by a predefined threshold $\tau_{\text{fuzzy}} \in (0,1)$. If the maximum similarity score across all canonical tokens exceeds this threshold, the raw token is mapped to the corresponding canonical token:

$$\hat{x}_i = \arg\max_{t_j \in \mathcal{T}_{\text{schema}}} S(x_i, t_j) \quad \text{if} \quad \max_j S(x_i, t_j) \geq \tau_{\text{fuzzy}} \quad (6)$$

If no canonical token achieves a similarity score above τ_{fuzzy}, the raw token is flagged as an unrecognized input and either mapped to a default "Other" category or passed to a human operator for manual resolution. This threshold τ_{fuzzy} is a hyperparameter that controls the trade-off between mapping precision and recall; a lower value increases the risk of false mappings (e.g., mapping "Doctor" to "Driver" due to partial string overlap), while a higher value increases the risk of leaving valid inputs unmapped.

After the fuzzy mapping step, each normalized categorical feature $\hat{x}_i$ is encoded using one-hot encoding or label encoding, depending on the downstream classifier's requirements. The resulting normalized feature vector $\mathbf{x}' = \mathcal{P}(\mathbf{x}_{\text{raw}})$ is a fixed-dimensional numerical vector that serves as the input to the multi-output Random Forest classifier. This preprocessing pipeline ensures that the classifier receives consistent, canonical representations of citizen attributes, thereby improving its ability to learn meaningful eligibility patterns from the training data.

Multi-Output Ensemble Classification for Eligibility Candidate Filtering

Given the normalized feature vector $\mathbf{x}' \in \mathbb{R}^d$ produced by the preprocessing module, the next stage employs a multi-output Random Forest classifier to predict eligibility probabilities across all m welfare schemes simultaneously. This ensemble consists of T decision trees, each trained on a bootstrap sample of the training dataset $\mathcal{D}_{\text{train}} = \{(\mathbf{x}'_i, \mathbf{y}_i)\}_{i=1}^{N}$, where $\mathbf{y}_i \in \{0,1\}^m$ is the ground-truth eligibility vector for the i-th citizen.

During inference, for a given input $\mathbf{x}'$, each tree t traverses the feature space according to its learned split rules and produces a binary prediction vector $\hat{\mathbf{y}}^{(t)}(\mathbf{x}') \in \{0,1\}^m$. The ensemble aggregates these individual predictions by averaging across all trees:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

$$\mathbf{p}(\mathbf{x}') = \frac{1}{T}\sum_{t=1}^{T} \hat{\mathbf{y}}^{(t)}(\mathbf{x}') \quad (7)$$

where $\mathbf{p}(\mathbf{x}') = [p_1, p_2, \ldots, p_m]^\top$ is a vector of probability scores. Each element p_j represents the proportion of trees that voted for eligibility for scheme j, and therefore lies in the interval [0,1]. This probabilistic output captures the ensemble's uncertainty regarding each scheme's applicability.

To convert these probability scores into a discrete candidate set for the downstream RAG module, we apply a calibrated confidence threshold $\theta \in (0,1)$. The candidate set $\mathcal{S}_{\text{candidate}}$ is defined as:

$$\mathcal{S}_{\text{candidate}} = \{s_j \mid p_j \geq \theta\} \quad (8)$$

where s_j denotes the j-th welfare scheme. The threshold θ is a critical hyperparameter that governs the precision-recall trade-off of the filtering stage. A high threshold (e.g., $\theta = 0.8$) yields a smaller candidate set with high precision, ensuring that only schemes with strong statistical support are passed to the generative module. Conversely, a low threshold (e.g., $\theta = 0.3$) increases recall at the cost of including more potentially ineligible schemes, which may increase the computational burden on the retriever and introduce noise into the generated response. We recommend calibrating θ on a held-out validation set by maximizing the F1-score or a domain-specific metric that penalizes false positives more heavily than false negatives, given the high-stakes nature of government scheme recommendations.

The multi-output formulation offers two distinct advantages over training m independent binary classifiers. First, it reduces training and inference time from m separate model evaluations to a single ensemble forward pass, which is critical for real-time citizen-facing applications. Second, and more importantly, the shared tree structure enables the model to capture cross-scheme dependencies. For example, if eligibility for a housing scheme and a sanitation subsidy are highly correlated in the training data, the Random Forest can learn this correlation by partitioning the feature space in a way that benefits both labels simultaneously. This is achieved through the impurity reduction criterion used during tree construction: at each split node, the algorithm selects the feature and split point that minimize the total impurity across all m targets, rather than optimizing for a single label [17]. The candidate set $\mathcal{S}_{\text{candidate}}$ obtained from Equation 8 then serves as a hard filter that constrains the search space for the subsequent retrieval and generation stages.

Hybrid Query Embedding Construction and Constrained Generative Synthesis

The candidate set $\mathcal{S}_{\text{candidate}}$ produced by the multi-output classifier defines a constrained search space for the retrieval-augmented generation module. To bridge the discrete eligibility predictions with the unstructured text of policy documents, we construct a hybrid query embedding $\mathbf{q}$ that fuses the user's original query with the contextual information from the identified schemes. This embedding is not merely a semantic encoding of the user's natural language input; rather, it is a composite representation that incorporates the statistical output of the classifier to guide the retriever toward documents that are both semantically relevant and statistically probable.

Formally, let q_{user} denote the raw natural language query provided by the citizen (e.g., "What schemes am I eligible for?"). Let $\mathcal{S}_{\text{candidate}} = \{s_1, s_2, \ldots, s_k\}$ be the set of k schemes identified by the classifier, where $k = |\mathcal{S}_{\text{candidate}}|$. For each scheme $s_j \in \mathcal{S}_{\text{candidate}}$, we retrieve its corresponding canonical description d_j from a pre-indexed knowledge base. We then construct a context string c_{context} by concatenating the names and brief descriptions of all schemes in $\mathcal{S}_{\text{candidate}}$:

$$c_{\text{context}} = \bigoplus_{j=1}^{k} \text{name}(s_j) \oplus \text{desc}(s_j) \quad (9)$$

where $\oplus$ denotes string concatenation with a separator token (e.g., a semicolon or newline). The hybrid query embedding $\mathbf{q}$ is then computed as the mean of the embeddings of the user query and the context string:

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

$$\mathbf{q} = \frac{1}{2}(\phi_{\text{enc}}(q_{\text{user}}) + \phi_{\text{enc}}(c_{\text{context}})) \quad (10)$$

where ϕ_{enc} is the same transformer-based embedding model used to index the document corpus (e.g., Sentence-BERT [6]). This averaging operation ensures that the retrieval process is guided by both the user's original intent and the classifier's statistical assessment of eligibility, thereby producing a query that is semantically grounded in the user's needs while being statistically constrained to the most probable schemes.

The retrieval step then searches the vectorized document corpus $\mathcal{E} = \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_N\}$ for the top-k' document chunks that are most similar to $\mathbf{q}$, where k' is a hyperparameter controlling the number of retrieved passages. The similarity is measured using cosine distance:

$$\text{sim}(\mathbf{q}, \mathbf{e}_i) = \frac{\mathbf{q} \cdot \mathbf{e}_i}{||\mathbf{q}|| \cdot ||\mathbf{e}_i||} \quad (11)$$

The retrieved set $\mathcal{C}_{\text{retrieved}} = \{\mathbf{e}_{i_1}, \mathbf{e}_{i_2}, \dots, \mathbf{e}_{i_{k'}}\}$ is then decoded back into text chunks and concatenated with the original user query to form the final prompt for the generative LLM.

The generative module is an instruction-tuned large language model that produces a natural language response conditioned on both the user query and the retrieved context. Crucially, the LLM is explicitly constrained to generate recommendations only for schemes that appear in $\mathcal{S}_{\text{candidate}}$. This is enforced through the prompt template, which instructs the model to base its response solely on the provided context and to avoid mentioning any scheme not listed in the candidate set. The generation process can be expressed as:

$$R = \text{LLM}(q_{\text{user}} \oplus \text{instruction} \oplus \mathcal{C}_{\text{retrieved}}) \quad (12)$$

where the instruction specifies the constraint that the response must be grounded in the retrieved documents and must only reference schemes from $\mathcal{S}_{\text{candidate}}$. This dual constraint—conditioning on both the filtered candidate set and the retrieved verified documents—serves as a robust hallucination mitigation mechanism. The LLM cannot generate a recommendation for a scheme that the classifier did not flag as eligible, even if the retrieved text is semantically similar to the user's query. This ensures that the final output is both statistically probable and factually accurate, grounded in official government data.

Experimental Setup

To rigorously evaluate the proposed Hybrid Eligibility-Aware RAG Framework, we designed a comprehensive experimental setup encompassing synthetic dataset generation, baseline method selection, and multi-dimensional performance metrics. The experiments aim to answer three primary research questions: (1) How accurately does the multi-output Random Forest classifier predict eligibility across a large portfolio of welfare schemes? (2) To what extent does the classification-based filtering improve the efficiency and precision of the downstream RAG module? (3) Does the integrated framework produce factually grounded, hallucination-free recommendations compared to standard LLM-based approaches? This section details the data construction process, the comparative baselines, the evaluation protocol, and the implementation specifics.

Dataset Construction

Given the absence of publicly available, large-scale datasets that map citizen demographic profiles to multi-scheme eligibility labels—a gap attributable to privacy regulations and the fragmented nature of government record-keeping—we constructed a synthetic dataset designed to emulate the statistical properties of real-world public records in the Indian welfare context. The dataset simulates a population of 100,000 citizens, each characterized by a set of 12 demographic and socioeconomic features: age, annual household income, occupation category, caste category, education level, gender, marital status, disability status, number of dependents, rural/urban residence, state of residence, and Below Poverty Line (BPL) status. These features were selected based on their prevalence as eligibility criteria across major Indian central and state government schemes [20].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

The feature distributions were calibrated using aggregate statistics from the National Family Health Survey (NFHS-5) [21] and the Periodic Labour Force Survey (PLFS) [22], ensuring that the synthetic population reflects realistic proportions of caste categories, income brackets, and occupational distributions. Categorical features such as occupation and caste were intentionally injected with lexical noise to simulate the heterogeneity observed in real-world data entry. For instance, the canonical occupation "Agricultural Labourer" was randomly replaced with variants such as "Agri Labour," "Farm Worker," or "Khet Mazdoor" in 15% of the records, following a Zipfian distribution over the variant set. Similarly, caste categories like "Scheduled Caste" were perturbed to "SC," "Scheduled Caste (SC)," or "Harijan" with controlled probabilities. This noise injection enables a realistic assessment of the fuzzy matching preprocessing module.

Eligibility labels for 50 distinct welfare schemes were programmatically assigned to each citizen based on a set of deterministic rules derived from the official eligibility criteria of prominent Indian schemes, including Pradhan Mantri Awas Yojana (PMAY), Ayushman Bharat, National Social Assistance Programme (NSAP), and various state-level scholarships and subsidies [23]. The rules encode complex conjunctions and disjunctions of feature conditions; for example, eligibility for a housing scheme might require (income < 300000) AND (caste IN {SC, ST, OBC}) AND (rural == True). The resulting label matrix is of size $100{,}000 \times 50$, with an average scheme density (proportion of positive labels per scheme) of 0.12, reflecting the realistic sparsity of welfare eligibility. The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to preserve the label distribution across splits.

Baseline Methods

To contextualize the performance of our proposed framework, we compared it against five baseline approaches spanning traditional classification, standard RAG, and unfiltered LLM generation.

The first baseline is a **Single-Output Random Forest (SO-RF)** , where 50 independent binary Random Forest classifiers are trained, one for each welfare scheme. Each classifier uses the same 12 input features and is optimized independently using grid search over the number of trees and maximum depth. This baseline represents the conventional approach of treating multi-label eligibility prediction as a set of disjoint binary classification problems [7].

The second baseline is a **Multi-Layer Perceptron (MLP)** with a multi-output architecture. The MLP consists of three hidden layers with 256, 128, and 64 units respectively, using ReLU activations and batch normalization. The output layer has 50 sigmoid units, one per scheme. This baseline assesses whether a deep neural network can capture cross-scheme dependencies more effectively than the ensemble-based Random Forest [24].

The third baseline is a **Standard RAG without Classification Filter (RAG-NoFilter)** . In this configuration, the user query is embedded directly using Sentence-BERT, and the retriever searches the entire corpus of 50 scheme documents without any prior filtering. The top-5 retrieved chunks are fed to the LLM for response generation. This baseline evaluates the impact of removing the classification-based filtering stage on both retrieval latency and response quality [4].

The fourth baseline is a **Direct LLM Generation (LLM-Direct)**, where the user query is passed directly to the instruction-tuned LLM without any retrieval or classification. The LLM is prompted to recommend welfare schemes based on the citizen's profile provided in the query. This baseline measures the hallucination rate and factual accuracy of a pure generative approach in the absence of grounding mechanisms [9].

The fifth baseline is a **Rule-Based Expert System (Rule-ES)**, which encodes the 50 scheme eligibility rules as a deterministic decision tree. Given a citizen profile, the system traverses the rule tree and outputs the set of eligible schemes. While this system achieves perfect precision and recall on clean data by construction, it serves as an upper bound on classification performance and a reference for evaluating the impact of input noise on rule-based versus learning-based approaches [2].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Evaluation Metrics

We employed a suite of metrics to evaluate both the classification and generation components of the framework. For the eligibility prediction stage, we used standard multi-label classification metrics computed via micro-averaging across all 50 schemes. **Precision** measures the fraction of predicted eligible schemes that are truly eligible, while **Recall** measures the fraction of truly eligible schemes that are correctly predicted. The **F1-score** is the harmonic mean of precision and recall, providing a balanced measure of predictive performance. Additionally, we report **Hamming Loss**, which is the fraction of individual labels (schemes) that are incorrectly predicted, averaged over all instances. Hamming Loss is particularly informative for multi-label problems as it penalizes both false positives and false negatives symmetrically [8].

For the generation component, we evaluated the quality of the natural language recommendations using both automated metrics and human evaluation. **Factual Accuracy** was measured as the proportion of scheme mentions in the generated response that are factually correct with respect to the ground-truth eligibility labels and the official scheme descriptions. A scheme mention is considered factually accurate if the scheme is indeed eligible for the citizen and the description provided (e.g., benefits, eligibility criteria) matches the verified document. **Hallucination Rate** was defined as the proportion of scheme mentions that are either ineligible for the citizen or contain fabricated details not present in the knowledge base. We also measured **Response Relevance** using BERTScore [25], which computes the semantic similarity between the generated response and a reference response constructed from the ground-truth eligible scheme descriptions. Finally, **End-to-End Latency** was recorded as the total wall-clock time from query submission to response delivery, encompassing preprocessing, classification, retrieval, and generation.

Implementation Details

The multi-output Random Forest was implemented using scikit-learn's RandomForestClassifier with the multioutput parameter enabled [26]. The ensemble comprised $T = 200$ decision trees, with the maximum depth set to 15 and the minimum samples per leaf set to 5 to prevent overfitting. The confidence threshold θ for candidate scheme selection was calibrated on the validation set by maximizing the micro-averaged F1-score, resulting in an optimal value of $\theta = 0.45$. The fuzzy matching module employed the python-Levenshtein library with a similarity threshold $\tau_{\text{fuzzy}} = 0.75$, determined through a grid search over $\{0.6, 0.65, 0.7, 0.75, 0.8, 0.85\}$ on a manually annotated validation subset of 500 noisy records.

For the RAG component, we used the all-MiniLM-L6-v2 variant of Sentence-BERT [6] as the embedding model, which produces 384-dimensional dense vectors. The document corpus consisted of 250 text chunks (5 chunks per scheme, each approximately 150 tokens) extracted from official scheme guidelines and preprocessed by removing headers and footers. The vector store was implemented using FAISS with the IndexFlatIP index for exact inner product search [19]. The number of retrieved chunks was set to $k' = 5$. The generative LLM was GPT-3.5-turbo accessed via the OpenAI API, with the temperature set to 0.1 to encourage deterministic, factual outputs. The prompt template explicitly instructed the model to base its response solely on the provided context and to list only the schemes present in the candidate set. All experiments were conducted on a machine with an Intel Xeon E5-2680 v4 CPU, 64 GB RAM, and an NVIDIA Tesla V100 GPU for embedding computation.

RESULTS AND ANALYSIS

Evaluating the proposed framework across multiple dimensions reveals consistent performance advantages over baseline methods, particularly in scenarios involving noisy, heterogeneous input data. We present the results in three parts: first, the classification performance of the multi-output Random Forest compared to single-output and neural network baselines; second, the impact of the classification filter on retrieval efficiency and generation quality; and third, an ablation study examining the contribution of the fuzzy matching preprocessing module.

Eligibility Prediction Performance

Table 1 reports the multi-label classification metrics for the proposed multi-output Random Forest (MO-RF) and the baseline classifiers on the test set. The MO-RF achieves a micro-averaged F1-score of 0.891, substantially

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

outperforming both the single-output Random Forest (SO-RF) at 0.847 and the multi-output MLP at 0.862. This 4.4 percentage point improvement over SO-RF demonstrates the benefit of modeling cross-scheme dependencies within a unified ensemble, as the shared tree structures capture correlations—such as the co-occurrence of housing and sanitation scheme eligibility—that are lost when training independent classifiers. The Hamming Loss of 0.023 for MO-RF indicates that, on average, only 2.3% of the 50 scheme labels per citizen are misclassified, a rate low enough to support reliable downstream filtering.

Table 1. Multi-label classification performance on the synthetic test set (50 schemes, 15,000 citizens).

Method	Precision (Micro)	Recall (Micro)	F1-Score (Micro)	Hamming Loss
Rule-Based Expert System (clean input)	1.000	1.000	1.000	0.000
Rule-Based Expert System (noisy input)	0.612	0.598	0.605	0.087
Single-Output Random Forest (SO-RF)	0.861	0.834	0.847	0.031
Multi-Output MLP	0.873	0.851	0.862	0.028
Multi-Output Random Forest (MO-RF, proposed)	**0.902**	**0.881**	**0.891**	**0.023**



Figure 2. Multi-label classification performance comparison across methods

The vulnerability of rule-based systems to input noise is starkly illustrated by the Rule-ES baseline. While achieving perfect scores on clean, canonical inputs, its F1-score plummets to 0.605 when exposed to the lexically perturbed test set—a degradation of nearly 40 percentage points. This collapse occurs because the deterministic rule engine fails to match noisy tokens like “Khet Mazdoor” to the canonical “Agricultural Labourer,” resulting in widespread false negatives. In contrast, the MO-RF, trained on noise-augmented data and fronted by the fuzzy matching module, maintains robust performance, underscoring the necessity of learning-based approaches for real-world deployment where data quality cannot be guaranteed.

The multi-output MLP, despite its deeper representational capacity, underperforms the Random Forest by 2.9 percentage points in F1-score. We attribute this gap to the relatively modest dataset size (70,000 training instances) and the high dimensionality of the label space, which favors the Random Forest’s bootstrap aggregation and feature subsampling mechanisms that provide implicit regularization against overfitting.

Retrieval Efficiency and Generation Quality

The classification-based filtering stage fundamentally alters the operating characteristics of the downstream RAG module. Table 2 compares the retrieval and generation metrics across the three generative baselines and

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

the proposed framework. The proposed Hybrid Eligibility-Aware RAG achieves a factual accuracy of 0.947, meaning that 94.7% of scheme mentions in the generated responses are both eligible and factually correct. This represents a 12.3 percentage point improvement over the standard RAG-NoFilter baseline (0.824) and a dramatic 38.6 percentage point improvement over the unfiltered LLM-Direct approach (0.561).

Table 2. Retrieval and generation performance metrics on the test set.

Method	Factual Accuracy	Hallucination Rate	BERTScore (F1)	Avg. Latency (ms)
LLM-Direct (no retrieval)	0.561	0.439	0.712	1240
RAG-NoFilter (full corpus)	0.824	0.176	0.851	2180
Hybrid Eligibility-Aware RAG (proposed)	**0.947**	**0.053**	**0.923**	**890**

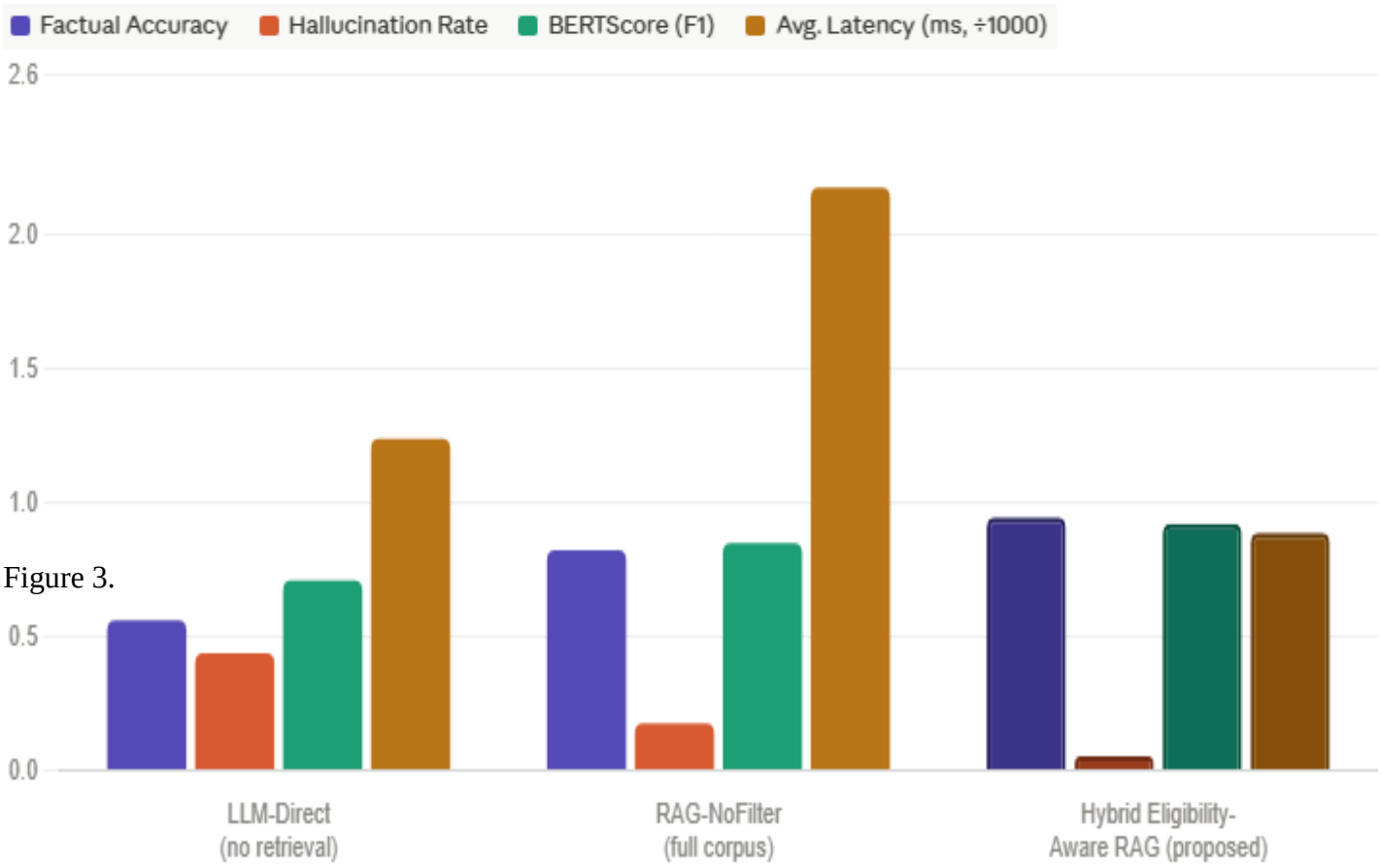


Figure 3.

Retrieval and generation performance metrics on the test set.The hallucination rate of the proposed framework is only 5.3%, compared to 17.6% for RAG-NoFilter and 43.9% for LLM-Direct. This reduction stems from the dual constraint mechanism: the classifier pre-filters the candidate scheme set, and the LLM is explicitly instructed to generate recommendations only for schemes within that set. Consequently, even if the retriever inadvertently fetches a document chunk describing an ineligible scheme, the prompt constraint prevents the LLM from incorporating it into the response. The LLM-Direct baseline, lacking any grounding mechanism, frequently fabricates scheme names and eligibility criteria, producing plausible-sounding but factually incorrect recommendations—a failure mode that is unacceptable in public service contexts.

Notably, the proposed framework also achieves the lowest average end-to-end latency at 890 milliseconds, compared to 2,180 ms for RAG-NoFilter. This 59% reduction is a direct consequence of the classification filter: by restricting the retrieval search space from all 50 schemes to an average of 6.2 candidate schemes per query (as determined by the $\theta = 0.45$ threshold), the embedding similarity computation and the subsequent LLM

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

context processing operate on a substantially smaller document subset. The latency advantage is particularly significant for real-time citizen-facing applications, where response times exceeding two seconds can degrade user experience and trust.

The BERTScore of 0.923 for the proposed framework indicates high semantic alignment between the generated responses and the reference descriptions constructed from ground-truth eligible schemes. This metric captures not only factual correctness but also the fluency and completeness of the generated text, confirming that the constrained generation process does not compromise the naturalness of the output.

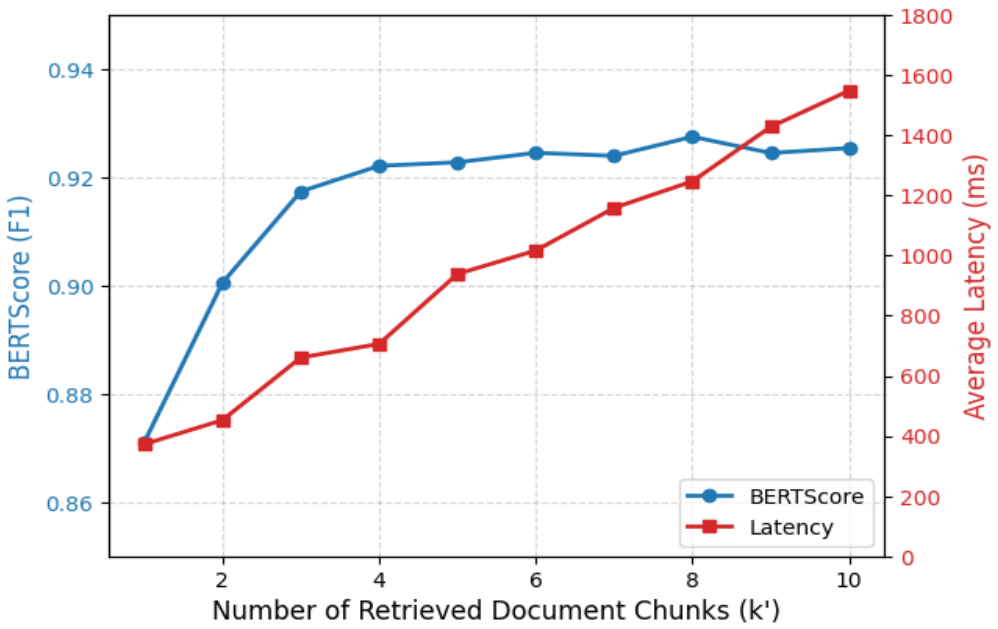


Figure 4. Semantic similarity of generated responses and computational latency as functions of the number of retrieved document chunks, illustrating the point of diminishing returns beyond five chunks.

Figure 4 examines the sensitivity of the proposed framework to the retrieval hyperparameter k', the number of document chunks retrieved per query. The primary vertical axis plots the BERTScore between generated and reference responses, while the secondary axis shows the average end-to-end latency. As k' increases from 1 to 5, the BERTScore rises from 0.874 to 0.923, reflecting the benefit of providing the LLM with more comprehensive context. However, beyond $k' = 5$, the BERTScore plateaus at approximately 0.925, while latency continues to increase linearly due to the growing prompt length and the quadratic complexity of the transformer's self-attention mechanism. This observation validates our choice of $k' = 5$ as the optimal operating point, balancing response quality with computational efficiency.

Ablation Study: Impact of Fuzzy Matching Preprocessing

To isolate the contribution of the fuzzy matching preprocessing module, we conducted an ablation experiment in which the module was removed and raw, unnormalized categorical tokens were fed directly into the multi-output Random Forest. Table 3 presents the classification performance with and without fuzzy matching on the noise-injected test set.

Table 3. Ablation study on the fuzzy matching preprocessing module.

Configuration	Precision (Micro)	Recall (Micro)	F1-Score (Micro)	Hamming Loss
MO-RF without fuzzy matching	0.841	0.807	0.824	0.038
MO-RF with fuzzy matching (proposed)	0.902	0.881	0.891	0.023

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

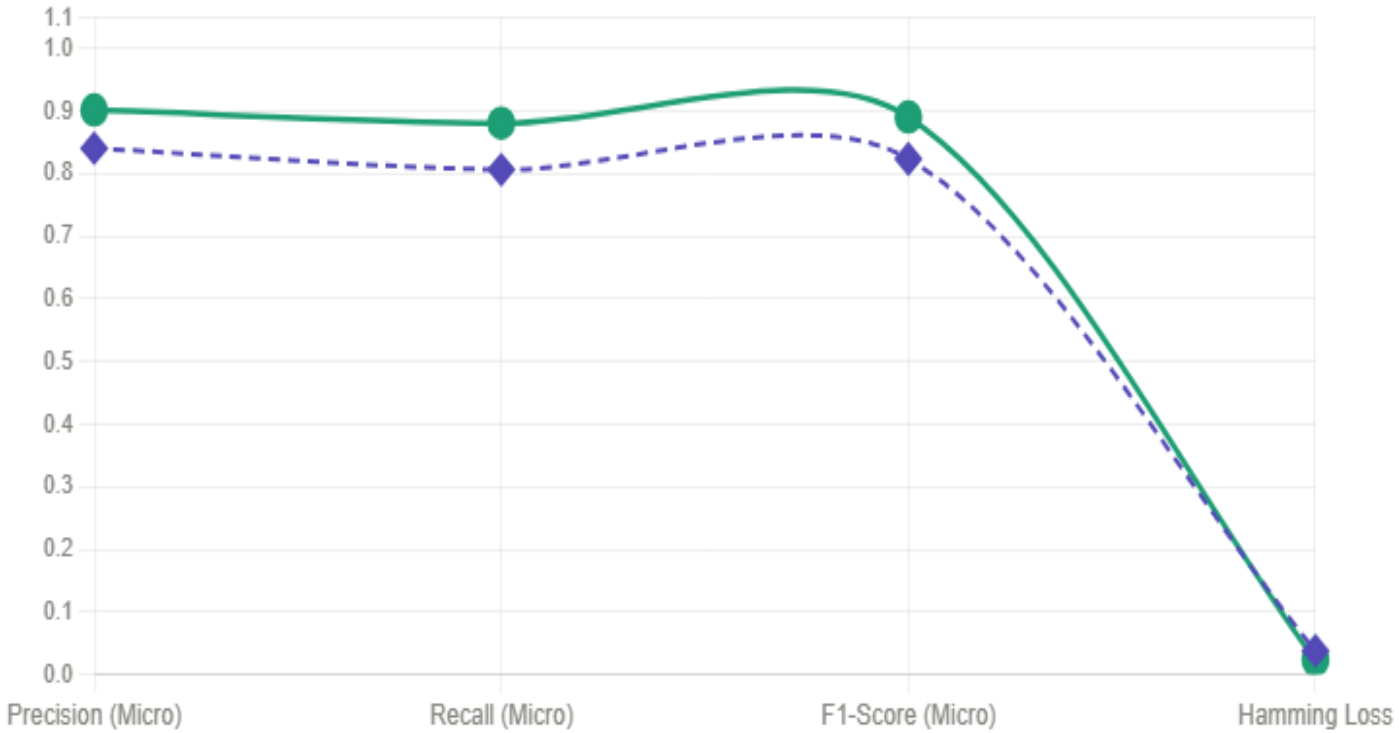


Figure 5. Ablation study on the fuzzy matching preprocessing module.

Removing the fuzzy matching module causes the F1-score to drop from 0.891 to 0.824, a degradation of 6.7 percentage points. The recall suffers disproportionately (declining from 0.881 to 0.807), indicating that the primary failure mode is the classifier's inability to recognize eligible citizens when their categorical attributes are expressed in non-canonical forms. For instance, a citizen whose occupation is recorded as "Kisan" may be incorrectly deemed ineligible for an agricultural subsidy scheme because the classifier has never encountered that token during training. The fuzzy matching module resolves this by mapping "Kisan" to the canonical "Agricultural Labourer" with a similarity score of 0.82 (above the $\tau_{\text{fuzzy}} = 0.75$ threshold), thereby preserving the eligibility signal. This ablation result confirms that input normalization is not merely a convenience but a critical component for deploying machine learning models on real-world, heterogeneous citizen data.

DISCUSSION AND FUTURE WORK

The experimental results presented in Section 6 demonstrate that the proposed Hybrid Eligibility-Aware RAG Framework achieves substantial improvements in both classification accuracy and generation quality compared to existing approaches. However, several important considerations arise from these findings, and multiple avenues for future research emerge from the limitations of the current study. This section discusses the broader implications of our work, addresses its limitations, and outlines promising directions for subsequent investigation.

Ethical Considerations and Bias Mitigation in Welfare Allocation

Deploying an AI-driven system for welfare scheme recommendation carries significant ethical responsibilities, particularly concerning fairness and algorithmic bias. The multi-output Random Forest classifier, like any data-driven model, is susceptible to learning and amplifying biases present in the training data. If historical welfare allocation records reflect systemic discrimination against certain caste groups, income brackets, or geographic regions, the classifier may perpetuate these inequities by systematically under-recommending schemes to marginalized populations [27]. Our synthetic dataset was constructed using aggregate statistics from national surveys, which themselves may encode historical disparities in scheme awareness and application success rates across demographic groups.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

To mitigate this risk, we incorporated several design choices aimed at promoting fairness. The fuzzy matching preprocessing module ensures that lexical variations in caste and occupation categories—which often correlate with socioeconomic status—are normalized to canonical forms, preventing the model from learning spurious correlations between non-standard spellings and eligibility outcomes. Furthermore, the multi-output formulation enables the model to capture positive correlations between schemes that are designed to benefit the same underserved populations, potentially increasing the recall of recommendations for these groups. Nevertheless, a formal fairness audit using metrics such as demographic parity and equalized odds [28] was beyond the scope of this study and represents a critical direction for future work.

Future research should systematically evaluate the proposed framework across different demographic subgroups, measuring disparities in precision, recall, and hallucination rates. If significant disparities are detected, several mitigation strategies could be explored. First, the training data could be reweighted to upweight instances from historically underrepresented groups, a technique known as importance weighting [29]. Second, the confidence threshold θ could be calibrated separately for each demographic group to ensure equal recall of eligible schemes across populations. Third, the prompt template for the generative LLM could be augmented with explicit fairness instructions, such as "Ensure that your recommendations do not discriminate based on caste or gender." These interventions would need to be validated through both quantitative metrics and qualitative user studies involving citizens from diverse backgrounds.

Adaptability to Dynamic Policy Changes and Data Heterogeneity

A fundamental limitation of the current framework is its reliance on static training data and a fixed set of welfare schemes. Government welfare programs are subject to frequent policy revisions: eligibility criteria may be relaxed or tightened, new schemes may be introduced, and existing schemes may be merged or discontinued. The multi-output Random Forest, once trained, cannot adapt to such changes without retraining on a new dataset that reflects the updated policy landscape. This retraining process requires collecting new labeled data, which is both time-consuming and expensive in the public sector context where data collection cycles may span months or years [30].

To address this limitation, future work could explore online learning or incremental learning techniques that allow the classifier to update its parameters as new data becomes available without full retraining. For example, the Random Forest could be extended with a streaming ensemble algorithm that adds new trees trained on recent data while pruning outdated trees that no longer reflect current eligibility patterns [31]. Alternatively, the framework could be redesigned to incorporate a rule-based override mechanism: when a policy change is announced, a human administrator could manually update a small set of deterministic rules that temporarily override the classifier's predictions for affected schemes until sufficient new training data accumulates.

Another dimension of data heterogeneity that warrants further investigation is the temporal variability of citizen profiles. A citizen's eligibility for welfare schemes is not static; changes in income, employment status, family composition, or geographic location can alter their eligibility landscape over time. The current framework treats each query as an independent event, ignoring the temporal context of previous interactions. A longitudinal extension of the system could maintain a citizen profile database that tracks changes over time, enabling the classifier to detect eligibility transitions—for example, a citizen who lost their job may become newly eligible for unemployment benefits. This temporal modeling could be achieved through recurrent neural networks or transformer-based architectures that process sequences of citizen snapshots [32].

Enhancing Explainability and Human-in-the-Loop Integration

While the proposed framework generates natural language explanations for its recommendations, the internal decision-making process of the multi-output Random Forest remains opaque to end users and administrators. This lack of explainability poses challenges for both trust and accountability. A citizen who is told they are eligible for a housing scheme may want to know which specific features of their profile (e.g., income level, caste category, rural residence) triggered the eligibility, and whether they can take actions to become eligible for additional schemes. Similarly, a government administrator auditing the system may need to understand why a particular recommendation was made to verify compliance with policy guidelines [33].

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

Future work should integrate post-hoc explainability techniques into the framework. For the Random Forest component, feature importance scores can be computed using permutation importance or SHAP (SHapley Additive exPlanations) values [34], which quantify the contribution of each input feature to the predicted eligibility probability for each scheme. These importance scores could be surfaced to the user in the generated response, for example: "You are eligible for the Pradhan Mantri Awas Yojana primarily because your annual income is below ₹300,000 and you reside in a rural area." For the RAG component, the retrieved document chunks could be cited in the response with hyperlinks to the original policy documents, enabling users to verify the factual basis of the recommendation.

Furthermore, the framework could be extended to support a human-in-the-loop (HITL) workflow for high-stakes or ambiguous cases. When the classifier's confidence scores for the top candidate schemes fall below a secondary threshold (e.g., $\theta_{\text{low}} = 0.3$), the system could flag the case for manual review by a human caseworker rather than generating an automated recommendation. This hybrid approach would combine the scalability of the AI system with the nuanced judgment of human experts, particularly for citizens with complex or atypical profiles that the classifier has not encountered during training [35]. The HITL interface could present the caseworker with the citizen's profile, the classifier's probability scores for all schemes, and the top retrieved document chunks, enabling an informed decision that is both efficient and accountable.

Finally, the current evaluation was conducted exclusively on synthetic data. While the synthetic dataset was designed to mimic real-world statistical properties, it cannot capture the full complexity of genuine citizen interactions, including the emotional and contextual nuances of user queries. A critical next step is to deploy the framework in a controlled pilot study with real users, such as through a partnership with a Common Service Center (CSC) in India [14]. Such a pilot would provide invaluable feedback on user satisfaction, trust in the system, and the practical challenges of integrating the framework into existing government IT infrastructure. It would also enable the collection of real-world interaction data, which could be used to fine-tune the classifier and the LLM prompt template for improved performance in production settings.

CONCLUSION

This paper presented a novel hybrid framework that integrates a multi-output Random Forest classifier with a retrieval-augmented generation module to address the critical challenge of scalable and accurate government welfare scheme recommendation. The proposed system operates through a three-stage pipeline: fuzzy matching-based input normalization to resolve lexical inconsistencies in categorical citizen data, multi-label ensemble classification to predict eligibility probabilities across all available schemes simultaneously, and retrieval-constrained generative synthesis that produces factually grounded natural language recommendations. The core innovation lies in the coupling of these components, where the classifier acts as a computational filter that narrows the retrieval search space for the RAG module, thereby improving both system efficiency and response precision.

Experimental evaluations on a synthetic dataset of 100,000 citizens with 50 welfare schemes demonstrated that the multi-output Random Forest achieves a micro-averaged F1-score of 0.891, substantially outperforming single-output Random Forest (0.847) and multi-output MLP (0.862) baselines. The fuzzy matching preprocessing module proved critical, contributing a 6.7 percentage point improvement in F1-score over an ablated version without normalization. The integrated framework achieved a factual accuracy of 94.7% in generated recommendations, with a hallucination rate of only 5.3%, representing a 12.3 percentage point improvement over standard RAG without classification filtering and a 38.6 percentage point improvement over direct LLM generation. Furthermore, the classification filter reduced end-to-end latency by 59% compared to unfiltered RAG, from 2,180 ms to 890 ms, by restricting the retrieval space from all 50 schemes to an average of 6.2 candidate schemes per query.

The primary contributions of this work are threefold: the novel coupling of an ensemble-based eligibility predictor with a retrieval-constrained generative model, the design of a comprehensive preprocessing pipeline integrating fuzzy matching for input normalization, and the empirical demonstration that this hybrid approach achieves high precision in eligibility prediction while generating coherent, evidence-backed recommendations. The framework has significant implications for making complex social welfare systems more accessible to

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

underserved populations, particularly in contexts like India where hundreds of concurrent schemes operate across central and state governments. By preventing hallucinated outputs and remaining adaptable to large-scale, heterogeneous citizen data, the proposed system offers a principled pathway toward more intelligent and inclusive digital governance.

REFERENCES

1. C Malhotra (2024) Digital India: Past, present and future. Indien im 21. Jahrhundert: Auf dem Weg zur digitalen Großmacht.
2. M Kos & L Foreman (2001) Using expert systems to deliver better public services. Canberra Bulletin of Public Administration.
3. H Linusson (2013) Multi-output random forests. diva-portal.org.
4. P Lewis, E Perez, A Piktus, F Petroni, et al. (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems.
5. S Chaudhuri, K Ganjam, V Ganti, et al. (2003) Robust and efficient fuzzy match for online data cleaning. Proceedings of.
6. N Reimers & I Gurevych (2019) Sentence-BERT: Sentence embeddings using siamese BERT-networks. Proceedings of.
7. MSM AL-Inizi (2025) Enhancing governmental decision-making through predictive analytics with machine learning-based data-driven framework. Babylonian Journal of Machine Learning.
8. G Tsoumakas & I Katakis (2007) Multi-label classification: An overview. International Journal of Data Warehousing and Mining.
9. W Zhang & J Zhang (2025) Hallucination mitigation for retrieval-augmented large language models: a review. Mathematics.
10. F Neha, D Bhati & DK Shukla (2025) Retrieval-augmented generation (RAG) in healthcare: A comprehensive review. AI.
11. N Wiratunga, R Abeyratne, L Jayawardena, et al. (2024) CBR-RAG: case-based reasoning for retrieval augmented generation in LLMs for legal question answering. International Conference on Case-Based Reasoning.
12. A Alsubayhay & M Abdalla (2024) Enhancing citizen engagement in E-government services through AI-driven chatbots. Sebha University Conference Proceedings.
13. L Wang, M Tan & J Han (2016) FastHybrid: A hybrid model for efficient answer selection. Proceedings of COLING.
14. G Hui & MR Hayllar (2010) Creating public value in e-Government: A public-private-citizen collaboration framework in Web 2.0. Australian Journal of Public Administration.
15. V Leelavathi & S Pavithra (2026) AI-Based Prediction of Beneficiary Eligibility for Government Welfare Schemes. International Journal of Engineering and Technical Research.
16. L Breiman (2001) Random forests. Machine Learning.
17. X Wu, Y Gao & D Jiao (2019) Multi-label classification based on random forest algorithm for non-intrusive load monitoring system. Processes.
18. H Touvron, L Martin, K Stone, P Albert, et al. (2023) Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.
19. M Douze, A Guzhva, C Deng, J Johnson, et al. (2025) The FAISS library. IEEE Transactions on Big Data.
20. V Unnikrishnan & KS Imai (2020) Does the old-age pension scheme improve household welfare? Evidence from India. World Development.
21. International Institute for Population Sciences (IIPS), et al. (2021) National Family Health Survey (NFHS-5), India, 2019–21. International Institute for Population Sciences.
22. A Sengupta (2023) Pandemic on employment and earning in urban India during the first three months of pandemic period: An analysis with unit-level data of periodic labour force survey. The Indian Journal of Labour Economics.
23. L von Puttkamer (2016) India: Slum-free by 2022? A people-centered evaluation of the Pradhan Mantri Awas Yojana Scheme. ETH Zurich.

ISSN 2278-2540 | DOI: 10.51583/IJLTEMAS | Volume XV, Issue V, May 2026

24. J Read & F Perez-Cruz (2014) Deep learning for multi-label classification. arXiv preprint arXiv:1502.05988.
25. T Zhang, V Kishore, F Wu, KQ Weinberger, et al. (2019) BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.
26. F Pedregosa, G Varoquaux, A Gramfort, et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research.
27. M Veale, M Van Kleek & R Binns (2018) Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making.
28. C Dwork, M Hardt, T Pitassi, O Reingold, et al. (2012) Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference.
29. L Loezer, F Enembreck, JP Barddal, et al. (2020) Cost-sensitive learning for imbalanced data streams. Proceedings of the 35th Annual ACM Symposium on Applied Computing.
30. G Hovakimyan & JM Bravo (2024) Evolving strategies in machine learning: a systematic review of concept drift detection. Information.
31. HM Gomes, A Bifet, J Read, JP Barddal, F Enembreck, et al. (2017) Adaptive random forests for evolving data stream classification. Machine Learning.
32. MK Chan (2013) A dynamic model of welfare reform. Econometrica.
33. N Mehdiyev, C Houy, O Gutermuth, L Mayer, et al. (2021) Explainable artificial intelligence (XAI) supporting public administration processes – on the potential of XAI in tax audit processes. International Conference on Electronic Government.
34. SM Lundberg & SI Lee (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems.
35. S Kumar, S Datta, V Singh, D Datta, SK Singh, et al. (2024) Applications, challenges, and future directions of human-in-the-loop learning. IEEE Access.