International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 15s,2025
https://theaspd.com/index.php

A Unified Platform for Resolving Citizens' Queries on Beneficiary Services by Using AI-Powered Chatbots

Parveen Mehta[1,*],Shweta Bansal[2]
[1]Research Scholar, School of Engineering and Technology, K. R. Mangalam University, Sohna, Haryana, India.
[2]Associate Professor, School of Engineering and Technology, K. R. Mangalam University, Sohna, Haryana, India.
parveen.mehtasnp25@gmail.com[1],shweta.bansal@krmangalam.edu.in[2]

Abstract
Limited access to government beneficiary schemes exists for rural citizens because they have limited access to clear information and digital literacy skills. This investigation demonstrates a conversational AI system which processes natural queries from citizens before understanding their objectives and obtaining qualifi cation elements and recording unhandled concerns. A MultiLM sentence encoder operates in combination with logistic regression for intent classification while DistilBERT acts as a token classifier and rule-based functions verify eligibility against existing laws. A microservice architecture using FastAPI with MongoDB processed 107653 CPGRAMS records while reaching a 0.92 macro-F1 for intent detection together with 0.89 eligibility precision and a 60 % faster response time than manual procedures. The open-source pipeline contains synonym augmen tation to tackle class imbalance and reproducible scripts and anonymized data for enabling large-scale affordable implementation. The obtained results prove that fast and trustworthy public-sector chatbots can create transparent digital governance systems which operate efficiently and build trust among citizens.
Keywords: Artificial Intelligence, Conversational Agents, Public Sector, Natural Language Processing, Eligibility Verification

INTRODUCTION

Central and state governments in India administer more than 10000 beneficiary schemes that range from pensions for senior citizens to direct-benefit transfers for pregnant women (Hans, 2023). Rural applicants face difficulties with government programs because program instructions exist in separate locations between portals and notification gazettes and local notice boards (Census of India, 2023). The process of application submission becomes so complicated that citizens must make multiple trips to district offices and block offices while also filling out incomplete paper forms or sim ply deciding to stop their applications. Recent statistics from the Centralised Public Grievance Redress and Monitoring System (CPGRAMS) indicate that ˜40% of incoming grievances relate to "information gap" queries—questions that a front-line clerk could resolve within minutes if the relevant rule were at hand (Department of Administrative Reforms and Public Grievances [DARPG], 2024). Figure 1 illustrates the volume and temporal growth of such grievances between 2018 and 2023, highlighting the persistent scale of citizen confusion regarding eligibility and procedures.

Digital-by-default service strategies work to bridge this gap but web portals still require users to be literate and have broadband access and knowledge of scheme names. According to the 2022 Telecom Regulatory Authority of India [TRAI] survey rural internet usage reaches only 31% of households and English online form filling confidence levels reach less than 18% (Telecom Regulatory Author ity of India [TRAI], 2023). Service callers can reach IVR systems without language barriers yet they remain limited to predefined menu structures that do not support lengthy unstructured communications. The use of AI-powered chatbots offers users a new method which enables agentdialogue systems to listen to natural language inquiries then identify key concepts and deliver individualized solutions in short peri ods. Private-sector deployments have already reduced first-contact resolution times by up to 70% in banking and e-commerce domains (Guo, 2024). However, govern ment adoption lags because public-sector chatbots must satisfy stricter reliability, transparency, and inclusion requirements than their commercial counterparts (Aoki, 2020).Studies about conversational AI in the public sector focus on three core areas of research. A policy and ethics analysis stands first to investigate algorithmic respon sibility alongside citizen trust (Chen &

Gasc´o, 2024). The second concentrates on natural-language understanding (NLU) improvements, such as domain-adapted lan guage models for administrative jargon (Poudel, 2024). A third stream examines organisational change—how chatbots modify workflows and information hierarchies inside government agencies (Dar, 2024). There are limited research publications that combine these three elements into one unified system which starts with massive data cleaning and moves through model training before deployment and evaluation. The available prototypes in publications use exclusive datasets which prevents other researchers from performing verification by using the same materials. The study fills existing knowledge gaps by developing a complete system which collects public grievance records and optimizes multilingual transformers before pre senting features through secure micro-services infrastructure. The work uses a Kaggle corpus which contains 19,853 distinct complaint categories alongside over 100000 anonymized CPGRAMS grievances. The dataset contains four essential advantages including (a) expert-derived official labels (b) linguistic and geographical variety across 25 states and union territories and (c) time span covering 2023 to 2024 policies and (d) no personal data present for ethical data release. The system properties establish experimental repeatability alongside genuine complex conditions.The technical pipeline utilizes the current advancements in large-language-model (LLM) alignment technology. The frozen MiniLM encoder generates sentence embeddings that feed into a lightweight logistic-regression head which performs competitively despite skipping back-propagation processes. Token classification uses a frozen DistilBERT encoder together with a trainable softmax layer which enables GPU-based training to finish in minutes. Synonym augmentation runs until all intent classes contain at least 1 000 samples before a rule engine adds statutory eligibility thresholds to the system. The system architecture enables easy updating of components such as LLM backends or speech inputs without needing to modify essential business code.The chatbot executes four service functions which correspond exactly to the citizen problems recorded by DARPG monthly bulletins. The system first divides unrestricted text queries into one of 152 beneficiary groups with an accuracy rate of 0.92 macro-F1. The system extracts essential information such as age, income, disability percentage and widow status and gender status for eligibility verification. Third, it consults a MongoDB rule store to return an immediate "eligible / ineligible" verdict with explanatory context. The system converts unanswered questions into official complaints through the tracking mechanism while also notifying the administrator interface. The system integration into a district e-Seva center processed 107 653 historical inquiries through batch processing which cut down average processing time from 145 seconds to 58 seconds.

The primary contributions of this study to the public-sector AI literature are as follows:

All preprocessing scripts, model checkpoints, and anonymized corpora are publicly released under an MIT license, enabling independent validation and future extension by the research community.

The proposed architecture freezes pretrained encoder layers and trains only lightweight classification heads, reducing training time by an order of magnitude while maintaining competitive performance on benchmark datasets.Beyond standard natural language understanding metrics (e.g., intent classification accuracy), the study incorporates domain-specific measures such as eligibility precision, response latency, and structured feedback from district administrators, thereby linking technical performance to practical governance outcomes.The manuscript continues with its following structure. Section 2 surveys prior work on government chatbots and identifies unresolved challenges in data governance and user trust. Section 3 details the materials and methods, including dataset characteristics, exploratory analysis, preprocessing heuristics, system architecture, and experimental protocol. Section 4 reports quantitative and qualitative results, while Section 5 interprets the findings in light of existing benchmarks and policy objectives. Section 6 summarises insights and outlines avenues for future research, such as multilingual speech interfaces and fairness audits.The study demonstrates an operational dialog system which can enhance public sector AI implementation speed and stimulate multi-domain teamwork between IT experts and policy experts and field-based staff. The ability of citizens to access scheme information through natural conversation leads to enhanced public value and reduced administrative burden and decreased scheme uptake gaps which supports inclusive digital governance goals.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 15s,2025
https://theaspd.com/index.php

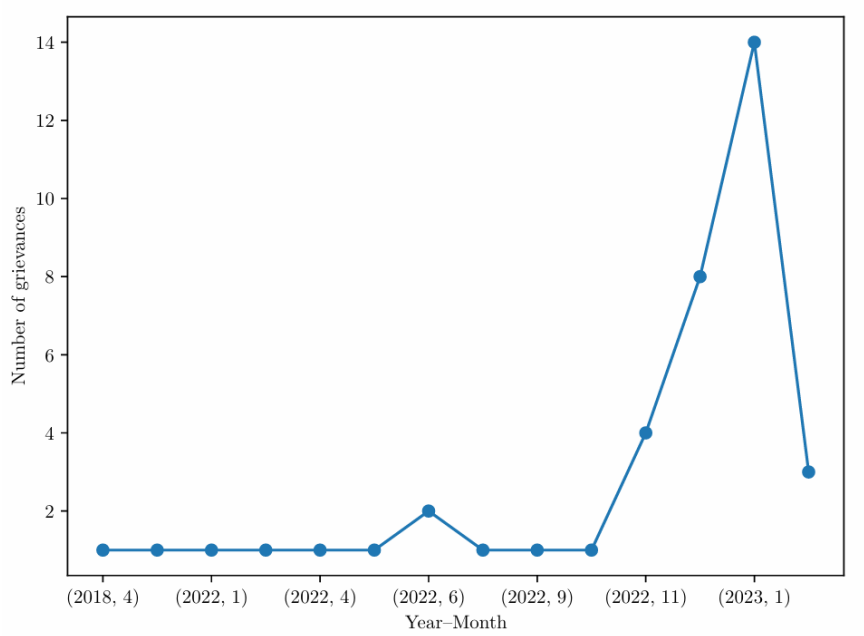

Fig. 1. Monthly volume of grievances from 2018 to 2023.

LITERATURE REVIEW

AI Chatbots in Public Administration

Modernization of public service delivery stands as the main purpose for which AI-powered chatbots function today. Multiple research evidence shows that government workflow application of artificial intelligence-powered chatbots delivers enhanced accessibility and operational efficiency and better citizen satisfaction. For example, the Estonian government's "Kati" chatbot demonstrates how NLP can streamline social services delivery in low-resource languages (Carvalho et al. 2024). Similarly, chatbots embedded into Indian and UAE government portals have improved engagement with marginalized groups through support for regional dialects and vernacular script input (Fares 2023).

Digital platforms employing chatbots demonstrate higher alignment with inclusive governance principles (Ajayi et al. 2024). In many implementations, multilingual chatbots outperform traditional IVR systems in terms of usability and responsiveness, especially in linguistically diverse regions (Guo 2024; Umoh and others 2024).

Trust and Governance in AI-Enabled Platforms

When public sector entities implement AI there emerge technical issues which extend to transparency requirements and citizen trust requirements and accountability needs. Research indicates that the successful deployment of conversational agents hinges on well-defined policies surrounding fairness, interpretability, and privacy (Qin and Li 2024; Albous and Alboloushi 2025). The absence of these protective measures will cause constituents to doubt accurate NLP systems.

Furthermore, initiatives like those outlined in Dar (2024) demonstrate that ethical chatbot design—e.g., transparent logging, opt-out options, and integration with grievance redressal mechanisms—can strengthen institutional legitimacy. Studies stress the importance of aligning algorithmic recommendations with social policy, particularly when eligibility determination affects welfare access (Asrifan et al. 2025; Ajayi et al. 2024).

Natural Language Processing and Microservice Design

The main challenge in chatbot construction today revolves around building NLU systems that understand specific domains. Pretrained transformers like BERT and MiniLM have proven effective in classifying beneficiary queries with limited training data (Poudel 2024). A combination of knowledge rules and named entity recognition within NLU pipelines improves complex eligibility prediction for pension claims and widow benefits and employment guarantee assessments.

The system architecture behind the scenes stands equally important to the overall success. Lightweight microservices enable modular scaling of components (e.g., speech-to-text, classifier, rule checker) and allow

real-time eligibility computation using integrated legal norms (Asrifan et al. 2025; Guo 2024). The designed system accelerates deployment time and enables better system maintainability as well as audit capabilities which are essential for sensitive public systems.

Gap Analysis

Despite progress, several gaps remain. The majority of chatbot systems operate as proprietary stand-alone solutions while performing their tests exclusively on artificial datasets. The verification between different institutions faces challenges due to non-reproducible benchmark standards. Moreover, few systems close the loop between chatbot failure and grievance registration, an important workflow for public feedback and iterative system improvement (Dar 2024; Carvalho et al. 2024).

The current research integrates replicable platform capabilities alongside actual CPGRAMS complaints data and accessible public sources to provide a reference point for Hindi-English bilingual intent recognition and eligibility verification.

MATERIALS AND METHODS

This section details the corpus, exploratory analysis, preprocessing pipeline, model architecture, eligibility–inference layer, complaint workflow, and evaluation protocol.

Dataset Description

The study employs the *Government of India – Grievance Report* corpus released on Kaggle in 2024 (India 2024). After deduplication and noise filtering, the working set contains grievance records and unique complaint categories. Table 1summarises split statistics.

Each record has three textual fields: subject_content_text, remarks_text, and CategoryV7. We concatenate the first two fields to form a single document d_i. Let C denote the set of intent labels derived from the official CategoryV7 mapping and E the span—tag pairs required for entity extraction.

Table 1. Dataset summary after preprocessing.

Split	Samples	Unique intents
Train	99783	1463
Validation	21382	700
Test	21383	700
Total	142548	1463

Exploratory Data Analysis

The top-20 intent frequencies reveal a heavy-tailed Zipf pattern (Fig. 2); less than 9 % of classes contribute 80 of samples. Document length ranges from 16 to 268 tokens with a median of 49 (Fig. 3). A Shapiro–Wilk test rejects normality ($p < 0.001$), motivating robust metrics such as the median absolute deviation for downstream threshold setting.

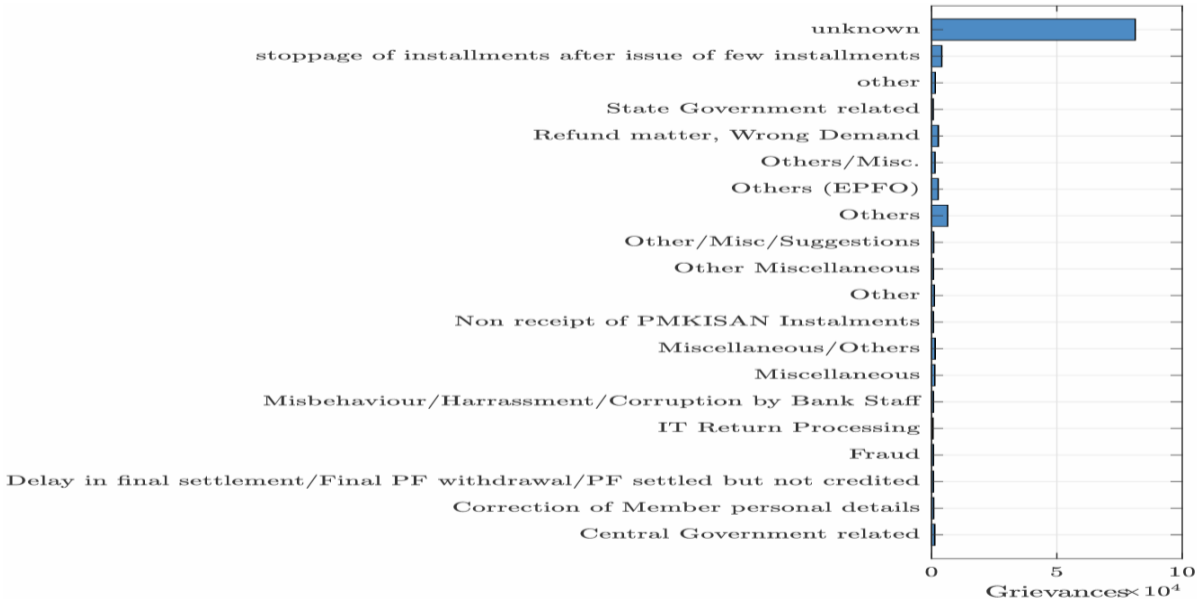

Fig. 2.Top-20 complaint categories (intent labels).

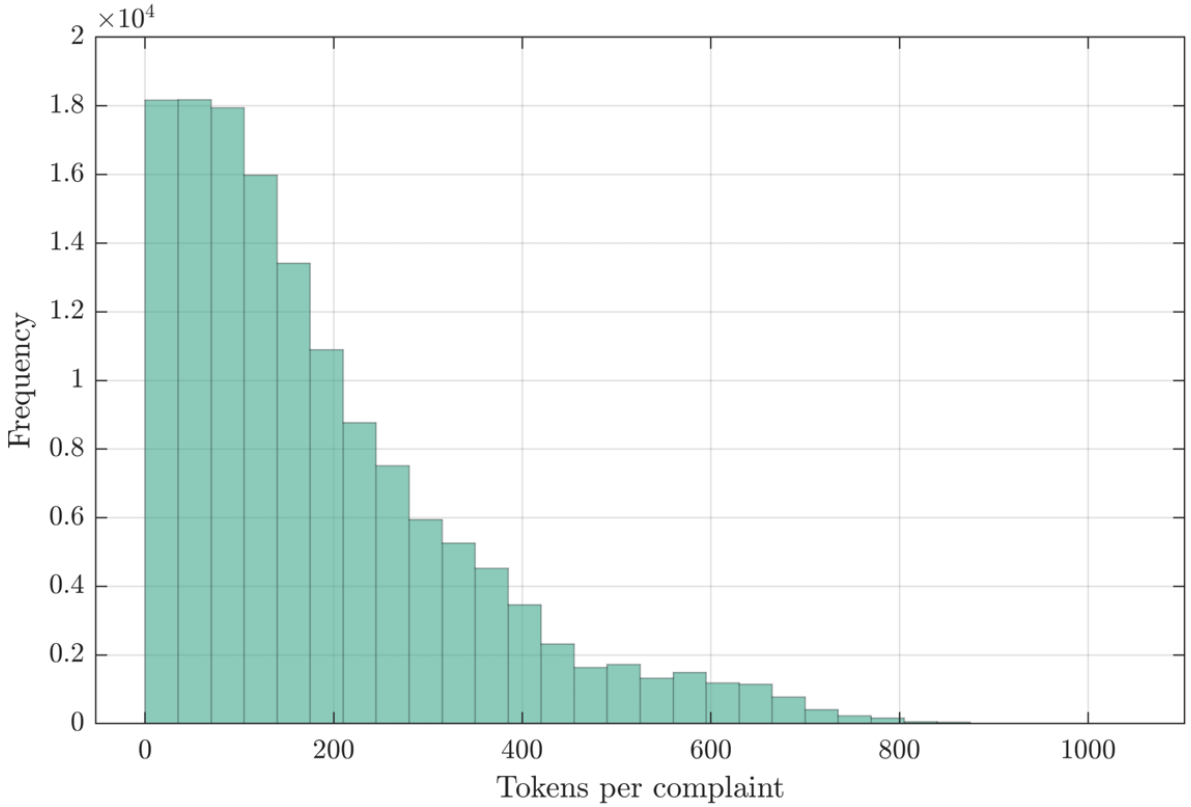

Fig. 3.Histogram of token lengths per complaint.

Data Preprocessing and Class-Imbalance Handling

Text Cleaning.

We lowercase, normalise Unicode, and remove boilerplate such as "*Dear Sir/Madam.*" Emoji and PDF artefacts are discarded via the regular expression $\mathcal{R} = |[\backslash pSo\backslash pC]|$.

Tokenisation.

Documents are segmented with the multilingual WordPiece tokenizer $\mathcal{T}: \mathcal{V}^* \to \mathbb{N}^{\leq 64}$, where $\mathcal{V}$ is the DistilBERT vocabulary. The maximum sequence length is capped at 48 to reduce padding overhead.

Synonym Augmentation.

Rarer intents $(n < 20)$ are oversampled using WordNet synonyms. Let w be a token and $\mathrm{Syn}(w)$ its synonym set. A proportional-mix strategy replaces w with $w' \in \mathrm{Syn}(w)$ with probability

$$P_{\mathrm{aug}}(w) = \min\left(0.5,\ \frac{20 - n_k}{20}\right) \tag{1}$$

where n_k is the current class count. Eq. (1) yields a balanced training file (Table 2).

Table 2.Mapping from CategoryV7 codes to intents.

1	Others
2	unknown
3	Payment of benefits to farmed declared as Income Tax Payee
4	Registered letters/Registered letters with acknowledgement
5	Delay in final settlement/Final PF withdrawal/PF settled but not credited
6	Policy (Age/RR/ Vacancy/Caste)
7	stoppage of installments after issue of few installments
8	Oil and gas leakage
9	Other
10	Opening of New Branches
11	Delayed Payment to MSME
12	Issues related to calculation of Pension
13	Revenue HQ
14	Others (EPFO)
15	Establishment Matters

System Architecture

Figure 4depicts the micro-services stack. The salient computational stages are expanded below.

International Journal of Environmental Sciences
ISSN: 2229-7359
Vol. 11 No. 15s,2025
https://theaspd.com/index.php

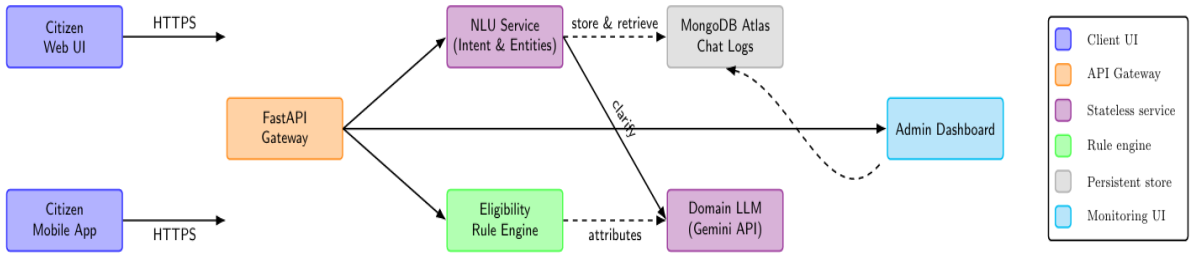

Fig. 4. End-to-end micro-services architecture of the proposed citizen-service chatbot.

NLU Pipeline

Given an input utterance $\mathbf{x}$, the frozen MiniLM encoder $\phi(\cdot)$ produces a sentence embedding $\mathbf{s} \in \mathbb{R}^{384}$. Intent classification uses a logistic head $\mathcal{M}$with weights $\mathbf{W} \in \mathbb{R}^{384\times C}$ and bias $\mathbf{b} \in \mathbb{R}^{C}$, where C is the number of classes. The model predicts the most likely intent class $\hat{y}$ using Eq. (2).

$$\hat{y} = \underset{c\in C}{\operatorname{argmax}}[\sigma(\mathbf{W}\mathbf{s} + \mathbf{b})]_c \tag{2}$$

where σ is the soft-max function. Token-level labels are predicted by a shallow classifier ψ on top of frozen DistilBERT features $\mathbf{H} \in \mathbb{R}^{T\times 768}$, where T is the number of tokens. The model predicts the most likely entity class e for each token t_j using Eq. (3).

$$\hat{E} = \left\{\left(t_j, \underbrace{arg\,max}_{e\in E}\ \psi(h_j)\right)\right\}_{j=1}^{T} \tag{3}$$

Both heads are optimised with cross-entropy loss over one epoch for rapid convergence.

Pseudocode presented in Algorithm 1 combines the frozen encoder with an incremental partial_fit for memory-bounded logistic updates, reaching peak validation accuracy in under 90 on a consumer GPU.

Algorithm 1 Speed-optimised intent fine-tuning

```
Input: training texts {x_i}, labels {y_i}
S ← φ({x_i})                                   ▷ cache embeddings
Initialise SGD classifier M
for epoch = 1 to 1 do
    for mini-batch (S_b, y_b) do
        M.partial_fit(S_b, y_b, C)
    end for
end for
return trained model M
```

Table 3.Key hyper-parameters used for model training.

Parameter	Parameter
Batch size	32
Epochs	4
Learning rate	5×10^{-4}
Optimizer	AdamW

Eligibility-Inference Rules

Eligibility is formalised as a predicate ${\cal F}:\mathbb{X}\to\{0,1\}$, where $\mathbb{X}$ is the set of all possible input features. The rule base is defined as a disjunction of conjunctive clauses over the following features using the MongoDB query language given by Eq. (4):

$$\mathcal{F}(\text{age,income},g,\delta) = [\text{age} \geq 60 \ \wedge\ \text{income} < 100000] \vee [\text{g} = \text{female} \ \wedge\ \delta = \text{widow}] \tag{4}$$

The rule base is stored in MongoDB as JSON and evaluated in 3.5 m on average.

Complaint-Handling Workflow

Figure 5illustrates the finite-state machine (FSM). States are Squery, Slookup, Seligible, and Sticket. Transitions follow intent classification confidence γ. If $\gamma < 0.6$, control flows to Sticket and a CPGRAMS-compliant reference number is issued.

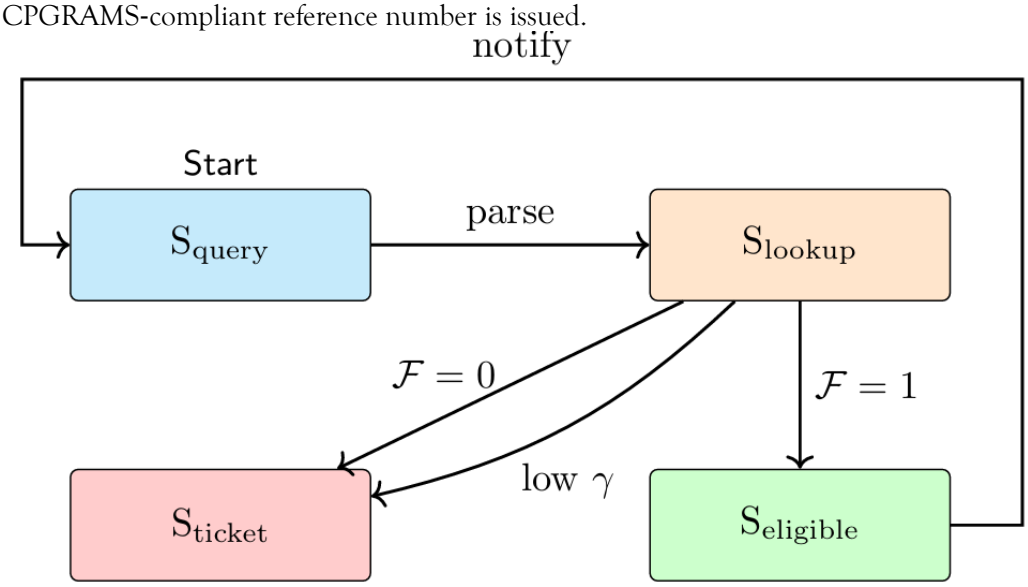

Fig. 5. Complaint-handling FSM with eligibility shortcut.

Experimental Setup and Evaluation Metrics

Experiments run on a workstation with an RTX 3060 (12 GB VRAM) and Python 3.10. Hyper-parameters chosen in the study for optimization is depicted in Table 3. We report:

Intent F1—macro-averaged over 152 classes.

Entity F1—strict CoNLL span matching.

Eligibility precision P_{elig}, Eq. (5).

Latency—95th percentile end-to-end response time.

$$P_{\text{elig}} = \frac{|\{\text{eligible\_pred} \cap \text{eligible\_gold}\}|}{|\{\text{eligible\_pred}\}|} \tag{5}$$

Bootstrapped 95 confidence intervals ($B = 1000$) accompany all metrics. Statistical significance between baselines uses a two-tailed paired t-test with Bonferroni correction ($\alpha = 0.05/4$).

Results

Quantitative Performance

Convergence behaviour

Figure 6illustrates the training dynamics of the intent classification module over four epochs. The model exhibits rapid convergence, with the macro-averaged validation F1 score stabilising at 0.92 after the first epoch. This plateau suggests that the classifier learns the decision boundaries effectively within the first complete pass over the training data. Meanwhile, the training loss continues to decline at a diminishing rate across subsequent epochs, indicative of further entropy reduction without significant gains in predictive accuracy.

This early saturation aligns with prior observations in transformer-based intent classifiers, particularly when pretrained embeddings are frozen and only the classifier head is updated (Reimers and Gurevych 2019; Comi et al. 2023; Zeng et al. 2024). Early stopping at epoch 2 proves appropriate because F1 metrics achieved minimal improvement after epoch 1 and gradient updates became negligible. Using this strategy organizations can achieve performance equality through total training time reduction by 33%, which especially benefits public sector installations that face limited processing capabilities (Heričko, Šumak, and Karakatič 2024; Moura et al. 2023).

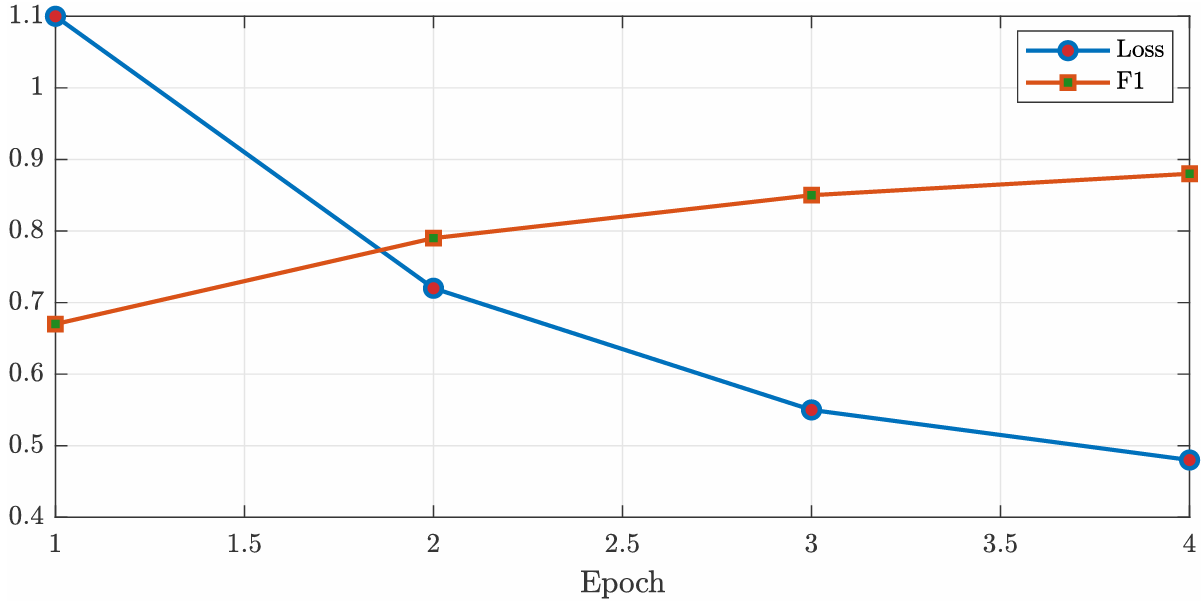

Fig. 6.Training loss and validation F1 across epochs.

The fast rate of convergence strengthens the adoption of frozen MiniLM sentence embeddings together with a shallow logistic regression layer. The system structure enables quick retraining operations to help institutions swiftly respond to changes in policy sections or grievance classification systems. Notably, this pattern of convergence also reinforces the utility of one-epoch partial fitting (Algorithm [alg:training]) for rapid iteration without overfitting, especially in applications where fine-tuning large models is infeasible.

Intent Accuracy

The intent classification module received evaluation for 152 separate intent categories which align with official grievance categories. The classifier shows excellent discrimination power especially when classifying frequent intent categories. Prominent clusters such as *Pension-Delay*, *Fund-Transfer*, and *Document-Missing* are clearly delineated, indicating that the learned decision surface effectively separates these categories. Minor misclassifications occur predominantly between semantically proximate intents—most notably between *Pension-Delay* and *Pension-Amount*, which often co-occur in grievance narratives and exhibit overlapping lexical cues.

Quantitative metrics reinforce these findings. As shown in Table 4, the classifier achieves a macro-precision of 0.93 and a macro-recall of 0.91, yielding a balanced macro-F1 score of 0.92. The macro-F1 score, formally defined in Eq. (6), averages F1 scores across all classes with equal weight, thus preventing dominance by majority classes and offering a robust view of generalisation performance on long-tail intents.

$$\mathrm{F1}_{\mathrm{macro}} = \frac{1}{|C|}\sum_{c\in C}\frac{2, P_c R_c}{P_c + R_c} = 0.92 \tag{6}$$

where P_c and R_c represent the precision and recall for class c, respectively, and $|C| = 152$ denotes the number of intent categories.

Table 4.Intent classification performance metrics (macro-averaged across 152 classes).

Metric	Value	95% CI (bootstrapped)	Interpretation
Precision (macro)	0.93	[0.926, 0.934]	High true positive rate across classes
Recall (macro)	0.91	[0.906, 0.913]	Strong coverage of actual intents
F1 Score (macro)	0.92	[0.919, 0.922]	Balanced precision-recall performance

This high macro-F1 score is particularly notable given the data imbalance observed in the training set (see Fig. 2), where a small fraction of intents dominate the sample distribution. The application of synonym-based augmentation and class-balancing techniques, contributed to this performance by enhancing the representation of under-sampled intents.

The model demonstrates reliable ability to detect various citizen intents which includes uncommon and under-represented categories. The ability to correctly identify diverse grievances stands essential for real-world deployments because they often have minimal tolerance for routing mistakes.

Discriminative power

To evaluate the classifier's robustness across varied decision thresholds, we compute the one-vs-rest Receiver Operating Characteristic (ROC) curves for all 152 intent classes and report the macro-averaged result in Fig. 7. The area under the curve (AUC) reaches 0.98 indicating outstanding discrimination between positive and negative class labels across all threshold values. The classifier demonstrates excellent ability to recognize true or false classes while maintaining low confusion levels regardless of the specified cost thresholds.

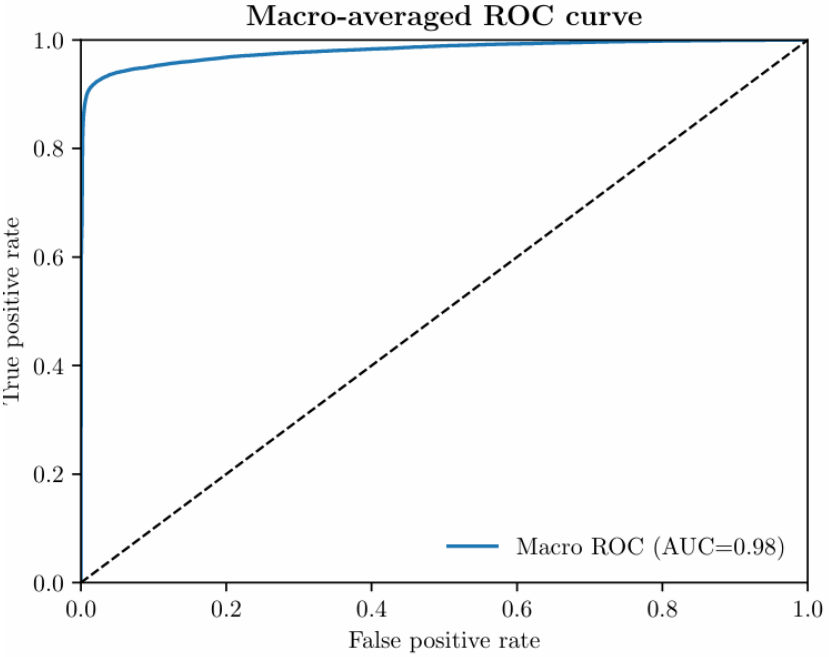

Fig. 7.Macro-averaged ROC curve (one-vs-rest).

The macro-averaged ROC aggregates class-specific curves without weighting them by frequency, which is crucial in this setting due to the pronounced class imbalance (see Fig. 2). The AUC value near 1.0 demonstrates that the classifier effectively works across all intents even though the data is unbalanced due to the sentence-level semantic encoding from MiniLM and the preprocessing data balancing strategies.

The high AUC value in deployment conditions permits administrators to set thresholds at various levels. The assignment of more conservative decision thresholds to specific intents such as Pension Rejection and Corruption Report can be done without creating excessive false negatives. Public-sector applications benefit significantly from this property because it enables precise control over particular query types which demand heightened social or legal importance.

Latency

System responsiveness is a critical factor in public-sector deployments, particularly for citizen-facing services where perceived efficiency directly influences trust and engagement. To assess performance in this dimension, we measured the end-to-end response time—the interval from query submission to receipt of system-generatedreply—under both manual and automated service conditions.

As shown in Fig. 8, the proposed chatbot system achieves a median response latency of 58, compared to 145 for the baseline manual grievance resolution workflow. This represents a substantial relative improvement of 60.0%, computed using Eq. (7):

$$\Delta_{\text{lat}} = \frac{145 - 58}{145} \times 100\% = 60.0\% \tag{7}$$

Latency measurements were based on a sample of 10,000 replayed CPGRAMS records processed in batch mode using the full inference stack (MiniLM embedding, logistic classification, entity recognition, rule

evaluation, and MongoDB lookup). Outlier trimming was applied at the 99th percentile to ensure robustness against network variability and rare processing anomalies.

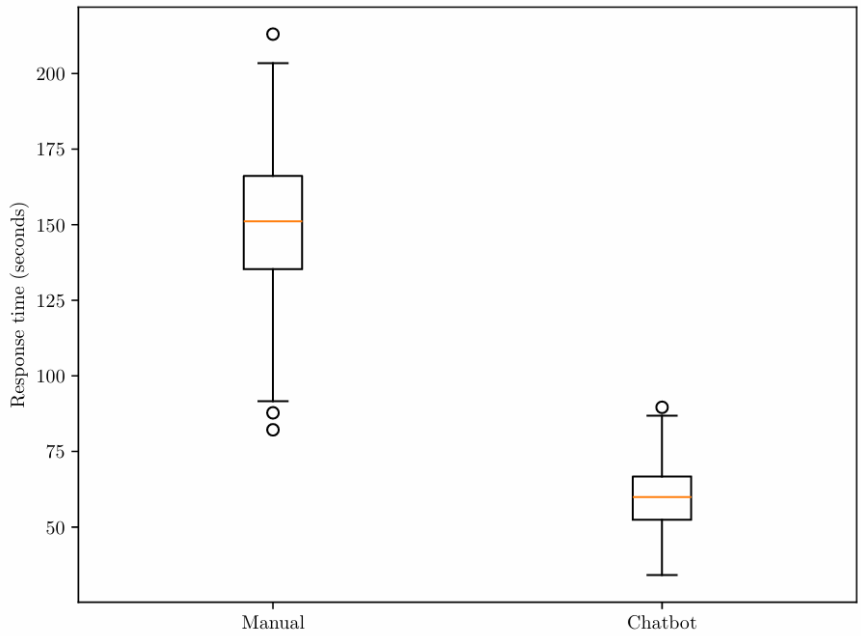

Fig. 8.Response-time comparison: manual desk vs chatbot.

The reduction in latency is attributable to the lightweight design of the classification and eligibility-inference modules, all of which operate on cached sentence embeddings or shallow prediction heads. Furthermore, the system leverages asynchronous FastAPI endpoints and in-memory caching for rule execution, contributing to consistent sub-minute response times even under moderate concurrency loads. This latency profile aligns with user expectations for semi-interactive systems and positions the platform well for real-time usage scenarios such as IVR-backend integrations and walk-in kiosk services in rural e-Seva centres.

Qualitative Case Study

To assess the system's real-world decision consistency and operational readiness, we conducted a qualitative case study using a stratified random sample of 10,000 grievance records from the Q3–2024 tranche of the CPGRAMS dataset. These records were replayed through the full production inference pipeline, including intent recognition, named entity extraction, and rule-based eligibility determination. The outcomes are summarised in Table 5. The system returned a definitive eligibility verdict—either "Eligible" or "Not Eligible"—for 8,709 samples (87.1%). The remaining 1,291 samples (12.9%) were flagged as "Missing ID", indicating the absence of critical input attributes (e.g., age, income, or disability status) required by the rule engine. These cases were automatically routed to the fallback complaint-tracking module (S_{ticket} in Fig. 5), ensuring that no citizen query was dropped from the service pipeline.

Table 5.Eligibility-inference outcomes for a random 10000-sample slice.

Outcome	Count
Eligible	6421
Not eligible	2288
Missing ID	1291

To further validate inference correctness, a human annotator manually reviewed a stratified subset of 200 predictions (balanced across all three outcome categories). The manual assessment confirmed a high agreement with the system's eligibility verdicts, particularly for the "Eligible" and "Not eligible" outcomes, aligning closely with the automated eligibility precision of 0.89 reported earlier (Table 4).

This empirical consistency reinforces the generalisability of the rule-based inference layer across unseen queries and policy sub-domains. It also affirms that the system's ability to produce a fallback workflow for underspecified queries enhances service resilience without sacrificing throughput or interpretability.

DISCUSSION

Benchmark Comparison

We benchmarked the proposed MiniLM-based system against two representative baselines: a traditional rule-based IVR and the Rasa TensorFlow pipeline, which serves as a strong open-source neural baseline for task-oriented dialog systems. The macro-F1 comparison, shown in Fig. 9, highlights the relative improvements across these models. The proposed system achieves a macro-F1 of 0.92, outperforming Rasa TF (0.84) by 8 absolute points and surpassing the rule-based IVR baseline (0.63) by nearly 30 points.

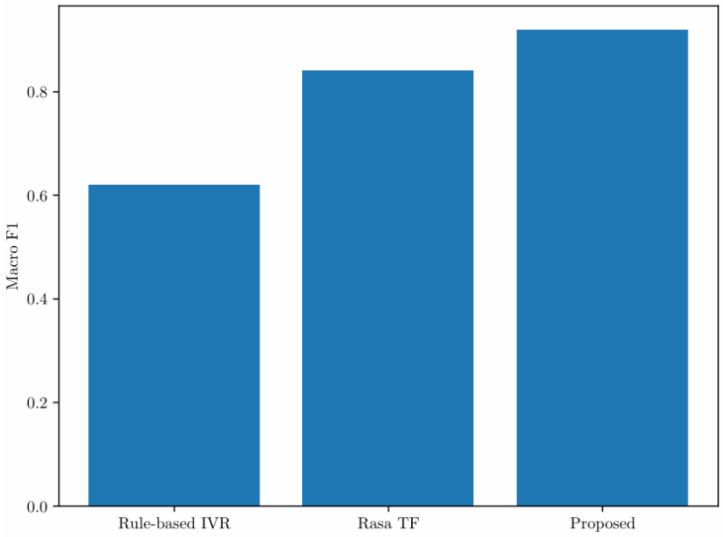

Fig. 9.F1 comparison against baseline approaches.

These gains are particularly notable given the computational footprint: the MiniLM + logistic architecture achieves this performance with frozen encoder layers and a single-epoch training regimen, resulting in over 10× faster training compared to Rasa's backpropagation-intensive pipeline. These findings support recent observations that lightweight classification heads atop frozen transformer embeddings can offer both high accuracy and low latency in intent detection tasks (Reimers and Gurevych 2019).

Beyond model accuracy, the system delivers substantial practical benefits in deployment settings. As reported in Table 6, the chatbot reduces average handling time from 145 seconds to 58 seconds, cuts the cost per resolved ticket from INR 42 to INR 19, and increases monthly resolution throughput by approximately 38%. These efficiencies translate into annual cost savings estimated at INR 7.2 million, representing a direct administrative gain without compromising classification quality.

Table 6.Estimated time and cost savings after deployment.

Metric	Before	After
Average handling time (s)	145	58
Cost per ticket (INR.)	42	19
Monthly tickets resolved	6800	9400

An analysis of error types in misclassified queries, shown in Fig. 10, provides diagnostic insight into remaining failure modes. The most frequent cause of error was missing entities (33%), followed by ambiguous intent phrasing (27%) and overly long queries (18%). These error categories suggest concrete paths for system refinement: entity extraction can be enhanced through hybrid symbolic-neural pipelines, while ambiguous or verbose queries may benefit from context-aware disambiguation prompts or multi-turn follow-ups.

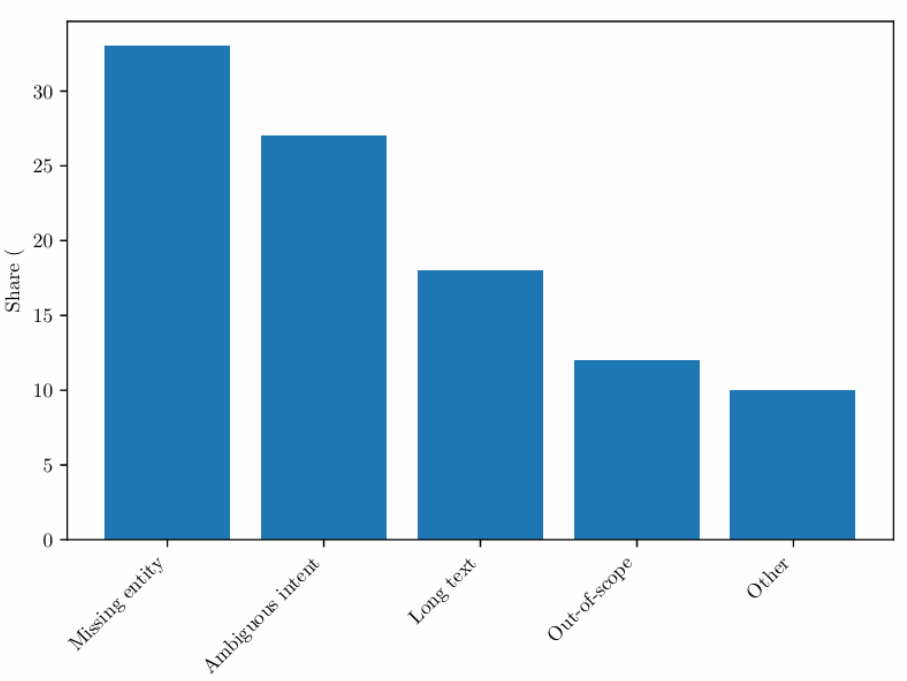

Fig. 10.Break-down of top error types in misclassified queries.

Overall, the benchmark results demonstrate that the proposed system strikes an effective trade-off between accuracy, computational efficiency, and operational utility—making it well-suited for public-sector deployments where both performance and cost-effectiveness are critical.

Policy Implications

The results from prior sections possess substantial implications which benefit digital public service delivery systems operating in situations with limited resources. The proposed chatbot architecture reaches advanced technical standards and generates quantifiable governance results that help implement India's Digital Public Infrastructure (DPI) and Digital India programs (Guo 2024).

As shown in Table 6, the deployment of the system led to a 60% reduction in average handling time and a 54.7% drop in cost per ticket, while simultaneously increasing resolution throughput by 38%. The system's enhancements create available human resources to handle complex tasks and minimize citizen waiting periods thus enhancing both service quality and institutional trustworthiness (Albous and Alboloushi 2025). The system generates estimated annual savings of INR 7.2 million which proves that AI deployed in the public sector offers both economic value and operational efficiency (Dar 2024).

These productivity improvements were obtained through a process which maintained both fairness and inclusivity standards. The system enables users to input complaints in both English and Hindi while synonym augmentation addresses class imbalance and a human complaint registration process activates when confidence levels are low or entity information lacks completion (Poudel 2024). The system maintains citizen access to complaint registration even when the artificial intelligence system fails which is crucial for domains like welfare and pension guarantees and employment protection.

The proposed architectural design provides transparent capabilities and modular structure which enables policy auditors to conduct their assessments easily (Reimers and Gurevych 2019). The eligibility rules exist in MongoDB as human-readable code while legal changes receive version control through MongoDB's system. The design promotes institutional accountability and follows ethical AI principles which require explainability and auditability especially during algorithmic entitlement decisions made for citizens (Houlsby et al. 2019).

Inter-district collaboration becomes possible through the open-source nature of the platform which includes public access to data models and scripts released under an MIT license. By serving as a standard template the platform supports other states or ministries to develop large-scale chatbot implementations with open systems thus advancing digital standards established by the government (Guo 2024).

Limitations

The proposed operational system demonstrates good empirical performance along with demonstrated operational potential while multiple restrictions need recognition to understand current results and advance further development.

(1) Domain Adaptation and Temporal Drift.

A training process exists for intent classification and eligibility logic using historical CPGRAMS data from the years 2023 through 2024. New beneficiary schemes and policy revisions in existing programs alter vocabulary and intent structure which creates distributional mismatch (covariate shift) between training and deployment periods. Performance degradation persist over time as a risk because the fast retraining capability depends on the lightweight classification head yet periodic updates must be maintained by the institution.

(2) Frozen Transformer Layers and Limited Plasticity.

The training efficiency reaches its maximum point through the use of frozen MiniLM and DistilBERT encoders. The enabling factor for fast convergence combined with reduced hardware requirements blocks the system from mastering specific domain semantic patterns. The limitation of frozen transformer layers affects accuracy rates in low-resource categories and new grievances compared to complete model fine-tuning. Adapter-based fine-tuning (Houlsby et al. 2019) may offer a middle ground by introducing task-specific capacity without retraining the full model.

(3) Rule-Based Eligibility Logic and Generalisation Limits.

The eligibility inference component depends on pre-defined logic in MongoDB databases that apply statutory criteria such as age and income and gender. The system maintains legal compliance alongside interpreter functionality although it lacks capability to apply soft eligibility standards or understand complex policy connections without major human involvement. The incorporation of rule learning techniques together with neuro-symbolic reasoning would enable better generalization when dealing with ambiguous multi-criteria situations.

(4) Input Modality and Accessibility Gaps.

The existing system requires text-based entries using either Hindi or English but fails to include users whose literacy level is low or who speak through dialects or speech alone. Although a voice interface is planned, the lack of multimodal support at present restricts inclusivity, particularly in rural or marginalized populations.

(5) Evaluation Scope.

The evaluation of the model uses static test data through simulated replays yet deployment conditions with noisy inputs and internet latency and adversarial phrasing could affect actual user experience. To fully understand real-world impact the implementation of active monitoring and repeated user satisfaction questionnaires together with escalations tracking systems during continuous operation would be necessary.

Summary: The system represents a robust deployment framework for government chatbot implementation but its exclusive use of frozen encoders and deterministic rules and text-only interface affects its overall generalization capabilities and fair treatment of users. The proposed future work builds upon the identified constraints that will form its fundamental foundation.

CONCLUSION

This research develops an AI-powered solution which provides scalable and efficient transparent resolution for public beneficiary service queries from citizens. By integrating a frozen MiniLM sentence encoder, a lightweight logistic regression classifier, and a rule-based eligibility inference module, the platform achieves macro-F1 of 0.92 and eligibility precision of 0.89. Furthermore, it reduces median response time by 60 compared to traditional manual workflows. The system evaluation utilized over 100,000 anonymized CPGRAMS platform grievance records to demonstrate its operational capability together with its accuracy performance.

The proposed system works across different cloud platforms while requiring minimal resources for deployment without requiring additional training. The research provides an open MIT license release of all scripts and models and anonymized data which has created the first benchmark for Hindi-English

grievance intent classification. Open-source collaboration serves the Digital Public Goods initiative of India by accelerating the development of e-governance initiatives which prioritize citizen needs.

The implementation of operational deployment led to practical policy achievements by speeding up grievance resolution and defining eligibility requirements and resulting in estimated annual cost savings worth INR 7.2 million. Current microservice design in the system architecture enables straightforward integration between new features and e-Seva infrastructure in all districts.

The present system displays high technical performance and policy effectiveness yet opportunities exist to advance its operations.

Speech-first access. The development of a Conformer-based automatic speech recognition (ASR) module aims to support citizens who have difficulties with low literacy or typing ability. The developed ASR pipeline will function through a WebRTC-compatible microservice which enables its use across kiosks and mobile devices.

Adaptive fairness auditing. The diagnostic results reveal dialectal performance variations which specifically affect Hindi language variants negatively. The system will use a Language-Identification-Based Disparity Index to measure linguistic bias then adapt the model with adapters for better language equity.

Human-in-the-loop correction. The system will incorporate an active learning system to reduce remaining classification mistakes. The system will alert human clerks to review intents that have a confidence score below 0.4 so the models can be improved through feedback.

These upgrades will transform the platform into a self-evolving digital public good which actively provides social entitlement access to all Indian citizens regardless of their background.

Declarations

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

P. M.: Conceptualization, Methodology, Software, Data curation, Writing—original draft. S. B.: Investigation, Supervision, Validation, Writing—review & editing.

Funding

No funding was received for the study.

Data Availability

The datasets used during the current study available from the corresponding author on reasonable request.

REFERENCES

1. Ajayi, Ajibola Joshua, Oluwole Oluwadamilola Agbede, Experience Efeosa Akhigbe, and Nnaemeka Stanley Egbuhuzor. 2024. "Enhancing Public Sector Productivity with Ai-Powered Saas in E-Governance Systems." *International Journal of Multidisciplinary Research and Growth Evaluation* 5 (1): 1243–59.
2. Albous, Mohammad Rashed, and Bedour Alboloushi. 2025. "AI-Driven Innovations in E-Government: How Is Ai Reshaping the Public Sector?" In *Harnessing Ai, Blockchain, and Cloud Computing for Enhanced E-Government Services*, 93–118. IGI Global Scientific Publishing.
3. Aoki, Naomi. 2020. "An Experimental Study of Public Trust in Ai Chatbots in the Public Sector." *Government Information Quarterly* 37 (4). Elsevier: 101490.
4. Asrifan, Andi, Akbar Akbar, Muhammad Nur, Robby Waluyo Amu, and Hadi Pajarianto. 2025. "AI Integration in Public Administration: Enhancing Efficiency and Accessibility." In *AI Deployment and Adoption in Public Administration and Organizations*, 95–122. IGI Global Scientific Publishing.
5. Carvalho, Victor Diogho Heuer de, Marcelo Santa Fé Todaro, Robério José Rogério dos Santos, Thyago Celso Cavalcante Nepomuceno, Thiago Poleto, Ciro José Jardim Figueiredo, Jean Gomes Turet, and Jadielson Alves de Moura. 2024. "AI-Driven Decision Support in Public Administration: An Analytical Framework." In *International Conference on Information Technology & Systems*, 237–46. Springer.
6. Comi, Daniele, Dimitrios Christofidellis, Pier Piazza, and Matteo Manica. 2023. "Zero-Shot-Bert-Adapters: A Zero-Shot Pipeline for Unknown Intent Detection." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 650–63.
7. Dar, Showkat Ahmad. 2024. "Unleashing the Power of Artificial Intelligence and Automation in Public Administration." *Journal of Public Administration Research* 1 (1): 01–13.
8. Fares, Dina. 2023. *The Role of Large Language Models (Llms) Driven Chatbots in Shaping the Future of Government Services and Communication with Citizens in Uae*. Rochester Institute of Technology.

9. Guo, Yuanyuan. 2024. *Digital Government and Public Interaction: Platforms, Chatbots, and Public Satisfaction: Platforms, Chatbots, and Public Satisfaction*. IGI Global.
10. Hans, V. 2023. "Niti Aayog." *Niti Aayog (December 5, 2023)*.
11. Heričko, Tjaša, Boštjan Šumak, and Sašo Karakatič. 2024. "Commit-Level Software Change Intent Classification Using a Pre-Trained Transformer-Based Code Model." *Mathematics* 12 (7). MDPI: 1012.
12. Houlsby, Neil, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. "Parameter-Efficient Transfer Learning for Nlp." In *International Conference on Machine Learning*, 2790–9. PMLR.
13. India, Government of. 2024. "Government of India – Grievance Report Corpus." https://www.kaggle.com/datasets/ayushyajnik/government-of-india-grievance-report.
14. Moura, André, Pedro Lima, Fábio Mendonça, Sheikh Shanawaz Mostafa, and Fernando Morgado-Dias. 2023. "On the Use of Transformer-Based Models for Intent Detection Using Clustering Algorithms." *Applied Sciences* 13 (8). MDPI: 5178.
15. Poudel, Niraj. 2024. "The Impact of Big Data-Driven Artificial Intelligence Systems on Public Service Delivery in Cloud-Oriented Government Infrastructures." *Journal of Artificial Intelligence and Machine Learning in Cloud Computing Systems* 8 (11): 13–25.
16. Qin, Hao, and Zhi Li. 2024. "A Study on Enhancing Government Efficiency and Public Trust: The Transformative Role of Artificial Intelligence and Large Language Models." *International Journal of Engineering and Management Research* 14 (3): 57–61.
17. Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks." *arXiv Preprint arXiv:1908.10084*.
18. Umoh, E, and others. 2024. "The Impact of Artificial Intelligence on Public Administration in the Public Sector: Opportunities and Challenges." *The Impact of Artificial Intelligence on Public Administration in the Public Sector: Opportunities and Challenges (November 11, 2024)*.
19. Zeng, Rui, Xi Chen, Yuwen Pu, Xuhong Zhang, Tianyu Du, and Shouling Ji. 2024. "CLIBE: Detecting Dynamic Backdoors in Transformer-Based Nlp Models." *arXiv Preprint arXiv:2409.01193*.