

Improving Access to Public Services through an Intelligent Assistant with RAG: Experimental Evidence in a Local Government of Lima

Jhair Collado¹, Kevin Tupac-Agüero¹, Fabrizio Del Carpio Delgado²,
Javier Gamboa-Cruzado¹, Mario Rodríguez Vasquez², Germán Machicado Lea²,
Ángel Núñez Meza³, Paola Del Pilar Condori Luis², Bella Citlali Martínez-Seis^{4,*}

¹ Universidad Nacional Mayor de San Marcos, Lima,
Peru

² Universidad Nacional de Moquegua, Moquegua,
Peru

³ Universidad Nacional Daniel Alcides Carrión, Pasco,
Peru

⁴ Instituto Politécnico Nacional, UPIITA, Mexico City,
Mexico

jhair.collado@unmsm.edu.pe, kevin.tupac@unmsm.edu.pe, delcarpiod@unam.edu.pe,
jgamboac@unmsm.edu.pe, mrodriguezv@unam.edu.pe, gmachicadol@unam.edu.pe,
anunezm@undac.edu.pe, pcondoril@unam.edu.pe, bmartinez@ipn.mx

Abstract. Timely and reliable access to information on public services remains a challenge in municipalities where traditional channels predominate, with restrictions in capacity, availability, and consistency of responses. The objective of this study was to develop and implement an Intelligent Assistant based on Retrieval-Augmented Generation (RAG) to optimize access to information and the management of queries on public services in Carabayllo, Lima. An applied study was conducted using an experimental post-test design with a control group and an experimental group, considering 60 observations distributed equally. The system was developed using Scrum and evaluated through four indicators: number of queries, accuracy of responses, waiting time, and user satisfaction level. The analysis included descriptive statistics, the Anderson-Darling test, t-tests for independent samples, and the Mann-Whitney U test. The results revealed statistically significant differences in favor of the experimental group across the four indicators: greater service capacity, higher response accuracy, shorter waiting time, and greater citizen satisfaction. Taken together, the findings show that incorporating RAG can improve operational and perceptual dimensions of municipal service. It is concluded that this architecture constitutes an effective

alternative for modernizing access to public information, although its performance depends on the quality of the corpus, the retrieval of context, and the implementation conditions.

Keywords. Retrieval-Augmented Generation, RAG, access to information, local governments, SCRUM methodology, LLMs.

1 Introduction

Currently, access to information and queries about services in the District of Carabayllo is managed predominantly through traditional channels, such as in-person offices and telephone lines.

This approach is limited and inefficient, since it relies on manual processes that generate long lines and prolonged waiting times, reaching approximately five days to obtain a response.

The main problem of these systems lies in the fragmentation of information and the restriction of office hours, which causes low accuracy in

responses and, consequently, a poor level of citizen satisfaction. This situation consumes valuable institutional resources and restricts the municipality's capacity to guarantee transparent and timely service in the face of high daily demand.

Chou et al. [13] evaluated a chatbot designed for older adults with depressive or anxiety disorders. Among 35 participants, more than 82% used it daily; the group aged 65 or older significantly reduced loneliness, and both groups reported favorable assessments of usability and satisfaction.

Alita [5] experimentally evaluated the use of LLMs in citizen participation for urban green space initiatives. The LLM demonstration significantly increased initial participation and involvement in value co-creation, although it reduced enthusiasm among citizens with higher self-efficacy.

Chiu et al. [11] developed MyEcoReporter, an LLM- and SMS-based chatbot to facilitate citizen reporting of environmental incidents to authorities. Pilot tests showed successful conversations, structured data storage, and rapid responses, reducing bureaucratic barriers and improving the accessibility of the process.

Haesevoets et al. [22] analyzed, across three studies, citizen perception of AI in public sector decisions. Participants preferred that AI have less decision-making weight than humans and act mainly as support, being more acceptable in decisions with a low ideological burden.

Asquer and Krachkovskaya [10] applied a chatbot-assisted case study to 18 graduate students for the design of public financial management systems. The results provided support for the use of the chatbot to explore problems, although they also revealed limitations in its pedagogical application. The advancement of information technologies drives their application in diverse domains; one study employed the Analytic Hierarchy Process under a neutrosophic environment to classify technological applications through multicriteria decision-making, evidencing the growing role of these decision-support tools [9].

Guo and Cugurullo [14] analyzed the integration of AI in a megahospital in China, identifying improvements in operational efficiency and patient care. However, they noted challenges

associated with data privacy, ethical considerations, and infrastructure, requiring a balance between technological innovation and public well-being.

Gao and Wang [19] proposed a four-layer IoT-based smart city architecture and integrated LLMs through TeachVQA. Evaluations on VQAv2 and OKVQA showed that the approach outperformed existing VQA methods, highlighting its potential to improve the effectiveness of intelligent educational systems.

Arslan et al. [45] reviewed RAG-LLM applications and developed a business information extraction case. With 120 manually curated questions, the RAG chatbot with Llama2 achieved 89% precision, 84.5% recall, and 85% accuracy, evidencing its usefulness for domain-specific queries.

Følstad and Bjerkreim-Hanssen [17] analyzed 2,663 dialogues from a chatbot used in six Norwegian municipalities. Of the responses, 68% were relevant and 66% of the dialogues provided help that was probably used; furthermore, brief and utilitarian messages were associated with better conversational outcomes.

Han et al. [23] integrated local knowledge with LLMs through embeddings, semantic similarity search, and dynamic retrieval. Compared with the isolated LLM, the LLM-LKB approach contributed evidence, specialized knowledge, and updatable information, improving the depth, precision, and regional relevance of responses on carbon neutrality.

David et al. [15] analyzed 26 local government policy documents on responsible AI. Participatory and deliberative considerations were the most frequent, while adaptability, explainability, sustainability, and accountability were the least present, evidencing an uneven incorporation of responsible AI criteria. In the Peruvian context, an experimental study applied Machine Learning (SVM) to automate sentiment analysis, evidencing significant reductions in time, cost, and personnel, in addition to greater user acceptance, which confirms the effectiveness of AI for optimizing service processes [27].

Fischer-Abaigar et al. [16] identified five sources of misalignment in public sector AI decision systems: distribution shifts, label bias, the influence of past decisions, competing objectives,

and human intervention. They proposed prioritizing decision outcomes and expert involvement, not only predictive accuracy. In specialized technical domains, artificial intelligence is consolidating as a dominant trend; a systematic review on malware detection in Android found that Machine Learning algorithms, led by Support Vector Machine (23.14%), concentrate its application, evidencing the maturity of these approaches [8]. The use of Machine Learning in specialized technical domains shows a growing trend; a systematic review on software-defined network management found that Support Vector Machine is the predominant algorithm, reflecting the consolidation of these approaches in recent research [37].

Androutsopoulou et al. [6] proposed a platform of public chatbots based on NLP, machine learning, and data mining, capable of integrating legislation, operational data, and social media. Validated through scenarios with three Greek agencies, it enabled government-citizen interactions in everyday language for queries of greater complexity, ambiguity, and uncertainty. Kalyuzhnaya et al. [26] evaluated a multi-agent system with 150 question-answer pairs for smart city management. It reached 94% to 99% accuracy in pipeline selection; upon incorporating RAG, response accuracy improved by 17% in strategic development queries and by 55% in questions about service accessibility.

Langgeng et al. [38] developed an LSTM chatbot for vocational registration information with 2,636 labeled questions. In testing, it obtained 87.6% accuracy, 88.4% precision, and 87.3% recall, showing consistent performance in the automation of informational queries. Kleiman and Barbosa [28] evaluated ChatPGD, a chatbot with GPT-3.5 for the Brazilian federal public sector. Over 15 days, 1,985 messages were recorded across 724 threads; 72% of respondents confirmed its usefulness, although 18% reported inaccurate or irrelevant information, 9% reported slowness, and 18% reported error messages.

Monteiro et al. [41] compared ten government chatbots through heuristic evaluation, five Brazilian and five foreign. They identified usability problems related to dialogue navigation, response time, conversation closure, guidance for formulating inputs, and standardization of

expressions, aspects that may condition citizen adoption.

Yun et al. [44] implemented RAG with LLMs to communicate public policies over more than 200 documents from regional governments in China and the United States. The average accuracy reached 85.58% and 90.67%, respectively; the failed cases were associated with queries that required greater context and the integration of complex information.

Wahidur et al. [43] developed LQ-RAG for legal queries, combining fine-tuning, hybrid retrieval, and recursive feedback. The system improved relevance by 23% over naive RAG and by 14% over RAG with fine-tuned LLMs; the absence of human expert feedback was recognized as a limitation.

Zaidani et al. [4] classified administrative requests in the Moroccan dialect using a dataset augmented with AugGPT. When comparing ML, DL, and LLMs, RNNs achieved the highest accuracy (93%), limited by the small initial dataset size and overfitting in transformers. Lee and Rew [30] developed a chatbot for LMS with augmented memory and confidence verification through RAG. The evaluations confirmed significant improvements in contextual consistency and satisfaction, limited by a small user sample of 30 and dependence on retrieval.

Choi et al. [12] optimized the retrieval of traffic laws through RAG with LoRA and multi-stage filtering. They managed to reduce the computational load by 18.7% and improve precision by 8.5%, limited by dependence on the quality of the corpus and the base models.

Mansurova et al. [32] evaluated LLMs and RAG for question answering in Kazakh through a three-stage framework. They found that RAG drastically improves the precision of local models compared with proprietary ones, although limited by mixed datasets and a restricted scope.

Madan and Ashok [7] explored the perception of AI benefits in Canadian administration through mixed methods (n = 272/38). They revealed that service demand drives early adoption more than technology, limited by the geographic context and a cross-sectional design.

Jiang et al. [25] developed ARGUS, a government QA system with hybrid retrieval (graph/vector). They achieved a ROUGE-1 of 0.68

and high semantic relevance through fine-tuning, limited by system latency and high computational deployment costs. In parallel, the critical role of electronic governance is emphasized as a driver of efficiency in public services, underscoring the urgency of establishing reference models that integrate AI and IoT to overcome technical barriers and guarantee data privacy [3]. When comparing methodologies for building knowledge bases in business processes, it is highlighted that LLMs offer superior performance in the extraction and structuring of information compared with the use of tools such as the Business Process Analysis Canvas (BPA Canvas) and standard NLP [33]. This study proposes a theoretical framework for the analysis of chatbots in Spanish public administration, arguing that the evaluation of these tools must transcend the effectiveness metric to address multidimensional challenges in the quality of public service [36].

In line with the United Kingdom's sustainability objectives toward a net-zero economy, the development of an "Energy Chatbot" is proposed for SMEs that combines LLMs and Information Retrieval techniques to support decision-making and reduce operating costs [31].

Finally, a predictive study applied to tourism in Torrevieja uses a combination of ARIMA and BERT models on Airbnb and Twitter data, achieving greater precision in the forecast of occupancy for more sustainable destination management [20].

Gesk and Leyer [40] evaluated, through experiments, the acceptance of AI in six public service scenarios with 329 participants. The results showed a preference for AI solutions in general public services, while certain specific services continued to be considered mainly the domain of human service.

Ahn and Chen [2] analyzed the willingness of U.S. public employees to use AI in government. Acceptance was associated with perceptions of greater efficiency and effectiveness, a positive view of AI, and prior experience with these technologies, highlighting the importance of training and familiarity.

Notwithstanding the reported advances, the available evidence remains heterogeneous and presents limitations related to factual accuracy, response relevance, latency, usability,

dependence on the quality of the corpus, and citizen acceptance.

Likewise, the studies are concentrated mainly in diverse international contexts, such as Norwegian municipalities [17], the Brazilian federal public sector [28], and governments of China and the United States [44], whereas experimental evidence on the implementation of RAG assistants in Latin American, and particularly Peruvian, local governments is limited. In addition, several investigations privilege specific technical metrics, such as latency, relevance, or retrieval performance [25, 43], without evaluating in an integrated manner their effects on the operational and perceptual dimensions of citizen service. Moreover, the persistence of accuracy and usability problems in government chatbots [28, 41] evidences that technological performance alone does not guarantee an improvement in service; consequently, the need remains open to experimentally validate the effect of RAG on service capacity, response accuracy, waiting time, and citizen satisfaction in real public service scenarios.

In view of these gaps, the research is justified by the need to provide rigorous empirical evidence on the effect of an Intelligent Assistant with RAG in a Peruvian municipal context, considering both the performance of the service process and the citizen's experience. Grounding responses in institutional documentation is especially relevant for improving their reliability and reducing problems associated with inaccurate information [22, 26]. The study is circumscribed to the district of Carabayllo, Lima, and evaluates, through an experimental post-test design with a control group and an experimental group, four indicators: number of queries, accuracy of responses, waiting time, and user satisfaction. Its scope does not comprise generalization to other jurisdictions, longitudinal analysis, economic or infrastructure evaluation, or other applications of public AI such as automated decision-making or policy formulation.

The objective of the research is to develop and implement an Intelligent Assistant that integrates the RAG (Retrieval-Augmented Generation) architecture and Language Models (LLMs) in order to optimize access to information and the management of public service queries in the

District of Carabayllo. The proposal seeks to increase the accuracy, immediacy, and efficiency of citizen service processes, responding to the need for modernization and operational optimization demanded by current public management.

Regarding the structure of the paper, it is organized into six main sections. The second presents the Theoretical Background with the technologies that underpin the work, focusing on Intelligent Assistants and the RAG architecture, while the third describes the applied methodological approaches, integrating the agile SCRUM methodology and an experimental design to guarantee rigorous scientific development. The fourth section details the case study and the technical implementation of the solution, grounded in a clean architecture and iterative development processes.

The experimental results and their contrast with previous research are presented in the fifth section, highlighting the statistical validation of the hypotheses and the differentiating contributions. Finally, the sixth section brings together the conclusions, where the findings are synthesized, the scope of the project is discussed, and recommendations oriented toward future studies on the subject are formulated.

2 Theoretical Background

2.1 Intelligent Assistants with RAG

To interpret the independent variable termed Intelligent Assistant with RAG, it is necessary to situate it within the field of large language models (LLMs), where Retrieval-Augmented Generation (RAG) constitutes an architecture oriented toward complementing the generative capabilities of these models. RAG should not be understood as an independent model, but rather as an architecture that strengthens LLMs by connecting them with external knowledge sources during the response generation process, favoring more precise, updated, and contextualized responses.

RAG combines information retrieval mechanisms with generative language models. Whereas models such as GPT generate responses from the knowledge acquired during

their training, a RAG architecture retrieves pertinent information from external sources in a dynamic and selective manner before producing the response, incorporating additional context that can improve its grounding and pertinence.

A chatbot is a conversational application capable of interpreting user inputs and generating responses through natural language processing techniques, providing immediate and consistent assistance [35].

In public services, Husain et al. [1] designed a knowledge-based chatbot prototype for business licenses, oriented toward providing real-time information, step-by-step guidance, form assistance, and referral to human agents. Retrieval-Augmented Generation complements LLMs by incorporating relevant external information, improving precision in domain-specific queries [45].

Arslan et al. [45] point out that LLMs may present accuracy limitations in specialized contexts and that the integration of RAG makes it possible to reduce this difficulty through the use of external information.

Likewise, Kalyuzhnaya et al. [26] identify hallucinations and knowledge obsolescence as limitations of LLMs and propose RAG as a mechanism to dynamically incorporate documentary context; experimentally, they reported improvements of 17% in queries on strategic development and 55% in questions related to service accessibility.

In language processing, Piizzi et al. [35] developed, for public administration, a model based on BERT, complemented with encoder/decoder modules and a highway network, which obtained better results than BIDAf in F1-score and Exact Match.

Cortés-Cediel et al. [36], after analyzing 33 chatbots described in the literature and 25 deployed in Spanish public administrations, identified that these tools are oriented mainly toward the search for information, documents, and government services, although limitations related to their evaluation and development persist.

For their part, Kruhlov et al. [29] found a positive relationship between the implementation of AI and various efficiency indicators of public services.

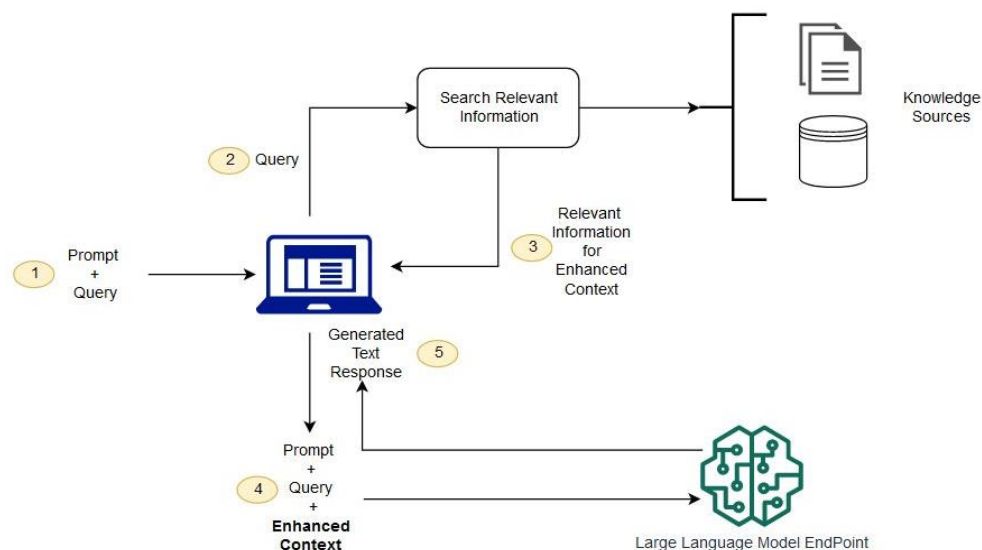


Fig. 1. RAG workflow with the LLM

Finally, Sievers and Russwinkel [39], in a project for the provision of public information through a social robot with language processing capabilities, warned that overestimating the system's capabilities can generate frustration and diminish acceptance of the service.

Figure 1 represents the joint operation of RAG and an LLM. Adapted from "What is RAG (retrieval-augmented generation)?", published by AWS in 2025.

2.2 Access to Information and Queries in Local Governments

The dependent variable Access to Information and Queries on Services refers to the capacity of citizens to obtain public information and resolve procedures effectively. Currently, this access can be limited by the inefficiency of traditional service channels, such as in-person offices and telephone lines, whose capacity is insufficient in the face of the volume of daily queries.

The problem of inefficiency of the traditional service channels is illustrated in Figure 2, taken from "Ayacucho: Defensoría del Pueblo reiterates its request to Reniec of Huamanga to improve service given the persistence of long lines and excessive delays for procedures," published by the



Fig. 2. Long lines and delays in citizen service

Defensoría del Pueblo in 2025, which evidences long lines and delays in citizen service.

The improvement of access seeks to modernize a limited and inconsistent process through a centralized technological solution available 24/7.

Citizen access to information and public services comprises the possibility of locating government information, identifying the required services, and interacting with institutions through digital channels.

Cortés-Cediel et al. [36], after analyzing 33 chatbots reported in the literature and 25 deployed in Spanish public administrations, identified that

these tools are oriented mainly toward the search for information, documents, and government services, as well as access to open data and citizen participation content.

Larsen and Følstad [21], through 34 in-depth interviews with citizens and public service providers, found that municipal chatbots especially improve the access and efficiency of services, although these benefits represent incremental improvements and not a radical transformation of public provision.

For their part, Vassilakopoulou et al. [42] evidenced that chatbots can complement human work in citizen service: during periods of high demand, the chatbot of the Norwegian Labour and Welfare Administration handled a volume equivalent to the capacity of 220 agents, while approximately one out of every five queries was transferred to human service.

Al-Besher and Kumar [3] hold that e-government services can expand access and reduce provision costs, proposing a citizen-centered framework that integrates AI and IoT to improve their efficiency, effectiveness, and ease of use, additionally considering privacy and security risks.

Finally, Papageorgiou et al. [34] proposed a multimodal framework with RAG and MLLMs to facilitate the access, interpretation, and analysis of Eurobarometer data, enabling focused queries and information synthesis for citizens, researchers, and public policy makers; however, the authors note that its effectiveness requires specific performance evaluations.

Therefore, four (4) indicators were selected:

- Number of queries: measures the number of queries made by citizens during a given period.
- Response accuracy: evaluates that the responses provided are correct, pertinent, and complete.
- Waiting time: measures the interval elapsed between the submission of the query and the final response of the assistant.
- Customer satisfaction level: evaluates the citizen's perception regarding the quality and usefulness of the assistant.

2.3 Articulation between Intelligent Assistants with RAG and Citizen Access to Public Services

The integration of an Intelligent Assistant with RAG transforms access to information and queries on public services by combining the semantic retrieval of institutional sources, vector databases, and language models capable of generating contextualized responses, making it possible to expand the availability, precision, and capacity of citizen service; this relationship is supported by studies that evidence improvements in access and efficiency of services through public chatbots [21], greater service capacity through hybrid human-AI schemes [42], and increases in accuracy when RAG incorporates specialized external knowledge [26, 45].

However, its effectiveness depends on the quality and updating of the corpus, the correct retrieval of context, the control of hallucinations, latency, privacy, and human supervision [3, 36]; therefore, the articulation between the two variables evolves toward more intelligent, scalable, and verifiable public service models, a conceptual foundation for the experimental design and technological implementation developed in the following sections.

3 Research Method

The research adopts an applied approach oriented toward the development and evaluation of an Intelligent Assistant based on Retrieval-Augmented Generation (RAG) to improve access to information and queries on public services. To achieve this objective, the Scrum methodology for system development is integrated with an experimental post-test design with a control group and an experimental group, in order to evaluate the effects of the proposed solution.

The methodological process comprises the operationalization of the dependent variable through four indicators: number of queries, response accuracy, waiting time, and user satisfaction level. The data are obtained from the traditional service channels and from the interactions with the RAG-based assistant, and subsequently analyzed through descriptive and

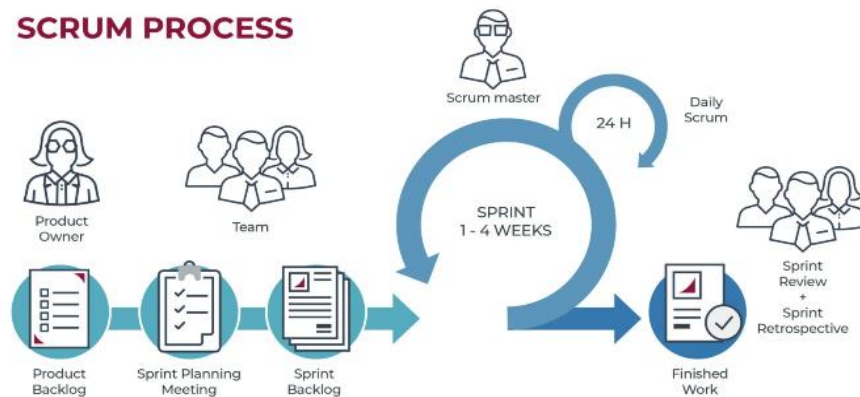


Fig. 3. SCRUM cycle

inferential statistical techniques to determine the significance of the observed differences.

3.1 SCRUM Methodology

Scrum is an agile framework oriented toward the management and development of complex projects, especially applicable to software development, based on effective collaboration, adaptation to change, and continuous improvement.

Gamboa-Cruzado et al. [18] describe its application through four main stages: Sprint Planning, Sprint, Sprint Review, and Sprint Retrospective. In Sprint Planning, the Product Owner presents to the team the prioritized elements of the Product Backlog that will be developed; during the Sprint, the team executes the selected tasks and holds daily meetings to coordinate progress and resolve obstacles; in the Sprint Review the obtained increment is presented and feedback is collected; finally, in the Sprint Retrospective the achievements and difficulties of the process are examined in order to define improvement actions.

Complementarily, Huayana-Salvador et al. [24] consider Scrum as one of the agile foundations integrated into the collaborative ASWAN metamodel, together with the Rational Unified Process (RUP) and Extreme Programming (XP). The authors characterize Scrum as a framework oriented toward generating value through

adaptable solutions in the face of complex problems, and identify the Scrum Master, Product Owner, and development team as the main participants.

Figure 3 presents a representation of the elements that make up the Scrum cycle, taken from "Scrum: what it is and why it has become one of the most popular agile methodologies," published by Ausum Cloud in 2024.

The first stage corresponded to the creation of the Product Backlog, where the vision of the assistant was defined as a unified 24/7 digital channel for municipal queries. The users and their needs were identified, and the Scrum roles (Product Owner, Scrum Master, and development team) were established. Thus, the Product Backlog was structured into Epics for the conversational interface, the RAG engine, and data ingestion.

During sprint planning, the work was organized into iterations, establishing the duration of each sprint and prioritizing the highest-impact requirements. A schedule was defined in which the backend and the RAG engine were developed first, subsequently the interface, and finally the integration and deployment. Each sprint had a Sprint Backlog with defined tasks and responsible parties.

During the execution of the sprints, the product was developed in an iterative and incremental manner, advancing from the core of the system (ingestion and semantic search) to the chat

Table 1. Operationalization of the dependent variable

Indicator	Index	Unit of Measure	Observation Unit
Number of queries	[80-110]	Consultations/day	Manual Review
Response accuracy of the query	[80-90]	Percentage (%)	Indirect Observation
Query waiting time	[5-60]	Minutes	Manual Review
Customer satisfaction level	[1 - Strongly disagree, 2 - Disagree, 3 - Neutral, 4 - Agree, 5 - Strongly agree]	Customer perception (% assigned per level)	Manual Review

interface in React and its full integration. Progress was managed through a digital Kanban board. The activities were adjusted according to the technical needs identified in each iteration.

In the sprint review, at the end of each iteration, the increments were presented to the involved stakeholders to validate their operation. The RAG engine, the interface experience, and the full query-response flow were evaluated. The feedback obtained made it possible to refine the Product Backlog and readjust priorities.

Finally, during the sprint retrospective, the team analyzed the satisfactory aspects and the problems identified. With this, improvement actions were defined, such as adjusting technical agreements and messages to the user. These actions optimized the application of Scrum and strengthened the quality of the developed solution.

3.2 Applied Research Methodology

3.2.1 Operationalization of Variables

Table 1 presents the operationalization of the dependent variable Access to Information and Queries on Public Services.

For its evaluation, indicators, units of measure, and units of observation were established that make it possible to objectively quantify the performance of the citizen service processes considered in the study.

3.2.2 Research Design

The research corresponded to an applied study with an experimental post-test design and random assignment, oriented toward evaluating the effect of the Intelligent Assistant based on Retrieval-Augmented Generation (RAG) on access to information and queries on public services. For this purpose, the results obtained between an

experimental group, in which the intelligent assistant was used, and a control group, in which the traditional service process was maintained, were compared:

RGe	X	O ₁
RGc	--	O ₂

Where:

R: random assignment of the observation units.

Ge: experimental group.

Gc: control group.

X: application of the Intelligent Assistant based on RAG.

O₁: post-test measurement of the dependent variable in the experimental group.

O₂: post-test measurement of the dependent variable in the control group.

The experimental group (**Ge**) was constituted by the processes of access to information and queries on public services carried out through the Intelligent Assistant with RAG, while the **control group (Gc)** was made up of equivalent processes attended to through the traditional channels. For each group, **30 measurements** were considered, selected randomly, with the purpose of comparing the behavior of the indicators defined for the dependent variable.

For the **number of queries** indicator, the number of queries attended to during the observation period was recorded in both groups. In the control group, the queries processed through the traditional channels were counted, while in the experimental group those managed through the intelligent assistant were recorded.

Regarding the **response accuracy** indicator, the correspondence between the responses provided and the official information used as a reference was evaluated. In the control group, the

responses issued through the traditional procedure were analyzed, while in the experimental group the responses generated by the RAG assistant from the institutional documentary corpus were verified.

For the **waiting time** indicator, in the control group the interval elapsed between the submission of the query and the obtaining of the response through the conventional channels was measured. In the experimental group, the time between the submission of the query to the assistant and the generation of the final response was recorded.

Finally, to measure the **satisfaction level**, Likert-scale questionnaires were applied to the users of both groups. In the control group, the perception regarding the traditional service was evaluated, while in the experimental group the interaction experience with the Intelligent Assistant based on RAG was assessed.

3.2.3 Universe and Sample

The study population was constituted by the processes of access to information and queries on public services recorded in the district of Carabayllo during the period established for the research. The unit of analysis corresponded to each citizen query process, considering the interactions carried out through the traditional service channels and through the Intelligent Assistant based on RAG.

For the experimental analysis, a sample of 60 observations was considered, distributed into two independent groups: 30 measurements corresponding to the control group (Gc), associated with the traditional service channels, and 30 measurements corresponding to the experimental group (Ge), in which the Intelligent Assistant with RAG was used. The observations were selected randomly from the available records, with the purpose of comparatively evaluating the indicators of number of queries, response accuracy, waiting time, and user satisfaction level.

3.2.4 Data Collection Procedures

Data collection was carried out through field, experimental, and documentary research techniques, using specific instruments in

accordance with the indicators and stages defined in the study.

In the field research, direct and indirect observation techniques were applied, using observation sheets, field diaries, photographic records, and municipal documentation. These instruments made it possible to characterize the operation of the traditional citizen service channels and to collect information related to the number of queries, the accuracy of the responses, and the waiting times recorded during the service process.

In the experimental phase, data corresponding to the control group and the experimental group were collected. In the control group, the measurements obtained through the traditional service channels were recorded, while in the experimental group the records generated during the interaction with the Intelligent Assistant based on RAG were used. The measurements obtained made it possible to compare the behavior of the indicators defined for both groups and subsequently to perform the corresponding statistical analysis.

Likewise, a documentary review of institutional, regulatory, and scientific information related to public services, citizen service, and intelligent systems was carried out. For this purpose, documentary analysis sheets, bibliographic records, and digital documents from institutional and academic sources were used.

Finally, to evaluate the satisfaction level, structured questionnaires with a Likert scale were applied to the participating citizens. Complementarily, interviews were conducted with the personnel in charge of citizen service and with systems specialists, with the purpose of collecting information on the operation of the current process and the technical aspects associated with the implementation of the Intelligent Assistant with RAG.

3.2.5 Statement of Hypotheses

In the present study, a general hypothesis is proposed that guides the development of the research and makes it possible to evaluate the influence of the proposed model. From this, specific hypotheses are derived that address the different aspects of the performance and quality of citizen service.

H1: The implementation of the assistant with RAG increases the number of queries attended to compared with the traditional channels.

H2: The use of the assistant with RAG improves the accuracy of the responses provided to the citizen.

H3: The adoption of the assistant with RAG reduces the waiting time to obtain a response compared with the traditional mechanisms.

H4: The interaction with the assistant with RAG increases the overall user satisfaction with the citizen service.

For the validation of the hypotheses, procedures were established that made it possible to directly analyze the indicators proposed in the research.

Each indicator was linked to specific measurement methods, enabling precise and objective comparisons between the experimental group and the control group:

μ_1 = Population mean (H_1 , H_2 , H_4) of the Post-tests of the Control Group (G_c),

μ_2 = Population mean (H_1 , H_2 , H_4) of the Post-tests of the Experimental Group (G_e),

where: $H_0: \mu_1 \geq \mu_2$ $H_a: \mu_1 < \mu_2$.

On the other hand:

μ_1 = Population mean (H_3) of the Post-tests of the Control Group (G_c),

μ_2 = Population mean (H_3) of the Post-tests of the Experimental Group (G_e),

where: $H_0: \mu_1 \leq \mu_2$ $H_a: \mu_1 > \mu_2$.

The analysis of the hypotheses comprised descriptive and inferential statistics. In the descriptive stage, measures of central tendency and dispersion were calculated to characterize the results of the post-tests of both groups.

For indicators I1, I2, and I3, the assumption of normality was evaluated through the Anderson-Darling test, considering a significance level of $\alpha = 0.05$. Since no sufficient evidence was found to reject normality ($p > 0.05$), the differences between the groups were contrasted through a t-test for independent samples.

For indicator I4, corresponding to the satisfaction level and measured through a Likert

scale, the non-parametric Mann-Whitney U test was used to compare the distributions of the scores between the Control Group and the Experimental Group.

In all tests, a significance level of $\alpha = 0.05$ was adopted; therefore, values of $p < 0.05$ were considered statistically significant evidence to reject the corresponding null hypothesis.

4 Case Study

For the development of the case study, the Scrum methodology was applied, structured into five execution phases that made it possible to organize, in an agile, iterative, and systematic manner, the process of development and implementation of the Intelligent Assistant based on RAG.

4.1 Creation of the Product Backlog

The construction of the Product Backlog began with the definition of a Product Vision oriented toward solving the inefficiency of municipal service. The objective was established to create "a unified digital channel that offers instantaneous and reliable responses 24 hours a day, reducing the bureaucratic burden and increasing transparency."

For the management of the project, the roles were assigned in accordance with the Scrum framework: Product Owner (Municipality), Scrum Master (Researcher), Development Team, and End Users (Citizens).

The user analysis identified a citizen profile characterized by the need to carry out administrative procedures with agility, limited by time restrictions and the complexity of the current processes.

Based on these needs, the functional requirements were broken down into three strategic Epics in Table 2.

4.2 Sprint Planning

The planning of the project was structured to guarantee an incremental evolution of the product, prioritizing the technical feasibility of the logical core before addressing the interaction layer. The

Table 2. Epics of the RAG system

EPIC	Description
EPIC 01 (Interface)	Development of an intuitive, responsive, and user-friendly chat interface.
EPIC 02 (RAG Generation)	Ensure the assistant understands natural language and provides accurate responses based solely on official documents.
EPIC 03 (Data Ingestion)	Ability to upload, process, and store municipal documents in PDF format for the knowledge corpus.

strategy was organized into a twelve-week schedule, divided into three functional sprints of four weeks each, designed to mitigate technical risks early.

During the first sprint (weeks 1-4), the effort was concentrated on the construction of the backend and the logic of the RAG engine.

At this stage, the configuration and data ingestion modules were developed, implementing the text extraction and chunking algorithms necessary to process the official documents. Likewise, the vectorization and storage tasks in the vector database were executed, ensuring that the system had a consultable and optimized knowledge base before developing any visual interface.

Subsequently, the second sprint (weeks 5-8) addressed the presentation layer. The team focused on the construction of the user interface, developing the chat components and state management to guarantee a fluid experience.

Visual feedback mechanisms and responsive design were implemented, ensuring that the interaction with the language model was intuitive and complied with the usability standards defined in the requirements.

Finally, the third sprint (weeks 9-12) was dedicated to end-to-end integration and system stabilization. The frontend was connected with the backend services, validating the complete data flow from the citizen's query to the generation of the response.

This phase culminated with the execution of functional and usability tests, as well as the preparation of the environment for production deployment, closing the development cycle with a comprehensive and validated solution.

4.3 Sprint Execution

The construction of the solution was grounded in a Clean Architecture pattern, structuring the backend into concentric layers (Domain, Application, Presentation, and Infrastructure) to decouple the business rules from external services.

The technical implementation was divided into three essential components:

4.3.1 Solution Architecture

The solution architecture integrates the components responsible for interaction with the citizen, query processing, retrieval of institutional information, and generation of responses through a language model.

As shown in Figure 4, the solution comprises the access, application, and orchestration layers, the RAG core, knowledge management, the language model, and output, articulated through a processing flow that makes it possible to retrieve relevant information and generate contextualized responses from institutional sources.

4.3.2 Logical Architecture and Design Pattern

With the purpose of guaranteeing the maintainability, scalability, and decoupling of the system, the solution was built under the Clean Architecture pattern (based on Hexagonal Architecture principles).

This design has as its fundamental objective the inversion of dependencies, ensuring a strict separation of responsibilities in which the business core does not depend on external details such as the database or the web interface.

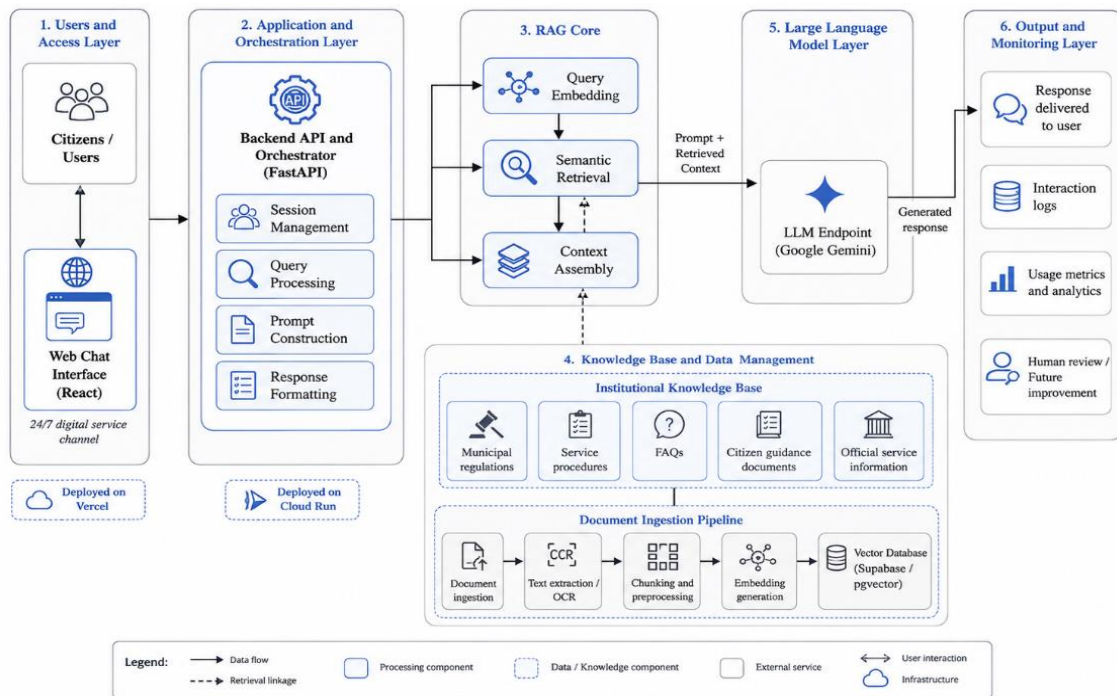


Fig. 4. Architecture of the intelligent assistant based on RAG

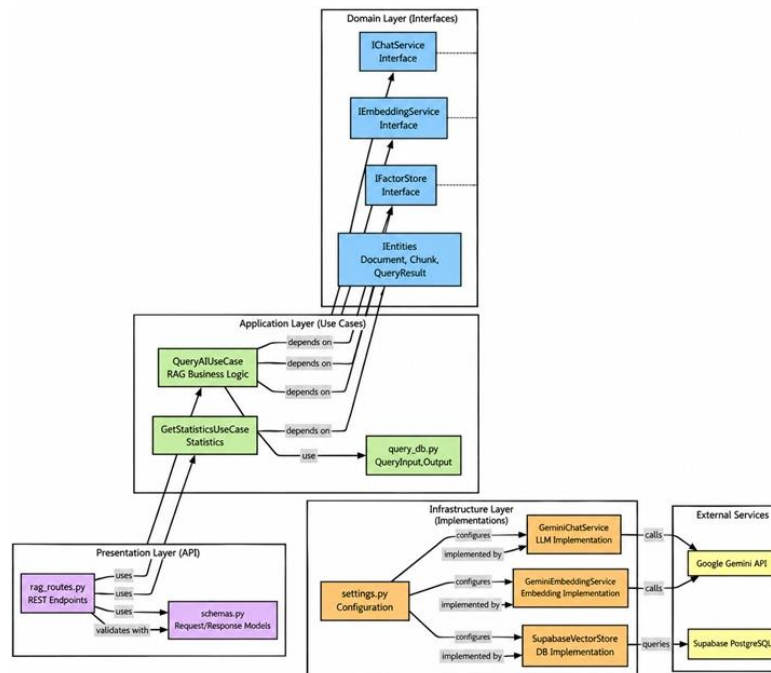


Fig. 5. Clean architecture pattern of the back-end solution

As illustrated in Figure 5, the system is structured into concentric layers with delimited responsibilities:

- **Domain Layer:** Contains the pure business rules and the interface definitions (IChat, IEmbed, IVector), remaining completely independent of external libraries.
- **Application Layer:** Orchestrates the system's use cases, such as QueryRAG, acting as an intermediary that depends only on the domain abstractions.
- **Presentation Layer:** Manages data input through API routes developed in FastAPI and Pydantic validation schemas.
- **Infrastructure Layer:** Implements the domain interfaces by connecting with concrete services, such as the Google Gemini API for chat or Supabase for vector storage.

4.3.3 Sprint 1: Construction of the "Brain" (Backend Core RAG)

This sprint focused on the development of the Retrieval-Augmented Generation (RAG) engine, considered the critical component of the system. The architecture was divided into two essential operational flows: Data Ingestion (offline) and the Query Flow (online), as can be seen in Figure 6.

An automated process was implemented to transform the official documents into a consultable knowledge base. An Intelligent Chunking strategy was applied that segments the text by semantic units (such as legal articles or paragraphs), maintaining an overlap to preserve the context between fragments.

Subsequently, each segment is converted into a vector embedding and stored in the database to enable semantic search. The response flow operates in real time: upon receiving a question, the system retrieves the most relevant fragments from the database based on their vector similarity. With this information, an enriched prompt is constructed that combines the user's query with the retrieved context, instructing the language model (LLM) to generate a natural response

grounded strictly in the processed official documentation.

4.3.4 Sprint 2: Construction of the "Face" (Frontend - React)

The second iteration was oriented toward the materialization of Epic 01, developing an accessible and responsive conversational interface through the React library.

The presentation logic was decoupled into two main components: App.jsx, which manages the transition of states between the welcome view and the active chat session, and ChatInterface.jsx, responsible for managing the message history and the visual loading feedback (isLoading).

To guarantee the readability of the responses generated by the RAG model, a secure rendering of HTML (lists, bold) was implemented, along with a clear visual differentiation between the user and assistant roles, as shown in Figures 7, 8, and 9.

4.3.5 Sprint 3: Integration and Deployment

The deployment architecture defines the physical distribution of the solution's components and the technologies used for their execution in production. As shown in Figure 10, the frontend developed in React is deployed on Vercel and communicates through HTTPS with the backend implemented in FastAPI on Google Cloud Run. The backend securely accesses the Google Gemini APIs through Google Secret Manager and connects with Supabase, which uses PostgreSQL and pgvector for the storage and retrieval of the vectors associated with the documentary corpus.

The final phase of the project consolidated the solution through the integration of services and its deployment on a modern and scalable cloud infrastructure.

First, communication between the client and the server was established through secure REST APIs, enabling a fluid exchange of data between the decoupled components. For the infrastructure, the frontend was hosted on Vercel with the objective of taking advantage of its global content delivery network (CDN) and guaranteeing low latency in citizen access.

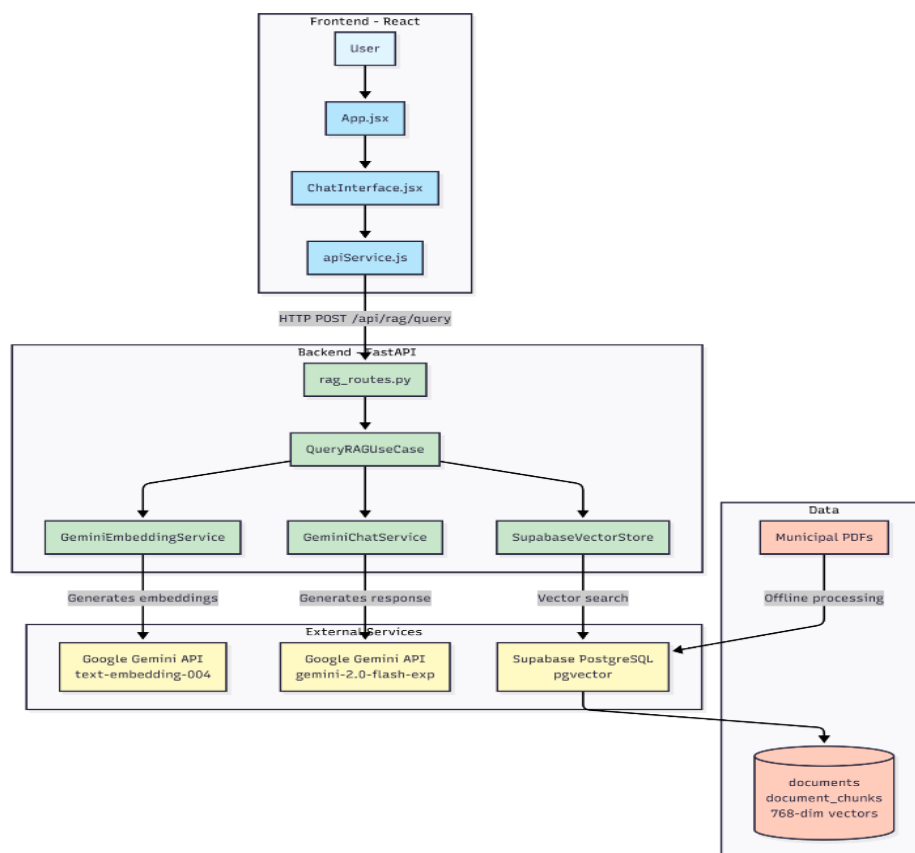


Fig. 6. Component architecture and data flow

For its part, the backend was deployed on Google Cloud Run through containers, using a serverless architecture that enables the automatic auto-scaling of instances according to operational demand.

Finally, the persistence of vector data was centralized in Supabase to enable semantic search, while the security of the critical credentials was delegated to a cloud secrets manager (Google Secret Manager) to guarantee their protection at execution time.

4.3.6 Testing Strategy and Quality Assurance

To ensure the stability of the system within the Scrum cycle, a continuous testing strategy was executed using Pytest. Mocking techniques were applied to simulate the external dependencies (Google Gemini API and Supabase), which made

it possible to validate the business logic without incurring network latency or credit consumption. As a validation metric, Code Coverage was analyzed at the end of development.

As evidenced in the coverage report in Figure 11, the system reached a global coverage of 95%, achieving 100% effectiveness in critical modules such as the domain entities (chat_message, document, chunk) and the data transfer objects (DTOs). This guarantees that the core of the RAG architecture is free of logical errors and ready for production.

4.4 Sprint Review

The validation of the increments was carried out through formal review sessions at the end of each cycle, enabling the inspection of the product by the stakeholders and the dynamic adaptation of the

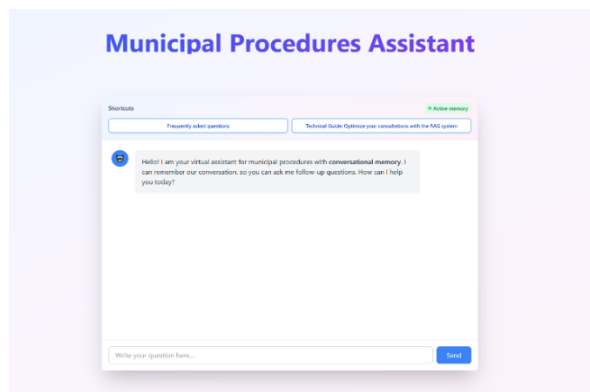


Fig. 7. Home page of the Intelligent Assistant

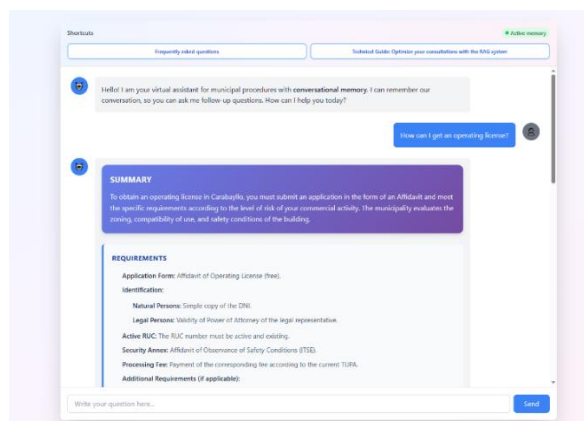


Fig. 8. Interaction flow of a citizen query in natural language

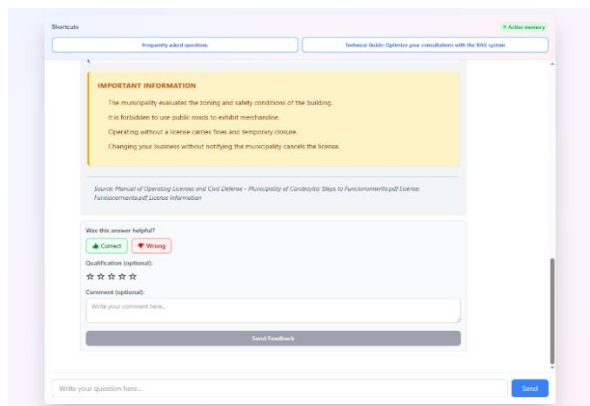


Fig. 9. Feedback provided by the user

requirements. In the first stage, the capacity of the ingestion engine to process the documentary

corpus and execute semantic searches was demonstrated, which made it possible to validate the technical feasibility of the RAG architecture before advancing to the visual development.

Subsequently, the evaluation of the user interface highlighted the clarity in the interaction, incorporating as an improvement the implementation of dynamic welcome messages to guide the citizen in their first contact with the system.

Finally, the comprehensive validation of the system confirmed the precision of the query and response flow; as a key finding, the need was identified to include direct links to the source documents together with the generated response, a fundamental requirement to guarantee administrative transparency in future versions.

4.5 Sprint Retrospective

The retrospective analysis made it possible to identify technical and process obstacles, facilitating the adaptation of the work plan to ensure the quality of the final product. In the initial phase, although the semantic segmentation strategy proved effective for digital documents, a critical limitation was detected in the processing of scanned files, which motivated the inclusion of an optical character recognition (OCR) solution in the technical roadmap to expand the coverage of the knowledge base.

Subsequently, during the development of the interface, integration frictions were identified due to the lack of a detailed technical specification of the exchange data; as a corrective measure, a formal API contract was established that significantly streamlines the final connection between the client and the server.

Finally, the evaluation of the integration evidenced that the system's error messages were excessively technical for the profile of the average citizen, prioritizing the implementation of friendly and clear notifications to optimize the user experience in the face of connectivity failures.

5 Results and Discussion

This section presents and analyzes the results obtained after the implementation of the Intelligent

Table 3. Post-test Results Gc and Ge

N°	I1: Quantity of consultations (consult./day)		I2: Accuracy of the response (%)		I3: Waiting time (min)		I4: Customer Satisfaction Level (Likert Scale)	
	Postest CG	Postest EG	Postest CG	Postest EG	Postest CG	Postest EG	Postest CG	Postest EG
1	64	101	86	95	6048	10	Disagree	Agree
2	62	91	88	93	6912	20	Disagree	Strongly agree
3	66	95	85	92	7200	7	Disagree	Agree
4	63	89	87	92	6480	30	Disagree	Agree
5	65	93	89	95	5904	40	Neutral	Strongly agree
6	67	96	84	91	7056	6	Strongly disagree	Neutral
7	62	92	90	94	6192	15	Disagree	Strongly agree
8	68	97	83	93	6768	35	Strongly disagree	Neutral
9	65	95	86	91	5760	13	Disagree	Agree
10	65	91	88	93	6624	25	Disagree	Strongly agree
11	62	96	87	94	6336	5	Neutral	Agree
12	66	97	85	95	6912	38	Neutral	Agree
13	66	91	89	94	6048	18	Disagree	Strongly agree
14	67	94	84	96	7200	32	Disagree	Neutral
15	65	95	91	93	5904	12	Strongly disagree	Strongly agree
16	62	98	82	90	7056	23	Disagree	Neutral
17	64	93	86	92	6480	36	Disagree	Agree
18	63	92	88	91	6192	9	Neutral	Strongly agree
19	65	92	85	91	6768	28	Neutral	Agree
20	67	91	87	93	5760	40	Disagree	Agree
21	63	93	89	96	6624	7	Strongly disagree	Strongly agree
22	61	92	84	90	6048	19	Disagree	Neutral
23	68	96	90	93	6912	29	Disagree	Strongly agree
24	64	97	83	91	5904	16	Disagree	Neutral
25	64	96	86	92	7056	39	Disagree	Agree
26	61	90	88	94	6336	6	Neutral	Strongly agree
27	66	98	85	93	7200	26	Neutral	Agree
28	67	98	87	95	6192	33	Disagree	Agree
29	65	94	89	96	6768	10	Disagree	Strongly agree
30	63	93	84	91	6480	22	Strongly disagree	Neutral

Assistant based on Retrieval-Augmented Generation (RAG), developed through the Scrum methodology. The analysis compares the performance of the Control Group, corresponding to the traditional service channels, with the Experimental Group, in which the intelligent assistant was used, with the purpose of evaluating the effect of the solution on access to information and queries on public services in the district of Carabayllo. The four indicators defined in the study are examined through descriptive and inferential statistics, and their results are contrasted with the objectives and previous research to determine the effectiveness and scope of the proposal.

5.1 Experimental Results

Table 3 presents the consolidated summary of the post-test measurements obtained in the Control Group and the Experimental Group. The data corresponding to the number of queries (I1), response accuracy (I2), and waiting time (I3) were collected through observation sheets and system records, while the user satisfaction level (I4) was obtained through surveys applied to the citizens.

These results make it possible to quantitatively compare the performance of the traditional service channels with the Intelligent Assistant based on RAG.

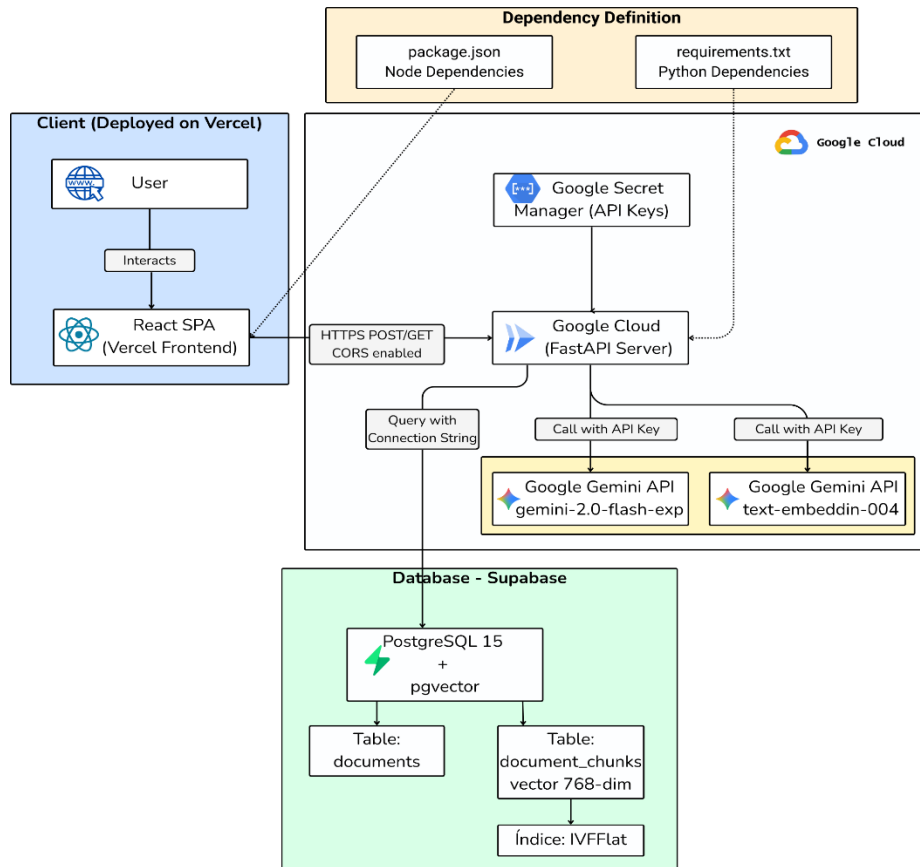


Fig. 10. Deployment architecture of the solution

```
----- coverage: platform win32, python 3.13.7-final-0 -----
```

Name	Stmts	Miss	Branch	BrPart	Cover	Missing
application__init__.py	0	0	0	0	100%	
application\dtos__init__.py	3	0	0	0	100%	
application\dtos\query_dto.py	16	0	0	0	100%	
application\dtos\stats_dto.py	11	0	0	0	100%	
application\use_cases__init__.py	3	0	0	0	100%	
application\use_cases\get_statistics.py	17	0	0	0	100%	
application\use_cases\query_rag.py	125	4	32	4	95%	78, 112, 179, 218
domain__init__.py	0	0	0	0	100%	
domain\entities__init__.py	7	0	0	0	100%	
domain\entities\chat_message.py	26	0	6	0	100%	
domain\entities\chat_session.py	42	0	6	0	100%	
domain\entities\chunk.py	18	0	0	0	100%	
domain\entities\document.py	24	0	0	0	100%	
domain\entities\feedback.py	82	0	28	0	100%	
domain\entities\query_result.py	27	0	2	0	100%	
domain\interfaces__init__.py	6	0	0	0	100%	
domain\interfaces\chat_service.py	8	2	0	0	75%	39, 55
domain\interfaces\chat_session_store.py	29	8	0	0	72%	39, 55, 71, 89, 110, 126, 142, 163
domain\interfaces\embedding_service.py	12	3	0	0	75%	33, 49, 70
domain\interfaces\feedback_repository.py	16	4	0	0	75%	30, 55, 73, 89
domain\interfaces\vector_store.py	15	4	0	0	73%	46, 59, 72, 90
TOTAL	487	25	74	4	95%	

Coverage HTML written to dir htmlcov

Fig. 11. Unit tests for coverage

For the quantitative analysis of indicator I2: Response accuracy, the following expression was used:

$$Accuracy (\%) = \frac{N_c}{N_t} \times 100,$$

where:

- N_c = number of correct responses,
- N_t = total number of evaluated responses.

This metric expresses, in percentage terms, the proportion of responses considered correct with respect to the total number of responses evaluated in each group.

Normality Test: Increase of I1, I2 and Reduction of I3

The assumption of normality of the data corresponding to the post-tests of the indicators number of queries (I1), response accuracy (I2), and waiting time (I3) was evaluated, with the purpose of determining the pertinence of using parametric tests in the inferential analysis.

For this, the Anderson-Darling test was applied to the results of the Control Group and the Experimental Group, considering a significance level of $\alpha = 0.05$. The results obtained are presented in Figures 12, 13, and 14.

The results of the Anderson-Darling test showed that, for indicators I1, I2, and I3, in both the Control Group (Gc) and the Experimental Group (Ge), the significance values were higher than the established level ($p > 0.05$).

Consequently, no sufficient statistical evidence was found to reject the assumption of normality of the distributions, which supports the application of parametric tests for the subsequent inferential analysis of these indicators.

5.2 Discussion of Results: Effect on Access to Information and Service Queries in the District of Carabayllo

In this section, the results obtained for the four indicators of the dependent variable are analyzed and interpreted, comparing the performance of the Control Group (Gc) and the Experimental Group (Ge).

The analysis integrates descriptive and inferential statistics with the purpose of determining the effect of the implementation of the

Intelligent Assistant based on RAG on access to information and queries on public services in the district of Carabayllo.

5.2.1 With Descriptive Statistics

This subsection presents the descriptive analysis of the post-test measurements corresponding to the Control Group (Gc) and the Experimental Group (Ge), with the purpose of characterizing the behavior of the indicators and examining the differences observed between the two groups. Table 4 presents, for the indicators number of queries (I1), response accuracy (I2), and waiting time (I3), the main descriptive measures, including the mean and the standard deviation (StDev). Likewise, the Anderson-Darling (AD) statistics and their respective p-values are included as a complement to the normality evaluation developed in Section 5.2.

These results provide the quantitative basis for interpreting the differences between the traditional service channels and the Intelligent Assistant based on RAG.

In addition, Table 5 presents the 95% confidence intervals for the measures of each indicator of the experimental group, together with the values of kurtosis, skewness, and third quartile (Q3).

For Indicator I₁ (Number of queries)

With $n=30$ observations per group, the Experimental Group reached a mean of 94.20 consultations/day (StDev = 2.870) compared with 64.53 consultations/day (StDev = 2.030) in the Control Group, which represents a descriptive increase of 45.97%; the Anderson-Darling tests showed no evidence to reject normality in Gc (AD = 0.456; $p = 0.250$) or in Ge (AD = 0.402; $p = 0.337$). In Ge, the mean was estimated with a 95% CI of 93.128 to 95.272 consultations/day, while the skewness = 0.272808 indicates a slight positive skewness and the kurtosis = -0.487699 a moderately platykurtic distribution, without pronounced deviations from symmetry.

The $Q_3 = 96.250$ consultations/day indicates that approximately 75% of the experimental observations were situated at values equal to or lower than this level, additionally evidencing a relatively stable concentration around the mean.

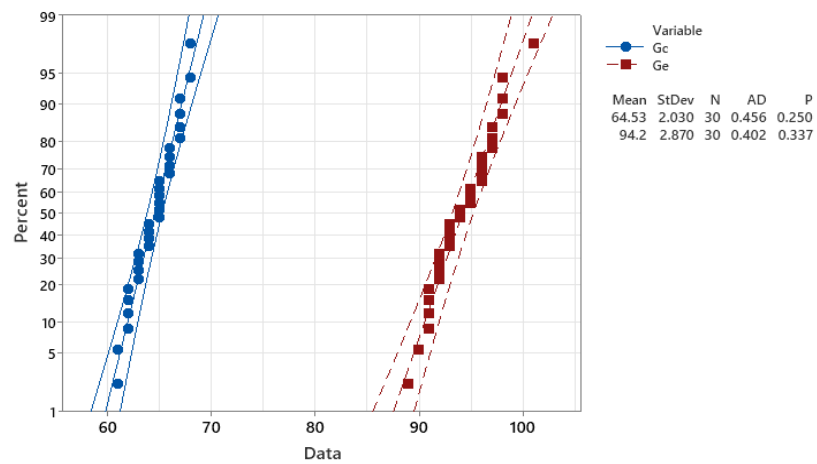


Fig. 12. Normality test of indicator 1: Number of queries

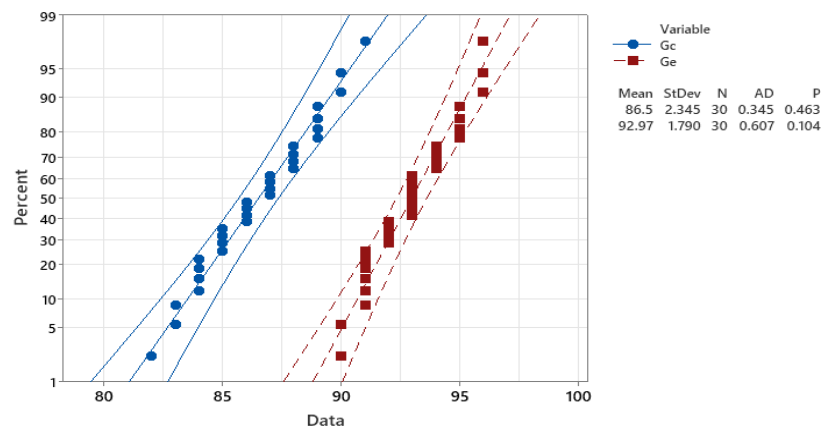


Fig. 13. Normality test of indicator 2: Accuracy of the response to a query

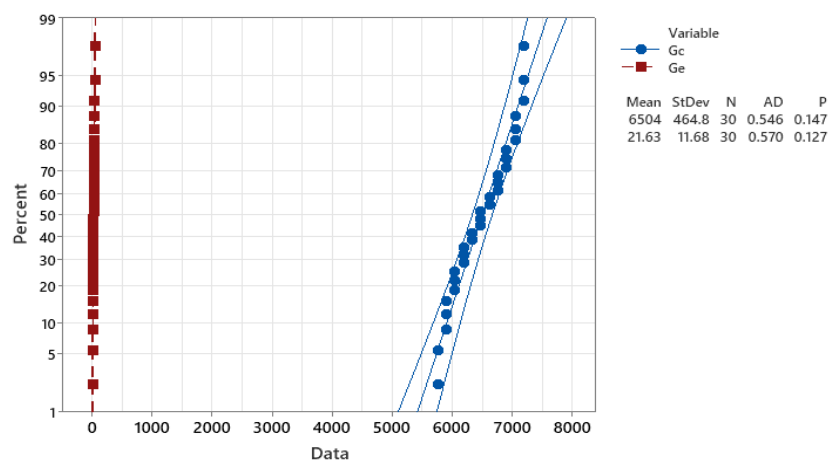


Fig. 14. Normality test of indicator 3: Query waiting time

Table 4. Results with descriptive statistics (1)

Sample	N	Mean	StDev	AD	p-Value
I1: Posttest (Gc)	30	64.53	2.030	0.456	0.250
I1: Posttest (Ge)		94.2	2.870	0.402	0.337
I2: Posttest (Gc)	30	86.5	2.345	0.345	0.463
I2: Posttest (Ge)		92.97	1.790	0.607	0.104
I3: Posttest (Gc)	30	6504	464.8	0.546	0.147
I3: Posttest (Ge)		21.63	11.68	0.570	0.127

Table 5. Results with descriptive statistics (2)

Sample	N	95% Confidence Intervals for the Mean	Kurtosis	Skewness	Q3
I1: Posttest (Ge)	30	93.128-95.272 consultations/day	-0.487699	0.272808	96.250
I2: Posttest (Ge)	30	92.298-93.635 %	-0.963536	0.130653	94.250
I3: Posttest (Ge)	30	17.274-25.993 min.	-1.35057	0.12914	32.250

The increase of 45.97% observed in the mean number of queries attended to is situated in the same direction as previous evidence on the capacity of chatbots to expand service in public contexts, although the differences in scale, observation period, and design prevent a direct comparison of the magnitudes.

It is worth noting that Følstad and Bjerkreim-Hanssen [17] analyzed 2,663 dialogues from six Norwegian municipalities and noted that Kommune-Kari handled around one million dialogues per year, evidencing its operational capacity in a consolidated municipal environment.

In contrast, Kleiman and Barbosa [28], in a much more limited period, recorded 1,985 messages distributed across 724 threads over 15 days, which shows that a considerable volume of citizen interaction can also be concentrated in short-duration implementations.

It is relevant that Madan and Ashok [7] identified service demand as a significant driver of the early adoption of AI in public administration; therefore, the increase recorded in Carabayllo could be explained partially because an automated channel, available and with immediate access, makes it possible to materialize queries that the traditional channels restrict by capacity, schedule, or personnel availability.

Notwithstanding this, Jiang et al. [25], through ARGUS and a hybrid retrieval scheme for government services, show the usefulness of more flexible retrieval mechanisms for answering heterogeneous queries; consequently, the combination of semantic retrieval and automated generation constitutes a plausible mechanism for sustaining greater service volumes.

Taken together, these antecedents support the direction of the observed effect, while the specific increase of 45.97% should be interpreted as a result particular to the experimental and operational conditions evaluated in the present study.

The observed increase suggests that the Intelligent Assistant based on RAG can expand the operational service capacity, by making it possible to process a greater volume of queries without depending exclusively on the immediate availability of municipal personnel.

A plausible mechanism is the simultaneous automation of the retrieval and generation of responses, which reduces the bottlenecks associated with the manual search for information and makes it possible to attend to queries continuously.

The reduced dispersion of Ge, its relatively narrow 95% CI, and the absence of important

Table 6. Hypothesis testing for the parametric indicators

Sample	n	Ho	t-value	p-value
I1: Posttest (Gc)	30	$\mu_1 \geq \mu_2$	-46.23	< 0.000
I1: Posttest (Ge)				
I2: Posttest (Gc)	30	$\mu_1 \geq \mu_2$	-12.00	< 0.000
I2: Posttest (Ge)				
I3: Posttest (Gc)	30	$\mu_1 \leq \mu_2$	76.36	< 0.000
I3: Posttest (Ge)				

Table 7. Hypothesis testing for the non-parametric indicator

Sample	n	Ho	w-value	p-value
I4: Posttest (Gc)	30	$\mu_1 \geq \mu_2$	489.50	< 0.000
I4: Posttest (Ge)				

deviations from normality further suggest that this increase in capacity was consistent across the analyzed observations and was not the result of a few extreme values.

For Indicator I₂ (accuracy of the response to the query)

For n=30 observations per group, the mean accuracy increased from 86.50% (StDev = 2.345) in Gc to 92.97% (StDev = 1.790) in Ge, equivalent to an improvement of 6.47 percentage points; the Anderson-Darling results were AD = 0.345, p = 0.463 for Gc and AD = 0.607, p = 0.104 for Ge, so no evidence was found to reject normality. In Ge, the 95% CI of the mean was 92.298%-93.635%, with skewness = 0.130653, close to zero, and kurtosis = -0.963536, indicative of a distribution flatter than the normal one and with less concentration around the center.

The Q3 = 94.250% establishes that approximately 75% of the observations presented accuracy values equal to or lower than 94.25%, while 25% reached values equal to or higher than that quartile.

The results obtained for I2 show a trend consistent with the previous literature, although the comparison must be made considering that the studies employ different architectures, domains, and evaluation metrics.

It is worth noting that the accuracy of 92.97% reached in the present study exceeds that reported by Langgeng et al. [38] (87.6%), Yun et al. [44]

(90.67% for U.S. policies and 85.58% for China policies), Choi et al. [12] (85% with RAG-LoRA compared with 78% of the base model), Mansurova et al. [32] (75.4% in End-to-End RAG and 86.7% in an ideal scenario), and Arslan et al. [45] (85% with RAG-Llama2), while it is practically equivalent to the 93% obtained through RNN by Zaidani et al. [4].

In contrast, Kalyuzhnaya et al. [26] reported relative improvements of 17% in strategic development queries and 55% in service accessibility upon incorporating RAG; although the 94%-99% reported by these authors corresponds to pipeline selection and not directly to the final accuracy of the responses, their results support the favorable effect of incorporating external knowledge retrieval.

It is relevant that other works evaluated dimensions related to but not equivalent to accuracy: Følstad and Bjerkreim-Hanssen [17] recorded 68% of relevant responses and 66% of dialogues with help probably used.

Wahidur et al. [43] obtained improvements of 13% in Hit Rate, 15% in MRR, and up to 23% in relevance; and Jiang et al. [25] reached 81% of semantic relevance and an F1-score of 0.75 through hybrid retrieval.

Notwithstanding these advances, Kleiman and Barbosa [28] found that 18% of users reported inaccurate or irrelevant information, evidencing that factual quality continues to be a relevant limitation in government chatbots.

Taken together, the improvement of 6.47 percentage points observed in this study, from 86.50% to 92.97%, is aligned with the evidence that the use of retrieval and contextualization mechanisms can increase the quality of the responses; nevertheless, its magnitude must be interpreted specifically within the institutional corpus, the municipal context, and the evaluation procedure used in the present research.

The greater accuracy observed in Ge is compatible with the functional mechanism of RAG, through which the query is linked with information retrieved from the institutional corpus before generating the response, reducing exclusive dependence on the parametric knowledge of the language model.

The lower StDev of Ge (1.790 compared with 2.345) further suggests greater homogeneity of the responses, a relevant aspect in public services where the consistency of the information provided to the citizen constitutes an operational requirement. The 95% CI fully situated at levels higher than 92% and the low skewness support the descriptive stability of the performance, although the improvement must be statistically attributed to the assistant only after considering the corresponding inferential test.

For Indicator I₃ (query waiting time)

With $n=30$ observations per group, the mean waiting time decreased from 6,504 min (StDev = 464.8) in Gc to 21.63 min (StDev = 11.68) in Ge, corresponding to an approximate descriptive reduction of 99.67%; the Anderson-Darling tests showed $AD = 0.546$, $p = 0.147$ for Gc and $AD = 0.570$, $p = 0.127$ for Ge, with no sufficient evidence to reject normality. For Ge, the 95% CI of the mean was 17.274-25.993 min, while the skewness = 0.12914 reflects a practically symmetric distribution and the kurtosis = -1.35057 evidences a markedly platykurtic distribution with respect to the normal one. The $Q_3 = 32.250$ min means that approximately 75% of the times recorded in Ge were equal to or lower than 32.25 min and that the remaining 25% were located above this value.

The reduction of 99.67% observed in the mean waiting time, from 6,504 min in Gc to 21.63 min in Ge, evidences a substantial improvement in the timeliness of service; however, its comparison with previous research requires distinguishing between

citizen waiting time and technical latency of the system.

It is worth noting that Kalyuzhnaya et al. [26] reported that LLM-based agents can reduce certain urban planning tasks from days to hours and, for configurations with a vector database and access through API, recorded response times of 9.2 to 14.6 s; although these measurements are not equivalent to those of the present study, both evidence that automation through retrieval and generation can considerably shorten processes that require the search for and processing of information.

In contrast, Wahidur et al. [43] showed that the incorporation of more advanced components increased latency from 7.2 s with Naive RAG to 14.6 s with LQ-RAG, revealing a trade-off relationship between architectural complexity, response quality, and speed; therefore, the reduction observed in Carabayllo should not be interpreted solely as a consequence of lower computational latency, but also of the elimination of stages inherent to traditional service, such as waiting for personnel availability and the manual search for information.

It is relevant, furthermore, that Monteiro et al. [41] identified response time as a usability problem in government chatbots, which reinforces the importance of the evaluated indicator.

Taken together, the antecedents support the direction of the improvement obtained, although the specific magnitude of the reduction corresponds to the comprehensive citizen service process evaluated in this study and cannot be directly equated with the technical execution times reported in other systems.

The magnitude of the reduction suggests that automation eliminates a substantial part of the times associated with lines, personnel availability, documentary search, and manual processing present in the traditional service channels.

The plausible mechanism is that the assistant executes, in an integrated manner, the semantic retrieval of institutional content and the generation of the response, decreasing the intermediate stages required to resolve a citizen query.

Nevertheless, the StDev = 11.68 min with respect to a mean of 21.63 min shows appreciable relative variability in Ge, so factors such as the complexity of the query, the volume of retrieved

context, the processing of the model, or connectivity could explain part of the observed differences and should be considered in the evaluation of the system's performance.

5.2.2 With Inferential Statistics: Hypothesis Testing

In this phase, the hypothesis testing was carried out with the purpose of determining whether the differences observed between the Control Group (Gc) and the Experimental Group (Ge) were statistically significant. For the parametric indicators number of queries (I1), response accuracy (I2), and waiting time (I3), whose data showed no sufficient evidence to reject the assumption of normality, parametric tests for independent samples were applied, considering a significance level of $\alpha = 0.05$.

Table 6 presents the results of these contrasts, while Table 7 reports the contrast of the satisfaction level (I4), evaluated through a non-parametric test due to its ordinal nature.

For Indicator I₁ (number of queries)

For the samples I1: Posttest (Gc) and I1: Posttest (Ge), with $n = 30$ observations per group, the null hypothesis $H_0: \mu_1 \geq \mu_2$ was contrasted, where μ_1 represents Gc and μ_2 represents Ge. The result $t = -46.23$; $p < 0.001$ makes it possible to reject H_0 and evidences that the number of queries attended to in Ge was significantly greater than in Gc. The negative sign of t is consistent with a mean of Gc lower than that of Ge.

The inferential effect observed for I1 ($t = -46.23$; $p < 0.001$) bears a conceptually consistent relationship, although not directly equivalent, with the findings of Madan and Ashok [7], who validated their hypothesis H1b by identifying service demand as a significant predictor in the adoption and perception of AI benefits in public administration. In contrast, the present study evaluates a directly observable operational result, evidencing that the Experimental Group attended to a significantly greater number of queries than the Control Group.

This difference is relevant, because the antecedent of Madan and Ashok [7] provides evidence on the role of demand pressure as a factor that favors the incorporation of AI, while our

results show that, once the solution is implemented, this technology can translate into a greater effective service capacity. Consequently, both findings are complementary: the first explains a factor that drives technological adoption and the second evidences a quantifiable effect on the volume of processed queries. Taken together, this convergence supports that AI-based solutions can respond to scenarios of high demand for public services, although the effect demonstrated in Carabayllo must be circumscribed to the experimental conditions evaluated.

The result indicates that the implementation of the Intelligent Assistant based on RAG produced a statistically significant increase in the query service capacity compared with the traditional process.

This effect can be explained by the automation of information retrieval and response generation, which reduces dependence on the immediate availability of personnel and makes it possible to process queries with greater continuity. Consequently, the results support hypothesis H1 and suggest that the RAG architecture constitutes an effective mechanism for expanding the operational capacity of the citizen service.

For Indicator I₂ (accuracy of the response to the query)

For the samples I2: Posttest (Gc) and I2: Posttest (Ge), with $N = 30$ observations per group, $H_0: \mu_1 \geq \mu_2$ was evaluated. The statistic obtained, $t = -12.00$; $p < 0.001$, provides sufficient evidence to reject H_0 and establish that the mean accuracy of the responses of Ge was significantly higher than that of Gc. The negative value of t reflects the expected direction of the difference, as Ge presents a higher level of accuracy.

The significant effect observed in I2 ($t = -12.00$; $p < 0.001$) presents a clear convergence with previous research that attributes improvements in response quality to the incorporation of RAG mechanisms, although the magnitudes and metrics used are not directly equivalent.

It is worth noting that Kalyuzhnaya et al. [26] reported increases of 17% in the accuracy of responses on strategic development and 55% in queries on service accessibility, results that coincide with the direction of the improvement of

6.47 percentage points obtained in the present study and reinforce the hypothesis that providing the model with pertinent external information favors more correct and contextualized responses. In contrast, Wahidur et al. [43] evaluated dimensions distinct from final accuracy, obtaining improvements of 13% in Hit Rate, 15% in MRR, and 23% in relevance of LQ-RAG compared with Naive RAG; therefore, their results should not be numerically compared with the 92.97% accuracy reached in Carabayllo, but they do provide complementary evidence that a more effective and relevant retrieval strengthens the contextual information available for response generation.

Taken together, the three results converge in the same direction: the incorporation of retrieval and specialized knowledge improves different components of the quality of RAG systems; notwithstanding this, the present study provides specific evidence in a municipal context by statistically demonstrating that this improvement translated into a significantly higher accuracy compared with the traditional service channels.

The significant difference supports that the use of the Intelligent Assistant with RAG improved the accuracy of the responses provided to citizens. A possible explanatory mechanism is the prior retrieval of relevant institutional content, which supplies specific context to the language model before generating each response and reduces its exclusive dependence on parametric knowledge. Consequently, the results support H2, although the effect must be circumscribed to the corpus, queries, and experimental conditions evaluated.

For Indicator I₃ (query waiting time)

For the samples I3: Posttest (Gc) and I3: Posttest (Ge), with N = 30 observations per group, H0: $\mu_1 \leq \mu_2$ was contrasted, given that the hypothesis proposed a reduction of the waiting time through the assistant. The result $t = 76.36$; $p < 0.001$ makes it possible to reject H0 and demonstrates that the mean waiting time of Gc was significantly greater than that recorded in Ge. The positive sign of t is consistent with a difference $Gc - Ge$ greater than zero.

The statistically significant reduction of the waiting time observed in the present study is consistent with previous research that recognizes

speed of response as a critical dimension in conversational systems, although the metrics employed are not directly equivalent.

Monteiro et al. [41], after evaluating ten government chatbots, identified response time as one of the usability aspects susceptible to improvement; in contrast, Kalyuzhnaya et al. [26] and Wahidur et al. [43] specifically quantified the technical latency of RAG-based architectures.

Kalyuzhnaya et al. [26] recorded times of 9.2–14.6 s in complete configurations with retrieval and access to services, while Wahidur et al. [43] reported 7.2, 11.2, and 14.6 s according to the complexity of the architecture, showing that greater sophistication of the system may imply increases in latency.

It is worth noting that these values correspond to the technical processing of the queries and not to the comprehensive citizen waiting time evaluated in Carabayllo; therefore, it is not methodologically valid to compare them directly with the 21.63 min recorded in Ge.

Taken together, the antecedents confirm the relevance of temporal performance in government chatbots, while the present study expands this evidence by demonstrating, through an inferential contrast, that the use of the Intelligent Assistant based on RAG was associated with a significant reduction of the waiting time compared with the traditional service process.

This result supports that the implementation of the Intelligent Assistant based on RAG significantly reduced the waiting time to obtain a response. The plausible mechanism lies in the automation of stages that, in traditional service, depend on personnel availability, documentary search, and the manual elaboration of the response, replacing them with semantic retrieval and automated generation.

Therefore, the results support H3 and evidence a significant improvement in the temporal efficiency of the evaluated process.

For Indicator I₄ (customer satisfaction level)

For the samples I4: Posttest (Gc) and I4: Posttest (Ge), with $n = 30$ observations per group, a non-parametric test was applied to contrast whether the satisfaction level of the Experimental Group was higher than that of the Control Group. The

result obtained, $W = 489.50$; $p < 0.001$, makes it possible to reject the null hypothesis and evidences a statistically significant difference in favor of Ge. Therefore, the satisfaction levels recorded after the use of the Intelligent Assistant based on RAG were significantly higher than those observed with the traditional channels.

The statistically significant increase in satisfaction observed in Ge ($W = 489.50$; $p < 0.001$) is consistent with antecedents that show a favorable assessment of conversational systems when these incorporate greater context, usefulness, and ease of interaction, although the differences in scales, populations, and domains prevent a direct comparison of the magnitudes.

It is worth noting that Lee and Rew [30] recorded an increase from 3.30 to 3.90 on a Likert scale of 1 to 5 upon incorporating memory into the chatbot, a result that converges with the direction of the effect obtained in the present study; complementarily, Chou et al. [13] reported mean satisfaction scores of 5.33 and 5.15 on a 7-point scale, in addition to favorable assessments of usability.

In contrast, Kleiman and Barbosa [28] addressed the experience from indicators of usefulness and operation: 72% of users considered ChatPGD useful and 73% reported no operation problems, although negative perceptions associated with inaccurate or irrelevant information (18%), slowness (9%), and error messages (18%) persisted, which evidences that satisfaction can also be conditioned by deficiencies in the quality and performance of the system.

It is relevant, likewise, that Monteiro et al. [41] identified navigation and response time as usability factors susceptible to improvement, while Yun et al. [44] observed favorable effects of the use of RAG-LLM on citizen participation in policy communication; although participation, usability, and satisfaction constitute different constructs, these findings provide complementary evidence on dimensions of the user experience linked with the acceptance of conversational systems.

Along these lines, Gesk and Leyer [40] and Haesevoets et al. [22] show that the acceptance of AI in public services is conditioned by the type of service and by the degree or function assigned to the technology; therefore, greater satisfaction

should not be attributed exclusively to the incorporation of AI or RAG, but rather interpreted in relation to the concrete characteristics of the interaction and the service provided. Taken together, these antecedents support the direction of the effect found, while the present study provides specific inferential evidence that, under the conditions evaluated, the satisfaction levels with the Intelligent Assistant based on RAG were significantly higher than those obtained through the traditional channels.

The significant increase in satisfaction suggests that the use of the assistant improved the citizens' perception of the service process. This effect may be related to mechanisms such as the reduction of waiting time, the greater availability of the service, and the delivery of more accurate responses, factors that decrease the effort required to obtain municipal information.

Consequently, the results support H4 and show that the operational improvements observed with RAG were accompanied by a significantly more favorable assessment on the part of the users.

6 Conclusions and Future Research

I1 - Number of queries. The implementation of the Intelligent Assistant based on RAG significantly increased the service capacity compared with the traditional channels, supporting H1.

This finding suggests that the value of AI in public services does not reside solely in automating responses, but in transforming demand pressure into a greater operational capacity; however, this expansion must be interpreted jointly with the quality, timeliness, and perception of the service. I2 - Response accuracy. In turn, the Intelligent Assistant significantly improved the accuracy of the responses, supporting H2 and evidencing the relevance of retrieving institutional information before generation.

In conceptual terms, the finding shows that performance does not depend exclusively on the parametric capabilities of the language model, but also on the pertinence of the retrieved knowledge; therefore, its scope remains conditioned by the corpus and the queries evaluated. I3 - Waiting time.

Complementarily, the use of the assistant significantly reduced the waiting time, supporting H3 and showing a substantive improvement in temporal efficiency. Nevertheless, it is fundamental to distinguish the comprehensive citizen service time from the technical latency of the system: the observed benefit derives from the automation of intermediate stages of the process, although the recorded variability evidences that the complexity of the queries, the retrieved context, and the technical conditions continue to be relevant factors. I4 - Customer satisfaction level.

Finally, the interaction with the Intelligent Assistant produced satisfaction levels significantly higher than those obtained through the traditional channels, supporting H4.

Taken together, this result shows that the operational improvements of the previous indicators were accompanied by a more favorable assessment of the service; however, satisfaction should not be attributed exclusively to RAG, since it is also conditioned by the accuracy, speed, usefulness, usability, and specific characteristics of the citizen-system interaction.

Future research should evaluate the behavior of these indicators in different types of public services and operational conditions, delving into the effect of the complexity of the queries, the retrieved context, and connectivity on performance.

Likewise, it is pertinent to examine the balance between response quality, latency, usability, and citizen acceptance in order to determine under what conditions the incorporation of RAG generates sustainable benefits in public service.

References

1. **Husain, H., Afandi, R., Sensuse, D. I., Lusa, S., Safitri, N., Elisabeth, D. (2024).** Designing a Knowledge-Based Chatbot to Elevate Business Licensing Services in Indonesia. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 8(5), 681–689. DOI: <https://doi.org/10.29207/resti.v8i5.6069>
2. **Ahn, M. J., Chen, Y.-C. (2022).** Digital transformation toward AI-augmented public administration: The perception of government employees and the willingness to use AI in government. *Government Information Quarterly*, 39(2), Art. no. 101664. DOI: <https://doi.org/10.1016/j.giq.2021.101664>
3. **Al-Besher, A., Kumar, K. (2022).** Use of artificial intelligence to enhance e-government services. *Measurement: Sensors*, 24, 100484. DOI: <https://doi.org/10.1016/j.measen.2022.100484>
4. **Zaidani, H., Koulali, R., Maizate, A., Ouzzif, M. (2025).** Augmentation and Classification of Requests in Moroccan Dialect to Improve Quality of Public Service: A Comparative Study of Algorithms. 1–25.
5. **Alita, L. (2025).** Enhancing citizen engagement in urban greening: The potential of large language models in value co-creation. *Technological Forecasting and Social Change*, 216, Art. no. 124134. DOI: <https://doi.org/10.1016/j.techfore.2025.124134>
6. **Androutsopoulou, A., Karacapilidis, N., Loukis, E., Charalabidis, Y. (2019).** Transforming the communication between citizens and government through AI-guided chatbots. *Government Information Quarterly*, 36(2), 358–367. DOI: <https://doi.org/10.1016/j.giq.2018.10.001>
7. **Madan, R., Ashok, M. (2025).** Making Sense of AI Benefits: A Mixed-method Study in Canadian Public Administration. *Information Systems Frontiers*, 27(3). DOI: <https://doi.org/10.1007/s10796-024-10475-0>
8. **Cardenas-Quispe, A., Vergaray-Mezarina, R., Gamboa-Cruzado, J. (2021).** Machine Learning para la Detección de Malware en Android: Revisión Sistemática de la Literatura. *Revista Ibérica de Sistemas e Tecnologías de Informação (RISTI)*, no. E45, 318–331.
9. **Gamboa-Cruzado, J., Morante-Palomino, E., Alzamora Rivero, C., Lima Bendezu, M., Fernandez, D. M. M. (2024).** Research on the Classification and Application of Physical Education Teaching Mode by Neutrosophic Analytic Hierarchy Process. *International Journal of Neutrosophic Science*, 23(3), 51–62. DOI: <https://doi.org/10.54216/IJNS.230305>

10. **Asquer, A., Krachkovskaya, I. (2022).** Designing public financial management systems: exploring the use of chatbot-assisted case studies. *Public Money & Management*, 42(7), 551–557. DOI: <https://doi.org/10.1080/09540962.2022.2069412>
11. **Chiu, W. A. (2025).** MyEcoReporter: a prototype for artificial intelligence-facilitated pollution reporting. *Journal of Exposure Science & Environmental Epidemiology*. DOI: <https://doi.org/10.1038/s41370-025-00747-5>
12. **Choi, Y., Kim, S., Cedric, Y., Bassole, F. (2025).** Enhanced Retrieval-Augmented Generation Using Low-Rank Adaptation. pp. 1–19.
13. **Chou, Y.-H., Lin, C., Lee, S.-H., Lee, Y.-F., Cheng, L.-C. (2024).** User-Friendly Chatbot to Mitigate the Psychological Stress of Older Adults During the COVID-19 Pandemic: Development and Usability Study. *JMIR Formative Research*, 8, Art. no. e49462. DOI: <https://doi.org/10.2196/49462>
14. **Guo, Z., Cugurullo, F. (2025).** Smart Urbanism through Artificial Intelligence (AI)-Megaprojects: The Case of China's Healthcare Services. *Public Administration and Development*, 1–17. DOI: <https://doi.org/10.1002/pad.2111>
15. **David, A., Yigitcanlar, T., Desouza, K., Li, R. Y. M., Cheong, P. H., Mehmood, R., Corchado, J. (2024).** Understanding local government responsible AI strategy: An international municipal policy document analysis. *Cities*, 155, Art. no. 105502. DOI: <https://doi.org/10.1016/j.cities.2024.105502>
16. **Fischer-Abaigar, U., Kern, C., Barda, N., Kreuter, F. (2024).** Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector. *Government Information Quarterly*, 41, Art. no. 101976. DOI: <https://doi.org/10.1016/j.giq.2024.101976>
17. **Følstad, A., Bjerkreim-Hanssen, N. (2024).** User Interactions With a Municipality Chatbot: Lessons Learnt From Dialogue Analysis. *International Journal of Human-Computer Interaction*, 40(18), 4973–4986. DOI: <https://doi.org/10.1080/10447318.2023.2238355>
18. **Gamboa-Cruzado, J., Blanco-Maslucan, R., Galindo-Arone, D., Jáuregui-Romero, H., Núñez Meza, A., Arauco-Esquivel, S., Núñez-Prado, C. J. (2025).** Augmented Reality as a Tool for Cultural Learning in Educational Institutions in Peru. *Computación y Sistemas*, 29(2). DOI: <https://doi.org/10.13053/CyS-29-2-5739>
19. **Gao, T., Wang, G. (2024).** Integrating IoT and visual question answering in smart cities: Enhancing educational outcomes. *Alexandria Engineering Journal*, 108, 878–888. DOI: <https://doi.org/10.1016/j.aej.2024.09.059>
20. **Gines, J., Manuel, G., Giner, J., De Lucia, P., Antonio, M., Bernabeu, C., Norberto, J., Lopez, M., Carlos, J., Escriba, C., Maria, J., Canales, C. (2024).** Advancing smart tourism destinations: A case study using bidirectional encoder representations from transformers-based occupancy predictions in Torrevieja (Spain). *IET Smart Cities*, 422–440. DOI: <https://doi.org/10.1049/smc2.12085>
21. **Larsen, A.-G., Følstad, A. (2024).** The impact of chatbots on public service provision: A qualitative interview study with citizens and public service providers. *Government Information Quarterly*, 41(2). DOI: <https://doi.org/10.1016/j.giq.2024.101927>
22. **Haesevoets, T., Verschuere, B., Van Severen, R., Roets, A. (2024).** How do citizens perceive the use of Artificial Intelligence in public sector decisions? *Government Information Quarterly*, 41(1), Art. no. 101906. DOI: <https://doi.org/10.1016/j.giq.2023.101906>
23. **Han, T., Cong, R.-G., Yu, B., Tang, B., Wei, Y.-M. (2024).** Integrating local knowledge with ChatGPT-like large-scale language models for enhanced societal comprehension of carbon neutrality. *Energy and AI*, 18, Art. no. 100440. DOI: <https://doi.org/10.1016/j.egyai.2024.100440>
24. **Huayana-Salvador, C., Gamboa-Cruzado, J., Núñez-Prado, C. J. (2024).** Innovation in Mobile Applications with the New Collaborative ASWAN Methodology for Sales Management of Apartments: A Case Study from the Real Estate Sector, Peru.

- Computación y Sistemas, 28(3), 1183–1200. DOI: <https://doi.org/10.13053/CyS-28-3-5005>
25. **Jiang, S., Xie, X., Tang, R., Wang, X. (2025).** ARGUS: Retrieval-Augmented QA System for Government Services.
 26. **Kalyuzhnaya, A., Mityagin, S., Lutsenko, E., Getmanov, A., Aksenkin, Y., Fatkhiev, K., Fedorin, K., Nikitin, N. O., Chichkova, N., Vorona, V., Boukhanovsky, A. (2025).** LLM Agents for Smart City Management: Enhancing Decision Support Through Multi-Agent AI Systems. *Smart Cities*, 8(1), Art. no. 19. DOI: <https://doi.org/10.3390/smartcities8010019>
 27. **Alegre-Veliz, R., Gaspar-Ortiz, P., Gamboa-Cruzado, J., Rodríguez Baca, L., Grandez Pizarro, W., Menéndez Mueras, R., Chavez Herrera, C. (2022).** Machine Learning for Feeling Analysis in Twitter Communications: A Case Study in HEYDRUI!, Peru. *International Journal of Interactive Mobile Technologies (iJIM)*, 16(24), 126–141. DOI: <https://doi.org/10.3991/ijim.v16i24.35493>
 28. **Kleiman, F., Barbosa, M. M. (2024).** Management and Performance Program Chatbot: A Use Case of Large Language Model in the Federal Public Sector in Brazil. *Digital Government: Research and Practice*. DOI: <https://doi.org/10.1145/3700141>
 29. **Kruhlov, V., Bobos, O., Hnylianska, O. (2024).** The Role of Using Artificial Intelligence for Improving the Public Service Provision and Fraud Prevention. 16(02), 913–928.
 30. **Lee, J., Rew, J. (2025).** Memory-Augmented Large Language Model for Enhanced Chatbot Services in University Learning Management Systems. 1–42.
 31. **Arslan, M., Mahdjoubi, L., Munawar, S. (2024).** Driving sustainable energy transitions with a multi-source RAG-LLM system. *Energy, Buildings*, 324, 114827. DOI: <https://doi.org/10.1016/j.enbuild.2024.114827>
 32. **Mansurova, A., Tleubayeva, A., Nugumanova, A., Shomanov, A. (2025).** A Systematic Evaluation of Large Language Models and Retrieval-Augmented Generation for the Task of Kazakh Question Answering. 1–34.
 33. **De Nicola, A., Formica, A., Mele, I., Missikoff, M. (2024).** A comparative study of LLMs and NLP approaches for supporting business process analysis. *Enterprise Information Systems*, 18(10), 1240–1268. DOI: <https://doi.org/10.1080/17517575.2024.2415578>
 34. **Papageorgiou, G., Sarlis, V., Maragoudakis, M. (2025).** A Multimodal Framework Embedding Retrieval-Augmented Generation with MLLMs for Eurobarometer Data.
 35. **Piizzi, A., Vavallo, D., Lazzo, G., Dimola, S., Zazzera, E., Tempo, S. R. L. (2024).** A Natural Language Processing Model for the Development of an Italian-Language Chatbot for Public Administration. 15(9), 54–58.
 36. **Cortés-Cediel, M. E., Segura-Tinoco, A., Cantador, I., Rodríguez Bolívar, M. P. (2023).** Trends and challenges of e-government chatbots: Advances in exploring open government data and citizen participation content. *Government Information Quarterly*, 40, 101877. DOI: <https://doi.org/10.1016/j.giq.2023.101877>
 37. **Aparcana-Tasayco, A. J., Gamboa-Cruzado, J. (2022).** Machine Learning for Management in Software-defined Networks: A Systematic Literature Review. *IEIE Transactions on Smart Processing and Computing*, 11(6), 400–411. DOI: <https://doi.org/10.5573/IEIESPC.2022.11.6.400>
 38. **Langgeng, Y. S. H., Setiawan, E. I., Imron, S., Santoso, J. (2023).** Long Short-Term Memory-Based Chatbot for Vocational Registration Information Services. *Journal of Applied Data Sciences*, 4(4), 414–430. DOI: <https://doi.org/10.47738/jads.v4i4.128>
 39. **Sievers, T., Russwinkel, N. (2024).** Project Report: Requirements for a Social Robot as an Information Provider in the Public Sector. *KI - Künstliche Intelligenz*, 38(3), 145–149. DOI: <https://doi.org/10.1007/s13218-024-00840-1>
 40. **Gesk, T. S., Leyer, M. (2022).** Artificial intelligence in public services: When and why citizens accept its usage. *Government Information Quarterly*, 39(3), Art. no. 101704.

DOI: <https://doi.org/10.1016/j.giq.2022.101704>

41. **Monteiro, M. de S., Batista, G. O. da S., Salgado, L. C. de C. (2023).** Investigating usability pitfalls in Brazilian and Foreign governmental chatbots. *Journal on Interactive Systems*, 14(1). DOI: <https://doi.org/10.5753/jis.2023.3104>
42. **Vassilakopoulou, P., Haug, A., Salvesen, L. M., Pappas, I. O. (2023).** Developing human/AI interactions for chat-based customer services: lessons learned from the Norwegian government. *European Journal of Information Systems*, 32(1), 10–22. DOI: <https://doi.org/10.1080/0960085X.2022.2096490>
43. **Wahidur, R. S. M., Kim, S., Choi, H., Bhatti, D. S., Lee, H.-N. (2025).** Legal Query RAG.

IEEE Access, 13, 36978–36994. DOI: <https://doi.org/10.1109/ACCESS.2025.3542125>

44. **Yun, L., Yun, S., Xue, H. (2024).** Improving citizen-government interactions with generative artificial intelligence: Novel human-computer interaction strategies for policy understanding through large language models. *PLOS ONE*, 19(12), Art. no. e0311410. DOI: <https://doi.org/10.1371/journal.pone.0311410>
45. **Arslan, M., Munawar, S., Cruz, C. (2024).** Business insights using RAG-LLMs: a review and case study. *Journal of Decision Systems*. DOI: <https://doi.org/10.1080/12460125>.

Article received on 06/05/2026; accepted on 21/08/2026.

**Corresponding author is Bella Citlali Martínez-Seis.*