International Journal of Integrated Knowledge (IJIK)
Volume: 01 Issue: 01 | April-2026

AI-POWERED CHATBOTS IN CITIZEN SERVICE DELIVERY: EFFECTIVENESS, TRUST, AND INCLUSION

A Multi-Dimensional Study of Government Chatbot Deployments in India with Evidence from DigiSeva, UMANG, and State eGovernance Portals

Krapali Sikarwar
Research Associate, e-Governance Studies
Ujjain,(M.P.)India
Email: sikarwarkrapali@gmail.com
ORCID: 0009-0005-0380-2599

ABSTRACT

Artificial intelligence-powered chatbots have emerged as one of the fastest-growing eGovernance innovations globally, with 43% of national government portals now deploying AI chatbot functionality as a primary citizen service interface. In India, chatbot deployments across central and state eGovernance platforms — including UMANG, DigiSeva, MADAD, and numerous state-level virtual assistants — represent significant investments in AI-mediated citizen-government interaction. Yet the dominant discourse around government chatbot adoption focuses almost exclusively on deployment metrics and cost efficiencies, while neglecting three dimensions of critical governance importance: the quality and accuracy of information delivered to citizens (effectiveness), the conditions under which citizens trust and rely on government chatbot responses (trust), and the differential impact of chatbot interfaces on digitally and socially marginalized citizen groups (inclusion). This paper addresses these gaps through a rigorous, multi-method empirical study drawing on primary data from 890 respondents across urban, peri-urban, and rural settings in five Indian states, supplemented by a systematic audit of query resolution quality across six government chatbot platforms and thematic analysis of 72 in-depth qualitative interviews. The study develops and validates an original Government Chatbot Quality Index (GCQI) — a composite instrument measuring technical performance, information accuracy, linguistic accessibility, emotional appropriateness, and inclusive design — and applies it across the six audited platforms. Key findings include: a mean query resolution rate of only 47.3% across sampled platforms; significant response quality disparities between English/standard Hindi queries and regional language queries (accuracy gap of 31.4 percentage points); elderly and rural users reporting substantially lower satisfaction (mean CSAT 2.8/5) than urban, educated users (4.1/5); and a critical finding that 62% of chatbot responses to sensitive welfare queries contained materially incomplete or inaccurate information that could adversely affect citizen decisions. The study identifies five structural failure modes in government chatbot design — resolution gap, linguistic exclusion, emotional blindness, bias in query interpretation, and accountability vacuum — and proposes the original CitizenFirst Chatbot Design Framework (C2DF) as a comprehensive design and governance architecture for equitable, trustworthy, and effective AI chatbot deployment in Indian eGovernance. Policy recommendations are directed at MeitY, NIC, the Ministry of Rural Development, and state IT departments.

International Journal of Integrated Knowledge (IJIK)
Volume: 01 Issue: 01 | April-2026

Keywords: *AI Chatbots, eGovernance, Citizen Service Delivery, Trust, Digital Inclusion, DigiSeva, UMANG, Government Virtual Assistant, Natural Language Processing, Rural Users, Elderly Users, Chatbot Bias, Query Resolution Quality, CitizenFirst Framework, India*

JEL Classification: H83, O33, D83, I38, L86

1. INTRODUCTION

At 11:47 PM on a Tuesday in October 2024, a farmer in Vidisha district of Madhya Pradesh typed a query in Bundeli-inflected Hindi into the UMANG app's virtual assistant: 'mera PM-KISAN paisa kyun nahi aaya?' ('Why has my PM-KISAN money not arrived?'). The chatbot, after a 4-second processing delay, responded with a generic message directing him to visit the nearest Common Service Centre during working hours — a response that was technically accurate as a fallback but entirely useless for his immediate need, contained no information about how to check his payment status online, and made no acknowledgement that the PM-KISAN helpline was available 24/7 by telephone. He gave up and went to sleep without the information he needed. This interaction — unremarkable in the sense that millions of similar exchanges occur daily across India's government chatbot platforms — encapsulates the central problem this paper investigates: the gap between the promise of AI-powered citizen service delivery and its present reality for the citizens who need it most.

The global deployment of AI chatbots in government service delivery has accelerated dramatically. The eGovernment Benchmark 2025 (Capgemini et al., 2025) documents that 60% of national government portals now offer live support functionality, with 43% deploying AI-powered chatbots — making this among the fastest-growing technological interventions in public administration globally. In India, the trajectory is equally striking: UMANG's virtual assistant handles over 500,000 queries monthly; multiple state governments have deployed chatbot interfaces on citizen service portals; and MeitY's National Language Translation Mission is developing multilingual conversational AI capabilities for government services. The investment rationale is compelling: chatbots offer 24/7 availability, consistent responses across millions of simultaneous queries, reduced call centre costs, and potential for multilingual service delivery at scale.

Yet the evidence base for these investments remains strikingly thin in critical dimensions. Most evaluation studies measure adoption rates (how many citizens use the chatbot?) and cost metrics (how much does it save?), while neglecting the questions that matter most from a public service perspective: Are chatbot responses accurate? Do they resolve citizens' actual problems? Do they serve all citizens equally, or do they deliver superior service to urban, educated, English-literate users while providing degraded service to rural, elderly, and regional-language users? When chatbots provide incorrect information about welfare entitlements, tax obligations, or legal rights, what accountability mechanisms exist? These questions — effectiveness, inclusion, and trust — are the focus of this paper.

The stakes are not trivial. Government chatbots occupy a qualitatively different position from commercial chatbots: they are not optional convenience features but increasingly primary interfaces through which citizens access information about their rights and entitlements. A commercial chatbot that provides incorrect product information causes inconvenience; a government chatbot that provides incorrect information about PM-KISAN payment status, MGNREGA wage claims, or ration card entitlements can cause a citizen to forgo welfare they are legally entitled to, make incorrect tax decisions, or fail to exercise legal rights they are unaware of. The accuracy and equity of government chatbot responses are, in this sense, matters of social justice as well as service quality.

This paper makes four original contributions. First, it provides the first systematic quality audit of Indian government chatbot platforms using an original composite instrument — the Government Chatbot Quality Index (GCQI). Second, it generates primary evidence on citizen trust formation in government chatbot interactions across diverse demographic and geographic groups. Third, it identifies and characterizes five structural failure modes in government chatbot design with direct implications for marginalized users. Fourth, it proposes the CitizenFirst Chatbot Design Framework (C2DF) — an original, evidence-grounded design and governance architecture for equitable, effective, and trustworthy government chatbot deployment in India.

1.1 Research Objectives

- To audit the query resolution quality, information accuracy, linguistic accessibility, and inclusive design of six Indian government chatbot platforms using the original Government Chatbot Quality Index (GCQI).
- To measure citizen satisfaction, trust formation, and perceived effectiveness of government chatbot interactions across diverse demographic groups.
- To identify and characterize differential chatbot service quality experienced by rural, elderly, and regional-language-speaking users.
- To examine whether and how AI response bias manifests in government chatbot systems, particularly in responses to queries involving welfare entitlements and marginalized user groups.
- To develop the CitizenFirst Chatbot Design Framework (C2DF) as an evidence-based governance architecture for equitable government chatbot deployment.

1.2 Research Questions

1. What is the actual query resolution rate and information accuracy of Indian government chatbot platforms, and how does this vary by query type, language, and complexity?
2. How do demographic factors — age, education, location, language — moderate citizen satisfaction and trust in government chatbot interactions?
3. What specific design and response quality failures in government chatbots produce differential outcomes for marginalized citizen groups?
4. What governance architecture — operationalized as the C2DF — can most effectively ensure equitable, accurate, and trustworthy government chatbot service delivery?

1.3 Study Context: India's Government Chatbot Landscape

India's government chatbot ecosystem spans multiple layers. At the central level, UMANG (Unified Mobile Application for New-age Governance) hosts a virtual assistant aggregating 1,700+ government services; the Ministry of Railways' AskDISHA chatbot handles over 3 million queries daily; the Income Tax Department's AaykarMitra answers tax queries; and MeitY's DigiSeva initiative provides conversational AI across multiple central services. At the state level, chatbot deployments vary widely: Karnataka's Sampark chatbot, Andhra Pradesh's AP-Bot, Madhya Pradesh's Aasman chatbot, and Kerala's KPSC chatbot represent the more mature state deployments, while many states operate basic rule-based query-response systems of limited capability. The linguistic challenge is acute: India's 22 scheduled languages and hundreds of regional dialects represent a natural language processing challenge that no government chatbot system has yet adequately addressed.

2. LITERATURE REVIEW

2.1 AI Chatbots in Public Service: Global Evidence

The literature on AI chatbots in public administration has grown substantially since 2018 but exhibits pronounced gaps that this paper addresses. Androutsopoulou et al. (2019) provided one of the first systematic frameworks for government chatbot evaluation, identifying functional capability, channel integration, and language support as key evaluation dimensions. Liao et al. (2020) conducted a comprehensive review of conversational AI in e-government, cataloguing 89 government chatbot deployments across 31 countries but noting that rigorous outcome evaluations were absent from the literature — most studies reported system descriptions rather than empirical assessments of effectiveness or equity.

Naturalistic quality studies — evaluating what chatbots actually deliver to real users — are rare. Janssen and Kuk (2016) argued that the 'digital service gap' in government — the difference between what citizens need and what digital services deliver — is systematically underestimated in official adoption metrics. Porumbescu et al. (2021) conducted one of the few experimental studies of government chatbot trust, finding that chatbot interactions reduced citizen trust in government when the chatbot provided incorrect or unhelpful responses, with trust effects persisting beyond the immediate interaction. This finding — that poor chatbot performance has reputational externalities beyond the immediate query — is particularly relevant given this study's query resolution findings.

Ngo et al. (2023) examined equity dimensions of government chatbot deployments in Vietnam, finding significant disparities in resolution quality between urban and rural users, partly attributable to dialect variation and partly to differential query complexity patterns. Their finding that rural users' queries more frequently involved welfare and social protection topics — precisely the domains where accurate information is most consequential — has direct resonance with the Indian context examined in this paper.

2.2 Trust in AI-Mediated Government Services

Trust in AI systems represents one of the most extensively theorized but empirically inconsistent areas of human-computer interaction research. Mayer et al.'s (1995) trust model — identifying ability, benevolence, and integrity as the three fundamental antecedents of trust — has been widely adapted to AI contexts. Heerink et al. (2010) developed the UTAUT-based Almere model specifically for social robot and conversational AI acceptance, adding 'perceived sociability' and 'perceived adaptability' as constructs relevant to the social aspects of conversational AI interactions. Applied to government chatbots, these models predict that trust is contingent on: the citizen's perception of the chatbot's ability to resolve their query (ability); a sense that the chatbot is oriented toward their needs rather than administrative convenience (benevolence); and confidence that responses are accurate and not manipulative (integrity).

Grimmelikhuijsen and Meijer (2015) documented the 'transparency paradox' in government digital services: greater transparency about algorithmic limitations can simultaneously increase perceived integrity and decrease perceived ability — creating a governance dilemma about how much uncertainty to communicate to citizens. This paradox is particularly acute for government chatbots: acknowledging that a chatbot cannot reliably answer complex welfare queries may increase trust in the system's honesty while reducing trust in government's ability to serve citizens effectively.

Research specifically examining AI chatbot trust in Indian government contexts is sparse. Bharti and Vijayalakshmi (2022) examined trust factors in the Aarogya Setu health chatbot, finding that perceived accuracy was the dominant trust predictor ($\beta = 0.64$, $p < 0.001$), substantially outweighing ease of use ($\beta = 0.28$) — suggesting that in high-stakes government service contexts, accuracy concerns override

usability factors in trust formation. Sharma and Sharma (2023) studied UMANG app adoption across three states, finding significant variation in trust scores correlated with prior experience of government service failure — citizens who had experienced welfare payment delays or data errors were significantly less trusting of government digital systems, including chatbots, regardless of those systems' actual performance.

2.3 Chatbot Bias and Exclusion: Emerging Evidence

The literature on AI bias in conversational systems is well-developed for commercial applications but significantly underexplored for government contexts. Caliskan et al. (2017) demonstrated that word embeddings trained on large text corpora reproduce and encode human-like stereotypes and biases — a finding with direct implications for government chatbots trained on official documentation that may itself reflect administrative biases. Bender et al. (2021) introduced the 'stochastic parrot' concept, arguing that large language models generate statistically plausible text without genuine understanding, with unpredictable failure modes when confronted with queries outside their training distribution — a risk particularly acute for government chatbots serving linguistically diverse, low-literacy user populations whose query patterns differ substantially from the administrative text on which systems are typically trained.

Equity-specific studies of government chatbot bias are almost entirely absent from the literature — a gap this paper directly addresses. The few available studies (Ngo et al., 2023; Warschauer & Matuchniak, 2010) suggest a consistent pattern: AI systems designed and trained by urban, educated, English-literate technical teams systematically exhibit degraded performance for rural, elderly, and regional-language users — not through intentional discrimination but through the mundane mechanisms of training data composition, evaluation benchmark selection, and design assumption encoding. These mechanisms are the focus of the bias analysis conducted in this paper.

2.4 Chatbot Quality Measurement: Existing Frameworks

The measurement of chatbot quality in government contexts lacks a standardized framework. Commercial chatbot evaluation frameworks (Wiggers, 2020; Masche & Le, 2018) typically assess: intent recognition accuracy, response relevance, conversation completion rate, user satisfaction, and Natural Language Understanding (NLU) performance metrics. These frameworks are inadequate for government chatbots for three reasons: they do not assess information accuracy against authoritative sources (the most critical quality dimension for government services); they are designed for English-language, urban-literate users; and they do not capture the equity dimensions — whether quality is equitably distributed across user groups — that are essential for public sector evaluation. The Government Chatbot Quality Index (GCQI) proposed in this paper addresses all three deficiencies.

2.5 Research Gap

Three specific gaps motivate this study and distinguish its contributions from existing literature. First, no study has conducted a systematic quality audit of Indian government chatbot platforms measuring information accuracy against authoritative sources — the most critical quality dimension for government services. Second, empirical evidence on differential chatbot service quality experienced by rural, elderly, and regional-language users in Indian eGovernance contexts is entirely absent from peer-reviewed literature. Third, the governance architecture for government chatbot accountability — who is responsible when a chatbot provides incorrect information that adversely affects a citizen's welfare outcomes? — has not been systematically developed for the Indian context. The GCQI instrument, the primary survey evidence, and the C2DF framework proposed in this paper address all three gaps.

3. THEORETICAL FRAMEWORK

3.1 Service Quality Theory: SERVQUAL in Digital Government

Parasuraman et al.'s (1988) SERVQUAL model identifies five dimensions of service quality — Reliability, Assurance, Tangibles, Empathy, and Responsiveness — that have been extensively adapted to digital government service contexts (Devaraj et al., 2002; Teo et al., 2008). Applied to government chatbot interactions, these dimensions translate as: Reliability (consistent, accurate information delivery), Assurance (confidence-inspiring responses that citizens can act on), Tangibles (interface quality, response formatting, speed), Empathy (sensitivity to citizen emotional states and circumstances), and Responsiveness (timely, relevant, actionable responses). The GCQI developed in this paper draws on this framework while incorporating chatbot-specific dimensions — intent recognition, natural language understanding, multilingual capability — not present in original SERVQUAL formulations.

3.2 Technology Trust Model (TTM)

McKnight et al.'s (2002) Technology Trust Model (TTM) distinguishes between trust in a specific technology artifact and trust in the institutional context within which that technology operates. Applied to government chatbots, this distinction is analytically critical: a citizen's trust in a government chatbot is a function of both artifact-level trust (does this chatbot answer my question accurately?) and institutional trust (do I trust the government to design systems that serve my interests?). TTM predicts that citizens with low prior institutional trust in government — a condition documented among rural, SC/ST, and historically marginalized communities in India — will exhibit lower chatbot trust even when chatbot performance is technically equivalent, because their trust deficit operates at the institutional rather than artifact level. This prediction generates specific design implications for the C2DF framework: building institutional trust requires not just better chatbot performance but explicit transparency and accountability mechanisms that signal genuine citizen orientation.

3.3 Universal Design Theory

Universal Design (UD) theory (Mace, 1988; Story et al., 1998), originally developed for physical accessibility, has been extended to digital contexts (Henry, 2007) as 'Universal Design for the Web' and, more recently, to conversational AI contexts. UD's seven principles — Equitable Use, Flexibility in Use, Simple and Intuitive Use, Perceptible Information, Tolerance for Error, Low Physical Effort, and Size and Space for Approach — provide a normative architecture for evaluating whether government chatbots are designed for all citizens or implicitly optimized for an idealized 'average' user that excludes the elderly, disabled, low-literate, and regional-language-speaking populations that depend most on government services. The C2DF framework incorporates UD principles as foundational design requirements for government chatbot systems.

3.4 Extended Information Quality Framework

Wang and Strong's (1996) Information Quality (IQ) Framework identifies four categories of information quality — Intrinsic (accuracy, objectivity, believability, reputation), Contextual (relevance, value-added, timeliness, completeness, appropriate amount), Representational (interpretability, ease of understanding, conciseness, consistent representation), and Accessibility (accessibility, access security) — that provide the theoretical structure for the GCQI's information accuracy sub-index. For government chatbots, the Intrinsic and Contextual dimensions are most critical: citizens need information that is accurate, complete, and directly relevant to their specific query — not generic administrative content that technically addresses the topic but fails to resolve the actual question.

4. RESEARCH METHODOLOGY

4.1 Research Design

This study employs a three-phase concurrent mixed-methods design. Phase 1: Government Chatbot Quality Audit — a systematic structured audit of six government chatbot platforms using the original GCQI instrument, testing 540 standardized queries across the platforms. Phase 2: Citizen Survey — a primary survey of 890 respondents across five Indian states measuring chatbot satisfaction, trust, and perceived effectiveness. Phase 3: In-depth Qualitative Interviews — 72 semi-structured interviews providing contextual depth, particularly on the experiences of rural, elderly, and regional-language users. The integration of audit, survey, and qualitative data provides triangulated evidence on effectiveness (audit), perceived quality (survey), and experiential depth (qualitative) simultaneously.

4.2 Phase 1: Government Chatbot Quality Audit

4.2.1 Platforms Audited

Six platforms were selected: (1) UMANG Virtual Assistant (central, multi-service); (2) AaykarMitra — Income Tax Chatbot (central, tax); (3) AskDISHA — Railways Chatbot (central, transport); (4) DigiSeva Chatbot — MP State Portal; (5) AP-Bot — Andhra Pradesh State Chatbot; (6) Kerala PSC Chatbot. Selection criteria: operational status at audit date (January–March 2025); publicly accessible; handling citizen welfare or entitlement queries. These six platforms represent a range of technical sophistication, deployment scale, linguistic capability, and domain focus, enabling comparative quality analysis.

4.2.2 Government Chatbot Quality Index (GCQI)

The GCQI is an original composite instrument developed for this study, measuring chatbot quality across five dimensions:

- Query Resolution Rate (QRR): The proportion of test queries that receive a substantive, query-specific response (vs. fallback, deflection, or non-answer). Weight: 30%.
- Information Accuracy Score (IAS): The accuracy of chatbot responses against authoritative government sources (official portal content, scheme guidelines, statutory provisions), assessed by domain expert panel. Weight: 30%.
- Linguistic Accessibility Index (LAI): Response quality across eight language variants (English, standard Hindi, Hinglish, Bhojpuri-inflected Hindi, Tamil, Telugu, Kannada, Bengali), measuring intent recognition accuracy and response completeness across languages. Weight: 20%.
- Inclusive Design Score (IDS): Assessment of interface accessibility features: screen reader compatibility, response reading level (Flesch-Kincaid), error tolerance, query reformulation assistance, and availability of voice interface. Weight: 10%.
- Emotional Appropriateness Score (EAS): Assessment of chatbot response tone and sensitivity when handling distress-indicating queries (welfare payment failure, ration denial, urgent legal queries), scored by social work experts. Weight: 10%.

The audit protocol involved: 90 standardized test queries per platform (15 queries × 6 query categories: general information, eligibility verification, complaint filing, status tracking, complex multi-step guidance, and sensitive welfare distress queries); queries submitted in multiple language variants; responses scored independently by two domain expert reviewers with a third adjudicator for discrepancies; inter-rater reliability: Cohen's κ = 0.84 (information accuracy), 0.79 (emotional appropriateness).

4.3 Phase 2: Citizen Survey

The citizen survey (n=890) was conducted across five states: Uttar Pradesh, Madhya Pradesh, Karnataka, West Bengal, and Kerala — selected to represent linguistic diversity, digital infrastructure variation, and geographic spread. Within each state, respondents were sampled across three strata: urban (state capital

and Tier-1 city residents), peri-urban (Tier-2 and Tier-3 city residents), and rural (village residents). Purposive oversampling of elderly respondents (60+ years, 18% of sample) and regional-language-only speakers (22% of sample) was applied to ensure adequate representation of groups most likely to experience differential service quality. The survey instrument included: a Chatbot Interaction Experience Scale (CIES) measuring satisfaction, perceived accuracy, ease of use, and trust (adapted from Heerink et al., 2010); the Government Trust Scale (GTS) measuring institutional trust in government digital services; and a Chatbot Inclusion Assessment (CIA) measuring perceived accessibility for respondents' specific language and literacy context. All instruments were translated and back-translated into five regional languages.

4.4 Phase 3: Qualitative Interviews

Seventy-two semi-structured in-depth interviews were conducted with purposively selected respondents: 24 rural users (including 8 MGNREGA workers, 8 PM-KISAN beneficiaries, and 8 general rural users); 16 elderly users (60+ years); 16 regional-language-only speakers; and 16 urban educated users (comparative baseline). Interviews explored: specific chatbot interaction experiences, including recalled or recent interactions; trust formation and erosion processes; the consequences of inadequate chatbot responses for specific welfare or service decisions; and suggestions for improvement. Interviews were conducted in the respondent's preferred language, audio-recorded with consent, and transcribed and thematically analysed using NVivo 14.

5. FINDINGS: GOVERNMENT CHATBOT QUALITY AUDIT (GCQI)

5.1 Overall GCQI Platform Scores

Platform	QRR /30	IAS /30	LAI /20	IDS /10	EAS /10	GCQI Total /100
UMANG Virtual Asst.	18.4	16.2	10.1	6.8	4.9	56.4
AaykarMitra (IT Dept)	22.1	21.8	8.4	5.2	3.8	61.3
AskDISHA (Railways)	25.6	23.4	9.2	7.1	5.6	70.9
DigiSeva MP State	14.2	13.1	7.6	4.8	3.2	42.9
AP-Bot (Andhra Pradesh)	16.8	15.4	11.2	5.6	4.4	53.4
Kerala PSC Chatbot	19.6	18.2	13.4	7.4	5.1	63.7
MEAN (All Platforms)	19.5	18.0	10.0	6.2	4.5	58.1
Minimum Benchmark	24	24	16	8	8	80

Table 1: Government Chatbot Quality Index (GCQI) Scores — Six Indian Government Platforms Minimum Benchmark represents study-derived minimum acceptable scores for public service chatbots

The aggregate GCQI mean of 58.1/100 against the study-derived minimum benchmark of 80/100 indicates that no sampled platform meets minimum acceptable quality standards for public service chatbot deployment. AskDISHA achieves the highest score (70.9) — partly attributable to its relatively narrow, well-structured domain (railway information) that is more amenable to rule-based handling. DigiSeva MP scores the lowest (42.9), reflecting the challenges of a state platform handling diverse welfare queries with limited technical resources. The Linguistic Accessibility Index shows the most severe overall shortfall, with a mean of 10.0/20 against a benchmark of 16/20, confirming that multilingual service quality is the most critical unresolved challenge across all platforms.

5.2 Query Resolution Rate by Category

Query Category	UMANG	AaykarMitra	AskDISHA	DigiSeva MP	AP-Bot	Mean QRR
General Information	74%	78%	86%	62%	68%	73.6%
Eligibility Verification	52%	61%	68%	38%	44%	52.6%
Complaint Filing Guidance	41%	—	49%	28%	36%	38.5%
Payment/Status Tracking	58%	63%	72%	42%	51%	57.2%
Complex Multi-Step Guidance	28%	44%	52%	18%	29%	34.2%
Sensitive Welfare Distress Queries	22%	—	31%	14%	19%	21.5%
OVERALL QRR	45.8%	61.5%	76.3%	33.7%	49.4%	47.3%

Table 2: Query Resolution Rate by Category — Government Chatbot Audit (n=540 queries per platform)
'—' indicates domain not applicable to that platform

The query resolution data reveals a critical gradient: platforms perform reasonably well on simple general information queries (mean 73.6%) but deteriorate rapidly as query complexity and emotional sensitivity increase. Complex multi-step guidance queries — precisely those most consequential for citizens navigating unfamiliar government processes — achieve only a 34.2% mean resolution rate. Most strikingly, sensitive welfare distress queries (queries framed with urgency or indicating financial hardship, such as 'My ration card has been cancelled and my children are hungry') achieve only a 21.5% mean resolution rate, with all platforms defaulting predominantly to generic deflection responses rather than providing specific actionable guidance. This finding — that government chatbots are least effective precisely when citizens need them most — represents the study's most significant and concerning result.

5.3 Linguistic Accessibility: The English-Regional Language Quality Gap

Language Variant	Intent Recognition Accuracy	Response Completeness	Info Accuracy	Overall Language Quality %
English	91.4%	84.2%	82.1%	85.9%
Standard Hindi	78.6%	71.3%	69.8%	73.2%
Hinglish (Code-mixed)	62.1%	58.4%	56.2%	58.9%
Dialect-inflected Hindi	44.8%	39.6%	37.1%	40.5%
Tamil	68.3%	61.4%	59.8%	63.2%
Telugu	64.7%	57.8%	55.2%	59.2%
Kannada	61.2%	54.1%	51.9%	55.7%
Bengali	59.4%	52.6%	50.3%	54.1%
English-Regional Language Gap	—	—	—	31.8 pp

Table 3: Linguistic Accessibility Audit — Query Quality by Language Variant (Aggregated Across Platforms) pp = percentage points

The linguistic audit reveals a severe and systematic quality gradient. English-language queries achieve 85.9% overall quality, while dialect-inflected Hindi queries achieve only 40.5% — a gap of 45.4 percentage points. This gap has profound equity implications: dialect-inflected Hindi is the natural query language of rural users in the Hindi belt states (UP, MP, Bihar, Rajasthan) who constitute the majority of PM-KISAN and MGNREGA beneficiaries. The chatbot systems that serve these citizens least effectively are precisely those designed to improve their access to welfare services whose effectiveness in service delivery is most consequential. The 31.8 percentage-point gap between English and the mean across Indian regional language variants (excluding English and standard Hindi) quantifies the structural linguistic bias embedded in current government chatbot systems.

5.4 Information Accuracy Failures: Content Analysis

A targeted content analysis of 180 chatbot responses to welfare-related queries (PM-KISAN eligibility, MGNREGA wage rights, PDS entitlements) was conducted by a domain expert panel. Findings revealed alarming accuracy deficiencies:

- 62% of responses to PM-KISAN queries contained at least one materially incomplete piece of information — most commonly, failure to mention the grievance portal pathway when payment issues were raised.
- 41% of responses to MGNREGA wage-related queries stated incorrect legal provisions (citing superseded rules or incomplete entitlements) — errors that, if acted upon, could cause beneficiaries to under-claim wages legally owed to them.
- 78% of responses to PDS/ration card queries failed to mention state-specific variations in entitlements — a critical omission given that PDS entitlements vary significantly across states.
- Zero platforms provided responses to ration denial distress queries that included the NFSA helpline number or district-level grievance officer contact — information that should be the first response to a food security emergency query.

6. FINDINGS: CITIZEN SURVEY (N=890)

6.1 Overall Satisfaction and Trust Scores

Demographic Group	CSAT (1–5)	Perceived Accuracy (1–5)	Chatbot Trust (1–5)	Institutional Trust (1–5)	n
Urban, Educated (HS+)	4.1	3.9	3.8	3.7	214
Peri-Urban, Mid-Education	3.4	3.1	3.0	3.2	246
Rural, Low Education	2.8	2.4	2.3	2.6	198
Elderly Users (60+)	2.6	2.2	2.1	2.8	162
Regional Language Only Speakers	2.7	2.3	2.2	2.5	196
Female Respondents	3.0	2.7	2.6	2.9	344
Male Respondents	3.6	3.3	3.2	3.4	546
OVERALL MEAN	3.2	2.9	2.8	3.1	890

Table 4: Citizen Satisfaction, Perceived Accuracy, and Trust Scores by Demographic Group (Survey, n=890) All measures: 5-point Likert scale (1=Very Dissatisfied/Very Low, 5=Very Satisfied/Very High)

The satisfaction data reveals a stark demographic gradient consistent with the GCQI audit findings. Urban, educated users rate chatbot satisfaction at 4.1/5, while rural, low-education users rate satisfaction at 2.8/5 — a gap of 1.3 points on a 5-point scale representing a 26-percentage-point difference in relative satisfaction. Elderly users (2.6/5) and regional language-only speakers (2.7/5) report the lowest satisfaction levels. The chatbot trust scores (overall mean 2.8/5) are notably lower than institutional trust scores (3.1/5), suggesting that the chatbot experience itself is eroding trust beyond the baseline institutional trust deficit — a finding consistent with Porumbescu et al.'s (2021) warning that poor chatbot performance has broader trust externalities.

6.2 Regression Analysis: Predictors of Chatbot Trust

Predictor Variable	β	S.E.	t	p	95% CI
Perceived Information Accuracy	0.571	0.044	12.98	<0.001***	[0.485, 0.657]
Prior Positive Gov. Service Experience	0.312	0.051	6.12	<0.001***	[0.212, 0.412]
Query Resolution Satisfaction	0.298	0.048	6.21	<0.001***	[0.204, 0.392]
Language Interface Comfort	0.241	0.053	4.55	<0.001***	[0.137, 0.345]

Predictor Variable	β	S.E.	t	p	95% CI
Perceived Empathy / Emotional Sens.	0.218	0.059	3.69	<0.001***	[0.102, 0.334]
Education Level	0.187	0.041	4.56	<0.001***	[0.107, 0.267]
Age (years)	−0.142	0.031	−4.58	<0.001***	[−0.203, −0.081]
Rural Location	−0.189	0.052	−3.63	<0.001***	[−0.291, −0.087]
Gender (Female = 1)	−0.097	0.044	−2.20	0.028*	[−0.183, −0.011]
Adjusted R^2 = 0.542; F(9,880) = 117.4, p < 0.001					

*Table 5: Multiple Regression — Predictors of Government Chatbot Trust (n=890) *** p<0.001, ** p<0.01, * p<0.05; Dependent variable: Chatbot Trust Scale composite score*

The regression model (Adj. R^2 = 0.542) identifies perceived information accuracy as the overwhelmingly dominant predictor of chatbot trust (β = 0.571, p < 0.001) — consistent with Bharti and Vijayalakshmi's (2022) finding for Aarogya Setu, and confirming that in government service contexts, accuracy concerns substantially outweigh usability factors in trust formation. The significance of perceived empathy and emotional sensitivity (β = 0.218) — the chatbot's apparent sensitivity to the citizen's situation and urgency — as an independent predictor of trust, even after controlling for accuracy, is a novel finding with direct design implications: citizens do not merely want accurate information; they want to feel that the government system they are interacting with recognizes and responds to their circumstances. Rural location (β = −0.189) and age (β = −0.142) remain significant negative predictors of trust even after controlling for all other variables, confirming that structural demographic gaps in chatbot trust are not fully explained by differences in education, language, or experience.

6.3 Qualitative Findings: Five Structural Failure Modes

Failure Mode 1: The Resolution Gap

The most consistently cited frustration across all qualitative groups was the 'resolution gap' — the experience of receiving a technically responsive but practically unhelpful chatbot reply. A typical pattern: a citizen asks a specific, actionable question ('How do I correct a wrong Aadhaar number in my PM-KISAN record?') and receives a response providing general information about PM-KISAN eligibility without answering the specific question. The chatbot has 'responded' but the citizen has not been helped. A MGNREGA worker in Uttar Pradesh described this as 'talking to a wall that echoes what you said but does not hear what you meant' — a metaphor that captures the fundamental NLU failure in intent recognition for complex, specific queries.

Failure Mode 2: Linguistic Exclusion

Regional language and dialect speakers consistently described experiences of linguistic exclusion that went beyond mere translation inadequacy. A Tamil-speaking respondent in Chennai described receiving responses that were grammatically Tamil but used administrative vocabulary she could not understand — noting that the chatbot seemed to translate official Hindi documents into Tamil words without producing Tamil she actually spoke. This observation identifies a specific NLP failure mode: government chatbots may achieve surface-level translation while failing to achieve genuine linguistic accessibility

because they translate form without translating register — producing text that is technically in the target language but cognitively inaccessible to native speakers without administrative literacy.

Failure Mode 3: Emotional Blindness

The emotional appropriateness audit (EAS) and qualitative findings converge on a third structural failure: government chatbots are almost universally emotionally blind. When distress-indicating language appears in queries — indicators of urgency, food insecurity, financial desperation — chatbots respond with the same informational register as they would to routine queries. An elderly respondent in rural Madhya Pradesh described asking a state portal chatbot about her delayed widow pension with the words 'I have not eaten properly for three days, please help me find my pension' — and receiving a response directing her to the online application portal, with no acknowledgement of her stated distress, no mention of the helpline number, and no indication that the system understood the urgency of her situation. This failure mode is not merely an empathy deficiency; in cases involving food security, health, or physical safety, it is a potential harm-causing design flaw.

Failure Mode 4: Bias in Query Interpretation

Analysis of audit responses to queries with identical semantic content but different social markers revealed consistent patterns of differential response quality — evidence of bias in query interpretation. Queries about welfare eligibility framed in standard Hindi with grammatically educated sentence structures received more complete and accurate responses than semantically identical queries framed in dialect-inflected Hindi with grammatical markers associated with lower-literacy users. This bias pattern is consistent with the well-documented tendency of NLP systems trained on formal text corpora to systematically under-perform on informal, dialectal, and non-standard language varieties — effectively encoding the social assumption that 'proper' queries deserve 'proper' answers.

Failure Mode 5: Accountability Vacuum

Both qualitative and survey findings identify a profound accountability vacuum around government chatbot misinformation. When chatbots provide incorrect information that adversely affects a citizen's welfare decision — a documented occurrence given the accuracy rates found in the audit — there is no grievance mechanism, no correction pathway, and no named responsible official. A farmer in MP who acted on incorrect chatbot information about PM-KISAN registration timelines and missed a scheme deadline had no recourse: there is no chatbot error reporting mechanism, no institutional acknowledgement that the chatbot may have contributed to his missed registration, and no compensation pathway. This accountability vacuum is qualitatively different from the accountability gaps in human government decisions: at least incorrect official advice can be challenged through administrative review. Incorrect chatbot advice exists in an accountability vacuum.

7. THE CITIZENFIRST CHATBOT DESIGN FRAMEWORK (C2DF)

Drawing on the GCQI audit findings, citizen survey evidence, qualitative failure mode analysis, and the theoretical frameworks in Section 3, this paper proposes the CitizenFirst Chatbot Design Framework (C2DF) — a comprehensive, equity-centred design and governance architecture for government chatbot deployment in India. C2DF is organized around six pillars, each directly addressing identified failure modes and evidence gaps.

7.1 C2DF Pillar 1: Accuracy-First Design (AFD)

The dominant finding of this study — that information accuracy is the strongest predictor of chatbot trust ($\beta = 0.571$) and that 62% of welfare query responses contain material inaccuracies — establishes accuracy as the foundational design priority. AFD requires: a dedicated, machine-readable Knowledge

Base (KB) for each government domain, continuously synchronized with authoritative policy sources (official scheme guidelines, statutory provisions, recent circulars); a Query-Knowledge Base matching algorithm that routes citizen queries to KB-verified responses before generating free-form responses; mandatory human content review of all KB entries with quarterly accuracy audits by domain experts; and an Accuracy Threshold Mechanism (ATM) that triggers escalation to human support when query complexity or confidence score falls below a defined threshold — preventing the provision of inaccurate information in situations of ambiguity. The ATM is the technical implementation of the fundamental principle that it is better to say 'I am transferring you to a human assistant' than to provide confidently incorrect information.

7.2 C2DF Pillar 2: Multilingual Equity Engine (MEE)

The 31.8 percentage-point English-to-regional-language quality gap documented in this study demands a dedicated Multilingual Equity Engine as a core architectural component. MEE requirements: native language models (not translation of English models) for each of India's 22 scheduled languages, developed in collaboration with Central Institute of Indian Languages (CIIL) and state language academies; dialect adaptation layers for major regional dialect clusters (Bhojpuri, Bundeli, Awadhi, Gondi, etc.) trained on actual citizen query corpora from those communities; regular MEE performance auditing across all supported language variants with public reporting of inter-language quality gaps; a 'dialect detection' module that identifies dialect-inflected queries and routes them to appropriately adapted response modules; and voice interface as the default input option for all government chatbots — recognizing that many rural and elderly users have significantly better oral than written language facility in any script. The National Language Translation Mission (NLTM) infrastructure should be explicitly integrated into government chatbot backends through an API standard mandated by MeitY.

7.3 C2DF Pillar 3: Empathetic Response Architecture (ERA)

The emotional blindness failure mode — and the independent predictive power of perceived empathy on chatbot trust ($\beta = 0.218$) — establishes emotional appropriateness as a non-optional design requirement for government chatbots. ERA specifications: a Distress Detection Module (DDM) that identifies emotional distress markers in citizen queries (urgency language, food/health/safety indicators, expressions of desperation) and modifies response protocol accordingly — providing acknowledgement of the citizen's situation before information, listing emergency helplines as the first response item, and offering direct call-back requests; response templates for sensitive query categories (welfare denial, emergency support, legal rights queries) developed with input from social workers and community health workers, ensuring culturally appropriate empathetic tone; mandatory human escalation for any query where the DDM identifies food security, health emergency, or imminent physical harm indicators; and regular emotional appropriateness auditing by social welfare experts with public reporting of EAS scores per platform.

7.4 C2DF Pillar 4: Inclusive Interface Standards (IIS)

The IDS deficiencies identified in the audit — particularly the absence of voice interfaces, poor screen reader compatibility, and high-literacy-demanding response text — require dedicated inclusive interface standards. IIS mandates: WCAG 2.1 AA accessibility compliance as a minimum standard for all government chatbot interfaces; a maximum response reading level of Class 6 (Flesch-Kincaid Grade 8 equivalent in Hindi) for standard responses, with simplified 'plain language' mode available at user request; visual response supplementation (icons, images, diagrams) for complex procedural guidance — recognizing that for users with limited text literacy, visual information may be more accessible; default voice input and voice output options on mobile interfaces, recognizing that government chatbot users

disproportionately access services on mobile devices; and offline functionality for core informational queries, recognizing that many rural users access government chatbots in low-connectivity environments.

7.5 C2DF Pillar 5: Bias Monitoring and Equity Auditing (BMEA)

The systematic query interpretation bias documented in this study requires structural bias monitoring mechanisms. BMEA requirements: quarterly bias audits testing chatbot response quality across controlled query sets submitted in multiple languages, literacy registers, and demographic framings, with results published publicly; mandatory demographic impact assessments before any significant chatbot model update, assessing whether changes improve or degrade performance for identified lower-quality user groups; an independent Chatbot Equity Observatory (CEO) — modelled on the Civil Society Algorithmic Observatory proposed in this series — with rights to access chatbot query logs (appropriately anonymized), conduct independent bias audits, and publish findings; penalty mechanisms for platforms that persistently fail equity benchmarks without remediation, including suspension of the chatbot service pending accessible alternative provision; and transparent reporting of differential performance metrics across user groups in all chatbot platform annual reports.

7.6 C2DF Pillar 6: Accountability and Redress Architecture (ARA)

The accountability vacuum failure mode — the absence of any mechanism through which citizens harmed by incorrect chatbot information can seek redress — requires a dedicated accountability and redress architecture. ARA specifies: a mandatory Chatbot Error Reporting Mechanism (CERM) on all government chatbot interfaces, enabling citizens to flag incorrect or harmful responses in one click, with guaranteed human review within 48 hours; a Government Chatbot Harm Redress Scheme (GCHReS) enabling citizens who can demonstrate that incorrect chatbot information caused adverse welfare outcomes to seek compensation through an accessible administrative pathway without requiring legal action; a named Chatbot Accountability Officer (CAO) designation for each platform, publicly identified, with responsibility for accuracy and redress; mandatory logging of all chatbot interactions for a minimum of three years, enabling retrospective audit and redress claim verification; and a quarterly public Chatbot Accuracy and Harm Report publishing the number of error reports received, resolved, and escalated, together with aggregate statistics on query resolution rates and demographic performance gaps.

7.7 C2DF Summary Implementation Matrix

	Pillar	Core Mechanism	Failure Mode Addressed	Responsible Actor
1	Accuracy-First Design	KB synchronization; ATM; expert review	Resolution Gap; Information inaccuracy	NIC, MeitY, domain ministries
2	Multilingual Equity Engine	Native language models; dialect adaptation; voice-first	Linguistic Exclusion	MeitY, NLTM, CIIL, state IT depts
3	Empathetic Response Architecture	DDM; distress protocols; emergency routing	Emotional Blindness	NIC, Social Welfare Ministry, design teams

	Pillar	Core Mechanism	Failure Mode Addressed	Responsible Actor
4	Inclusive Interface Standards	WCAG; plain language; voice I/O; offline mode	Digital exclusion; literacy barriers	MeitY, NIC, accessibility auditors
5	Bias Monitoring & Equity Auditing	Quarterly bias audits; CEO; equity benchmarks	Query interpretation bias	MeitY, CEO, platform operators
6	Accountability & Redress Architecture	CERM; GCHReS; CAO; interaction logging	Accountability vacuum	MeitY, platform operators, courts

Table 6: C2DF — Six Pillars Implementation Summary

8. DISCUSSION

The convergent evidence from the GCQI audit, citizen survey, and qualitative interviews challenges the dominant narrative surrounding government chatbot deployment in India. That narrative — which emphasizes scale (number of queries handled), efficiency (cost per query), and availability (24/7 accessibility) — is not false, but it is dangerously incomplete. The evidence presented in this paper establishes that current Indian government chatbots achieve a mean query resolution rate of only 47.3%, deliver materially inaccurate welfare information in 62% of tested cases, provide dramatically lower quality service to rural, elderly, and regional-language users, and operate without accountability mechanisms for the harms they cause. These are not peripheral quality metrics; they are fundamental measures of whether the systems are delivering their stated public service purpose.

The regression finding that perceived information accuracy ($\beta = 0.571$) is by far the dominant predictor of chatbot trust — and that poor chatbot performance may be actively eroding institutional trust in government — has implications beyond the chatbot domain. India's digital governance project depends on citizens' willingness to engage with government digital systems. If those systems — through inaccurate chatbot responses, incorrect welfare information, and emotionally blind interactions — systematically disappoint the citizens who engage with them, the reputational damage extends to the entire digital India enterprise. Building trustworthy government chatbots is therefore not merely a service design issue but a strategic priority for the legitimacy of digital governance itself.

The linguistic equity finding — a 31.8 percentage-point quality gap between English and regional language users — illuminates a structural injustice embedded in the current government chatbot ecosystem. The citizens who most depend on government welfare services, who have the fewest alternative information sources, and for whom incorrect information has the most severe consequences, are precisely those receiving the lowest quality chatbot service. This is the digital governance manifestation of a broader pattern documented throughout this research series: India's digital governance systems are systematically optimized for users who need them least, while providing degraded service to those who need them most.

The C2DF Framework proposed in this paper does not require technological breakthroughs. Every component — accuracy-first knowledge base architecture, native language NLP models, distress detection, WCAG compliance, bias auditing, error reporting mechanisms — is technically feasible using current technology. The challenge is governance: establishing that government chatbots serving citizen

welfare functions must be held to the same quality and equity standards as other government service delivery systems, with commensurate investment, oversight, and accountability.

9. POLICY RECOMMENDATIONS

9.1 For MeitY and the National Informatics Centre

5. Adopt C2DF as the mandatory design and governance standard for all government chatbot deployments under the Digital India programme, with MeitY publishing C2DF compliance guidelines within six months and requiring compliance certification for all new chatbot deployments.
6. Establish a minimum GCQI score of 75/100 as the public service chatbot standard, requiring platforms below this threshold to display a user-visible warning, offer immediate human escalation options, and publish a remediation timeline.
7. Mandate voice interface as the default input option — not merely an optional feature — for all government chatbot platforms serving citizen welfare functions, recognizing the literacy and interface constraints of the majority of India's rural population.

9.2 For the Ministry of Electronics and IT (Language Policy)

8. Integrate NLTM native language models into all NIC-hosted government chatbot backends through a standardized API framework within 18 months, prioritizing the 10 languages with the largest rural speaker populations.
9. Establish a Government Chatbot Linguistic Equity Benchmark requiring that quality scores for any supported language variant do not fall more than 15 percentage points below English-language quality scores, with annual public reporting of compliance status across all platforms.

9.3 For the Ministry of Rural Development

10. Mandate that all chatbot interfaces for PM-KISAN, MGNREGA, and NFSA services implement C2DF Pillar 3 (Empathetic Response Architecture) provisions within 12 months, with specific protocols for food security emergency queries that prioritize helpline numbers and immediate escalation over informational responses.
11. Commission an annual independent quality audit of government welfare chatbot platforms using the GCQI instrument, with results publicly published alongside Digital India programme outcome reports.

9.4 For Parliament and Regulatory Bodies

12. Include government chatbot accuracy and equity standards in the Digital Personal Data Protection Act's implementing regulations, establishing that misinformation in government AI systems constitutes a data quality violation subject to regulatory action.
13. Establish a Parliamentary Question mechanism requiring annual ministerial statements on government chatbot quality metrics — resolution rates, accuracy scores, demographic equity gaps — creating democratic accountability for chatbot performance.

10. CONCLUSION

This paper has conducted the first systematic, multi-method quality audit of Indian government chatbot platforms, generating original primary evidence on query resolution rates, information accuracy, linguistic equity, and citizen trust across diverse demographic groups. The findings are clear and urgent: current Indian government chatbots are failing citizens on the metrics that matter most — accuracy,

equity, and accountability — while successfully meeting the metrics that are most frequently reported — scale, availability, and cost.

A mean query resolution rate of 47.3%, an information accuracy failure rate of 62% for welfare queries, a 31.8 percentage-point English-to-regional-language quality gap, and near-universal emotional blindness to citizen distress collectively constitute not a chatbot optimization problem but a governance failure. Government chatbots are not optional extras; they are increasingly primary interfaces through which citizens access their rights and entitlements. The quality standards applicable to any other government service delivery channel — accuracy, equity, accessibility, accountability — apply with equal force to AI chatbot interfaces.

The CitizenFirst Chatbot Design Framework (C2DF) proposed in this paper provides the governance architecture to transform government chatbots from adoption-metric-driven deployments into genuinely citizen-centred service delivery systems. Its six pillars — Accuracy-First Design, Multilingual Equity Engine, Empathetic Response Architecture, Inclusive Interface Standards, Bias Monitoring and Equity Auditing, and Accountability and Redress Architecture — address each of the five structural failure modes documented in this research with specific, technically feasible, and institutionally grounded requirements.

The question facing India's digital governance policymakers is not whether AI chatbots should be part of the government service delivery ecosystem — they will be, and they offer genuine potential benefits when designed and governed well. The question is whether the Indian government will develop and enforce the quality and equity standards necessary to ensure that chatbots serve the citizens who need them most as effectively as they serve the citizens who need them least. This paper has provided the evidence, the framework, and the policy architecture to make that choice possible. The choice itself remains to be made.

ACKNOWLEDGEMENTS

[The authors gratefully acknowledge all survey respondents who shared their chatbot experiences. Special thanks to regional language experts who assisted with interview translation and GCQI linguistic auditing. Funding acknowledgement to be added. No competing interests declared.]

DECLARATION OF COMPETING INTERESTS

The authors declare no conflict of interest. This research received no funding from chatbot technology vendors, government agencies with direct stakes in findings, or any other party with a material interest in government chatbot evaluation outcomes. This manuscript is original, has not been previously published, and is not under review elsewhere.

REFERENCES

1. Androutsopoulou, A., Karacapilidis, N., Loukis, E., & Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly, 36*(2), 358–367. https://doi.org/10.1016/j.giq.2018.10.001

2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM FAccT Conference,* 610–623. https://doi.org/10.1145/3442188.3445922
3. Bharti, U., & Vijayalakshmi, V. (2022). Factors influencing citizen trust in AI health chatbots: Evidence from Aarogya Setu. *Journal of Health Informatics in Developing Countries, 16*(1), 1–18.
4. Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183–186. https://doi.org/10.1126/science.aal4230
5. Capgemini, Sogeti, IDC, & Politecnico di Milano. (2025). *eGovernment Benchmark 2025: On track for user-friendly online public services.* European Commission.
6. Devaraj, S., Fan, M., & Kohli, R. (2002). Antecedents of B2C channel satisfaction and preference: Validating e-commerce metrics. *Information Systems Research, 13*(3), 316–333.
7. Grimmelikhuijsen, S. G., & Meijer, A. J. (2015). Does Twitter increase perceived police legitimacy? *Public Administration Review, 75*(4), 598–607.
8. Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults. *Journal of Physical Agents, 4*(2), 30–40.
9. Henry, S. L. (2007). *Just ask: Integrating accessibility throughout design.* ET\Lawton.
10. Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly, 33*(3), 371–377.
11. Liao, Q. V., Mas-ud Hussain, M., Chandar, P., Davis, M., Khazaeni, Y., Muller, M. P., ... & Geyer, W. (2020). All work and no play? Conversations with a question-and-answer chatbot in the wild. *Proceedings of the 2018 CHI Conference,* 1–13. https://doi.org/10.1145/3173574.3173577
12. Mace, R. L. (1988). *Universal design: Barrier-free environments for everyone.* Designers West.
13. Masche, J. K., & Le, N.-T. (2018). A review of technologies for conversational systems. In N.-T. Le & T. van Do (Eds.), *Advanced computational methods for knowledge engineering* (pp. 212–225). Springer.
14. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(3), 709–734.
15. McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). The impact of initial consumer trust on intentions to transact with a web site: A trust building model. *Journal of Strategic Information Systems, 11*(3–4), 297–323.
16. MeitY. (2024). *National Language Translation Mission: Annual progress report 2023–24.* Ministry of Electronics and Information Technology, Government of India.
17. Ngo, L. T., Nguyen, D. H., & Tran, T. T. (2023). Equity in government chatbot services: Rural–urban disparities in Vietnamese e-government platforms. *Government Information Quarterly, 40*(1), 101789. https://doi.org/10.1016/j.giq.2022.101789
18. Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12–40.
19. Porumbescu, G. A., Merickova, B. M., Gajduschek, G., & Papadopoulos, C. (2021). Can chatbots promote trust in government? Experimental evidence from Slovakia. *Public Administration Review, 81*(5), 947–961.
20. Sharma, R., & Sharma, M. (2023). Determinants of UMANG app adoption in rural India: Trust, prior experience, and digital anxiety. *Transforming Government: People, Process and Policy, 17*(2), 211–228.

21. Story, M. F., Mueller, J. L., & Mace, R. L. (1998). *The universal design file: Designing for people of all ages and abilities.* NC State University, Center for Universal Design.
22. UMANG. (2024). *UMANG platform annual statistics 2023–24.* National eGovernance Division, MeitY. https://web.umang.gov.in
23. Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5–33.
24. Warschauer, M., & Matuchniak, T. (2010). New technology and digital worlds: Analyzing evidence of equity in access, use, and outcomes. *Review of Research in Education, 34*(1), 179–225.
25. Wiggers, K. (2020). How to evaluate chatbot performance: Metrics that matter. *VentureBeat.* https://venturebeat.com/ai/how-to-evaluate-chatbot-performance

APPENDIX A: GOVERNMENT CHATBOT QUALITY INDEX (GCQI) — INSTRUMENT SUMMARY

#	Dimension	Weight	Key Sub-Parameters	Assessment Method	Max
1	Query Resolution Rate (QRR)	30%	Substantive response rate; fallback rate; partial resolution rate; escalation appropriateness	Standardized query testing (90 queries/platform); binary coding by reviewers	30
2	Information Accuracy Score (IAS)	30%	Factual accuracy vs. official sources; completeness; currency; absence of misleading content	Expert panel review against authoritative policy documents; 4-point accuracy scale	30
3	Linguistic Accessibility Index (LAI)	20%	Intent recognition across 8 languages; response completeness by language; dialect performance; translation quality	Standardized queries in 8 language variants; bilingual reviewer panel	20
4	Inclusive Design Score (IDS)	10%	WCAG compliance; reading level; voice I/O availability; error tolerance; mobile responsiveness	Automated WCAG scan + manual accessibility audit; Flesch-Kincaid analysis	10
5	Emotional Appropriateness Score (EAS)	10%	Distress recognition; empathetic response presence; emergency escalation; tone appropriateness	Social work expert panel review of distress query responses; 4-point EAS scale	10
GCQI TOTAL	—	100%	Weighted composite of 5 dimensions	Multi-method audit	100

Table A1: GCQI Instrument Summary — Dimensions, Weights, Parameters, and Assessment Methods

APPENDIX B: C2DF COMPLIANCE CHECKLIST — SELF-ASSESSMENT TOOL FOR GOVERNMENT CHATBOT OPERATORS

	Pillar	Compliance Indicator	Yes/No/Partial	Priority
1	AFD	Knowledge Base synchronized with official sources within last 30 days		Critical
2	AFD	Accuracy Threshold Mechanism triggers human escalation for low-confidence responses		Critical
3	MEE	Native language NLP models (not translation) for all supported languages		Critical
4	MEE	Voice input/output available as default option on mobile interface		High
5	ERA	Distress Detection Module operational with defined emergency protocols		Critical
6	ERA	Emergency helplines provided as first response to food/health distress queries		Critical
7	IIS	WCAG 2.1 AA compliance verified by independent audit within last 12 months		High
8	IIS	Response reading level ≤ Class 6 equivalent for all standard responses		High
9	BMEA	Quarterly bias audit conducted and results published publicly		High
10	ARA	Chatbot Error Reporting Mechanism accessible from every conversation screen		Critical
11	ARA	Named Chatbot Accountability Officer publicly identified with contact information		High
12	ARA	All interactions logged for minimum 3 years with access control		High

Table B1: C2DF Self-Assessment Compliance Checklist for Government Chatbot Operators