frontiers | Frontiers in Political Science

TYPE Methods
PUBLISHED 01 October 2025
DOI 10.3389/fpos.2025.1601440

OPEN ACCESS

EDITED BY
Alberto Asquer,
SOAS University of London, United Kingdom

REVIEWED BY
Mosab Alrashed,
American International University, Kuwait
Balázs Hohmann,
University of Pécs, Hungary

*CORRESPONDENCE
Theodoros Papadopoulos
✉ t.papadopoulos@aegean.gr

RECEIVED 27 March 2025
ACCEPTED 04 September 2025
PUBLISHED 01 October 2025

CITATION
Papadopoulos T, Alexopoulos C and Charalabidis Y (2025) Evaluating chatbot architectures for public service delivery: balancing functionality, safety, ethics, and adaptability. *Front. Polit. Sci.* 7:1601440. doi: 10.3389/fpos.2025.1601440

Evaluating chatbot architectures for public service delivery: balancing functionality, safety, ethics, and adaptability

Theodoros Papadopoulos*, Charalampos Alexopoulos and Yannis Charalabidis

Department of Information and Communication Systems Engineering, School of Engineering, University of the Aegean, Samos, Greece

The increasing integration of AI-driven interfaces into public service channels has catalyzed a vibrant discourse on the interplay between technological innovation and the traditional values of public governance. This discussion invites a critical exploration of how emerging chatbot architectures can be aligned with ethical principles and resilient public sector practices. While there is research assessing the potential benefits of integrating chatbots in service delivery, existing evaluation approaches often lack specificity to the unique context of public administration, failing to adequately balance technical performance with crucial ethical considerations, safety requirements, and core public service principles like transparency, fairness, and accountability. This research addresses this critical gap by developing and applying *a structured evaluation framework specifically designed for assessing diverse chatbot architectures within the public sector*. The methodology offers actionable insights to guide the selection and implementation of chatbot solutions that enhance citizen engagement, streamline government services, and uphold key public service values. A key contribution is the introduction of *fifteen pre*-assessed evaluation criteria, encompassing areas such as input understanding, error handling, legal compliance, safety, and personalization, which are applied to *four* distinct chatbot architectures. Our findings indicate that while no single architecture is universally optimal, hybrid retrieval-augmented generation (RAG) systems emerge as the most balanced approach, effectively mitigating the risks of pure generative models while retaining their adaptability. Ultimately, this work provides actionable guidance for policymakers and researchers, supporting informed decisions on the responsible use of chatbots and emphasizing the critical balance between innovation and public trust.

KEYWORDS

politics of technology, chatbots, public service delivery, evaluation framework, AI ethics, LLMs

1 Introduction

In recent decades, digitization and automation are promoted as key drivers in the long-established efforts to streamline and automate public service delivery, that had already begun in previous decades by the New Public Management dictates. The intensity of these efforts has intensified and reinforced by a wide combination of factors, among which, technological advancement, the wide adoption of digital interfaces in market and commerce, the latest economic and health crises, and recently, the growing uptake and normalization of the applications of artificial intelligence (AI).

Chatbot systems that had already found their way in the commercial sector and markets as a means for enhancing customer service and reducing costs, experienced an explosive increase in their adoption with the emergence of generative AI, soon finding their way to the public sector. These systems transform how citizens interact with public administrations, providing constant support, automating the response to citizens' queries or providing interactive interfaces for online service delivery. By facilitating communication and improving access to services, chatbots are bridging the gap between citizens and the state, allowing for more responsive and efficient public services. However, the underlying technologies supporting these chatbots vary significantly, affecting their performance, security, user experience and crucially, their alignment with public values. As governments increasingly integrate these technologies, understanding their capabilities and limitations becomes critical to ensure effective implementation, informed decision-making, and maintaining public trust.

1.1 Research focus and objectives

This research addresses the challenge of integrating chatbots into public services in a responsible way. While chatbots present significant opportunities for modernizing citizen-state interactions—personalizing processes, improving accessibility and citizen engagement—their underlying architectures differ in critical aspects, presenting distinct advantages and disadvantages related in particular to key parameters of public services. Nevertheless, there is a significant gap in the current literature on how to systematically assess these aspects in the context of public service delivery. This gap consists of two parts. First, existing evaluations tend to focus narrowly on technical performance measures, overlooking the rigorous assessment required to align a chatbot's capabilities with the fundamental principles of public administration. Moreover, comparative analyses across the full range of potential chatbot architectures relevant to public sector development remain rare.

To address this gap, we propose and apply a comprehensive framework aimed at evaluating chatbot architectures, particularly for public service delivery. A key contribution is the multidimensional structure of the framework, which contains fifteen distinct evaluation criteria, selected to reflect public service priorities. These criteria include technical dimensions (e.g., input comprehension, error handling, scalability, etc.), user-centric aspects (e.g., multilingual support, accessibility, personalization, etc.), as well as broader parameters such as cross-sectoral adaptability and ethical compliance, so as to ensure evaluations relevant to the public service context. We analyze and evaluate these criteria across four prevalent chatbot architectural paradigms—rule-based, retrieval-based, generative and the increasingly important hybrid retrieval-augmented generation (RAG) systems. By integrating these multifaceted factors, the research offers a structured methodology for decision makers and managers to select chatbot systems that not only enhance service delivery but also support the core values of public services and mitigate potential risks.

The remainder of this paper is structured as follows. Section 2 establishes the foundational context by reviewing the evolution of conversational AI, outlining the ethical and legal landscape for its use in public administration, and synthesizing the key literature that informs our evaluation framework. Section 3 then presents a detailed typology of the four primary chatbot architectures that are the focus of this study. Building upon this, Section 4 outlines the multi-phase comparative methodology developed for this research, detailing the process for framework development, architectural profiling, and heuristic evaluation. Section 5 presents the comprehensive results of this evaluation, followed by Section 6, which discusses the implications of these findings and offers key recommendations for policymakers and public managers. Finally, Section 7 acknowledges the limitations of this study and suggests directions for future research, before Section 8 offers the concluding remarks.

2 Background

2.1 The evolution of conversational AI

Chatbots can be defined as conversational agents that employ natural language to interact with users, replacing the traditional interfaces often used by digital service providers. Chatbots can operate as standalone software or be embedded into physical equipment such as speakers, screens or serving desks.

Chatbots have evolved significantly since the introduction of early systems like ELIZA in 1966 (Weizenbaum, 1966). Constrained by the computational and technological limitations of their time chatbots of that era were based on simple script-based interactions, operating within a narrow parameter space. The turn of the 21st century with the emergence of the internet and the exponential availability of digital data marked a transformative shift for the abilities and the interfaces of chatbots. The appearance of virtual assistants like Siri and Alexa, integrating speech recognition with Natural Language Understanding (NLU), significantly enhanced chatbot functionalities and their potential use cases. Chatbots soon found their way into social media and messaging platforms and gradually expanded to customer service, marketing, and e-commerce applications, by companies seeking automation of routine inquiries, reduced costs and response times and improved customer satisfaction.

In the past decade, advances in Machine Learning (ML) and Natural Language Processing (NLP) have enabled chatbots to handle ever more complex context-aware interactions, and perform diverse tasks, such as multi-lingual communication, translation and problem solving. The introduction of Transformers architecture (Vaswani et al., 2017) has led to the development of sophisticated language models such as GPT (Radford et al., 2018). These innovations, combined with improvements in computer hardware and rapidly increasing availability of digitized text, have given rise to the "era" of Large Language Models (LLMs), a new generation of language models, pre-trained in enormous amounts of text data that provide the foundational capabilities needed by chatbots to support multiple use cases and applications. Pre-trained LLM-driven chatbots can generate more dynamic and contextually aware responses, excel in maintaining dialogues and addressing complex queries, while exhibiting better contextual understanding and adaptability.

Recent advances in multi-modal AI architecture, together with much more powerful computing systems have further improved their capabilities, introducing a new era of chatbots that not only are able to understand and respond to user requests, but can reason, plan, act, and adapt their learning autonomously. This new era of AI agents enables chatbots to plan and execute complex tasks, adapting their behavior to achieve specific goals without requiring constant human oversight. This autonomous operation signifies a departure from traditional, purely reactive chatbot paradigms, toward goal-oriented, intelligent agents, and is poised to revolutionize numerous industries, restructuring established workflows, and redefining the dynamics of human-AI collaboration.

2.2 The ethical and legal landscape for AI in public administration

During the last decade, chatbots have become increasingly popular across various industries playing a significant role in digital transformation. This is particularly evident in the provision of public services where an increasing number of chatbots are being deployed, transforming public sector interactions by serving as an intuitive user interface for citizens interactions. Public service chatbots generally fall into two main categories. The first category consists of informative chatbots, which provide policy information, emergency alerts, real-time updates and public service guidelines. The second category includes service-completion chatbots, designed to facilitate administrative tasks such as appointment scheduling, form submission, and application processing. By automating routine transactions, these bots help streamline operations, reduce operational costs and the burden on traditional service channels.

According to a 2022 study, chatbots represent the most frequently employed AI-based tool within the European Union, constituting 22.8 percent of all use cases (Van Noordt and Misuraca, 2022). A meta-analysis of 30 studies (Ma'rup et al., 2024), indicates that chatbots have a significant positive effect on the efficiency of public services, reducing response time and increased user satisfaction.

However, while the integration of AI chatbots in public administration is a change that improves the quality and accessibility in citizen-state interactions, their deployment also introduces several technical, legal, and ethical risks that require careful consideration, robust legal frameworks, interdisciplinary oversight, and risk mitigation mechanisms, to ensure safe and effective implementation. The following sections identify and discuss how these emerging risks affect established legal norms and ethical imperatives, highlighting the pressing need for mitigation strategies to reconcile the capabilities of modern chatbots with the foundational requirements of public service governance.

2.2.1 Risks associated with the deployment of chatbots in public administration

In recent years, the adoption of chatbot systems in public services has been driven and accelerated by the emergence of LLM models. These models allow even more natural and fluent interactions and have reduced user concerns and resistance, which has led to growth in adoption. This increased integration, however, is now highlighting several risks that need to be addressed to preserve the principles of good governance. Modern LLMs are based on generative AI and operate probabilistically, generating responses based on the statistical likelihood of word sequences. This lack of determinism, along with their reliance on training data, introduces risks such as inconsistencies, factual inaccuracies, unpredictable outputs, and inherited biases, reducing reliability and posing challenges for their use in high-stakes domains of public services and governance.

Generative AI models tend to produce inaccurate or even completely fabricated outputs, often referred to as "hallucinations", due to their probabilistic nature. Such risks may lead to incorrect legal guidance or misleading information, thus exposing public organization to liability risks. Dahl et al. (2024) showed that LLMs hallucinate at least 58% of the time in legal tasks and concluded by cautioning against the rapid and unsupervised integration of popular LLMs into high-stakes environments. A notorious example is the Mata v. Avianca case, in which attorney Steven A. Schwartz used ChatGPT for legal research and filed a federal brief containing fabricated cases, citations, and holdings—an oversight that led to sanctions from United States District Court (2023). This case underscores the significant challenges in detecting and mitigating hallucinations, particularly in public-service deployment scenarios where legal accuracy and accountability are critical.

In addition, reliance on vast amounts of training data means that generative chatbots can inadvertently amplify biases present in their training datasets. While hallucinations refer to instances when a model produces factually incorrect or entirely fabricated content, biases represent systematic distortions in outputs that mirror pre-existing prejudices or stereotypes hidden in the data. Such biases, stemming from skewed or unrepresentative datasets, may result in discriminatory outcomes, harmful stereotypes and even incite hate speech. Skewed datasets, whether representatively valid or not, produce skewed models, which in turn can affect the output of the chatbots. Moreover, biases may also emerge during fine-tuning processes like Reinforcement Learning from Human Feedback (RLHF) where subjective interpretations of "human values" can pass from evaluators to the model. Combined with the opaque nature of deep learning architectures, these issues hinder transparency and accountability, complicating efforts to identify and mitigate the origins of erroneous outputs (O'Neil, 2016; Biggio and Roli, 2018).

Adversarial attacks present an additional risk to chatbot systems, particularly those based on large language models. In these attacks, malicious actors craft subtle hard-to-detect input modifications—often referred to as "jailbreaks" or "prompt injections"—that exploit vulnerabilities and design glitches in the model to bypass safety mechanisms and trigger inappropriate or harmful outputs. Such manipulations can cause the chatbot to generate misleading or damaging responses, potentially revealing sensitive data and undermining public trust in government services. Adversarial techniques have been used to force chatbots to reveal confidential details or produce content that violates established guidelines (Liu et al., 2023; Zhuo et al., 2023). The integration of models of other modalities (voice, image) into

LLMs may introduce additional risks of adversarial attacks (Ye et al., 2023). The combination of generative models with retrieval capabilities further increases the risks -blurring the line between data and instructions- permitting attacks that can remotely affect other users' systems by strategically injecting the prompts into data likely to be retrieved at inference time (Greshake et al., 2023).

The integration of LLM-based chatbots within public services also raises significant data privacy and security concerns. Chatbot systems interacting with citizens, receive and process sensitive personal information, including personal details, and even health records. Potential vulnerabilities in the chatbot's architecture, data storage, or APIs can be exploited by malicious actors to gain unauthorized access to this data, leading to identity theft, financial fraud, and other harmful consequences. The increasing sophistication of cyberattacks targeting AI systems, as highlighted in recent research (Biggio and Roli, 2018), underscores the urgent need for robust security measures. Furthermore, reliance on third-party LLMs and commercial cloud infrastructure limits public service organizations' control over data usage. User queries and other related data may be utilized for model training or fine-tuning purposes, in ways that conflict with privacy regulations such as the General Data Protection Regulation (GDPR). To address these challenges, it is crucial for public administration to adopt privacy-by-design frameworks and enforce strict contractual data usage agreements.

2.2.2 Ethical requirements

These technical and operational risks associated with chatbots stand in stark conflict with the ethical and legal standards traditionally upheld in public service delivery. The deployment of chatbot introduces complexities that can undermine core principles of public administration such as accountability, transparency, and fairness, eroding public trust and exposing public organizations to legal liability.

Fairness as a fundamental principle of public sector mandates the impartial, equitable, and unbiased delivery of government services and policies. Legal frameworks, ethical codes, and administrative protocols require public servants to act objectively, transparently, with respect to the rights and dignity of all citizens. But this principle is directly engaged in the case of chatbots, by the manifestation of the bias risk, as prejudices present in training data may yield discriminatory outputs that contravene the fairness mandate. A notable example, as described by Lippens (2024), involves ChatGPT, which, during simulated job application assessments, exhibited discriminatory tendencies against certain ethnic and gender groups. The model assigned lower suitability scores to applicants from specific backgrounds, reflecting societal stereotypes embedded within the training data. Additionally, chatbot interfaces can create accessibility disparities for individuals with disabilities, while the digital divide may further marginalize those lacking technological access or digital literacy.

Transparency, another foundational requirement in public administration, requires that government processes and decision-making mechanisms remain open and comprehensible to citizens. Applied in the context of chatbots deployment for public service delivery, this principle mandates clear communication regarding the nature of automated interactions, including disclosure of the chatbot's operational methods, training data sources, and inherent limitations, as well as, clear explanation of any decision and information provided. But, the opaque, "black box" architecture and the probabilistic nature of generative models obscure the reasoning behind chatbot responses, impeding citizens' ability to scrutinize and challenge responses, decisions or recommendations. This lack of explainability directly challenges transparency, potentially eroding trust and hindering engagement with public services. To address this, public administrations must implement measures that demystify the underlying algorithms, establish accessible channels for feedback and fallback mechanisms to human representatives for unresolved issues. Felzmann et al. (2019) argue that for AI systems to be considered trustworthy, transparency requirement should be tailored to the stakeholder more broadly, including developers, users, regulators, deployers, and society in general. This is echoed in a broader analysis by Hohmann (2021), which examines how major intergovernmental organizations interpret transparency, reinforcing its status as a fundamental principle in any public-facing system.

The principle of accountability, a cornerstone of responsible governance, is also directly challenged by the integration of chatbot technology into public services. Demanding that public institutions and their representatives are answerable for their actions, decisions, and outcomes, this principle requires clear lines of responsibility, robust oversight mechanisms, and accessible avenues for redress when failures occur or harm is inflicted. But the complexity and obscurity of LLMs, coupled with the potential for automated decision-making processes, impedes the identification of who is responsible in case of misinformation or wrong decisions. This can erode public trust and expose organizations to legal liability, as demonstrated by a ruling against Air Canada, where the company was held accountable and liable for misrepresentations made by its chatbot regarding bereavement travel policies (Civil Resolution Tribunal, 2024).

Respect for citizen's privacy and data protection constitute a major set of ethical and legal obligations of public administration. Given the chatbots' role in handling personal, and in some cases even sensitive information, risks of violation of those principles can materialize. Vulnerabilities in chatbot systems can expose sensitive data during breaches and cyberattacks. For instance, ChatGPT publicly admitted that its famous chatbot system leaked chat history of users due to vulnerabilities in the Redis client open-source library (OpenAI, 2023). The volume, velocity, and variety of data collected, coupled with the potential for automated analysis and profiling, increases the risks for user privacy and data security. To ensure respect for user privacy and robust data protection within chatbot deployments, public administrations must prioritize adherence to established data protection principles. This entails obtaining informed consent prior to data collection, minimizing data collection to what is strictly necessary for specified purposes, implementing robust security measures to safeguard against unauthorized access or disclosure, and affording individuals the rights to access, rectify, and erase their personal data. Adherence to these tenets is critical for sustaining public trust, upholding constitutional rights, and ensuring compliance with data protection regulations.

2.2.3 Ethical and regulatory frameworks

Deploying AI in public services introduces risks that not only undermine the integrity of public interactions but also challenge the adequacy of current regulatory frameworks designed to protect citizens and ensure equitable service delivery. The integration of emerging technologies into public administration has strained traditional mechanisms for enforcing principles of ethical governance, underscoring the need for new legal and administrative frameworks adapted to this evolving landscape.

In the past 5 years alone, nearly a hundred different non-legally binding ethical codes or statements have been adopted by public, private, and non-governmental organizations, all promoting similar principles, such as transparency, fairness, respect for human autonomy, and privacy (Maclure and Morin-Martel, 2025). Often articulated by international organizations and expert bodies, those frameworks provide principles, ethical compliance standards and guidelines for addressing crucial ethical challenges and uphold the rights of citizens. For example, the OECD's "Recommendations on AI" establish internationally recognized principles for trustworthy AI, emphasizing transparency, accountability, fairness, and robustness (OECD, 2019). Similarly, the European Commission's "Ethics Guidelines for Trustworthy AI" outline essential requirements for AI systems to be lawful, ethical, and robust, stressing human agency and oversight, technical robustness, and data governance (EC, 2019). UNESCO's "Recommendation on AI Ethics", adopted in November 2021, also promotes a human-centered approach, prioritizing inclusivity, human dignity, and accountability (UNESCO, 2021).

However, there is growing skepticism regarding the potential of ethical principles to help enact responsible development of AI technologies (Maclure and Morin-Martel, 2025). Most of these ethical frameworks are largely conceptual; although they offer guidance and promote ethical awareness, their abstract, general, and voluntary nature limits their practical effectiveness in protecting citizens' rights. The realization that existing legal rules were insufficient to protect people's rights against AI's risks, and that the promulgation of nonbinding AI ethics guidelines did not provide a satisfactory solution either (Smuha, 2025), ultimately led to the introduction of the first enforceable regulatory measures. In the European context, the advancement of AI technologies has prompted the European Union to proactively establish regulatory frameworks that ensure the ethical deployment of AI. A cornerstone of these efforts is the Artificial Intelligence Act (AI Act), which categorizes AI applications based on their risk levels. Representing a significant milestone in the regulation of artificial intelligence, the Act aims to drive the development, deployment, and use of AI across Europe, with a substantial focus on safety, protection of fundamental rights. It introduces a risk-based classification system, assigning obligations proportional to the risks posed by various AI applications, with a particular emphasis on high-risk and unacceptable use-cases. Under the provisions of the Act, chatbots are generally categorized as limited-risk systems, but generative Large Language Models, depending on their scale and societal impact, can be classified as General Purpose AI (GPAI) systems, for which special obligations are introduced. These include preparing detailed technical documentation for submission to the AI Office upon request, creating deployer-oriented guidelines, ensuring compliance with EU copyright law, and providing summaries of training data sources. For high-impact GPAI models with systemic risks (identified as those involving computational resources exceeding $10^{\wedge}25$ FLOPs), additional measures include continuous risk assessment, model evaluations, cybersecurity safeguards, and mandatory notifications to the European Commission regarding new qualifying systems. Traditional rule and retrieval based chatbots, in contrast, face fewer regulatory burdens under the AI Act due to their limited scope and deterministic nature. While they must comply with general data protection standards, including the GDPR, they are exempt from the extensive risk management and transparency obligations imposed on generative LLMs.

However, policymakers should exercise caution in relying solely on the EU AI Act's provisions for assessing chatbot risks. As Smuha and Yeung (2025) argue, while the Act aims to address "systemic risks" from General Purpose AI (GPAI) models, its focus on computational metrics and data size thresholds for identifying such risks is problematic as this threshold is rather arbitrary, potentially excluding influential GPAI models that fall below it, particularly as the industry shifts toward smaller, more potent models. Moreover, limiting "systemic risks" to GPAI models overlooks the potential for even traditional rule-based AI systems to pose significant risks. A more comprehensive risk assessment should encompass diverse model types and consider potential impacts on public health, safety, fundamental rights, and society, beyond mere computational resource criteria.

2.3 Literature review: evaluating chatbots in public service

This section reviews relevant empirical and theoretical literature to inform the development of a practical evaluation methodology. While the preceding discussion outlined the broad risks and governance principles, a focused review of empirical studies on existing chatbot implementations is necessary to identify the specific parameters and criteria that determine their effectiveness and alignment with public values. This section, therefore, surveys the empirical and theoretical literature and serves as the direct foundation for the evaluation framework proposed in Chapter 4, ensuring that the selected evaluation criteria are substantiated by real-world challenges and scholarly insights into the deployment of chatbots in the public sector.

Recent empirical research provides a granular account of the challenges inherent in chatbot implementation within public organizations. Chen et al. (2024) examined chatbot adoption across 22 U.S. state governments, distinguishing between the drivers of technology adoption and the determinants of successful implementation. In their study Chen et al. (2024) highlight real-world benefits such as 24/7 service and multilingual support, noting that chatbots can "reduce staff workloads" and improve access across service lines when properly deployed and designed. The authors identify knowledge-base creation and maintenance, managing technology skills, securing adequate

resources, navigating safety regulations, and meeting citizen expectations as the most crucial determinants of a chatbot's success. These findings suggest that technical performance is intertwined with organizational capacity, ethical considerations, and user-centric design, supporting the need for a multidimensional evaluation approach. The need for a multidimensional and domain-specific approach is further corroborated by recent research in other high-stakes sectors. Gupta et al. (2025) likewise propose an evaluation framework for financial-sector chatbots that assesses cognitive intelligence, user experience, operational efficiency, and ethical compliance. Their findings reinforce the inadequacy of generic evaluations and the need for a domain-specific, multidimensional approach. Their framework, which assesses chatbots across cognitive intelligence, user experience, operational efficiency, and ethical compliance, reinforces the validity of a holistic evaluation model. The emergence of such specialized frameworks underscores a critical consensus: generic, one-size-fits-all evaluations are insufficient. Meaningful assessment requires a multi-dimensional methodology tailored to the unique operational realities, user expectations, and ethical obligations of the specific sector.

Government chatbots are expected to provide accurate, reliable, and consistent information, but studies highlight challenges in maintaining data quality and handling complex citizen queries (Ma'rup et al., 2024; Cortés-Cediel et al., 2023). The risk of "hallucinations" (factually incorrect outputs) is well documented in LLMs, especially in high-stakes domains such as healthcare and legal services (Dahl et al., 2024). Rule-based systems, by contrast, provide consistency and predictability but lack the flexibility of LLMs. Hybrid retrieval-augmented generation (RAG) architectures, which ground responses in a curated knowledge base, are emerging as an effective approach to reduce hallucinations. For example, Vemulapalli (2025) finds that RAG systems can reduce hallucination rates by roughly 70% compared to pure generative models. These findings emphasize the importance of criteria such as Input Comprehension (how well the system decodes varied language) and Response Accuracy and Factual Integrity (checking answers against official data) for comprehensive architecture selection.

While accuracy and reliability define the baseline for performance, literature consistently shows that user perception and trust ultimately determine whether such systems succeed in practice. Beyond functional performance, the success of a digital government service is contingent on public trust and a positive user experience. A seminal experimental study by Aoki (2020) directly investigated the drivers of initial public trust in government chatbots. Her research revealed that trust is highly context-dependent, with the public showing significantly less trust in chatbots for service areas requiring empathy and complex situational judgment, such as parental support, compared to more procedural tasks like waste separation. This crucial finding empirically demonstrates that "performance" in the public sector is judged not only on technical accuracy but also on the perceived ability to handle nuanced human needs. Furthermore, Aoki's discovery that communicating citizen-centric purposes—such as ensuring "uniformity in response quality" or "24-h, 365-day, timely responses"—measurably enhances public trust, even if the effect sizes are small, provides a direct link between strategic communication and user acceptance. These findings are consistent with research by Abbas et al. (2023), which highlights the importance of clarity and effective error recovery, and shows that government chatbot users demand not only efficiency but also trustworthiness, ease of use, and nuanced conversation handling. They report that a positive user experience "is heavily dependent on the chatbot's ability to provide clear, understandable responses and to offer effective recovery mechanisms when errors occur". Recent experimental work reinforces this: Zhou et al. (2025) find that empathetic chatbot communication significantly enhances user trust and satisfaction in digital public services. In short, peer-reviewed evidence shows that fluent, context-aware dialogue and robust recovery from misunderstandings are essential for user satisfaction, and thus must be explicitly evaluated for the selection of chatbot architecture. These insights motivate our User Experience and Communication category, which includes criteria like Conversational Fluency (grammatical, coherent responses), Error Handling and Recovery (fallbacks and clarification), and Response Timing (speed and consistency).

A growing body of work has begun to categorize and analyze the risks related to generative or RAG implementations of chatbot systems. The literature highlights numerous ethical and safety dimensions that public agencies must consider. Gan et al. (2024) present a comprehensive survey that addresses the expanding security, privacy, and ethical challenges associated with LLM–based agents, focusing on implementations that use transformer models as control hubs to perform complex tasks. The authors propose a novel taxonomy that organizes threats by both their sources—such as malicious inputs, model misuse, or data vulnerabilities—and their impacts across different agent modules and operational stages. Yu et al. (2025) propose taxonomies for threats across agent components and stages. Cui et al. (2024) developed a risk framework for LLM models, focusing on the risks of four LLM modules: the input module, language model module, toolchain module, and output module. Ammann et al. (2025), have conducted research focusing on the vulnerabilities of RAG pipelines, and outlining the attack surface from data pre-processing and data storage management to integration with LLMs. The identified risks are then paired with corresponding mitigations in a structured overview. Greshake et al. (2023) provided an overview of safety threats and design vulnerabilities of LLMs, including prompt engineering and hallucinations, and highlighted the risks associated with their inclusion in chatbots. Bommasani et al. (2021) surveyed some of the risks that accompany the widespread adoption of foundation models, ranging from their technical underpinnings to their societal consequences. Their research highlights important risks related to the use of LLMs in chatbots, such as the handling of output liability and discrimination and noting that the use of such models by governmental entities—at a local, state or federal level—necessitates special considerations. A survey by Chu et al. (2024) delivers a structured and comprehensive taxonomy of research on fairness in LLMs. A notable insight stressed by the study is that LLMs can produce accurate outputs grounded in flawed rationale, thereby amplifying discriminatory patterns despite surface-level correctness. This underscores the complex

challenge of ensuring both fairness and transparency in LLM-generated text—highlighting the need for rigorous, multi-stage evaluation and diverse mitigation strategies. These findings directly inform our Ethics and Safety dimension, which includes sub-criteria such as Bias Mitigation and Fairness, Ethical Compliance, and Data Protection and Privacy, grounded in the literature. By embedding these ethical requirements into the framework, we address the "expanding security, privacy, and ethical challenges" identified in scholarly surveys.

Practical deployments demand that chatbots be adaptable across domains and scalable to workload. Studies of conversational AI architectures underscore this need: for example, Mechkaroska et al. (2024) demonstrate that as user demand grows, maintaining responsiveness and low latency requires both vertical and horizontal scaling of the system. This justifies criteria for Scalability and Resource Efficiency in our framework, ensuring a chatbot can handle large user loads without degrading performance. Additionally, public service chatbots can adapt to different domains and offer full-service delivery, when designed to allow integration with existing information systems and data sources. This architectural paradigm was formally introduced in the foundational work of Lewis et al. (2020), who proposed an end-to-end model that combines a pre-trained retriever to find relevant documents with a pre-trained generator to synthesize an answer. The key advantage of this approach for adaptability is that it decouples the model's linguistic capabilities from its domain-specific knowledge. To adapt the system to a new public service domain, administrators can update or replace the external knowledge base without the need for costly and time-consuming model retraining. These observations support our Adaptability criteria: the chatbot's ability to incorporate new data domains, connect to APIs, and be updated by non-experts. In sum, the literature suggests that evaluating domain versatility, integration flexibility, and update efficiency is crucial for public-sector chatbots, since they must scale out to varied tasks and growing usage without undue cost or complexity.

Finally, the principles of digital equity and seamless service integration emerge as foundational themes in the literature. Public sector chatbots cannot be considered successful if they fail to serve all members of the public, including those with disabilities, limited digital literacy, or different language needs. Scholarship on the "digital divide" warns that new technologies can inadvertently exacerbate inequalities if not designed with universal access in mind (Helsper, 2021). Therefore, a chatbot's value is intrinsically tied to its inclusivity, justifying a direct evaluation of its accessibility (e.g., compliance with WCAG standards) and multilingual capabilities. Moreover, the literature on digital government transformation emphasizes a shift from siloed information provision to integrated, end-to-end service delivery (Wirtz et al., 2018). The ultimate goal is to automate entire service workflows, not just answer queries. Research indicates that a chatbot's true value is realized when it is seamlessly integrated into existing government processes, allowing citizens to perform tasks such as scheduling appointments or checking application statuses directly within the conversational interface (Androutsopoulou et al., 2019). Practitioner-focused reports, such as those from the Center for Technology in Government at the University at Albany, echo this, noting that chatbots can significantly "improve citizens' access to public services" by breaking down language and time barriers. This body of work provides a clear mandate for our "Inclusivity and Service Delivery" dimension, which, with its criteria of "Multilingual Support and Accessibility" and "Service Delivery and Process Automation," ensures that the evaluation captures a chatbot's alignment with the core public values of universal service and administrative efficiency.

The evaluation framework presented in Chapter 4 is directly informed by these findings. Our framework, therefore, synthesizes these empirically and theoretically derived dimensions into a structured and comprehensive methodology which includes multi-dimensional criteria for assessing chatbot architectures in the public sector.

3 A typology of chatbot architectures

A prerequisite for evaluating alternative approaches to chatbot systems is the establishment of a clear and coherent typology of chatbots that defines the application scope of the methodology. For practical usability, this typology should encompass all major chatbot implementation techniques while filtering out unnecessary details and variations. Chatbots can be classified on a multitude of parameters, such as domain knowledge, service type, interaction modality (text, voice, or multimodal), architectural design, personalization abilities, cognitive capabilities, or learning adaptability. Among these, the underlying architectural design is a fundamental parameter which examines the technologies used by chatbots to process user input and generate responses. A typical chatbot architecture consists of at least three components: natural Language Understanding (NLU) component, Natural Language Generation (NLG) component and User Interface (UI). From these three components NLU and NLG are the most important for both the classification of chatbots, but also, for their actual performance and capabilities. NLU is responsible for interpreting and processing user inputs, whereas NLG focuses on constructing appropriate responses. But details of operation, mixing of approaches and variations in comprehension and response synthesis, allows distinguishing subcategories and introduction of hybrid approaches such as modern RAG implementations that integrate external knowledge with AI-driven responses.

To ensure a comprehensive and practical evaluation, our methodology adopts a typology based on four architectural categories of chatbots, namely rule-based, retrieval-based, generative and hybrid. These categories effectively account for key differences in comprehension, response synthesis, and underlying knowledge base, making them the most suitable for the application of the evaluation. The following sections will outline the core characteristics and operational principles of these four categories, which will serve as the primary focus of the evaluation methodology.

3.1 Rule-based architectures

Rule-based chatbots are among the earliest paradigms of conversational AI. Typically operating using predefined sets of rules and patterns, are implemented using decision trees

and finite-state machines. Systems in this category trigger responses based on simple keyword detection rules and predefined conversation flows. Scripting languages such as AIML (Artificial Intelligence Markup Language), RiveScript, and ChatScript, are often being used to facilitate the definition of keyword pattern-matching rules, response templates and decision flows.

The NLU component in rule-based chatbots is deterministic; it relies on predefined rulesets to identify keywords and intent, without any contextual understanding capabilities. As a result, these chatbots require highly structured input to function effectively, lacking adaptability to unexpected queries or paraphrased expressions. Their NLG component relies on template-based responses, selecting pre-written outputs based on the identified input. While this guarantees predictability and reliability, it also limits their flexibility, making them more effective for narrowly defined, structured, and repetitive interactions, such as automated customer support, FAQ systems, and basic transactional workflows.

Variations among rule-based chatbots stem from their structural complexity and the sophistication level of their rulesets. Simpler approaches rely on detecting specific words or phrases in user inputs to trigger predetermined responses hardcoded in the source code. Decision tree based chatbots follow a pre-defined tree structure in which each branch represents a distinct dialogue path, while template-based use pre-written scripts or conversation templates, often implemented in languages such as AIML.

3.2 Retrieval-based architectures

While rule-based chatbots rely on a predetermined set of rules and response templates to guide their interactions, retrieval-based chatbots use existing data to generate responses. Instead of relying on rulesets and keyword detection, these systems leverage semantic similarity techniques to identify the user intent, matching input vectors with relevant responses residing in locally stored knowledge bases, FAQ, document corpora, and/or knowledge graphs. This allows for more flexibility and adaptability than rigid rule-based approaches.

The NLU component in retrieval-based chatbots typically employs statistical similarity techniques, such as TF-IDF, cosine similarity, or word embeddings (e.g., Word2Vec, GloVe). For NLG, retrieval-based systems rely on response selection rather than generation, ranking and retrieving the most contextually similar response from their underlying knowledge base. Overall, retrieval-based chatbots improve over rule-based, with their ability to handle more complex and paraphrased user queries while maintaining factual accuracy.

Retrieval-based chatbots can be further subdivided by the underlying retrieval techniques. More recent approaches employ dense embeddings from models like BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and ELECTRA (Clark et al., 2020), to capture semantic similarity instead of using traditional methods based on simple vector space techniques such as TF-IDF. These models improve on the intent and entities recognition of syntactically complex inputs, leading to better contextual understanding. Additionally, retrieval-based chatbots can be classified as static or dynamic depending on how they process and rank responses. Static retrieval systems select the best-matching response from a fixed, pre-indexed corpus using a single-pass assessment. In contrast, dynamic retrieval systems continuously refine response ranking in real time, employing techniques such as context-aware re-ranking, query expansion, and neural search models to enhance response relevance based on conversational history and user intent. Other variations include personalization-enhanced models, which adapt responses based on historical user interactions, and systems that integrate structured knowledge—such as knowledge graphs—to improve the relevance and precision of retrieved information.

3.3 Generative-based architectures

In contrast to retrieval-based chatbots, where the response is retrieved from predefined sources, generative-based chatbots produce responses dynamically relying on deep learning techniques and pre-trained Large Language Models. A milestone for their success was the advent of Transformer models (such as GPT), which, based on decoder-only architectures[1] enabled context-aware self-adaptive text generation, dramatically improving the fluency and naturalness of the generated text and revolutionizing the chatbot landscape.

The NLU component in these systems is inherently integrated within the generative process, where a single model both interprets and produces text in a single pipeline. For NLG, generative models use single-step decoding, predicting each token based on prior context in an autoregressive manner. This allows the generation of coherent, contextually appropriate, and novel responses, making these models highly effective for open-ended dialogues and creative outputs. However, this generative capacity often produces "hallucinations",[2] as responses are generated probabilistically rather than retrieved from a validated source.

Generative-based chatbots can be further classified into open-domain models, trained on massive and diverse datasets, capable of engaging in unconstrained conversations, and closed-domain models, fine-tuned on specific datasets to ensure domain relevance and enhanced factual accuracy. Recent research approaches introduce Transformer variations with memory, that integrate long-term contextual awareness and dialogue history, allowing for more coherent multi-turn dialogue interactions (Wu et al., 2022; Bulatov et al., 2024).

1 Also known us causal decoder architectures.

2 The term has captured the popular imagination. "Hallucinate" secured its position as the word of 2023 (https://dictionary.cambridge.org/editorial/word-of-the-year/2023) while Dictionary.com noted a 46% surge in searches for the term over the past year (https://www.dictionary.com/browse/hallucinate).

3.4 Hybrid approaches using retrieval-augmented generation (RAG)

Hybrid approaches address the risks of purely generative models by integrating retrieval-based knowledge with generative response synthesis. Retrieval augmented generation is an architectural approach for optimizing the performance of generative models by connecting them with external knowledge bases. RAG helps large language models (LLMs) deliver more relevant responses of higher quality, combining the factual accuracy of retrieval-based systems with the fluency and creativity of generative models.

The NLU component in RAG-based chatbots first retrieves contextually relevant documents, knowledge snippets, or structured data from an external corpus (e.g., a vector database, a knowledge graph), leveraging semantic similarity techniques like those used in retrieval-based chatbots. The NLG component then synthesizes responses using a generative model, leveraging the retrieved knowledge as additional context. This approach enhances factual consistency, reduces the risk of hallucinations associated with purely generative models, and allows the chatbot to provide more informative and grounded responses.

Similarly to static and dynamic retrieval systems, two primary hybrid approaches are based on the flow of interaction between the retriever and the generator components. Passive RAG operates through a single interaction in which the retriever supplies data to the generator, and the generator produces a response without further feedback. Conversely, in active RAG, a two-way exchange between the retriever and the generator, is taking place, during which, the retriever continuously refines its data selection based on the evolving text generated, while the generator can request additional information to clarify uncertainties. This process enhances the integration of context and improves the overall accuracy and relevance of the responses generated. More variations of hybrid RAG models can stem from their retrieval strategies (e.g., vector search, knowledge graph traversal), retrieval source (differentiating between offline document-based RAG, which rely on pre-indexed, static corpora and online internet-based RAG, leveraging dynamic web content via APIs and search engines to incorporate real-time information), the granularity of the retrieved information (e.g., full documents, paragraphs, sentences), and the generative constraints applied during response synthesis (e.g., prompt engineering, fact verification).

4 Methodology: a framework for comparative architectural evaluation

To address the research gap, we developed and applied a multi-stage comparative analytical methodology designed to systematically evaluate the four primary chatbot architectures focusing particularly on the needs, particularities and use cases of public administration. This targeted approach ensures that the proposed framework not only evaluates the operational performance of chatbot architectures, but also their alignment with values and ethical standards for the provision of public services, providing guidance toward robust, citizen-focused AI deployments. The methodology was executed in three distinct phases, as illustrated in Figure 1.

4.1 Phase 1: development of the evaluation framework

The foundation of this research is a robust evaluation framework derived directly by the key challenges and requirements identified in the ethical and legal considerations survey (Section 2.2) and the literature review (Section 2.3). This phase translated the identified key challenges into a set of structured, measurable criteria.

1. **Criterion derivation**: An initial long-list of potential evaluation criteria was derived from the key themes identified in the literature: functional efficacy, user-centricity, ethical governance, and inclusivity. The process ensured that the criteria were grounded in documented real-world challenges (Chen et al., 2024), user expectations (Abbas et al., 2023; Aoki, 2020), technical benchmarks (Vemulapalli, 2025), and ethical principles (Gan et al., 2024; Chu et al., 2024). By grounding the framework in this extensive literature, we ensure that its foundational dimensions are robustly justified by both theory and empirical evidence on public-sector chatbot deployments.
2. **Criterion refinement and validation**: The initial list was systematically refined into a final set of fifteen criteria by applying four rigorous selection principles:

    - **Universal applicability**: Ensuring each criterion is relevant to all four architectures. Criteria depending on architecture-specific features are excluded to ensure that comparisons remain valid and unbiased.
    - **Differentiability**: Selecting criteria that highlight meaningful performance variations. A criterion is considered differentiable if it provides measurable variation across architectures rather than yielding uniform or indistinct results.
    - **Granular measurability**: Disaggregating broad concepts into specific, assessable components.
    - **Contextual relevance and value alignment**: Prioritizing criteria that reflect the unique needs and values of the public sector.

3. **Detailed criterion structuring**: To ensure each criterion was unambiguous, contextually relevant, and assessable, every refined criterion was then formally operationalized using a four-part structure:

    - *Focus*: Defines the specific aspect of the chatbot being evaluated (e.g., the ability to interpret user input).
    - *Objective*: Outlines the intended role and importance of this criterion within the public service context (e.g., to ensure citizens' needs are correctly understood).
    - *Key determinants*: Identifies the underlying technical and architectural elements that influence performance for that criterion (e.g., the sophistication of the NLU model).

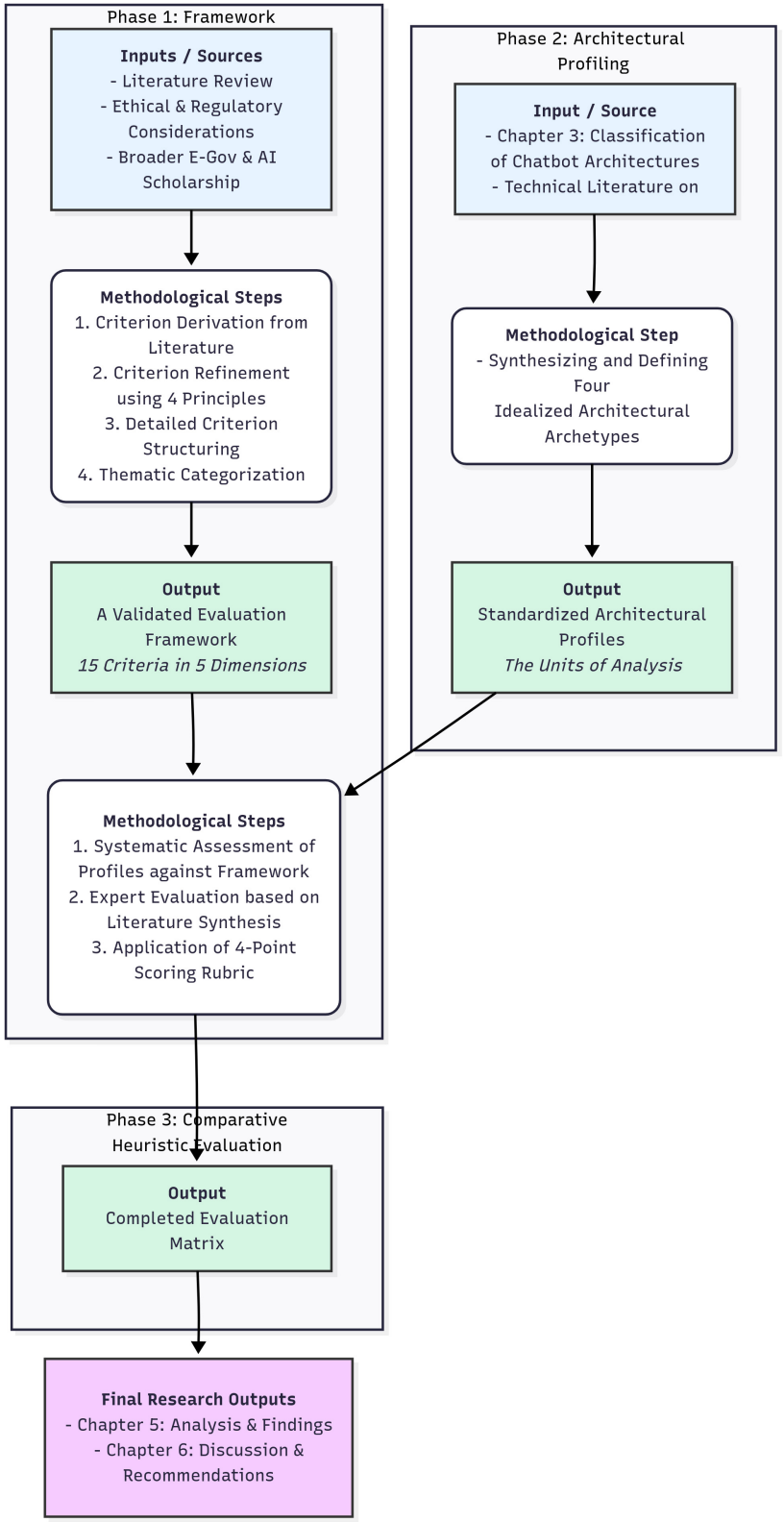

FIGURE 1
Methodology flow diagram.

- *Assessment methods*: Proposes potential methods for how the criterion could be empirically measured, adding a layer of practical applicability to the framework.

This structured definition process adds analytical rigor and ensures that every evaluation is based on a clear, multi-faceted understanding of the criterion.

4. **Thematic categorization**: For analytical clarity, the final fifteen validated criteria were organized into five distinct thematic dimensions: (I) Core Functionality and Understanding, (II) User Experience and Communication, (III) Ethics and Safety, (IV) Adaptability and Scalability, and (V) Inclusivity and Service Delivery, ensuring a holistic and well-balanced assessment.

4.2 Phase 2: architectural profiling

To ensure a fair and consistent comparison, we first established standardized, idealized profiles for each of the four architectures, as detailed in Section 3. Without this architectural profiling, the research would be vulnerable to a major methodological flaw: comparing idiosyncratic, real-world implementations that have too many confounding variables. These profiles—Rule-Based, Retrieval-Based, Generative, and Hybrid (RAG)—served as the consistent units of analysis for the evaluation. Each profile was defined by a consistent set of five core characteristics derived from the technical literature:

- *Core operational logic*: How does it process input and generate a response? (e.g., deterministic keyword matching vs. probabilistic token prediction).
- *Underlying technology*: What are the key components? (e.g., decision trees, finite-state machines vs. Transformer models, vector databases, a combination of underlying technologies).
- *Knowledge management*: How does it store and use information? (e.g., hard-coded scripts vs. a curated document corpus vs. parametric knowledge learned during pre-training).
- *Inherent strengths*: What is the architecture naturally good at? (e.g., Rule-Based systems excel at consistency and predictability).
- *Inherent weaknesses*: What are its systemic risks? (e.g., Generative models have an inherent risk of hallucination).

This step prevents the comparison of specific, idiosyncratic implementations and instead focuses the analysis on the fundamental capabilities of each architectural paradigm.

4.3 Phase 3: comparative heuristic evaluation and scoring

The core of the methodology involved a comparative heuristic evaluation where the four architectural profiles were systematically assessed against the fifteen criteria. This heuristic approach was selected as it is ideally suited for a conceptual, comparative study focused on evaluating the inherent capabilities and systemic risks of idealized architectural types, rather than the empirical performance of a single, specific implementation,

1. **Evaluation process**: The scoring was conducted through an expert evaluation, based on a comprehensive synthesis of the evidence presented in the literature review. For each criterion, we assessed the inherent capacity of each architecture to meet its demands. The judgment was based not on a single implementation but on the documented potential and systemic risks associated with each architectural approach. For instance, the scoring for "Response Accuracy" considered both the high risk of hallucination in pure generative models (Dahl et al., 2024) and the documented mitigation effect of RAG architectures (Vemulapalli, 2025).
2. **Scoring rubric**: To ensure consistency, a 4-point ordinal scale was used, with each level having a clear operational definition:

   - *Limited*: The architecture has inherent design characteristics that make it fundamentally ill-suited to meet the criterion without significant and unnatural modifications.
   - *Moderate*: The architecture can meet the criterion but has significant constraints or requires substantial effort, and performance may be inconsistent.
   - *High*: The architecture is naturally suited to meet the criterion, and it represents a core strength.
   - *Very high*: The architecture represents the state-of-the-art for the criterion; its design is optimized to excel in this area.

The output of this phase is the comprehensive evaluation matrix presented in the Results section (Table 1), which provides the empirical basis for the subsequent analysis, discussion, and policy recommendations.

4.4 Criteria for evaluating chatbot architectures in public service delivery

Building on the literature review and the preceding framework operationalization, this section outlines the criteria used to evaluate chatbot architectures in public service contexts. The criteria are organized into five overarching dimensions: core Functionality and Understanding, User Experience and Communication, Ethics and Safety, Adaptability and Scalability, and Inclusivity and Service Delivery. Each dimension reflects a cluster of concerns repeatedly emphasized in the literature, ranging from technical accuracy to broader sociotechnical values such as fairness, accessibility, and democratic legitimacy.

- **Core functionality and understanding** focuses on a chatbot's capacity to provide accurate, reliable, and contextually appropriate responses. This includes not only response accuracy but also the system's ability to correctly interpret diverse citizen queries.

TABLE 1 Overview matrix with evaluation results across all criteria.

Criterion	Rule-based	Retrieval-based	Generative	Hybrid (RAG)
I. Core functionality and understanding				
Input comprehension and intent recognition	Limited	Moderate	High	Very High
Response accuracy and factual integrity	High	High	Moderate	Very High
Consistency and reliability	Very High	High	Moderate	High
II. User experience and communication				
Conversational fluency and contextual awareness	Limited	Moderate	Very High	Very High
Error handling and recovery	Limited	Moderate	Moderate	High
Response timing and responsiveness	Very High	High	Limited	Moderate
III. Ethics and safety				
Bias mitigation and fairness	Very High	High	Limited	Moderate
Ethical compliance and liability safety	High	High	Limited	Moderate
Data protection and privacy	High	High	Moderate	High
IV. Adaptability and scalability				
Domain versatility and integration	Limited	Moderate	High	Very High
Scalability and resource efficiency	Very High	High	Limited	Moderate
Maintainability and update efficiency	Moderate	High	Limited	Moderate
V. Inclusivity and service delivery				
Multilingual support and accessibility	Limited	Moderate	High	Very High
Personalization and contextualization	Limited	Moderate	High	Very High
Service delivery and process automation	Limited	Moderate	Moderate	High

- **User experience and communication** addresses how effectively the chatbot interacts with users. Dimensions such as conversational fluency, error handling, and responsiveness are critical in shaping trust and satisfaction with digital public services.
- **Ethics and safety** captures the governance-related risks of deploying AI-driven systems. Criteria in this category encompass alignment with ethical and regulatory standards such as data protection, fairness, bias mitigation.
- **Adaptability and scalability** evaluates the ability of chatbot architectures to transfer across service domains and to operate efficiently under varying workloads. This dimension reflects the need for public-sector systems to remain sustainable and flexible in rapidly changing administrative environments.
- **Inclusivity and service delivery** ensures that chatbot deployments advance digital equity and enhance service integration. Criteria include accessibility for citizens with diverse needs and the ability to support multilingual interactions while embedding chatbots into end-to-end service processes.

To ensure transparency and replicability, each criterion is further elaborated through four analytical layers: focus, Objective, Key Determinants, and Assessment Methods. This structure translates high-level evaluation concerns into operational guidance, specifying both the intended role of each criterion and the mechanisms by which it can be empirically assessed.

4.4.1 Core functionality and understanding

Input comprehension and intent recognition

- *Focus*: The chatbot's ability to accurately interpret user inputs, encompassing natural language variations, implicit needs, complex or ambiguous queries, and underlying intent, including context maintenance across multi-turn conversations.
- *Objective*: Ensure the chatbot correctly comprehends user queries and their underlying intent to enable relevant and accurate responses, while effectively managing conversational context.
- *Key Determinant:* The architecture's sophistication in employing Natural Language Understanding, including techniques such as named entity recognition, intent classification, semantic analysis, and leveraging conversational history for contextual awareness and fallback handling.
- *Assessment Methods*: A/B testing with varied question complexity, human evaluation using predefined complex queries, automated testing with diverse input datasets, confusion matrix analysis, NLU score evaluation.

Response accuracy and factual integrity

- *Focus*: The chatbot's capacity to provide factually correct, current, and reliable information pertinent to the user's request and specific to the public service domain, including the consistent referencing of official sources.
- *Objective*: Ensure the factual accuracy and currency of chatbot responses, based on reliable data, official sources, and verifiable information.
- *Key Determinant*: The robustness and currency of the knowledge base, the integrity and completeness of rule logic,

and the quality and verification processes of the training datasets, including automated updating mechanisms.
- *Assessment Methods*: Manual expert review, automated verification against curated databases, error analysis, precision/recall metrics.

Consistency and reliability

- *Focus*: The chatbot's ability to deliver consistent, predictable, and reliable outputs across repeated or similar queries, ensuring stability under varying conditions and over time.
- *Objective*: Build trust by ensuring uniform responses for similar user inputs, minimizing confusion and guaranteeing uniform service delivery.
- *Key Determinant*: The architecture's ability to maintain response uniformity through structured data, predefined workflows, robust logic, version control, and quality control processes. This includes mechanisms to detect and address inconsistent behavior.
- *Assessment Methods*: Regression testing with predefined queries, stress testing under variable loads, statistical variance analysis, stability metrics, automated consistency scoring.

4.4.2 User experience and communication

Conversational fluency and contextual awareness

- *Focus*: The ability to produce responses that are fluent, grammatically correct, coherent, and contextually appropriate within ongoing dialogue, including the management of multi-turn conversations and conversational transitions.
- *Objective*: Foster user confidence and engagement through intuitive, natural, and fluid conversational interactions that effectively track conversation history and intent.
- *Key Determinant*: Reliance on advanced Natural Language Generation models that produce contextually relevant and adaptive responses, while also managing conversational memory and transitions.
- *Assessment Methods*: User surveys, human evaluation scales, conversational log analysis, multi-turn dialogue simulation.

Error handling and recovery

- *Focus*: The chatbot's ability to gracefully handle errors (e.g., misunderstandings, invalid inputs, technical issues) and guide users toward resolution or alternative paths, including effective communication when an answer cannot be found.
- *Objective*: Prevent user frustration and maintain smooth interactions through clear feedback, alternative suggestions, and robust recovery mechanisms.
- *Key Determinant*: The architecture's robustness in error detection and handling, supported by fallback strategies and recovery mechanisms.
- *Assessment Methods:* Simulated user tests with flawed inputs, error log analysis, fault injection tests, post-error user feedback, evaluation of fallback mechanisms.

Response timing and responsiveness

- *Focus*: The speed at which the chatbot provides responses, including consistency under varying conditions.
- *Objective*: Enhance user experience by ensuring timely answers, especially in high-pressure or emergency scenarios.
- *Key Determinant:* The efficiency of the underlying architecture in processing queries, data access, and response delivery with minimal latency.
- *Assessment Methods*: Automated timing measurements, load testing, throughput analysis, time-to-response distribution analysis, user perception studies.

4.4.3 Ethics and safety

Bias mitigation and fairness

- *Focus*: The chatbot's ability to avoid and mitigate biases in its outputs, ensuring equitable treatment of all users, regardless of demographic characteristics (e.g., race, gender, religion, etc.).
- *Objective*: Guarantee non-discriminatory service delivery and promote fairness, equity, and inclusion.
- *Key Determinant*: The architecture's inherent design elements that may introduce bias, alongside the practicability of implementation of bias detection mechanisms, debiasing techniques, and regular fairness audits.
- *Assessment Methods*: Bias audits of training data and outputs, A/B testing with diverse inputs, statistical bias analysis, fairness metric computations.

Ethical compliance and liability safety

- *Focus*: Avoidance of offensive or misleading outputs extending beyond factual accuracy to include the prevention of harm.
- *Objective*: Protect users and maintain ethical integrity and liability safety in public service delivery.
- *Key Determinant*: The nature of output generation (deterministic vs. probabilistic), combined with safeguards and compliance protocols embedded within the architecture.
- *Assessment Methods*: Legal compliance reviews, scenario-based liability simulations, red-teaming exercises and content filtering effectiveness analysis, documentation analysis, audit trail evaluation.

Data protection and privacy

- *Focus*: The inherent design characteristics and operational mechanisms of each chatbot architecture that contribute to specific vulnerabilities in, handling, storage, and processing of user data.
- *Objective*: Identify and mitigate unique vulnerabilities arising from each architecture's inherent design, ensuring user data is protected and handled responsibly.
- *Key Determinant*: Features that influence data breaches, leaks, or privacy violations, including data storage practices, external data dependencies, and the risk of unintended data regurgitation.

- *Assessment Methods*: Security audits, data flow analysis, penetration testing, privacy impact assessments, vulnerability scanning.

4.4.4 Adaptability and scalability

Domain versatility and integration

- *Focus*: The chatbot's capacity to operate effectively across multiple public service areas and integrate seamlessly with existing systems and databases.
- *Objective*: Ensure scalable performance across diverse domains and interoperability with public service infrastructures.
- *Key Determinant*: Flexibility in incorporating domain-specific datasets, workflows, APIs, and external systems, along with the ability to transition between domains.
- *Assessment Methods:* Pilot testing in new domains, API integration tests, cross-domain scenario simulations, case studies

Scalability and resource efficiency

- *Focus*: The ability to scale and manage large user volumes and complex requests, while maintaining high performance and efficient resource use.
- *Objective*: Deliver cost-effective service that avoids bottlenecks, optimizes resources, and minimizes environmental impact.
- *Key Determinant*: Efficient resource utilization, optimization strategies, and load balancing techniques that enable high concurrency without performance degradation..
- *Assessment Methods*: Load testing, resource monitoring, cost-benefit analysis, performance benchmarking, stress testing

Maintainability and update efficiency

- *Focus*: The ease with which the chatbot can be updated and maintained in response to evolving data, procedures, and legal changes, while minimizing effort, energy and need of specialized skills.
- *Objective*: Minimize operational overhead while keeping the system relevant, accurate, and compliant with evolving requirements and standards.
- *Key Determinant*: The architecture's modularity, ease of use in its design and adaptability to incremental updates, version control practices, and the presence of automated model retraining protocols, along with user-friendly maintenance interfaces.
- *Assessment Methods*: Code and documentation reviews, version control analysis, update frequency tracking, automated regression testing, system log analysis.

4.4.5 Inclusivity and service delivery

Multilingual support and accessibility

- *Focus*: The ability to process and respond to queries in multiple languages, ensuring equitable access for users with diverse abilities, literacy levels, cultural backgrounds and technological access.
- *Objective*: Ensure inclusivity and consistent service quality across various languages and interfaces.
- *Key Determinant*: Advanced NLP for multilingual support, adherence to accessibility standards (e.g., WCAG), and support for diverse interaction modes considering device and bandwidth constraints.
- *Assessment Methods*: Multilingual user testing, accessibility audits, cross-device testing, user surveys across diverse groups

Personalization and contextualization

- *Focus*: The chatbot's capacity to tailor interactions based on individual user needs, contexts, and preferences while maintaining privacy and fairness.
- *Objective*: Enhance service relevance and user satisfaction through dynamic, personalized interactions.
- *Key determinant*: Utilization of adaptive models, contextual data services, and user segmentation techniques that enable dynamic adjustment of responses.
- *Assessment methods*: A/B testing personalized versus generic responses, user studies, analysis of user logs, context retention testing.

Service delivery and process automation

- *Focus*: The ability to support structured workflows, integrate with other public service channels, and automate routine tasks, such as form filling and application submissions.
- *Objective*: Enhance efficiency and reduce administrative burdens in handling complex, multi-step public service procedures.
- *Key determinant*: The integration capabilities with external systems, API support, workflow management, and the robustness of process automation features.
- *Assessment methods*: User-centered design tests, workflow analysis, API integration testing, A/B testing of user interfaces, process completion rate measurement.

5 Results

This chapter presents the findings from the comparative evaluation of four chatbot architectures using our proposed framework. An overview of the results is provided in Table 1 and Figure 2, presenting the assigned score (ranging from Limited to Very-High) for each architectural profile and evaluation criterion. Note that for presentation purposes this table contains only the assigned score (Limited to Very-High) for each architecture and criterion combination.

The next section offers a summary of the results. Detailed explanations and justifications for each criterion category are then provided in subsequent sections. These analyses, organized across the five evaluation categories, highlight how different architectures perform and offer key policy recommendations for decision-makers in public administration. For each category, the detailed

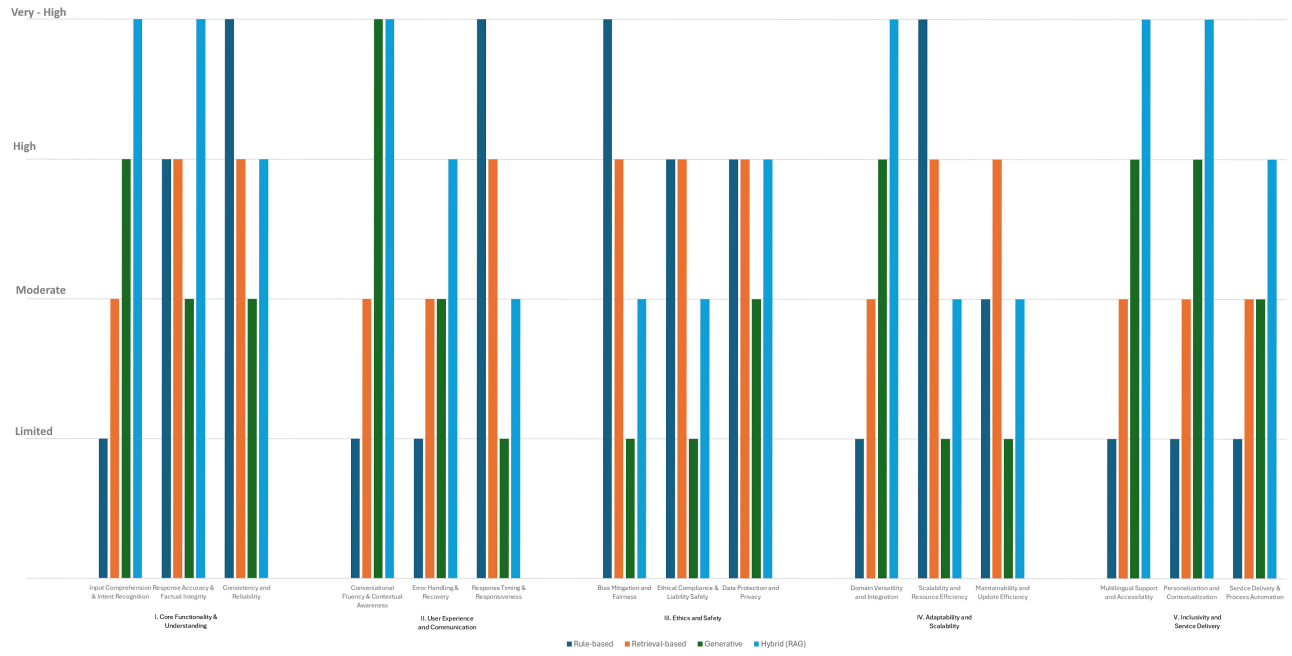

FIGURE 2
Results of the evaluation on the 15 criteria.

TABLE 2 Evaluation results for core functionality and understanding.

I. Core functionality and understanding				
Criterion	Rule-based	Retrieval-based	Generative	Hybrid (RAG)
Input comprehension and intent recognition	**Limited:** Rigid patterns limit understanding of varied or ambiguous inputs.	**Moderate:** Can process a wider range of inputs but relies on curated pre-existing data.	**High:** Handles varied inputs well, though occasional misinterpretations can occur.	**Very-High:** Combines fluent interaction with factual grounding for contextual relevance and input understanding.
Response accuracy and factual integrity	**High**: Provides accurate responses within a fixed scope.	**High**: Reliably retrieves content from controlled sources, minimizing risks of inaccuracies. Strong dependance on the quality and currency of the retrieval corpus	**Moderate**: Context-dependent; prone to hallucinations that can compromise factual correctness.	**Very-High**: Integrates verified retrieval with tailored, accurate responses improving accuracy overall.
Consistency and reliability	**Very-High**: Consistent and predictable due to predefined rules.	**High**: Generally reliable with minor variations due to data dependency.	**Moderate**: While contextually coherent, there is some probability of inconsistent outputs due to probabilistic nature.	**High**: Combines stable retrieval with flexible generation for improved consistency.

evaluation results are presented in both a table format (Tables 2–6) and through radar charts (Figures 3–7), where the length of each spoke is proportional to the magnitude of the assigned value (after quantification of discrete values and scaling up to the highest value).

5.1 Overview

Overall, the results, highlight that each architecture exhibits distinct strengths and limitations aligned with their inherent design principles. Rule-based systems demonstrate exceptional consistency and reliability due to their deterministic, predefined workflows, yet they are constrained by limited flexibility and contextual sensitivity. Retrieval-based chatbots offer enhanced factual accuracy by leveraging curated databases, though their performance is closely tied to the currency and comprehensiveness of their underlying data sources. Generative models excel in conversational fluency and contextual engagement, providing dynamic and human-like interactions; however, they are occasionally susceptible to errors and biases inherent in probabilistic outputs. Notably, the hybrid retrieval-augmented generation (RAG) approach emerged as the most balanced architecture. By integrating the robustness of retrieval-based

TABLE 3 Evaluation results for user experience and communication.

II. User experience and communication				
Criterion	Rule-based	Retrieval-based	Generative	Hybrid (RAG)
Conversational fluency and contextual awareness	**Limited:** Responses are static and lack natural flow, conversational richness and flexibility.	**Moderate:** Provides fluent responses but may miss nuanced contextual shifts.	**Very-High:** Generates engaging, fluid responses with effective context handling.	**Very-High:** Delivers both highly fluent and context-aware interactions by blending both approaches.
Error handling and recovery	**Limited**: Predefined fallback responses work but are limited and lack dynamic adaptability.	**Moderate**: Can handle some errors by retrieving alternative paths or providing fallback responses.	**Moderate**: Unpredictable outputs sometimes hinder effective error recovery.	**High**: Combines fallback mechanisms with the ability to re-query the knowledge base when an error occurs to provide a better answer.
Response timing and responsiveness	**Very-High**: Fast and efficient due to straightforward rule processing.	**High**: Response times can vary with data search and ranking delays.	**Limited**: Computational demands may result in occasional latency.	**Moderate**: Yields moderate response times because the NLG component only has to generate answers based on the corpus that is retrieved.

TABLE 4 Evaluation results for ethics and safety.

III. Ethics and safety				
Criterion	Rule-based	Retrieval-based	Generative	Hybrid (RAG)
Bias mitigation and fairness	**Very High:** Minimal risk of bias due to deterministic nature and reliance on explicitly defined rules. Inflexibility in adapting to diverse contexts may inadvertently perpetuate static biases.	**High:** Curated data and explicit controls promote better bias mitigation.	**Limited:** Greater risk of biases from training data without targeted debiasing.	**Moderate:** By grounding generated responses with contextually relevant, curated data, significantly reduce the propagation of biases. The generative component still carries some inherent risk for producing biases.
Ethical compliance and liability safety	**High**: Deterministic output enhances predictability and reduces liability risks. However, limited intent recognition capabilities may result in misleading responses.	**High**: Provides reliable, fact-based responses that support legal compliance. A high-quality dataset will allow safe performance.	**Limited**: Unpredictable outputs increase risks of misinformation and liability.	**Moderate**: While the retrieval component provides grounding, the generative aspect can still introduce unpredictability.
Data protection and privacy	**High**: Typically handle minimal user data, reducing the risk of data breaches and privacy violations. Data handling is deterministic.	**High**: Performance highly depends on whether the system collects and stores user data. Controlled data retrieval processes help protect user privacy.	**Moderate**: Complex model behavior can lead to higher risk of data leakage.	**High**: A secure retrieval component enhances security, but the generative component can introduce some privacy risks.

methods with the adaptability of generative models, the hybrid approach achieves superior performance across multiple evaluation criteria, including accuracy, scalability, and ethical safeguards.

In the subsequent sections, detailed analyses of individual criteria will further illuminate these findings and guide the formulation of targeted policy recommendations for the effective and responsible deployment of chatbot technologies in public service delivery.

5.2 Findings by evaluation dimension

5.2.1 Core functionality and understanding

The analysis of the results in this category reveals that rule-based systems offer high consistency due to their fixed logic and workflows but exhibit limitations in handling complex, ambiguous inputs, whereas retrieval-based systems improve factual accuracy by leveraging curated data; however, their effectiveness depends largely on the currency and completeness of their underlying knowledge bases. Generative systems show superior adaptability in interpreting diverse queries and maintaining context over multi-turn interactions, yet, their probabilistic output can lead to occasional inaccuracies or hallucinations. LLMs can produce consistent answers, particularly in the case of short and relatively simple queries. However, for more complex input, they are generally unable to produce the same answer to the same query over time. Hybrid chatbot architectures stand out by combining the strengths of retrieval and generative approaches—delivering enhanced input comprehension and reliable response generation with solid consistency.

Recommendations for core functionality:

- Adopt hybrid architectures for services requiring both adaptability and factual accuracy, particularly in complex administrative tasks.

TABLE 5 Evaluation results for adaptability and scalability.

IV. Adaptability and scalability				
Criterion	Rule-based	Retrieval-based	Generative	Hybrid (RAG)
Domain versatility and integration	**Limited:** Limited adaptability restricts performance across varied domains.	**Moderate:** Works effectively within curated domains but less adaptable to new contexts.	**High:** Generalizes well across diverse domains with sufficient training. They can be adapted to new domains by fine-tuning them on domain-specific data.	**Very-High:** Excels in integrating domain-specific data with flexible adaptation across diverse services. Integration of structured knowledge with dynamic response generation makes them highly suitable for multi-domain applications
Scalability and resource efficiency	**Very-High**: Lightweight design ensures high scalability and minimal resource use.	**High**: Generally efficient, though resource usage may increase with dataset size.	**Limited**: High computational requirements impact scalability and resource efficiency.	**Moderate**: Balances resource demands, offering moderate scalability compared to simpler architectures.
Maintainability and update efficiency	**Moderate**: Simple rule-based systems are easy to update and maintain but as complexity grows, managing large rule sets becomes increasingly cumbersome.	**High**: Updating knowledge bases is straightforward without need for altering the chatbot's core logic.	**Limited**: Complex retraining and updating processes hinder maintainability.	**Moderate**: Modular design allows for updates, though integration complexity adds moderate overhead.

TABLE 6 Evaluation results for inclusivity and service delivery.

V. Inclusivity and service delivery				
Criterion	Rule-based	Retrieval-based	Generative	Hybrid (RAG)
Multilingual support and accessibility	**Limited:** Predefined language scripts offer limited multilingual and accessibility support.	**Moderate:** Supports multiple languages when datasets are available, though flexibility is moderate.	**High:** Trained in diverse corpora, offering robust multilingual support and accessibility features.	**Very-High:** Integrates extensive multilingual capabilities with high accessibility enhanced by combining retrieval and generative capabilities for contextual accuracy.
Personalization and contextualization	**Limited**: Limited by static rules, offering minimal personalization.	**Moderate**: Provides some degree of personalization based on curated data.	**High**: Dynamically adapts responses based on user context for personalized interactions.	**Very-High:** Combines adaptive models with contextual analysis for advanced personalization.
Service delivery and process automation	**Limited**: Capable of supporting structured workflows with limited flexibility.	**Moderate**: Effective for automating routine tasks using retrieval and predefined logic However, their ability to handle complex workflows or integrate with external systems is limited.	**Moderate**: Supports basic process automation but may lack consistency in execution.	**High**: Integrates dynamic response generation with robust workflow automation for efficient process handling.

- Supplement rule-based systems with retrieval components to enhance flexibility without sacrificing consistency in areas where strict compliance is essential.
- Regularly update knowledge bases when deploying retrieval-based systems to maintain high accuracy and relevance.

5.2.2 User experience and communication

Findings in this category reveal that conversational quality and depth is highest in generative and hybrid systems, which produce fluid, natural language and maintain context over extended dialogues. But generative systems can be significantly slow, due to the computational demands of generating the text. In contrast, rule-based systems, though exceptionally fast, often yield simplistic, less adaptive exchanges that lack conversational depth, while retrieval-based systems provide moderate fluency, primarily due to reliance on pre-constructed responses. Error handling is more effective in hybrid systems, which integrate dynamic recovery mechanisms and fallback strategies, whereas generative models, while capable of generating creative error messages, may suffer from hallucinations and offer unreliable guidance. Rule-based systems demonstrate even lower adaptability when facing errors and unexpected inputs.

Overall, the results underscore that while rapid response is characteristic of rule-based systems, the superior engagement and adaptability of generative and hybrid models make them more suited for delivering a user-centric experience in public services.

Recommendations for user experience:

- Prioritize user-centric designs that integrate hybrid systems to enhance conversational quality and ensure dynamic error recovery.

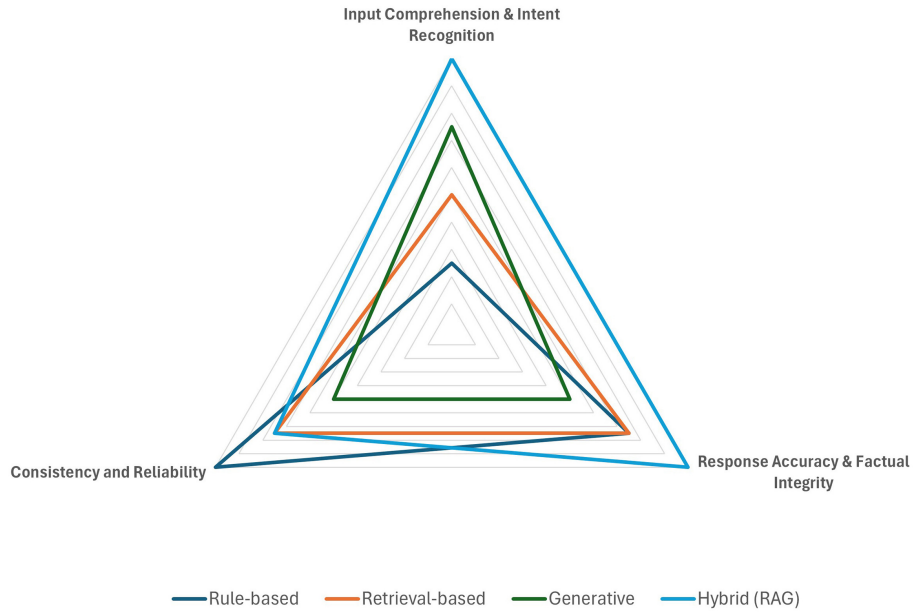

FIGURE 3
Core functionality and understanding.

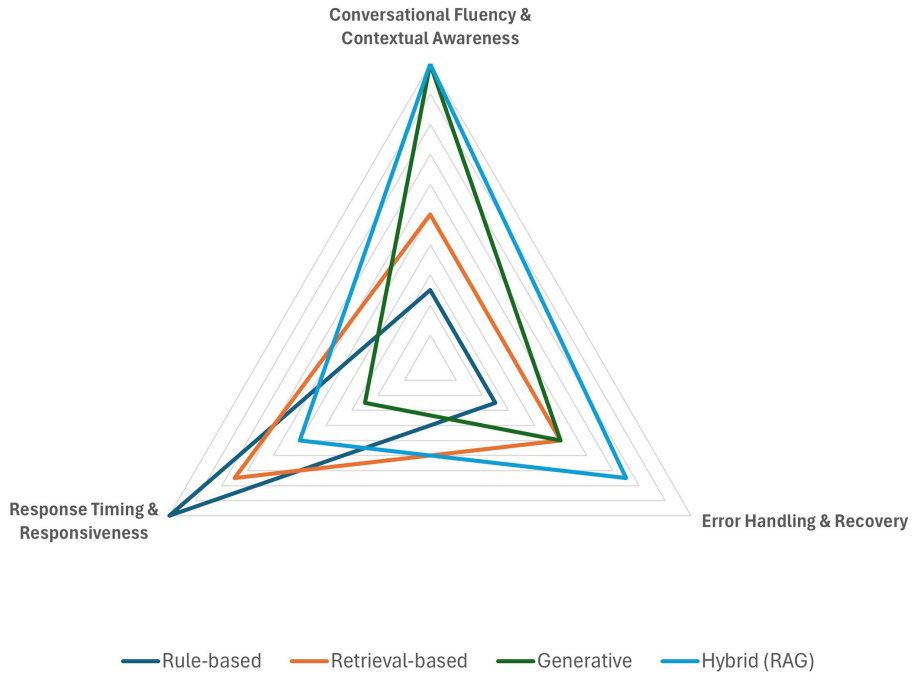

FIGURE 4
User experience and communication.

- Implement continuous user feedback mechanisms to refine error handling and response timing, especially in critical or emergency service scenarios.
- Tailor interface designs to specific public service contexts, ensuring that systems meet the unique interaction needs of diverse citizen groups.

5.2.3 Ethics and safety

Ethical compliance and safety are paramount, given the sensitive nature of public service interactions. The evaluation demonstrates that rule-based systems and retrieval-based models benefit from deterministic outputs and curated content, which minimizes risks of bias and unpredictable behavior.

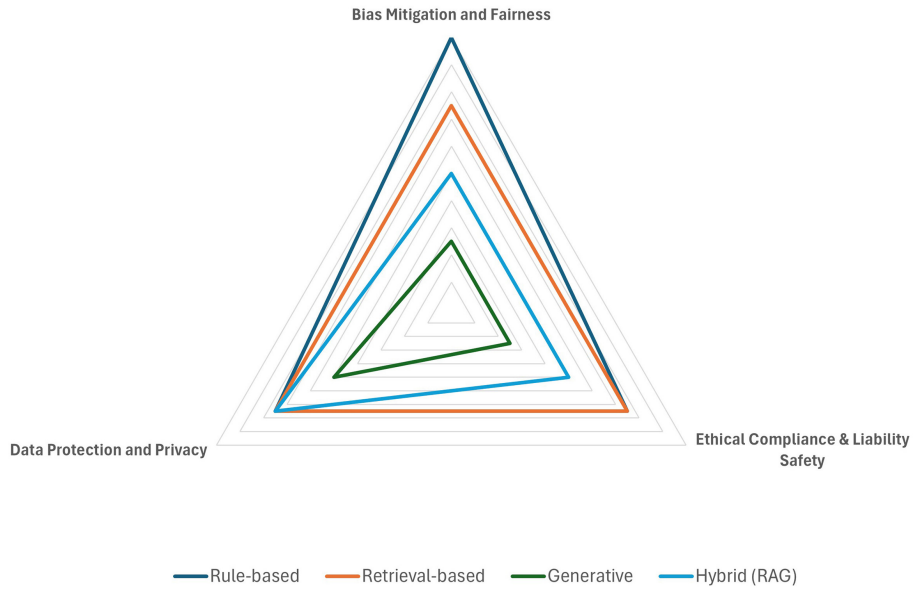

FIGURE 5
Ethics and safety.

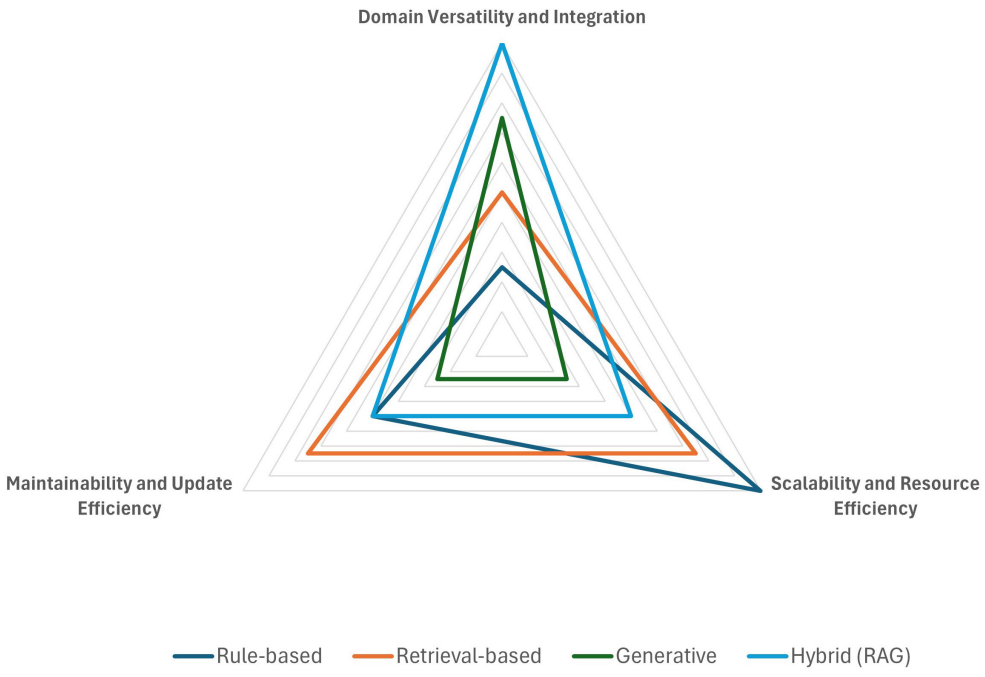

FIGURE 6
Adaptability and scalability.

However, limited intent recognition capabilities may result in misleading responses. Conversely, pure generative models tend to be more unpredictable, posing higher risks of bias, misinformation, and liability due to their probabilistic nature. Hybrid systems attempt to mitigate these challenges by grounding generative outputs with verified data but still require additional safeguards. Evaluation on data protection and privacy highlights that while rule-based, retrieval-based, and hybrid models can be designed to uphold strong privacy standards, pure generative systems require additional safeguards to mitigate data security risks.

Collectively, these findings reveal that traditional rule-based approaches are more effective at upholding ethical and safety requirements, and highlight the need for robust bias detection,

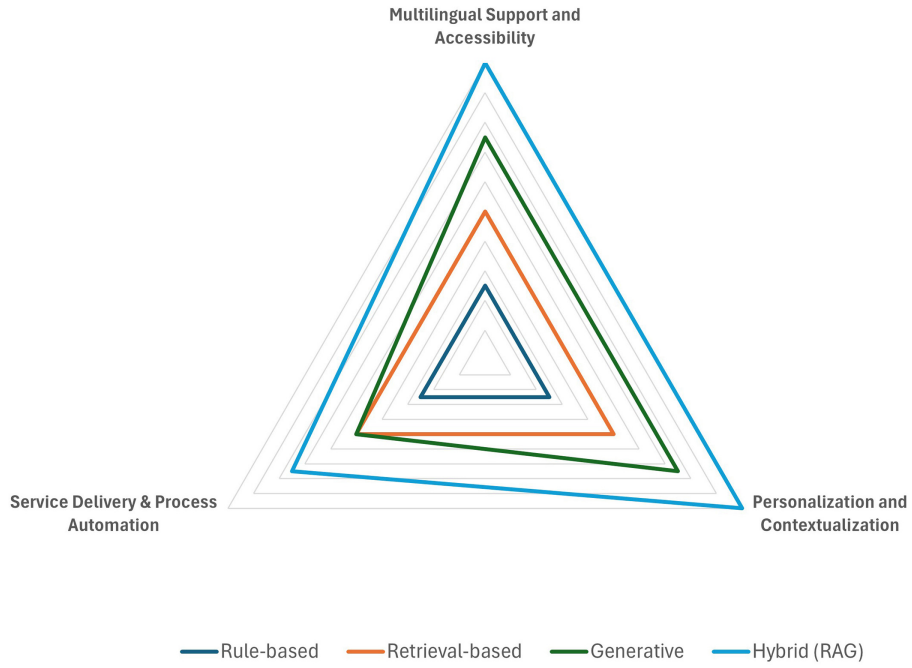

FIGURE 7
Inclusivity and service delivery.

continuous audits, and transparent safety measures to ensure that dynamic chatbot architectures meet public administration's ethical and legal standards. In every case, regular audits and strict adherence to established regulatory frameworks (e.g., GDPR, EU AI Act) are essential for maintaining ethical integrity.

Recommendations for ethics and safety:

- Implement robust bias mitigation strategies and regular audits across all architectures to ensure fairness and equity.
- Prefer rule-based solutions in high-risk applications or adopt hybrid solutions to balance adaptability with ethical safeguards.
- Establish clear legal and compliance frameworks that guide the deployment of chatbot technologies in sensitive areas such as healthcare and legal services.

5.2.4 Adaptability and scalability

Evaluation results indicates a trade-off between simplicity and flexibility. Rule-based systems, with their streamlined design, offer excellent scalability but suffer from limited adaptability across diverse domains. Retrieval-based models provide moderate versatility that is closely tied to the robustness of their knowledge bases. Generative systems excel in dynamic adaptation, demonstrating high domain generalization, yet their resource demands can hinder scalability. Fine-tuning or retraining LLMs is computationally expensive and time-consuming. Hybrid (RAG) architectures, by combining the responsive nature of generative methods with retrieval-based accuracy, achieve exceptional domain versatility and integration while balancing resource efficiency and maintainability, though periodic updates remain essential.

Recommendations for adaptability and scalability:

- Invest in hybrid architectures to ensure both high adaptability and scalable performance across multiple public service domains.
- Optimize resource allocation by integrating efficient load balancing and resource monitoring practices.
- Encourage continuous innovation and modular system design to facilitate timely updates and seamless integration with existing public service infrastructures.

5.2.5 Inclusivity and service delivery

Inclusivity and effective service delivery are fundamental to the mission of public administration. The analysis shows that advanced generative and hybrid systems outperform their rule-based and retrieval-based counterparts. These models demonstrate robust multilingual support, greater accessibility, and enhanced personalization by dynamically adapting to diverse user contexts. The hybrid approach leverages both flexible response generation and data grounding to enhance accessibility, streamline process automation, and foster equitable service delivery across various public service channels. In contrast, rule-based systems, with static response templates, and retrieval-based systems, limited by fixed datasets, tend to offer less adaptability and customization.

Overall, in terms of inclusivity and effective service delivery, advanced generative and hybrid systems are the preferred options. However, maximizing their benefits and meeting public service

mandates requires careful management of user data and adherence to accessibility standards.

Recommendations for inclusivity and service delivery:

- Prioritize the deployment of hybrid architectures to maximize inclusivity and offer comprehensive multilingual support.
- Incorporate advanced personalization features to tailor interactions to diverse user demographics and contexts.
- Ensure compliance with accessibility standards (e.g., WCAG) to guarantee that all users, regardless of ability or socio-economic status, can access public services efficiently.

6 Discussion and recommendations

The comparative analysis of chatbot architectures demonstrates that no single approach satisfies all public service requirements. Instead, each one has distinct advantages and limitations when applied to public service delivery and thus, decision-makers must adopt a tailored, domain-specific strategy that balances user experience, technical efficiency, ethical integrity, and regulatory compliance and aligns with the specific needs and criticality of the applied public service domain.

Rule-based systems, while offering high consistency, reliability, and ethical compliance, are limited in their ability to handle complex or nuanced user inputs and lack adaptability. Therefore, they are most appropriate for applications requiring strict determinism and low computational overhead, such as FAQs or basic regulatory information retrieval. At the other end of the spectrum, generative models, despite their performance in conversational fluency, contextual engagement and adaptability, should be deployed with caution in high-stakes domains due to their inherent unpredictability and their inherited safety risks. Retrieval-based architectures are less risky, and present improved factual accuracy by relying on curated data, but their performance is closely tied to the quality, currency and relevance of the underlying datasets. Hybrid architecture emerges as the optimal solution for many public service applications. Their balanced performance across evaluation categories —particularly in terms of input comprehension, response accuracy, ethical safeguards, and scalability—suggests that under a context-dependent selection strategy, hybrid models are particularly well-suited for complex and high-stakes public sector applications.

6.1 Recommendations for architecture selection and deployment

Based on these findings, the following key policy recommendations emerge for public administrations considering chatbot selection and deployment:

6.1.1 Adopt a context-dependent, risk-based approach

As highlighted throughout the evaluation, no single architecture is universally optimal. Public administrations should adopt a context-dependent, domain-specific and risk-based approach when selecting chatbot architectures. The decision should be made by a thorough assessment of:

- **Service complexity**: simple information retrieval tasks can be conveniently served by rule-based or retrieval-based systems, while complex administrative processes or nuanced citizen interactions often necessitate hybrid architectures.
- **User interaction needs**: the expected nature of user interactions is a defining factor that influences the need for fluency and contextual understanding. For open-ended and conversational dialogs, the use of a solution incorporating generative components is needed. Structured and predictable conversational flows can be served by classical approaches.
- **Domain-specific risks**: the criticality and sensitivity of the service domain are paramount. High-stakes domains (e.g., legal advice, healthcare interactions, benefits determination) demand higher levels of accuracy, reliability, and ethical safeguards, often favoring more controllable architectures or heavily scrutinized hybrid systems. Lower-risk domains (e.g., general FAQs, community event information) might allow for greater flexibility.
- **Data sensitivity**: the type of data the chatbot will process dictates security and privacy requirements, influencing the choice between cloud, on-premises, or edge deployments, which in turn, can affect the selection of the architecture.

6.1.2 Prioritize accuracy and reliability in high-stakes domains

Where factual accuracy, consistency, and reliability are non-negotiable (e.g., providing legal information, details on eligibility criteria, emergency instructions), architectures that offer greater control over outputs should be prioritized.

- **Classical approaches for determinism:** for simpler, high-risk applications requiring strict determinism and minimal ambiguity, classical rule-based or retrieval-based systems remain the most prudent choice. Their predictable nature minimizes the risk of factual errors or unpredictable behavior inherent in probabilistic models. Rule-based solutions are suitable for highly structured tasks, with a clearly defined domain and predictable interactions—such as procedural guidance or FAQs—where consistency and ease of audit are paramount. Retrieval-based architectures excel when a well-maintained, curated knowledge base is available, and are particularly suitable for public services where verified, standardized information is essential—such as public service catalogs or frequently asked questions.
- **Hybrid (RAG) for grounded generation:** when generative capabilities are desired for enhanced user experience in high-stakes areas, hybrid (RAG) architectures are strongly preferred over purely generative models. The retrieval component grounds the generated output in verified knowledge sources, significantly reducing the risk of hallucinations and enhancing factual integrity. Regular updating and auditing of the retrieval knowledge base are essential.

6.1.3 Leverage hybrid architectures for balanced performance

For a wide range of public service applications that involve complex queries, nuanced language, or dynamic information, but where absolute determinism is not strictly required, hybrid (RAG) architectures offer the most compelling balance. They combine the enhanced comprehension and conversational fluency of generative models with the factual grounding and reliability of retrieval systems, addressing key criteria across functionality, user experience, and safety. In multi-data environments investment in robust RAG systems should be encouraged.

6.1.4 Carefully consider generative model deployment options

When deploying chatbots incorporating generative components (standalone or hybrid), the choice of model type and deployment method carries significant implications. A more detailed explanation of the available options is given below.

6.2 Special considerations for generative AI deployment

When public service contexts demand open-ended, creative, and context-aware chatbot interactions, generative AI components, particularly Large Language Models (LLMs), are often essential. Implementing these components involves critical decisions regarding model selection, training methodology, and deployment environment, each presenting distinct trade-offs. Decision-makers must carefully weigh options such as developing proprietary models from scratch vs. fine-tuning existing pre-trained models and deploying systems via cloud services vs. local infrastructure.

6.2.1 Selection of model type

Public sector bodies should carefully evaluate their resources and needs before committing to a particular model. Developing proprietary LLMs, while offering maximum control, typically requires both substantial financial investment and a high level of expertise, potentially exceeding the capacity of many administrations. Most modern popular and powerful LLMs (like OpenAI's GPT, Google's Gemini Pro, Anthropic's Claude 3) are "closed-weight." This means the companies that created them do not publicly release their actual numerical weights.[3] Users can usually only interact with these models through an API, sending requests to the company's servers. "Open-weight" models on the other hand, are those whose trained parameters (weights) are publicly released, allowing anyone to download, run, and study them. But simply releasing a model's weights while keeping training methodology and data proprietary is not enough for the model to be truly considered open source and limits transparency, inspectability, reproducibility, and customization. In contrast, releasing a model as open source would entail providing the full source code and information required for retraining the model from scratch. This includes the model code, training methodology and hyperparameters, the original training dataset, documentation, and other relevant details.

In the case of public bodies, "open-weight" models which are also released under a permissive license (e.g., Apache 2.0, MIT and various BSD licenses) alongside their code and any other important details, is the best approach for ensuring safety, transparency and accountability.

6.2.2 Training options

When considering the issue of training a large language model in the context of specific domai, there are again different options. The most common approaches include the training of closed-models or the finetuning of existing models. While the terms are related and often overlap in practice, there's a key distinction between a closed-domain generative LLM and the process of fine-tuning an LLM: a closed-domain system is defined by its restricted scope, while fine-tuning an LLM is a common method used to improve the performance or adapt an LLM to specific tasks. Closed-models relate to the development of a proprietary LLM from scratch, while fine-tuning starts with a pre-trained LLM (often provided by a cloud platform or an open-source model) and is further trained using proprietary domain-specific datasets.

Developing proprietary closed-domain models trained from scratch on data owned by the public body is the preferred option for applications requiring simultaneously high safety, accuracy and reliability, such as in legal, medical, or technical domains. But recent research (e.g., Bommasani et al., 2021) indicates that despite decreasing costs, the required computational resources and specialized expertise continue to present significant barriers for many public administrations.

The alternative of fine-tuning pre-trained models provided by major LLM vendors is a simpler strategy, that can yield domain-specific accuracy without the need for excessively large computational and expertise requirements. However, these should be "open-weight" models, to be able to fine-tune them on specific datasets and adapt for tasks or domains. In addition, the level of safety provided by this option in terms of information accuracy is comparatively lower than that achieved by proprietary developed closed-models.

6.2.3 Deployment options

Using commercial cloud platforms (e.g., Microsoft Azure AI, Google Cloud AI Platform, AWS SageMaker) to deploy chatbots is a typical and relatively easy approach. These platforms provide comprehensive, end-to-end services for hosting and accessing LLMs and often include tools for fine-tuning. This takes a huge weight off the shoulders of public organizations when it comes to managing infrastructure, scaling, and maintenance, allowing them to focus on the design and development of the chatbot. However, using cloud infrastructure raises concerns about data sovereignty, residency, and security, especially when processing sensitive public sector information. Furthermore, it incurs significant, recurring

3 When a large language model is trained, it learns patterns and information from vast amounts of data. This learned knowledge is stored as billions of numerical parameters (called weights), which essentially constitute the trained model.

operational costs during the lifespan of a deployed model, and can potentially lead to vendor lock-in.

An alternative approach is on-premised deployment using servers in the organization's data centers or in a dedicated private cloud. This gives the organization greater control over the data security and compliance that is crucial when handling sensitive information. But it also requires substantial infrastructure investment and a high level of in-house technical expertise to manage the systems and components effectively.

Combining elements of on-premises/private cloud infrastructure with public cloud services offers a flexible middle ground. This hybrid approach allows organizations to strategically balance scalability, cost, and control. For instance, sensitive data processing or core model components could reside on-premises, while leveraging the public cloud's computational power for less critical tasks, model training, or handling peak loads. While offering adaptability, this model introduces greater architectural complexity and necessitates robust integration management.

Another viable strategy involves deploying edge models that run locally on user devices, thereby enhancing data privacy and reducing latency. Nonetheless, these models typically offer limited functionality and may amplify biases through compression techniques (Hendrycks et al., 2020).

6.3 Organizational and governing principles and recommendations

Selecting the right technology is only half the job when it comes to successfully deploying chatbots in public contexts. The design of technological systems, such as conversational AI agents, should be a process that considers both social, organizational and technical factors that influence their operation and usage. A sociotechnological approach, recognizing chatbots not simply as technical tools but also as mediation agents embedded within complex social, organizational, and political contexts, is essential for public institutions in order to enhance digital service delivery, foster citizen trust, and ensure that AI technologies contribute constructively to public governance. This approach implies both organizational requirements within the public administration and a broadening of the design scope to include social and ethical components.

6.3.1 Organizational principles

1. **Organizational readiness**: Adequate training and support should be provided to public sector employees who will be interacting with or managing chatbot systems. This encompasses re-skilling and targeted training programs in digital literacy and AI management. Adaption or redesign of current workflows is also an essential requirement to effectively harness the full potential of these technologies.
2. **Implement robust data governance and bias mitigation**: Regardless of the chosen architecture, rigorous data governance practices are essential. This comprises continuous update, careful curation and auditing of knowledge bases, implementation of privacy-enhancing technologies, and continuous monitoring for bias in training data and model outputs.
3. **Ensure transparency and explainability**: Public administrations should prioritize transparency in chatbot deployments, clearly informing users when they are interacting with an AI system and how decisions or recommendations are reached. This is particularly important for generative and hybrid models, where the reasoning process can be opaque.
4. **Establish clear lines of accountability**: Mechanisms for human oversight and intervention should be considered when designing chatbot systems. Providing clear pathways for users to escalate issues or challenge automated decisions, helps to ensure accountability and build public trust.
5. **Invest in ongoing monitoring and evaluation**: Continuous monitoring of chatbot performance, including accuracy, user satisfaction, and ethical compliance, is crucial. Periodic audits and evaluations should be conducted to help identify areas for improvement and adapt to evolving societal needs and technological advancements.
6. **Foster Inter-agency collaboration**: Public administrations should actively promote and invest in collaborative initiatives between different agencies and levels of government focused on chatbot deployment. Common evaluation standards and best practices should be established, to ensure interoperability and encourage innovation. Establishing collaborative platforms for sharing deployment experiences, challenges, lessons learned, and potentially even curated datasets or code can accelerate innovation, optimize resource allocation across the public sector, and build crucial collective expertise in managing AI technologies responsibly and effectively.
7. **Adopt an AI risk management strategy**: Given the multifaceted risks inherent in chatbot technologies (Cortés-Cediel et al., 2023), a comprehensive and rigorous risk management strategy, drawing upon established international frameworks such as the OECD AI Principles (OECD, 2019) and the NIST AI Risk Management Framework (NIST, 2023) is essential. This strategy must integrate proactive technical audits (to assess performance, security, and reliability), thorough legal reviews (to ensure compliance with all relevant laws and regulations), and critical ethical assessments (to evaluate alignment with public values, fairness principles, potential societal impacts, and fundamental rights).

6.3.2 Governance principles

The furious pace of technological progress, due to centralization, raises powerful concerns that demand the attention of humanists and social scientists in addition to technologists. Public administrations should not rely on *post-hoc* audits of ethical and social consequences, conducted only after the technical architecture and deployment decisions have been made. Instead, there is need to infuse social considerations and ethical design principles deeply into the technological development of governmental services and their surrounding ecosystem from the start. Such an approach emphasizes the following principles.

1. Prioritize user-centricity and human agency: design chatbot systems essentially around the needs, rights, preferences, and

diverse capabilities of citizens. This goes beyond ensuring intuitive interfaces, multilingual support, and accessibility. It requires conscious consideration of how technology mediates the citizen-state relationship and potentially shapes notions of citizenship (Jasanoff, 2016). Design choices must actively support, rather than inadvertently undermine, human agency. Especially in complex, sensitive, or emotionally charged situations where empathy and nuanced understanding are most important, clear paths for smooth escalation to human representatives must be freely available and easily accessible.

2. Implement inclusive and diverse stakeholder engagement: throughout the whole chatbot lifecycle—from initial conception and design to deployment and continuous evaluation—engage with a wide range of stakeholders in a proactive, inclusive, and continuous manner. This includes not only citizens (especially those who may be marginalized or at risk), civil society organizations, and domain experts, but also public sector employees who will interact with the system, ethicists, social scientists, and technologists. This engagement should aim not only to gather functional requirements but also to understand the complex social, cultural, institutional context, and power relations within which the chatbot will operate.
3. Embed continuous sociotechnical learning and adaptation: treat chatbot deployment as a continuous process of learning and adaptation within a dynamic sociotechnical context rather than as a static, one-time implementation. Establish robust, transparent mechanisms for collecting and systematically responding to user feedback, continuously monitoring system performance (including accuracy, fairness, bias metrics, and user satisfaction), tracking relevant technological advancements, and proactively reevaluating alignment with changing social norms, legal requirements, and ethical standards. This iterative approach allows for proactive adjustments and mitigation of unforeseen negative consequences, ensuring the chatbot remains effective, equitable, and consistent with democratic values and public service principles.

7 Limitations and future research

This research presented a comprehensive framework for evaluating chatbot architectures in the context of public service delivery, yet several limitations merit consideration. The evaluation framework relies on predefined criteria that, despite careful selection, may not capture the complete spectrum of diverse real-world applications. Domain-specific challenges in public administration may require further refinement of certain criteria to better reflect operational realities to specific domains, such as healthcare, education, or law enforcement. Moreover, although the criteria are designed for universal applicability, individual systems may include design details or unique approaches, not fully captured by the framework's generalized ratings.

Another limitation lies in the reliance on theoretical analysis and simulated scenarios for independent criteria evaluations, which introduces a degree of subjectivity in the results. The assignment of qualitative values (Limited to Very High) is based on expert judgment and the interpretation of available evidence and is using an average "typical" system as a baseline reference for each chatbot architecture, ignoring variances, sub-categories or specialized implementations. Subsequent research should explore the development of more specific, objective and quantitative metrics to enhance the assessment of chatbot performance.

The dynamic nature of AI and chatbot technologies also presents an inherent challenge. Rapid advancements in natural language processing and chatbot architectures may render parts of the evaluation framework obsolete unless continuously updated. This also applies to emerging ethical and safety concerns, which require ongoing reassessment in light of evolving regulatory standards and societal expectations.

Furthermore, the proposed framework is primarily tailored to text-based chatbots, thereby limiting its applicability to multimodal AI systems that incorporate voice, image, or video. Such systems can introduce additional challenges, unaddressed by the framework's text-centric evaluation, including technical aspects like speech recognition accuracy (e.g., handling dialects, ambient noise) and visual processing reliability (e.g., object detection in uploaded images), as well as additional ethical risks, such as privacy breaches from biometric data (e.g., voiceprints) or biases in image recognition algorithms. This gap can lead to systematic misrepresentations of performance in real scenarios, underscoring the need for expanded criteria to address modality-specific technical, ethical, and operational demands.

Looking ahead, future developments can focus on several key improvements. Empirical validation through user studies and field deployments would provide richer insights and help calibrate the framework more precisely. Developing and including more objective and quantitative metrics for evaluating performance, such as bias and natural language generation quality metrics, user satisfaction scores, and explainability ratings, could improve the usefulness and applicability of the framework. Enhancing the granularity of the evaluation criteria, especially for ethical and safety aspects could also improve applicability across different domains. Additionally, expanding the framework to address emerging challenges such as multimodal interactions, privacy-preserving techniques, and explainability in AI could significantly bolster its relevance in the rapidly evolving landscape of public chatbot deployment.

8 Conclusion

This research has presented a framework for evaluating chatbot architectures in the context of public service delivery, emphasizing the critical need to balance technological capabilities with ethical considerations and the core values of public administration. Our analysis of rule-based, retrieval-based, generative, and hybrid architectures across fifteen criteria, categorized into five key dimensions, demonstrates that no single architecture is universally superior. Instead, the optimal choice depends on the specific application, the resources available, and the ethical and operational priorities of the deploying organization.

The findings highlight the strengths and limitations of each approach, suggesting that for many public service scenarios, a hybrid approach represents the most promising path forward.

However, the selection of appropriate architecture is only the first step. Successfully deploying chatbots in public administration and effectively managing risk requires a holistic, sociotechnological approach. This involves prioritizing user-centered design, engaging stakeholders, ensuring organizational readiness, and establishing robust mechanisms for data governance and bias mitigation. Continuous monitoring, evaluation, and adaptation are essential to ensure that chatbot systems remain aligned with evolving needs, technological advancements, and ethical standards.

Ultimately, the aim of integrating AI-driven chatbots into public service should not be simply to automate tasks or reduce costs, but to enhance the quality, accessibility, and responsiveness of government services, fostering greater citizen engagement and trust. By adopting a thoughtful, ethical, and citizen-centered approach, public administrations can harness the transformative potential of chatbot technology while upholding the fundamental principles of good governance. Still, the ongoing evolution of AI demands a continuous and critical assessment of its implications for the public sector, ensuring that these powerful technologies are deployed in a way that serves the public good. Continued research and interdisciplinary collaboration will be essential to advance state-of-the-art and address the evolving challenges in this dynamic field.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

TP: Writing – original draft, Visualization, Conceptualization, Writing – review & editing, Methodology, Investigation. CA: Writing – review & editing. YC: Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abbas, N., Følstad, A., and Bjørkli, C. A. (2023). "Chatbots as part of digital government service provision - a user perspective," in *Chatbot Research and Design. CONVERSATIONS 2022. Lecture Notes in Computer Science, vol 13815*, ed. A. Følstad, et al. (Cham: Springer). doi: 10.1007/978-3-031-25581-6_5

Ammann, L., Ott, S., Landolt, C. R., and Lehmann, M. P. (2025). Securing rag: a risk assessment and mitigation framework. *arXiv*. doi: 10.1109/SDS66131.2025.00024

Androutsopoulou, A., Karacapilidis, N., Loukis, E., and Charalabidis, Y. (2019). Transforming the communication between citizens and government through AI-guided chatbots. *Gov. Inf. Q.* 36, 358–367. doi: 10.1016/j.giq.2018.10.001

Aoki, N. (2020). An experimental study of public trust in AI chatbots in the public sector. *Gov. Inf. Q.* 37:101490. doi: 10.1016/j.giq.2020.101490

Biggio, B., and Roli, F. (2018). "Wild patterns: ten years after the rise of adversarial machine learning," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY: Association for Computing Machinery). 2154–2156. doi: 10.1145/3243734.3264418

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Sydney, V. A., et al. (2021). On the opportunities and risks of foundation models. *arXiv*. doi: 10.48550/arXiv.2108.07258

Bulatov, A., Kuratov, Y., Kapushev, Y., and Burtsev, M. (2024). "Beyond attention: breaking the limits of transformer context length with recurrent memory," in *Proceedings of the AAAI Conference on Artificial Intelligence* (Palo Alto, CA: AAAI Press), 38, 17700–17708. doi: 10.1609/aaai.v38i16.29722

Chen, T., Gascó-Hernandez, M., and Esteve, M. (2024). The adoption and implementation of artificial intelligence chatbots in public organizations: evidence from U.S. state governments. *Am. Rev. Public Adm.* 54, 3–19. doi: 10.1177/02750740231200522

Chu, Z., Wang, Z., and Zhang, W. (2024). Fairness in large language models: a taxonomic survey. *ACM SIGKDD Explorations Newsletter*, 26, 34–48. doi: 10.1145/3682112.3682117

Civil Resolution Tribunal, C. R. T. (2024). *Decision Moffatt v. Air Canada, 2024 BCCRT 149*. Available online at: https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do (Accessed March 15, 2025).

Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D. (2020). "ELECTRA: Pre-training text as discriminators rather than generators," in *8th International Conference on Learning Representations (ICLR 2020)* (Addis Ababa).

Cortés-Cediel, M. E., Segura-Tinoco, A., Cantador, I., and Rodríguez Bolívar, M. P. (2023). Trends and challenges of e-government chatbots: advances in exploring open government data and citizen participation content. *Gov. Inf. Q.* 40:101877. doi: 10.1016/j.giq.2023.101877

Cui, T., Wang, Y., Fu, C., Xiao, Y., Li, S., Deng, X., et al. (2024). Risk taxonomy, mitigation, and assessment benchmarks of large language model systems. *arXiv*. doi: 10.48550/arXiv.2401.05778

Dahl, M., Magesh, V., Suzgun, M., and Daniel E. Ho, (2024). Large legal fictions: profiling legal hallucinations in large language models. *J. Legal Anal.* 16, 64–93. doi: 10.1093/jla/laae003

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1* (Minneapolis, MN), 4171–4186.

EC (2019). *European Commission: Ethics Guidelines for Trustworthy AI*. Brussels: European Commission's High-Level Expert Group on AI.

Felzmann, H., Villaronga, E. F., Lutz, C., and Tamò-Larrieux, A. (2019). Transparency you can trust: transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data Soc.* 6. doi: 10.1177/2053951719860542

Gan, Y., Yang, Y., Ma, Z., He, P., Zeng, R., Wang, Y., et al. (2024). Navigating the risks: a survey of security, privacy, and ethics threats in LLM-based agents. *arXiv*. doi: 10.48550/arXiv.2411.09523

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M., et al. (2023). "Not what you've signed up for: compromising real-world llm-integrated applications with indirect prompt injection," in *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (New York, NY: Association for Computing Machinery). 79–90. doi: 10.1145/3605764.3623985

Gupta, S., Ranjan, R., and Singh, S. N. (2025). Comprehensive framework for evaluating conversational AI chatbots. *arXiv*. doi: 10.48550/arXiv.2502.06105

Helsper, E. J. (2021). *The Digital Disconnect: The Social Causes and Consequences of Digital Inequalities*. London: Sage. doi: 10.4135/9781526492982

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., et al. (2020). Aligning AI with shared human values. *arXiv*. doi: 10.48550/arXiv.2008.02275

Hohmann, B. (2021). Interpretation the concept of transparency in the strategic and legislative documents of major intergovernmental organizations. *Közigazgatási és Infokommunikációs Jogi PhD Tanulmányok* 2, 48–68. doi: 10.47272/KIKPhD.2021.1.4

Jasanoff, S. (2016). *The Ethics of Invention: Technology and the Human Future*. New York, NY: W. W. Norton & Company.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems, Vol. 33*, eds H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Curran Associates, Inc.), 9459–9474. Available online at: https://arxiv.org/abs/2005.11401

Lippens, L. (2024). Computer says 'no': exploring systemic bias in ChatGPT using an audit approach. *Comp. Hum. Behav. Artif. Hum.* 2:100054. doi: 10.1016/j.chbah.2024.100054

Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., et al. (2023). Jailbreaking chatgpt via prompt engineering: an empirical study. *arXiv*. doi: 10.1145/3663530.3665021

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv*. doi: 10.48550/arXiv.1907.11692

Maclure, J., and Morin-Martel, A. (2025). AI ethics' institutional turn. *Digit. Soc.* 4:18. doi: 10.1007/s44206-025-00174-x

Ma'rup, M., Tobirin, and Ali Rokhman. (2024). Utilization of artificial intelligence (AI) chatbots in improving public services: a meta-analysis study. *Open Access Indonesia J. Soc. Sci.* 7, 1610–1618. doi: 10.37275/oaijss.v7i4.255

Mechkaroska, D., Domazet, E., Feta, A., and Shikoska, U. R. (2024). "Architectural scalability of conversational chatbot: the case of ChatGPT," in *Advances in Information and Communication. FICC 2024. Lecture Notes in Networks and Systems, vol 919*, ed. K. Arai (Cham: Springer). doi: 10.1007/978-3-031-53960-2_5

NIST (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. doi: 10.6028/NIST.AI.100-1.jpn

OECD (2019). *Recommendation of the Council on Artificial Intelligence, OECD/LEGAL/0449*. Paris: OECD.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.

OpenAI (2023). *March 20 ChatGPT Outage: Here's What Happened*. San Francisco, CA: OpenAI.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training* (San Francisco, CA).

Smuha, N. A. (2025). "The use of algorithmic systems by public administrations: practices, challenges and governance frameworks," in *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, ed. N. A. Smuha (Cambridge: Cambridge University Press), 383–410. doi: 10.1017/9781009367783.023

Smuha, N. A., and Yeung, K. (2025). "The European Union's AI act: beyond motherhood and apple pie?," in *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*, ed. N. A. Smuha (Cambridge: Cambridge University Press), 228–258. doi: 10.1017/9781009367783.015

UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO.

United States District Court, S.D (2023). *New York. Sanctions Order, Mata v. Avianca Inc., June 22, 2023, 22-cv-1461 (PKC)*. Available online at: https://cases.justia.com/federal/district-courts/new-york/nysdce/1:2022cv01461/575368/54/0.pdf?ts=1687525481 (Accessed March 15, 2025).

Van Noordt, C., and Misuraca, G. (2022). Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Gov. Inf. Q.* 39:101714. doi: 10.1016/j.giq.2022.101714

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17)* (Red Hook, NY: Curran Associates Inc.), 6000–6010.

Vemulapalli, V. K. C. (2025). Enterprise generative AI chatbot architecture: from natural language understanding to scalable deployment. *J. Comp. Sci. Technol. Stud.* 7, 668–678. doi: 10.32996/jcsts.2025.7.7.75

Weizenbaum, J. (1966). ELIZA-a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 36–45. doi: 10.1145/365153.365168

Wirtz, B. W., Weyerer, J. C., and Geyer, C. (2018). Artificial intelligence and the public sector-Applications and challenges. *Int. J. Public Adm.* 42, 596–615. doi: 10.1080/01900692.2018.1498103

Wu, Y., Rabe, M. N., Hutchins, D., and Szegedy, C. (2022). Memorizing transformers. *arXiv*. doi: 10.48550/arXiv.2203.08913

Ye, W., Ou, M., Li, T., Chen, Y., Ma, X., Yanggong, Y., et al. (2023). "Assessing hidden risks of LLMs: An empirical study on robustness, consistency, and credibility," in *Findings of the Association for Computational Linguistics: ACL 2023* (Association for Computational Linguistics), 15001–15018. doi: 10.18653/v1/2023.findings-acl.953

Yu, M., Meng, F., Zhou, X., Wang, S., Mao, J., Pang, L., et al. (2025). A survey on trustworthy LLM agents: threats and countermeasures. *arXiv*. doi: 10.1145/3711896.3736561

Zhou, M., Liu, L., and Feng, Y. (2025), Building citizen trust to enhance satisfaction in digital publicservices: the role of empathetic chatbot communication. *Behav. Inf. Technol.* 44, 1–20. doi: 10.1080/0144929X.2025.2451763

Zhuo, T. Y., Huang, Y., Chen, C., and Xing, Z. (2023). Red teaming chatgpt via jailbreaking: bias, robustness, reliability and toxicity. *arXiv*. doi: 10.48550/arXiv.2301.12867