Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Designing a Performance Metrics Evaluation Framework for NLP-Driven Chatbots in Local Government Unit

Aristotel Aaron C. Agpaoa [1*], Thelma D. Palaoag [2]
[1, 2] *Department of Information Technology, College of Information Technology and Computer Science, University of the Cordilleras, Baguio City, 2600, Philippines.*
Email: [1]arisaaronagpaoa@gmail.com, [2] tpalaoag@gmail.com
Orchid Id: [1] 0009-0000-5073-1512, [2]0000-0002-5474-7260
**Corresponding Author: Aristotel Aaron C. Agpaoa*

ARTICLE INFO

Received: 18 Dec 2024

Revised: 10 Feb 2025

Accepted: 20 Feb 2025

ABSTRACT

The integration of Natural Language Processing (NLP) in AI-driven chatbots offers a transformative solution to enhance service delivery and citizen engagement within the Local Government Units (LGUs). The primary goal of this study is to identify and evaluate relevant performance metrics for NLP-driven chatbot and proposes a comprehensive performance metrics evaluation framework tailored to LGUs. A systematic review of relevant academic literature was conducted to identify key performance aspects which were then organized into a multi-perspective framework. The proposed framework includes five perspectives: User Experience, Information Retrieval, Linguistic Quality, Technology Efficiency, and Public Service. These perspectives address critical aspects of chatbot performance, including task completion rates and response accuracy to linguistic coherence and public trust. The multi-perspective approach of the framework specifically addresses LGU challenges by incorporating bilingual support and inclusivity that ensures alignment with the unique needs of diverse citizens. Future work includes pilot testing with LGU-specific datasets to empirically validate the framework, refine its metrics and enhance its practical applicability.

Keywords: AI-powered Chatbots, NLP-driven Chatbots, Performance Metrics, Local Government Units, Framework Evaluation.

INTRODUCTION

Natural Language Processing (NLP) has revolutionized the development of intelligent systems, enabling chatbots to perform tasks like customer service, information retrieval and citizen assistance. Local Government Units (LGUs) play a pivotal role in delivering a wide array of public services. However, LGUs often face significant challenges due to limited resources which leads to inefficiencies in service delivery.

The integration of technology, specifically AI-powered chatbots, presents a promising solution to enhance service delivery and citizen engagement. Chatbots can streamline communication between LGUs and their constituents which provides timely information and assistance, increasing access and collaboration while reducing human staff workload [1] Chatbots, particularly those powered by Natural Language Processing (NLP), offer the potential to automate routine inquiries and provide citizens 24/7 access to information. This capability can significantly lessen the workload on LGU personnel while improving the overall service accessibility [2]. However, the success of the implementation of the NLP-driven chatbots requires an important understanding of both the specific needs of local governments and the expectation of their citizens. Despite its potential, the successful deployment of NLP-driven chatbots requires a clear understanding of relevant performance metrics and appropriate design considerations tailored to the specific needs of LGUs.

This study primarily aims to evaluate performance metrics and to design a chatbot performance metrics evaluation framework.

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

While chatbots have been increasingly adopted by local governments to improve service delivery, there is a notable absence of standardized evaluation frameworks that cater to the unique requirements of LGUs. Several studies emphasize the need for comprehensive assessment methodologies for conversational AI systems [3] [4]. Current frameworks focus primarily on quantitative metrics like BLEU, accuracy, and F1-score, but they often fail to address qualitative aspects such as public trust, multilingual inclusivity, and citizen participation [5]. This study bridges these gaps by proposing a tailored framework that evaluates NLP-driven chatbots from five key perspectives: User Experience, Information Retrieval, Linguistic Quality, Technology Efficiency, and Public Service. By addressing these features, the framework ensures practical relevance to LGUs and supports the goal of enhancing citizen engagement through more inclusive and effective AI solutions

METHODS AND METHODOLOGY

This study used the **PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework** to maintain transparency in the selection and evaluation of relevant academic literature. The systematic review aimed to identify performance metrics used in NLP-driven chatbots, specifically in the context of Local Government Units (LGUs). To achieve this, researchers conducted a comprehensive search across three academic databases: Scopus, Web of Science, and the Directory of Open Access Journals (DOAJ). The search utilized various combinations of keywords including "NLP architecture model," "chatbot," "Natural Language Processing," "Government," "Local Government Unit," and "performance metrics," refined using Boolean operators AND and OR.

The study established inclusion and exclusion criteria to guarantee the relevance and quality of the selected literature. Studies discussing chatbot performance metrics in public service contexts, particularly those addressing NLP technologies in chatbot design and evaluation, were included. Peer-reviewed journal articles published in English were prioritized. Articles focusing solely on private-sector or non-NLP chatbots, those lacking empirical data or measurable performance metrics, and non-peer-reviewed sources were excluded.

The screening process involved four stages:

- **Identification:** Initially, 120 records were identified through database searches, with an additional 20 sourced from citations within relevant papers.
- **Screening:** Duplicates were removed, and titles and abstracts were screened for relevance, excluding numerous articles based on the abstract review.
- **Eligibility:** Full texts of the remaining studies were meticulously reviewed against the pre-defined inclusion and exclusion criteria.
- **Inclusion:** Ultimately, 10 studies that met all the criteria were included in the systematic review.
- This rigorous selection process, visualized in Figure 1 (PRISMA Flow Diagram), resulted in a final set of 10 highly relevant papers for analysis and discussion.

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

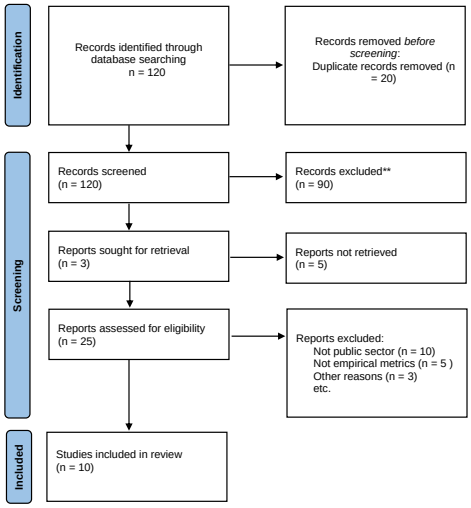

Figure 1. PRISMA Flow Diagram of the Study Selection Process

RESULTS AND DISCUSSION

The refinement resulted with 10 papers marked as relevant and their result is briefly discussed below. According to (Peras, 2018) it is mentioned in his study that previous research on evaluation of chatbots is a mix of a mostly qualitative and metrics and a few quantitative metrics [11].

The performance evaluation metrics for NLP chatbots in public service, particularly within local government units, have been reviewed across several studies. Commonly utilized metrics include accuracy, F1-Score, BLEU, recall, precision, and human evaluation, reflecting a focus on both quantitative and qualitative assessments of chatbot effectiveness [3]. The adoption of chatbots in local governments primarily aims to enhance information provisioning, citizen engagement, and service transactions, with ethical concerns surrounding accuracy and accountability being prominent [12]. Furthermore, the integration of NLP techniques such as sentiment analysis and machine learning are crucial for improving chatbot functionality and user experience [7]. Challenges identified include the need for comprehensive evaluations beyond usability, emphasizing the importance of public value generation and citizen participation in the design and implementation of these systems [5]. Most metrics employed to evaluate chatbot performance are human-based followed by bilingual evaluation metrics and accuracy metrics [8]. The study on the BERT-based virtual assistant for local government units utilized a variety of performance metrics to evaluate its effectiveness and overall quality [6]. Key metrics included functionality, which assesses the accuracy and completeness of the assistant's responses, ensuring that it provides relevant and correct information to user queries. Reliability was another critical metric, measuring the system's ability to deliver consistent and correct responses over time, which is essential for maintaining user trust and satisfaction. Usability was evaluated through user feedback and satisfaction surveys, highlighting how user-friendly the virtual assistant is in practice.

Efficiency was also a significant focus, encompassing metrics related to response time and resource usage, indicating how quickly and effectively the system can handle queries. Accuracy was specifically measured using a confusion matrix, which helped identify the types of errors made by the system, providing insights into how well it interprets user inquiries. Additionally, precision, recall, and F1-score were employed to assess the effectiveness of the NLP model in retrieving information from the knowledge base, offering a comprehensive view of the model's performance in terms of relevant results.

The satisfaction index was developed to gauge user satisfaction and the ease with which users could perform queries, complementing the response time metric to provide a holistic view of user experience. Finally, the overall evaluation framework for the virtual assistant was based on the ISO 9126 standard, which encompasses various software quality attributes, including functionality, reliability, usability, and performance. In the study Foundation Metrics: Quantifying Effectiveness of Healthcare Conversations powered by Generative AI introduced a comprehensive set of evaluation metrics specifically designed for healthcare chatbots. These metrics are essential for assessing the

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

performance of conversational models in a healthcare context, where accuracy and user experience are paramount. The first category of metrics focuses on accuracy, which evaluates how correctly the chatbot can provide information or perform healthcare-related tasks. This is crucial for ensuring that users receive reliable and valid responses. The second category is trustworthiness, which assesses the reliability of the chatbot in delivering information and support, as users need to feel confident in the advice they receive.

Another significant metric is empathy, which measures the chatbot's ability to understand and respond to users' emotional states. Lastly, the study includes computing performance metrics, which focus on the technical aspects of the chatbot's operation, such as response time and resource utilization. This is important for ensuring that the chatbot operates efficiently and provides timely responses to users.

The authors emphasize that these tailored evaluation metrics are necessary because existing metrics for generic large language models often overlook critical aspects relevant to healthcare, such as empathy and user-centered considerations. By establishing these metrics, the study aims to enhance the reliability and quality of healthcare chatbot systems, ultimately improving the patient experience [4]. Overall, the literature underscores the necessity for robust evaluation frameworks to ensure the effectiveness and accountability of chatbots in public service contexts [9].

Table 1. Summary of the Systematic Review

Focus Area	Key Findings	Cited Studies
Evaluation Metrics for Chatbots in Public Service	Common metrics: accuracy, F1-score, BLEU, recall, precision, human evaluation. Emphasis on quantitative and qualitative metrics for effectiveness.	Suhaili et al. (2021), Abeer (2022), Casas et al. (2024)
Challenges in Chatbot Evaluation	Need for bilingual support, public trust, and citizen participation. Focus on public value generation and accountability.	Cortes-Cediel et al. (2023)
NLP Techniques and Advances	Sentiment analysis and machine learning improve functionality and user experience. Importance of diverse metric evaluation.	Jiang et al. (2023)
BERT-based Virtual Assistant for LGUs	Metrics used: functionality, reliability, usability, response time. Based on ISO 9126 standard; identified errors using confusion matrix.	Casas et al. (2024)
Healthcare Chatbots and Generative AI	Tailored healthcare chatbot metrics: accuracy, trustworthiness, empathy, and computing performance. Improved patient experience and reliability.	Abbasian et al. (2023)

Proposed Chatbot Performance Metrics Evaluation Framework

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

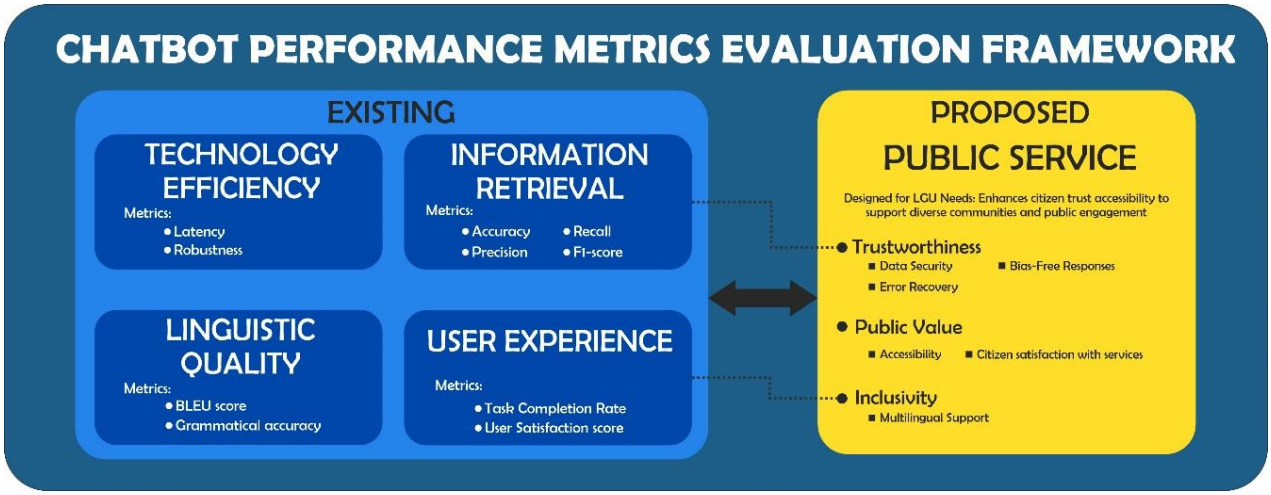

Figure 2. Chatbot Performance Metrics Evaluation Framework

The proposed framework systematically organizes identified metrics into evaluation perspectives, ensuring comprehensive assessment and practical relevance for government services. The framework provides five perspectives, like the approach proposed of T. Russell-Rose [10]. The framework includes the following five perspectives:

A. User Experience Perspective

The User Experience Perspective focuses on how well the chatbot meets citizen needs and ensures an intuitive, satisfactory interaction. This is critical for Local Government Unit (LGU) chatbots as citizens rely on these systems for essential services like inquiries about tax deadlines, permit applications, or disaster-related updates. Metrics such as task completion rate and user satisfaction score are central to this perspective. The task completion rate measures the success of users in completing tasks, such as retrieving specific information or resolving an issue. A high completion rate reflects a well-designed system capable of addressing user needs effectively. On the other hand, the user satisfaction score, typically gathered through post-interaction surveys, provides insights into how citizens perceive the system's ease of use, reliability, and helpfulness. This perspective ensures that the chatbot is user-friendly, functional, and widely accepted, driving trust and repeated usage among citizens.

B. Information Retrieval Perspective

The Information Retrieval Perspective evaluates the chatbot's ability to retrieve accurate, relevant, and complete responses. This is particularly significant for LGUs, where citizens depend on precise information for critical services like permits, fees, and local regulations. Metrics like accuracy, precision, recall, and F1-score ensure that the chatbot meets these expectations. Accuracy measures how often the chatbot provides the correct response, while precision focuses on the relevance of the information provided. Recall assesses the completeness of the responses, ensuring the chatbot does not omit crucial details. F1-score balances precision and recall, offering a comprehensive metric for evaluating response quality. By prioritizing accurate information retrieval, this perspective ensures that LGU chatbots maintain public trust and effectively support citizen inquiries.

C. Linguistic Quality Perspective

The Linguistic Quality Perspective addresses the clarity, coherence, and grammatical accuracy of the chatbot's responses, which are crucial for effective communication. For LGU chatbots operating in multilingual contexts, such as those supporting both Ilocano and English, linguistic quality is a critical component of user experience. Metrics like BLEU score and grammatical accuracy help evaluate this perspective. The BLEU score measures how closely the chatbot's responses resemble human-authored reference answers, ensuring that the language is natural and coherent. Grammatical accuracy evaluates the correctness of sentence structure, spelling, and syntax, enhancing the chatbot's professionalism and clarity. This perspective ensures that citizens can easily understand and trust the chatbot's responses, fostering effective communication in LGU applications.

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

D. Technology Efficiency Perspective

The Technology Efficiency Perspective assesses the technical performance of the chatbot, focusing on its ability to provide fast and reliable responses under various conditions. In LGU contexts, efficiency is vital during high-demand situations, such as emergencies or service deadlines. Metrics like latency and robustness are key indicators of this perspective. Latency measures the response time of the chatbot, with faster response times contributing to user satisfaction, especially for time-sensitive queries. Robustness evaluates the system's ability to handle unexpected inputs, such as typos or ambiguous queries, ensuring it remains functional and adaptable. This perspective highlights the importance of creating a chatbot that is not only responsive but also resilient, maintaining service quality even under challenging conditions.

E. Public Service Perspective

The Public Service Perspective is unique to LGU chatbots, focusing on inclusivity, trustworthiness, and public value generation. This perspective ensures that the chatbot aligns with the goals of public service by addressing the diverse needs of citizens and maintaining ethical standards. Metrics like accessibility and trustworthiness are central to this perspective. Accessibility ensures that the chatbot is inclusive, offering multilingual support and being easy to use for individuals with varying levels of digital literacy. Trustworthiness measures citizens' confidence in the chatbot's responses, ensuring the information provided is reliable, unbiased, and secure. This perspective underscores the broader role of LGU chatbots in fostering transparency, public engagement, and accountability, ultimately enhancing their value as tools for citizen support.

The proposed framework offers a more comprehensive approach compared to existing models in the field by integrating five key perspectives: User Experience, Information Retrieval, Linguistic Quality, Technology Efficiency, and Public Service. Unlike traditional evaluation methods that focus primarily on technical metrics such as accuracy and precision, this framework emphasizes inclusivity, multilingual support, and public trust—features that are critical for Local Government Units (LGUs). Additionally, it builds on standards like ISO 9126 but expands them to address the unique challenges faced by LGUs, such as handling bilingual inquiries and ensuring equitable access for citizens with varying digital literacy levels.

CONCLUSION

This study proposes a comprehensive performance metrics evaluation framework specifically designed for NLP-driven chatbots implemented within Local Government Units (LGUs). The framework is built upon five key perspectives that ensure relevance and practical applicability in LGU contexts: User Experience, Information Retrieval, Linguistic Quality, Technology Efficiency and Public Service.

Unlike traditional evaluation methods that primarily focus on technical metrics, this framework incorporates the unique challenges and requirements of LGUs, including multilingual support, inclusivity, and public trust. The Public Service perspective, a novel addition to the framework, emphasizes the chatbot's role in fostering transparency, public engagement, and accountability, reinforcing its value as an essential tool for citizen support.

Future work will focus on pilot testing this framework with LGU-specific datasets to empirically validate its effectiveness and refine the proposed metrics. This validation process will further enhance the framework's practical utility and contribute to the development of more inclusive and effective AI solutions for LGUs.

Funding Statement:

The authors did not receive financing for the development of this research.

Data Availability:

No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflict of interest:

The authors declare that there is **no conflict of interest.**

Journal of Information Systems Engineering and Management
2025, 10(43s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

REFERENCES

[1] A. Kumar, R. Singh, and S. Gupta, "Chatbots in public administration: A systematic review," Journal of Public Affairs, vol. 20, no. 2, p. e1984, 2020.
[2] A. Gonzalez, C. Rojas, and R. Torres, "The impact of chatbot technology on public service delivery: A case study from Chile," International Journal of Public Administration, vol. 42, no. 8, pp. 705–713, 2019. Available: https://doi.org/10.1080/01900692.2018.1523807.
[3] M. Sinarwati, N. Suhaili, M. Salim, and N. Jambli, "Service chatbots: A systematic review," Expert Systems With Applications, vol. 184, p. 115461, 2021. Available: https://doi.org/10.1016/J.ESWA.2021.115461.
[4] M. Abbasian, E. Khatibi, I. Azimi, et al., "Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI," npj Digital Medicine, vol. 7, p. 82, 2023. Available: https://doi.org/10.1038/s41746-024-01074-z.
[5] M. E. Cortés-Cediel, A. Segura-Tinoco, I. Cantador, and M. P. Rodríguez-Bolívar, "Trends and challenges of e-government chatbots: Advances in exploring open government data and citizen participation content," Government Information Quarterly, vol. 40, no. 1, p. 101877, 2023. Available: https://doi.org/10.1016/j.giq.2023.101877.
[6] J. Casas, R. Villafuerte, and J. Paragas, "Virtual assistant for local government units using BERT-based natural language processing," Cognizance Journal of Multidisciplinary Studies, vol. 4, no. 8, pp. 299–305, 2023. Available: https://doi.org/10.47760/cognizance.2023.v04i08.019.
[7] Y. Jiang, P. Pang, D. Wong, and H. Y. Kan, "Natural language processing adoption in governments and future research directions: A systematic review," Applied Sciences, vol. 13, no. 22, p. 12346, 2023. Available: https://doi.org/10.3390/app132212346.
[8] A. Abeer and L. Alhenaki, "English and Arabic chatbots: A systematic literature review," International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, 2022. Available: https://doi.org/10.14569/ijacsa.2022.0130876.
[9] A. Li, Z. Wang, E. Mendes, D. M. Le, W. Xu, and A. Ritter, "ChatHF: Collecting rich human feedback from real-time conversations," in Proc. 2024 Conf. Empirical Methods in Natural Language Processing: System Demonstrations, 2024, pp. 270–279. Available: https://aclanthology.org/2024.emnlp-demo.28.
[10] T. Russell-Rose, "A framework for chatbot evaluation," Information Interaction, Jan. 24, 2017. Available: https://isquared.wordpress.com/2017/01/24/a-framework-for-chatbot-evaluation/.
[11] D. Peras, "Chatbot evaluation metrics: Review paper," in Economic and Social Development (Book of Proceedings), R. Veselica, G. Dukić, and K. Hammes, Eds. Varazdin Development and Entrepreneurship Agency, 2018, pp. 89–97. Available: https://www.proquest.com/docview/2176212638.
[12] S. Senadheera, Y. Tan, K. C. Desouza, et al., "Understanding chatbot adoption in local governments: A review and framework," Journal of Urban Technology, vol. 30, no. 1, 2023. Available: https://doi.org/10.1080/10630732.2023.2297665.