



Development and Query Analysis of an LLM-Powered Police Chatbot with Expert Accuracy Evaluation

Nabila Putri Shalehah¹, Mohammad Givi Efgivia²

^{1,2}Faculty of Industrial Technology and Informatics, Universitas Muhammadiyah Prof. Dr. Hamka, Indonesia

ARTICLE INFO	ABSTRACT
Published Online: 15 July 2026 Corresponding Author: Mohammad Givi Efgivia	Although Conversational AI significantly improves the delivery of public information, Large Language Models (LLMs) frequently face challenges in adhering strictly to Standard Operating Procedures (SOPs) and rarely leverage interaction logs for strategic analysis. To overcome these limitations, this study presents a novel police public service chatbot built on the Laravel framework, integrating the Google Gemini API with Information Retrieval (IR) methodologies and an official SOP database. The system processes user inquiries through text preprocessing, TF-IDF weighting, and Cosine Similarity to extract relevant documents, guaranteeing responses that are both contextually accurate and SOP-compliant. Furthermore, it systematically records conversation logs to analyze public needs and trends. Comprehensive evaluation via Black Box testing, IR metrics, and qualitative reviews demonstrated superior performance. The proposed system achieved an accuracy of 95.42%, outperforming rule-based (82.14%), standard TF-IDF (89.72%), and baseline Gemini (93.41%) approaches. It also received excellent satisfaction ratings from domain experts (4.83/5) and a cohort of 100 citizens (4.73/5). In conclusion, this architecture effectively combines the conversational fluency of LLMs with the strict procedural precision necessary for robust public service administration.
KEYWORDS: Large Language Model, Public Service Chatbot, Query Analytics, Artificial Intelligence, Expert Based Evaluation.	

I. INTRODUCTION

AI-driven digital transformation, particularly through Large Language Models (LLMs), is revolutionizing public service delivery by outperforming traditional rule-based chatbots. The public reacts positively to transparent and responsive government chatbots [1]. However, success depends primarily on robust knowledge bases and inter-agency collaboration [2]. Aligning with this, Indonesia’s Electronic-Based Government System (SPBE) prioritized intelligent public service chatbots for 2025–2027. Despite this momentum, operational challenges persist. Accountability, trust, and misinterpretation issues often hinder chatbot effectiveness. Consequently, many institutions including law enforcement heavily rely on manual information dissemination for routine procedures like obtaining Police Clearance Certificates (SKCK). This manual reliance creates service bottlenecks, increases administrative workloads, and restricts after-hours

availability, draining resources better suited for strategic policing duties. Furthermore, existing literature reveals three research gaps. First, systems using Retrieval-Augmented Generation (RAG) are rarely anchored directly to official Standard Operating Procedures (SOPs), creating risks of procedural inaccuracies a vulnerability observed in specialized RAG frameworks struggling with factual accuracy [3]. Second, interaction logs are rarely analyzed to uncover public demand patterns. Third, system evaluations predominantly depend on computational metrics and user satisfaction rather than rigorous validation by domain experts, which is critical in regulated environments. This research urgently addresses these gaps within police services, an under-scrutinized sector handling overwhelming repetitive inquiries. Given the high stakes of legal accuracy, AI systems in policing require meticulous reliability testing [4]. Thus, this study develops an LLM-powered chatbot

synergizing Google Gemini AI, information retrieval, Query Frequency Analytics, and Expert-Based Evaluation, strictly grounded in authorized police documentation. Academically, this introduces an evaluation framework combining technical metrics with domain-expert validation, countering automation bias in government AI [5]. Practically, this system alleviates administrative burdens, provides 24/7 access, and extracts data-driven insights to enhance police services.

II. LITERATURE REVIEW

A. Artificial Intelligence in Public Service

Artificial Intelligence (AI) serves as a catalyst for digital transformation within government sectors by improving operational efficiency and citizen-centric services via sophisticated data analytics and natural language processing [6],[7]. Conversational agents have become especially prevalent, providing round-the-clock access to public data through natural dialogue, which consequently lowers administrative burdens and increases systemic transparency [8][9],[10]. Nevertheless, strictly regulated organizations, such as law enforcement agencies, demand that these technologies rigidly comply with official Standard Operating Procedures (SOPs). Given that contemporary chatbots frequently miss the mark on institutional validation and conversational analytics, combining AI capabilities with expert assessments is imperative for ensuring dependable public service delivery [11].

B. Large Language Model

Large Language Models (LLMs) including GPT, Gemini, and Llama have significantly propelled conversational AI forward. By analyzing massive volumes of text, they grasp intricate contexts and user intents much more effectively than conventional, rule-based systems [12],[13]. Relying on a unified Transformer architecture, these models expertly handle a wide array of linguistic functions to facilitate fluid and adaptable user interactions [14],[15]. To minimize the occurrence of "hallucinations" instances where the AI produces plausible yet factually flawed statements contemporary methodologies employ Retrieval-Augmented Generation (RAG). By integrating LLMs with Information Retrieval (IR) techniques and verified official databases, these frameworks can deliver the precise, contextually grounded, and reliable answers that are vital in public service scenarios [16].

C. Public Service Chatbot

Within the realm of e-government, AI-powered chatbots enable uninterrupted, tailored access to essential administrative functions, such as licensing and law enforcement inquiries. This drastically reduces the need for in-person interactions while simultaneously bolstering accountability [17],[18],[19]. Although LLMs are vastly superior to legacy systems at deciphering varied user intents, numerous public-sector chatbots still operate without proper

institutional validation frameworks and miss opportunities to utilize chat logs strategically. Fusing LLMs with Information Retrieval, query analytics, and expert oversight yields a robust solution capable of supporting data-informed, procedurally compliant public administration [20].

D. Query Analytics Using TF IDF

Dialogue histories contain essential clues regarding user behavior and the specific demands placed on public services [21],[22]. To analyze this information effectively, the Term Frequency-Inverse Document Frequency (TF-IDF) method is used to translate user inquiries into numerical weighted vectors [23],[24]. After initial text preprocessing steps, TF-IDF works in tandem with Cosine Similarity to precisely locate and extract the most applicable institutional records [25]. Moving beyond mere document retrieval, Query Analytics converts these chat logs into practical organizational intelligence, spotting emerging trends and Frequently Asked Questions (FAQs) to guide evidence-based policy and decision-making.

E. Expert Based Evaluation

Relying solely on quantitative metrics like Precision and F1-score is insufficient to verify a chatbot's practical value within an institution, because excellent technical scores do not necessarily equate to strict SOP adherence [26]. In sectors such as law enforcement where procedural exactness regarding permits or Police Clearance Certificates carries critical legal weight Expert-Based Evaluation is an absolute necessity to guarantee that system outputs are both accurate and secure [27],[28]. Additionally, human oversight remains crucial for judging the conversational flow, reliability, and overall qualitative nuances of LLM-generated text [29],[30]. Uniting this expert validation with advanced analytics and LLM technology creates a public service framework that is technologically proficient, legally sound, and strategically beneficial.

III. METHODOLOGY

A. Research Workflow

Utilizing a Design and Development Research (DDR) framework, this study develops an intelligent chatbot for the Pino Police Sector. As depicted in Figure 1, the research unfolds across the development, implementation, and evaluation phases. The foundational knowledge base was formulated by collecting requirements through interviews, field observations, and SOP reviews. The chatbot was subsequently built as a Laravel web application, fusing the Google Gemini API with TF-IDF and Cosine Similarity to retrieve relevant documents. To ensure reliability, the system was thoroughly tested using Black Box Testing, Information Retrieval metrics, Expert Judgment, and human evaluation.

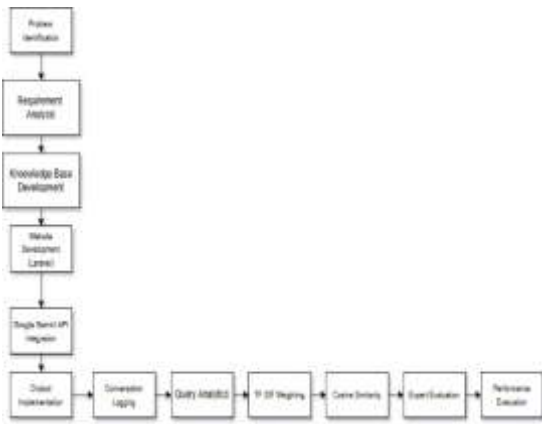


Figure 1. Research Workflow

In addition to delivering accurate replies, the system capitalizes on interaction logs to extract actionable insights, facilitating data-driven decisions and ongoing service optimization.

B. Requirement Analysis

A comprehensive requirement analysis was conducted to establish both the functional and non-functional specifications of the intended chatbot. To gather data, we observed public service operations at Polsek Pino directly, conducted structured interviews with police personnel, and examined official Standard Operating Procedures (SOPs). The findings indicated a high volume of citizen inquiries regarding standard procedures, including Police Clearance Certificates (SKCK), lost property reports, event permits, complaint mechanisms, and general operational details. Given their repetitive nature, these questions currently drain substantial administrative time and resources. Consequently, key functional requirements were established. The system is required to deliver automated answers in natural language by extracting precise details directly from authorized SOPs. Furthermore, to facilitate ongoing enhancements, the chatbot must log all user conversations in a database. This allows for the evaluation of public information needs and feeds data into analytical dashboards for strategic decision-making. These functional objectives are supported by essential non-functional criteria such as system scalability, usability, information security, response precision, and rigorous institutional alignment which collectively informed the subsequent system architecture and evaluation methods.

C. Knowledge Base Development

The success of applications driven by LLMs depends significantly on the caliber of domain specific data supplied during the generation of responses. To guarantee that the chatbot provides dependable and precise information, a structured Knowledge Base was constructed using official police records.

This repository was assembled using the Standard Operating Procedures (SOPs) of Polsek Pino, encompassing administrative rules, operational directives,

and public service protocols. The compiled documents were categorized into six main themes:

- Police Clearance Certificate (SKCK)
- Lost Property Report
- Public Event Permit
- Public Complaint Services
- Operational Information
- Office Contact Information

Prior to system integration, each document was subjected to preprocessing steps such as text cleaning, normalization, deduplication, and structural formatting to maintain uniformity and maximize the efficiency of Information Retrieval. In contrast to traditional chatbots that depend on pre-scripted, static replies, this architecture dynamically extracts pertinent institutional documents before crafting a response through the Google Gemini API. This methodology guarantees that the generated outputs strictly adhere to official policies while significantly reducing factual errors.

D. System Architectur

The chatbot was developed as a web application utilizing the Laravel framework, selected for its robust RESTful API integration features and scalable MVC (Model View Controller) structure. The framework functions on a modular client-server architecture, ensuring seamless interaction among various system components.

The process initiates at the user interface, allowing citizens to engage with the chatbot using natural language. The Laravel application server processes these incoming requests, serving as the central control unit. It oversees user authentication, orchestrates the retrieval of data from the SOP knowledge base, and interfaces directly with the Google Gemini API to produce contextually relevant answers.

Concurrently, a dedicated database meticulously records all interaction details, including user queries, AI responses, and corresponding timestamps. This stored data is channeled into an analytics dashboard, equipping administrators with real-time visualizations of trending topics, frequently accessed services, and general usage patterns. In essence, this modular configuration promotes fluid communication between the institutional data and AI components while securing the system's long-term scalability.

E. Google Gemini API Integration

Functioning as the primary language model, the Google Gemini API is responsible for interpreting user prompts and generating contextually appropriate responses. To prevent the model from depending exclusively on its pre-trained, generic data, Gemini was firmly grounded in the institutional knowledge base. This integration is essential for preserving procedural uniformity and factual correctness.

The process of generating responses utilizes a smooth, retrieval-augmented pipeline. Upon receiving a natural

language question from a user, the input is subjected to standard text preprocessing, which involves case folding, tokenization, stopword filtering, and stemming. Subsequently, the framework employs TF-IDF weighting and Cosine Similarity to identify and pull the most applicable SOP records from the knowledge base. These selected documents are then combined with the users initial query to form a structured prompt, which is transmitted to the Gemini API. By merging the users request with authoritative institutional context, Gemini constructs a precise, policy compliant answer. Ultimately, the reply is presented to the user while the complete interaction is simultaneously archived in the database. This strategic method drastically reduces factual hallucinations and guarantees rigorous compliance with law enforcement protocols.

F. Query Analytics

In addition to producing answers, the system archives comprehensive interaction data such as user queries, system responses, and timestamps to yield strategic insights for the organization. As illustrated in Figure 2, these raw conversation logs undergo continuous analysis to produce three primary metrics: frequently asked questions, the most sought-after services, and temporal usage trends. This analytical capability transforms the chatbot from a simple automated responder into a sophisticated decision-support mechanism, empowering police administrators to track public demand and formulate evidence-based operational decisions.

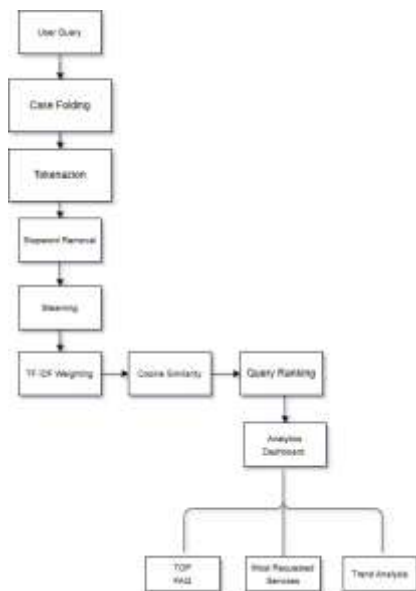


Figure 2. Query Analytics Workflow

The diagram demonstrates how user interactions serve a dual purpose: they are utilized to supply immediate answers to the public and to extract valuable analytical insights for continuous service enhancements.

G. Term Frequency Inverse Document Frequency (TF IDF)

To accurately match user inquiries with institutional records, this research utilizes Term Frequency-Inverse Document Frequency (TF-IDF) for term weighting. TF-IDF excels in Information Retrieval by algorithmically diminishing the impact of common words while emphasizing domain-specific terminology. By integrating Term Frequency (TF) which measures how often a word appears in a specific document with Inverse Document Frequency (IDF), which lowers the weight of words found broadly across the dataset, the framework produces precise numerical vectors. These vectors are subsequently processed through a Cosine Similarity function to extract the most applicable Standard Operating Procedure (SOP) documents for the chatbot.

$$TF(t, d) = f(t, d)$$

(TF(t,d)) = Term Frequency of term (t) in document (d)

(f(t,d)) = Number of occurrences of term (t) within document (d)

Since relying solely on term frequency is insufficient for filtering out common vocabulary, it is combined with Inverse Document Frequency (IDF) to give precedence to unique, highly relevant words.

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

(N) = Total number of documents

(df(t)) = Number of documents containing term (t)

Words that emerge across numerous documents are assigned lower IDF scores, as they offer less distinguishing value during the retrieval process.

$$TFIDF(t, d) = TF(t, d) \times IDF(t)$$

H. Cosine Similarity

After the TF-IDF vectorization process, Cosine Similarity is utilized to evaluate the semantic correlation between the user's input and the knowledge base. The document achieving the highest similarity ranking is extracted and integrated into the prompt for the Google Gemini API. This retrieval-augmented methodology anchors the AI's answers in authenticated SOPs, guaranteeing exceptional contextual accuracy and rigorous adherence to procedures.

$$\cos(Q, D) = \frac{Q \cdot D}{\|Q\| \|D\|}$$

- (Q) = User query vector
- (D) = Knowledge base document vector
- (Q · D) = Dot product between vectors
- (|Q|) = Magnitude of query vector
- (|D|) = Magnitude of document vector

The core innovation of this framework lies in its retrieval-augmented generation of responses. As illustrated in Figure 4, instead of depending exclusively on the LLMs inherent pre-trained data, the architecture leverages TF-IDF and Cosine Similarity to pull applicable SOPs prior to querying the Google Gemini API. Anchoring the model's replies in

these official documents drastically enhances factual reliability and curtails AI hallucinations.



Figure 3. Information Retrieval Workflow

As demonstrated in Figure 5, the Query Analytics module evaluates post-response engagements through the continuous assessment of conversation logs. This capability converts standard user questions into valuable strategic intelligence, allowing administrators to track service trends and improve public service operations based on empirical data.



Figure 4. Algorithm Chatbot Response Generation

In contrast to traditional conversational agents, this system actively mines chat logs, empowering administrators to drive service enhancements based on genuine public needs. A major advantage is its retrieval-augmented design, which pairs Information Retrieval techniques (TF-IDF and Cosine Similarity) with the Google Gemini API. By embedding pertinent SOP documents as validated context rather than trusting only the LLM's baseline training the framework reduces factual errors and guarantees rigid policy adherence. Following every response, all user interactions are recorded to power the analytics engine, yielding a dual advantage of precise answers and strategic intelligence that distinguishes this architecture from conventional LLM platforms.

I. Evaluation

To evaluate the system's effectiveness beyond basic computational statistics, this research established a multi-faceted evaluation framework that measures functional accuracy, document retrieval efficiency, output quality, and adherence to institutional protocols. Through the combination of Black Box Testing, Information Retrieval (IR) metrics, expert assessments, and human feedback, this evaluation guarantees that the application is not only

technically sound and highly practical but also strictly compliant with organizational requirements.

1) Black-box Testing

We conducted Black Box testing to ensure that every system component meets the established functional specifications. Without reviewing the internal code structure, this phase exclusively focuses on the system's functional outcomes. Each module is tested through a series of predefined input scenarios to ensure that the system consistently generates accurate outputs that align with expectations under real operational conditions.

The modules evaluated include:

- User Authentication
- Chatbot Interface
- Google Gemini API Integration
- Knowledge Base Access
- Conversation Database
- Query Analytics Dashboard

$$\text{Success Rate} = \frac{\text{Number of successful tests}}{\text{Total test case}} \times 100\%$$

2) Information Retrieval Evaluation

To measure the extent to which the system is capable of retrieving relevant institutional documents before generating a response, we evaluated the chatbot using standard Information Retrieval (IR) metrics: Accuracy, Precision, Recall, and F1-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Akurasi} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

- TP = True Positive
- FP = False Positive
- TN = True Negative
- FN = False Negative

3) Expert Based Evaluation

Given the sensitive nature of police information, computational metrics alone are insufficient to guarantee institutional validity. To ensure that the chatbot's responses are fully aligned with Standard Operating Procedures (SOPs), we engaged in expert-based evaluations conducted by two specialists in the field:

- Chief of Police (Kapolsek)
- Public Service Operator

Each expert assessed the chatbot's responses using a five-point Likert scale, ranging from 1 (Very Poor) to 5 (Very Good), based on the following criteria:

- Information Accuracy
- SOP Compliance
- Completeness
- Response Clarity

$$\bar{x} = \frac{\sum xi}{n}$$

- $\bar{x}$ = Average expert score
- x_i = Individual evaluation score
- n = Number of evaluators

4) Research Variables

To systematically evaluate the chatbot's architecture and performance, we defined specific independent and dependent variables as presented in Table 1. These variables provide a structured framework for measuring all aspects of the system, from technical data retrieval capabilities to the final conversational output.

Table 1. Research Variable

Variabel	Indicator	Measurant Method
Large Language Model (LLM)	Response Generation	Response relevance and contextual quality
TF-IDF	Keyword Weighting	TF-IDF Score
Cosine Similarity	Document Matching	Cosine Similarity Value
Expert Judgment	SOP Validation	Likert Scale (1–5)
Chatbot Performance	Accuracy, Precision, Recall, F1-score	Information Retrieval Metrics

These independent variables represent the core technological components of our framework, while the chatbot's overall performance serves as the dependent variable. Collectively, these variables facilitate a holistic evaluation that encompasses technical, organizational, and user perspectives.

5) Research Evaluation Framework

Figure 6 presents the comprehensive evaluation

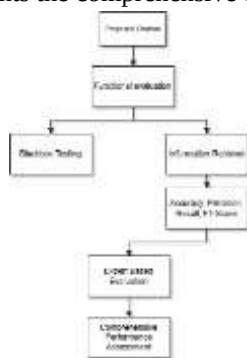


Figure 5. Research Evaluation Framework

This integrated methodology incorporates the Retrieval-Augmented Generation (RAG) workflow, which leverages TF-IDF, Cosine Similarity, and Google Gemini into our analytical evaluation module. Through a series of comprehensive assessments (including black-box testing, IR metrics, and expert evaluation), we ensure that the resulting

chatbot possesses technical robustness, procedural compliance, and tangible value for public service.

IV. RESULT AND DISCUSSION

A. System Implementation

The Laravel based public service chatbot for Polsek Pino serves a dual purpose: acting as a citizen information interface and an administrative analytics tool. Behind the scenes, the system processes every user query through an Information Retrieval pipeline (using TF IDF and Cosine Similarity) to ensure the Google Gemini API generates accurate responses grounded in official police SOPs. Beyond answering inquiries, all interactions are automatically logged in an analytics dashboard that highlights usage trends, empowering police administrators with data-driven insights to continuously improve service quality.

B. Black Box Testing

To ensure real-world reliability, we conducted Black Box testing across all primary modules, covering operational scenarios from user authentication and conversations to API integration and analytics. As summarized in Table 2, all tested components functioned exactly as intended, confirming the overall functional stability of the chatbot.

Table 2. Black Box Testing Results

No Module	Status
1. Login Authentication	Successful
2. Chatbot Interface	Successful
3. Google Gemini API	Successful
4. Knowledge Base Retrieval	Successful
5. Conversation Database	Successful
6. Query Analytics Dashboard	Successful
Success Rate :	100%

The testing phase achieved a 100% success rate across all essential modules, including authentication, query processing, API communication, and analytics. These flawless results confirm the system's architectural stability, establishing a reliable foundation for subsequent performance, expert, and human evaluations.

C. Information Retrieval Performance

Having established functional reliability, we evaluated the retrieval mechanism's effectiveness using a dataset of 1,236 conversational queries. To ensure information accuracy, the retrieved documents were manually cross-referenced against official SOPs. These outcomes were then quantified using standard IR metrics (Accuracy, Precision, Recall, and F1-score), as detailed in Table 3.

Table 3. Information Retrieval Performance

No Metric	Value (%)
1. Accuracy	95.42
2. Precision	94.87
3. Recall	96.13
4. F1-score	95.49

D. Query Analytics

Beyond generating responses, the system transforms raw logs from 1,236 queries into strategic intelligence. By extracting dominant topics via TF-IDF, the analysis yields clear statistical insights into FAQs and usage trends. This continuous, data-driven overview of public demand empowers police administrators to evaluate and enhance service quality based on real citizen interactions.

E. Top Frequently Asked Questions

By mining the accumulated conversation logs, we mapped the primary information needs of the citizens. Table 4. Top Frequently Asked Questions, outlines the ten most common inquiries submitted during the evaluation period, highlighting the core public services that drive community engagement with the police department.

Table 4. Top Frequently Asked Questions

Rank	User Query	Frequency
1	How to apply for a Police Clearance Certificate (SKCK)?	172
2	Requirements for obtaining an SKCK?	145
3	Police office operating hours?	118
4	How to report lost property?	107
5	Public event permit procedures?	95
6	SKCK service fees?	83
7	SKCK renewal procedure?	74
8	How to report theft?	63
9	Police contact number?	59
10	Police station address?	54

The analysis revealed that inquiries regarding Police Clearance Certificates (SKCK), operating hours, and lost property reports dominate public information needs. Automating these repetitive queries highlights the chatbot's effectiveness in significantly reducing administrative overhead while ensuring 24/7 service. Furthermore, identifying these recurring topics equips police administrators with a strategic, data-driven roadmap to refine service procedures and optimize public communication based on actual community needs.

F. Most Requested Services

To gain a broader view of public demand, we categorized the conversational logs by service type. As summarized in Table 5. Most Requested Public Services, this distribution analysis helps clarify which administrative areas require the most attention and resource allocation within the Polsek Pino office.

Table 5. Most Requested Services

No Service Category	Total Requests
1. Police Clearance Certificate (SKCK)	392
2. Lost Property Report	286
3. Public Event Permit	218
4. Public Complaint Service	194

No Service Category	Total Requests
5. Operational Information	146

Data reveals that administrative services such as Police Clearance Certificates (SKCK), Lost Property Reports, and Event Permits dominate interactions, indicating citizens primarily rely on the chatbot for procedural guidance. This usage pattern provides a clear mandate for police administrators to prioritize the digitalization and continuous updating of high-demand services (particularly SKCK). By leveraging this analytical data, resource allocation and service planning can be conducted strategically and based on solid evidence, moving away from guesswork to directly address actual citizen needs.

G. Trend Analysis

To understand how chatbot utilization evolved over time, we tracked conversation statistics on a weekly basis throughout the experimental period. Table 6. Weekly Chatbot Utilization Trends, summarizes these usage patterns, providing a chronological view of how the public engaged with the system.

Table 6. Weekly Chatbot Usage

Week	Total Queries
Week 1	180
Week 2	246
Week 3	317
Week 4	493
Total	1,236

As shown in Figure 7. Weekly Chatbot Usage Trend, the consistent rise in interaction volume reflects growing public trust in this AI-assisted service. Chatbot usage surged dramatically from 180 queries in the first week to a peak of 493 queries by the fourth week. This notable 174% increase clearly underscores the rapid public adoption of the chatbot for police information services.

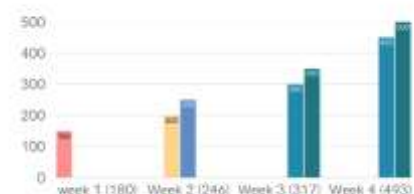


Figure 6. Weekly Chatbot Usage Trend

The upward usage trend indicates that citizens increasingly trust the chatbot as a reliable and stable channel for police information. Beyond simple automation, this highlights the strategic value of the Query Analytics framework. By continuously monitoring interaction patterns, police administrators can transform raw conversation logs

into organizational intelligence. This data-driven approach allows them to anticipate resource needs and prioritize service improvements, elevating the system far beyond the limitations of standard, passive chatbots.

H. Expert Based Evaluation

Since computational metrics cannot fully assess institutional policy compliance, two domain experts (the Chief of Police and a Public Service Officer) evaluated the chatbot using a 5-point Likert scale. As shown in Table 7. Expert Based Evaluation Results, they assessed four key dimensions: Information Accuracy, SOP Compliance, Completeness, and Response Clarity. This expert evaluation ensures that the chatbot's responses are not only technically accurate but also strictly adhere to official police SOPs, making them institutionally valid for real-world deployment.

Table 7. Expert Based Evaluation Results

<i>Evaluation Aspect</i>	<i>Mean Score</i>
Information Accuracy	4.82
SOP Compliance	4.91
Completeness	4.74
Response Clarity	4.86
Overall Mean	4.83 / 5.00
Category:	Very Good

The expert evaluation achieved an overall score of 4.83/5 ("Very Good"). SOP Compliance earned the highest rating (4.91), confirming strict procedural consistency, followed by strong scores in Information Accuracy (4.82) and Response Clarity (4.86). While the Completeness score (4.74) highlights an opportunity to expand the knowledge base, these results ultimately validate that grounding Google Gemini in official police documents creates a trustworthy system ready for real-world public service.

I. Comparison with Existing Methods

To contextualize our results, we compared the chatbots performance against three baseline approaches commonly cited in conversational AI research. Table 9. Performance Comparison summarizes how our framework stacks up against these benchmarks, highlighting the efficacy of our retrieval augmented architecture in delivering superior institutional accuracy and user satisfaction.

Table 8. Performance Comparison

<i>Method</i>	<i>Accuracy (%)</i>
Rule Based Chatbot	82.14
TF IDF Retrieval	89.72
Google Gemini Only	93.41
Proposed Method	95.42

The proposed framework consistently outperformed the baseline methods, achieving an accuracy of 95.42%. While the rule-based approach was hampered by rigid, predefined paths and the TF IDF method lacked conversational depth, our retrieval-augmented architecture combined the best of both worlds. Compared to standalone Google Gemini which is prone to hallucinations our system

grounds every response in verified SOP documents, ensuring that conversational flexibility never comes at the cost of factual accuracy.

J. DISCUSSION

The results confirm that a hybrid approach integrating LLMs with IR and analytics is essential for public service AI. By grounding Google Gemini in official SOPs, we successfully mitigated the inherent risks of LLM hallucinations while maintaining natural interaction. Beyond serving as a virtual assistant, the frameworks Query Analytics module turns passive conversation logs into actionable intelligence. For police administrators, this means moving away from speculative service planning toward an evidence based approach, where resource allocation is driven by real time citizen demand. Furthermore, the high scores in expert validation (4.91/5) for SOP compliance and user satisfaction (4.73/5) demonstrate that technical accuracy, institutional validity, and user experience can be achieved simultaneously in a single, scalable architecture.

V. CONCLUSION

This research demonstrates that a unified AI framework combining Google Gemini, information retrieval, and query analytics effectively modernizes police public services. The system achieved a 100% functional success rate, 95.42% retrieval accuracy, and high marks in both expert compliance (4.83/5) and user satisfaction (4.73/5).

A key contribution of this work is the transformation of conversational data into a decision support tool. By identifying high demand services and trending inquiries, the system enables administrators to optimize resource distribution and service delivery continuously. While this study focuses on a specific police jurisdiction, the framework is inherently scalable. Future development will focus on transitioning to advanced RAG (Retrieval Augmented Generation) pipelines, incorporating multimodal capabilities, and implementing real time knowledge synchronization to further enhance the systems interoperability across broader government sectors.

REFERENCES

1. M. Salar Khan and E. Arledge, "How to Promote AI in the US Federal Government: Insights from Policy Process Frameworks," 2023.
2. T. Chen, M. Gascó-Hernandez, and M. Esteve, "The Adoption and Implementation of Artificial Intelligence Chatbots in Public Organizations: Evidence from U.S. State Governments," *Am. Rev. Public Adm.*, vol. 54, no. 3, pp. 255–270, Apr. 2024, doi: 10.1177/02750740231200522.
3. D. Baur, J. Ansorg, C. E. Heyde, and A. Voelker, "Development and Evaluation of a Retrieval-Augmented Generation Chatbot for Orthopedic and Trauma Surgery Patient Education: Mixed-Methods Study," *JMIR AI*, vol. 4, 2025, doi: 10.2196/75262.

4. E. Halford and A. Webster, “Using chat GPT to evaluate police threats, risk and harm,” *Int. J. Law Crime Justice*, vol. 78, Sep. 2024, doi: 10.1016/j.ijlcj.2024.100686.
5. S. Alon-Barkat and M. Busuioc, “Human-AI Interactions in Public Sector Decision Making: ‘Automation Bias’ and ‘Selective Adherence’ to Algorithmic Advice,” *Journal of Public Administration Research and Theory*, vol. 33, no. 1, pp. 153–169, Jan. 2023, doi: 10.1093/jopart/muac007.
6. S. Mellouli, M. Janssen, and A. Ojo, “Introduction to the Issue on Artificial Intelligence in the Public Sector: Risks and Benefits of AI for Governments,” *Digital Government: Research and Practice*, vol. 5, no. 1, Mar. 2024, doi: 10.1145/3636550.
7. T. S. Gesk and M. Leyer, “Artificial intelligence in public services: When and why citizens accept its usage,” *Gov. Inf. Q.*, vol. 39, no. 3, Jul. 2022, doi: 10.1016/j.giq.2022.101704.
8. D. H. Hagos, R. Battle, and D. B. Rawat, “Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives,” Aug. 2024, [Online]. Available: <http://arxiv.org/abs/2407.14962>
9. A. Androniceanu, “Generative artificial intelligence, present and perspectives in public administration,” *Administratie si Management Public*, vol. 2024, no. 43, pp. 105–119, 2024, doi: 10.24818/amp/2024.43-06.
10. M. Babšek, D. Ravšelj, L. Umek, and A. Aristovnik, “Artificial Intelligence Adoption in Public Administration: An Overview of Top-Cited Articles and Practical Applications,” Mar. 01, 2025, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ai6030044.
11. A. F. Vatamanu and M. Tofan, “Integrating Artificial Intelligence into Public Administration: Challenges and Vulnerabilities,” *Adm. Sci.*, vol. 15, no. 4, Apr. 2025, doi: 10.3390/admsci15040149.
12. S. Wang *et al.*, “GPT, large language models (LLMs) and generative artificial intelligence (GAI) models in geospatial science: a systematic review,” 2024, *Taylor and Francis Ltd*. doi: 10.1080/17538947.2024.2353122.
13. H. Naveed *et al.*, “A Comprehensive Overview of Large Language Models,” Oct. 2024, [Online]. Available: <http://arxiv.org/abs/2307.06435>
14. C. Munyao and J. Ndia, “Natural Language Processing with Transformer-Based Models: A Meta-Analysis,” *Journal on Artificial Intelligence*, vol. 7, no. 1, pp. 329–346, 2025, doi: 10.32604/jai.2025.069226.
15. G. Feretzakis, K. Papaspyridis, A. Gkoulalas-Divanis, and V. S. Verykios, “Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review,” Nov. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/info15110697.
16. E. Jo, D. A. Epstein, H. Jung, and Y. H. Kim, “Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention,” in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, Apr. 2023. doi: 10.1145/3544548.3581503.
17. S. Kumar, “Smarter Governance: How Business AI Is Shaping The Future Of Public Services,” 2023.
18. A. Nikiforova, A. P. Rodriguez Müller, L. Tangi, and J. Martin-Bosch, “Proactive Public Services in the Age of Artificial Intelligence: Towards Post-Bureaucratic Governance,” in *Lecture Notes in Computer Science*, Springer Science and Business Media Deutschland GmbH, 2026, pp. 395–413. doi: 10.1007/978-3-032-01589-1_25.
19. T. Yigitcanlar, R. Y. M. Li, T. Inkinen, and A. Paz, “Public Perceptions on Application Areas and Adoption Challenges of AI in Urban Services,” *Emerging Science Journal*, vol. 6, no. 6, pp. 1199–1236, Dec. 2022, doi: 10.28991/ESJ-2022-06-06-01.
20. A. A. Chinchamatpure and S. B. Khandare, “AI-Driven Digital Transformation in Local Governance: Towards Sustainable and Paperless Operations,” 2025.
21. M. Das, S. Kamalanathan, and P. Alphonse, “A Comparative Study on TF-IDF feature Weighting Method and its Analysis using Unstructured Dataset,” 2020.
22. L. Heryawan, D. Novitaningrum, K. R. Nastiti, and S. N. Mahmudah, “Medical Record Document Search with TF-IDF and Vector Space Model (VSM),” vol. 14, no. 3, 2024.
23. D. Soyusiawaty, D. Hilmawan, and R. Wolley, “Hybrid Spelling Correction and Query Expansion for Relevance Document Searching,” 2021. [Online]. Available: www.ijacsa.thesai.org
24. S. Dai *et al.*, “AI-based NLP section discusses the application and effect of words models and TF based NLP section discusses the application and effect of words models and TF-IDF in NLP tasks based NLP section discusses the application and effect of bag-of,” 2024. [Online]. Available: <http://creativecommons.org/liJournalofArtificialIntelligenceGeneralScienceJAIGShttps://ojs.boulibrary.com/index.php/JAIGS/ojs.boulibrary.com/index.php/JAIGShttp://creativecommons.org/licenses/by/4.0>

25. F. V. Krasnov, “Embedding-based retrieval: measures of threshold recall and precision to evaluate product search,” *Business Informatics*, vol. 18, no. 2, pp. 22–34, 2024, doi: 10.17323/2587-814X.2024.2.22.34.
26. F. Diaz and A. Ferraro, “Offline Retrieval Evaluation Without Evaluation Metrics,” in *SIGIR 2022 - Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Association for Computing Machinery, Inc, Jul. 2022, pp. 599–609. doi: 10.1145/3477495.3532033.
27. P. Palai, K. Agrawal, D. P. Mishra, and S. R. Salkuti, “Text grouping: a comprehensive guide,” *IAES International Journal of Artificial Intelligence*, vol. 12, no. 3, pp. 1476–1483, Sep. 2023, doi: 10.11591/ijai.v12.i3.pp1476-1483.
28. E. Ben George, G. Jeba Rosline, N. Balasupramanian, and N. R. Wilfred Blessing, “Efficient Document Retrieval System using Locality Sensitive Hashing Nearest Neighbor Algorithm and Weighted Jaccard Distance for Retrieving Closest Personalities,” *Jurnal Kejuruteraan*, vol. 36, no. 4, pp. 1535–1543, Jul. 2024, doi: 10.17576/jkukm-2024-36(4)-19.
29. J. Tussupov *et al.*, “Analysis of Short Texts Using Intelligent Clustering Methods,” *Algorithms*, vol. 18, no. 5, May 2025, doi: 10.3390/a18050289.
30. F. Lechtenberg, J. Farreres, A. L. Galvan-Cara, A. Somoza-Tornos, A. Espuña, and M. Graells, “Information retrieval from scientific abstract and citation databases: A query-by-documents approach based on Monte-Carlo sampling,” *Expert Syst. Appl.*, vol. 199, Aug. 2022, doi: 10.1016/j.eswa.2022.116967.