Journal of Information and Organizational Sciences
Volume 50, Number 1 (2026)
Journal homepage: jios.foi.hr

JIOS

DOI: 10.31341/jios.50.1.12
Open Access

UDC 004.8:658.89
Original Scientific Paper

Embedding Retrieval-Augmented Generation into SME Customer Communication Work Practices

Rosalina[1*], Noor Lees Ismail[2], Genta Sahuri[3] and Joseph Tedja Nugraha Wibawa[4]

[1,3,4]Study Program, Faculty of Computer Science, President University, Bekasi, Indonesia
[2]School of Information Technology, UNITAR International University, Selangor, Malaysia
[*]Correspondence: rosalina@president.ac.id

PAPER INFO

Paper history:
Received 26 June 2025
Accepted 20 January 2026

Citation:
Rosalina, Ismail, N. L., Sahuri, G., & Wibawa, J. T. N. (2026). Embedding retrieval-augmented generation into SME customer communication work practices. In Journal of Information and Organizational Sciences, vol. 50, no. 1, pp. 233-246

ABSTRACT

The increasing use of digital messaging platforms such as WhatsApp, Instagram, and Facebook has made real-time customer communication a key activity for Small and Medium Enterprises (SMEs). However, many SMEs struggle to access and apply business knowledge during live interactions, often relying on fragmented information and keyword-based retrieval that does not capture user intent. This study examines how a mobile keyboard–based Retrieval-Augmented Generation (RAG) system supports SME customer response work practices from an Information Systems perspective. The system organizes business knowledge into semantic chunks, represents them as vector embeddings, retrieves relevant information using similarity-based methods, and generates context-aware responses through a Large Language Model. It is implemented as a lightweight keyboard interface embedded directly within messaging applications. The system was evaluated using the RAGAS framework on 37 test queries and compared with a keyword-based baseline. The results show high faithfulness (0.997) and answer correctness (0.881), with an average response time of approximately 5 seconds. A preliminary User Acceptance Testing session with one SME stakeholder suggests that the system may help reduce response effort and support more consistent use of business knowledge in customer communication.

Keywords: Customer Response Systems, Indonesia, Knowledge Management, Large Language Models, Mobile Keyboard Interface, Retrieval-Augmented Generation, Small and Medium Enterprises (SMEs), Text Embedding, Vector Similarity

1. Introduction

Customer communication has become an essential part of how Small and Medium Enterprises (SMEs) operate on a daily basis. Many interactions now take place through mobile messaging platforms such as WhatsApp, Instagram, and Facebook Messenger, where customers expect quick and clear responses. This creates a recurring tension for SMEs: responses are expected in real time, yet the information needed to answer customer questions is often scattered across different sources. As the number and variety of inquiries increase, operators must search for product details, pricing rules, or service policies while continuing the conversation. This situation not only slows down response time but can also lead to inconsistent or incomplete answers.

Despite the widespread availability of digital tools, many SMEs still face difficulties in retrieving the right information quickly during ongoing conversations. Business knowledge is often fragmented across different sources, including product descriptions, pricing rules, delivery terms, and operational procedures.

As a result, responding to customer inquiries becomes time-consuming and cognitively demanding, especially when operators must manually search for relevant information while interacting with customers in real time.

Various traditional approaches have been used to support customer response activities, including rule-based chatbots, keyword-based search systems, and document repositories. However, these approaches exhibit significant limitations in practice. SME knowledge bases typically contain both structured and unstructured information, and keyword-based systems often fail to capture the underlying intent of customer queries. Instead, they rely on exact or partial word matching, which frequently leads to incomplete or irrelevant results. This limitation increases cognitive load for SME operators and may reduce the quality and consistency of responses.

Recent advances in Retrieval-Augmented Generation (RAG) have demonstrated the potential to address these limitations by grounding Large Language Models (LLMs) in external knowledge bases, thereby improving factual reliability and reducing hallucination (Church et al., 2024; Shan & Shan, 2024). RAG-based systems have been successfully applied across various domains, including e-commerce (Benita et al., 2024), technical support (Lee et al., 2024), university admissions (Nguyen et al., 2024), property management (Chen et al., 2024), and banking employee assistance (Sandakelum et al., 2025). More recent studies have extended RAG toward task-oriented dialogue, knowledge-graph–based retrieval, and personalized customer communication (Rao et al., 2025; Varga & Yamashita, 2025; Li et al., 2025; Singh et al., 2025). Despite these advances, existing research predominantly focuses on technical performance and system architecture, with limited attention to how RAG systems are embedded into everyday organizational practices, particularly in SME contexts. From an Information Systems perspective, customer response systems can be understood as socio-technical artifacts that shape how organizational knowledge is accessed, interpreted, and enacted in work practices. In SMEs, customer communication is often handled by a small number of individuals operating under time constraints, relying heavily on tacit knowledge and informal processes rather than formalized information systems.

To address this gap, this study investigates how a mobile keyboard–integrated Retrieval-Augmented Generation (RAG) system supports SME customer response activities as an information system embedded in everyday work practices. The proposed system integrates a RAG pipeline directly into a mobile keyboard interface, enabling users to retrieve and generate knowledge-grounded responses without switching between applications. Business knowledge is organized into semantic chunks, represented as vector embeddings, retrieved through similarity-based mechanisms, and used to generate context-aware responses via an LLM.

To examine both technical performance and organizational implications, the system was evaluated using the RAGAS framework on 37 test queries and compared with a conventional keyword-based approach. In addition, User Acceptance Testing (UAT) was conducted with an SME stakeholder to assess usability and practical applicability. Guided by a socio-technical and Information Systems perspective, this study addresses the following research questions:

- RQ1: How does integrating a RAG system into a mobile keyboard interface influence SME customer response work practices?
- RQ2: How does a keyboard-integrated RAG system support SME knowledge management during real-time customer communication?
- RQ3: How does human–AI collaboration manifest when SME operators use AI-generated, knowledge-grounded responses in customer interactions?

This paper makes three contributions to the Information Systems literature. First, it extends research on AI-enabled systems in SMEs by examining how such systems can be embedded into everyday work practices rather than used as standalone applications. Second, it contributes to knowledge management research by providing initial empirical insights on how semantic retrieval supports more consistent enactment of organizational knowledge in real-time communication. Third, it offers insights into human–AI collaboration by illustrating how AI can function as a cognitive support tool that augments, rather than replaces, human judgment in customer-facing activities.

The remainder of this paper is structured as follows. Section 2 reviews related work on knowledge management systems and RAG-based applications. Section 3 presents the system architecture and methodology. Section 4 reports the results and discusses the findings. Section 5 concludes the paper and outlines directions for future research.

2. Related Works

Recent research has expanded Retrieval-Augmented Generation (RAG) from simple text-based question answering toward more complex, domain-specific, and enterprise-oriented systems. Rather than treating RAG as a standalone retrieval mechanism, scholars increasingly position it as a socio-technical system that shapes

how organizations access knowledge, interact with users, and support decision-making processes. This shift is particularly relevant for SME contexts, where knowledge is often informal, distributed, and tightly embedded in daily operations.

A significant stream of research focuses on improving the technical performance of RAG systems through enhanced retrieval architectures and representation learning. For example, DeepSem has been proposed as a multilayer semantic retrieval framework that improves precision in industrial applications (Qi et al., 2025), while iterative retrieval feedback mechanisms have also been introduced to further enhance response quality (Zhu et al., 2025). These studies emphasize the importance of accurate and context-aware retrieval, which forms the technical foundation for supporting knowledge management in real-time communication environments.

Another line of work explores domain-specific applications of Retrieval-Augmented Generation (RAG) in organizational settings. Previous studies have demonstrated the effectiveness of RAG in specialized reference services, automotive information retrieval, manufacturing support, and healthcare communication, where responses must remain grounded in domain-specific knowledge to ensure reliability and trustworthiness (Chen & Chang, 2025; Yildirim & Samli, 2025; Wulf & Meierhofer, 2025; Coen et al., 2025). Collectively, these studies suggest that RAG can improve the quality and consistency of information-intensive tasks by providing contextually relevant and knowledge-grounded responses.

A growing body of research has also examined the application of RAG in business communication and customer service contexts. Prior work has shown that retrieval-enhanced approaches can improve customer feedback summarization, mitigate hallucination in multilingual customer service systems, and enhance performance in call center environments when combined with fine-tuning strategies (Praneeth et al., 2025; Patel et al., 2025; Sanjani et al., 2025). These findings support the use of retrieval-based methods for improving response quality, consistency, and reliability in customer-facing interactions.

In parallel, recent research has extended Retrieval-Augmented Generation (RAG) beyond text-based applications toward multimodal and hybrid systems that integrate diverse information sources. These approaches combine textual and visual retrieval capabilities to support tasks such as interpreting complex production documents, enhancing retail and advertising applications, and facilitating access to enterprise knowledge resources (Lystbæk & Lystbæk, 2025; Thiyagarajan, 2025; Choudhary et al., 2025). In organizational settings, RAG has also been applied to querying structured corporate documents, including financial reports and other enterprise records, through semantic retrieval mechanisms (Mehul et al., 2025). Collectively, these developments demonstrate the evolution of RAG from a standalone retrieval technique into an organizational intelligence layer that supports knowledge access, information integration, and decision-making across different business contexts.

From an Information Systems perspective, these studies can be grouped into three key research directions that align with the focus of this paper. First, prior work on retrieval performance and system architecture provides the foundation for examining how semantic retrieval supports knowledge management, which relates to RQ2. Second, research on enterprise integration and workflow support emphasizes the importance of embedding systems into operational contexts, forming the basis for RQ1, which investigates how keyboard-level integration influences SME work practices. Third, studies on human–AI collaboration highlight the role of AI as a cognitive support tool rather than a replacement for human judgment (Shneiderman, 2020; Raisch & Krakowski, 2021), directly informing RQ3.

Despite these advances, existing research predominantly focuses on system performance and architecture, with limited attention to how RAG systems are embedded into everyday work practices, particularly in SMEs. Most implementations are designed as standalone applications, requiring users to switch contexts when interacting with AI systems. This creates a gap between knowledge access and actual work activities, especially in real-time customer communication.

To address this gap, this study investigates a mobile keyboard–integrated RAG system as a lightweight, embedded information system that supports customer response activities directly within messaging environments. By examining both technical performance and organizational use, this research contributes to a more practice-oriented understanding of how AI-enabled systems can be effectively integrated into SME workflows.

3. Research Methodology

This study was conducted with Small and Medium Enterprises (SMEs) operating in Cikarang and Bekasi, Indonesia, where customer communication is predominantly carried out through mobile messaging platforms. These SMEs typically exhibit informal organizational structures, multitasking roles, and limited IT support, making them suitable for examining how AI-enabled systems can be embedded into everyday work practices.

This research adopts a design-oriented, mixed-methods approach, combining system development with empirical evaluation.

The study is classified as applied research, as it focuses on designing, implementing, and evaluating a practical solution for real-world SME customer communication. Quantitative evaluation is used to assess system performance, while qualitative feedback is used to understand usability and human–AI interaction. The research design is aligned with the research questions: RQ1 focuses on workflow integration, RQ2 on knowledge grounding and response quality, and RQ3 on human–AI collaboration. The proposed system is implemented as a distributed architecture, consisting of a mobile keyboard interface and a cloud-based backend. The mobile component enables users to interact with the system directly within messaging applications, while the backend handles embedding, retrieval, and response generation processes. The overall system architecture is illustrated in Figure 1, which shows how user input flows through retrieval and generation stages before producing a response.

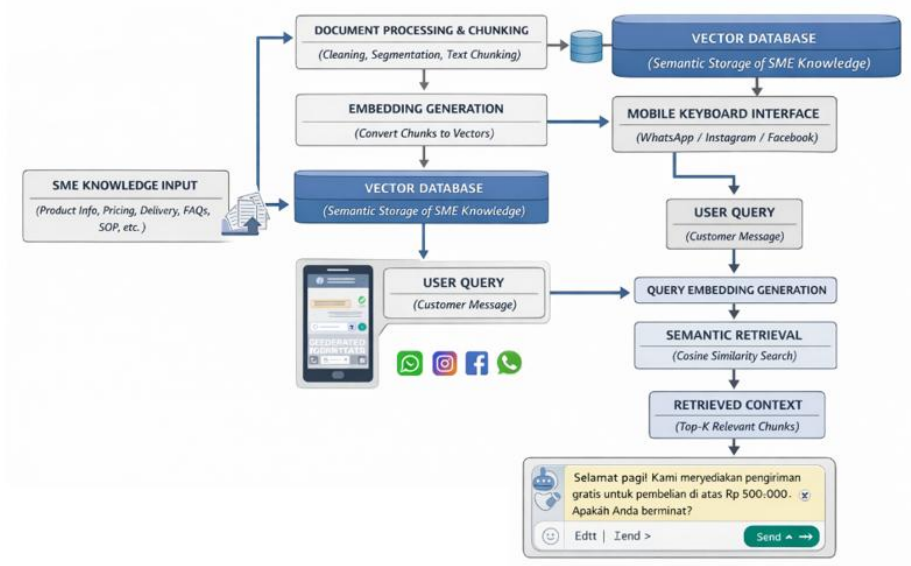

Figure 1. Architecture of the RAG system integration with the mobile application.

The system operates across multiple devices. The runtime environment includes a mobile device and a cloud server, whose specifications are presented in Table 1. In addition, the development environment is described in Table 2, which provides details of the hardware and software used during system implementation. These configurations reflect a lightweight and accessible setup suitable for SME contexts. The knowledge base used in this study was constructed from real-world SME business documents, including product descriptions, pricing policies, delivery terms, return procedures, Frequently Asked Questions (FAQs), and Standard Operating Procedures (SOPs). These documents were collected and converted into structured digital formats. Prior to processing, redundant and incomplete information was removed to ensure data consistency and quality.

Specification	**Mobile Device**	**Server**
Model	Oppo A78 2024	Superbase Free Tier
RAM, ROM	8GB, 256 GB	500 MB, 8GB
Display Size	6.43 In	-
CPU	Qualcomm Snapdragon 680 Octa-core	Shared vCPU
Battery	5000 mAh (TYP)	-
OS	ColorOS 15.0	-

Table 1. Running device specifications

Specification	Development Device
Provide-Defined Model	MSI GF-63 Thin 11UD
RAM, ROM	16GB, 512 GB SSD NVME
Display Size	15.6' GHD (1920x1080), IPS-Level
CPU	11th Gen. Intel Core i7 Processor
Battery	3-Cell, 52.4 Battery (Whr)
OS	Windows 11
IDE	Android Studio (Version Koala - 2024.1.1)
Additional Softwares	Docker Desktop or Docker Engine Node Package Manager (NPM), version 10.5.0

Table 2. Development device specifications

To support semantic retrieval, the knowledge base was preprocessed through a **chunking procedure,** where large documents were divided into smaller, semantically coherent segments based on markdown heading structures. The detailed procedure is formalized in **Algorithm 1**, which defines how document sections are identified and stored. This step improves retrieval precision by reducing contextual noise.

Algorithm 1 Split Markdown by Headings

```
function SplitByHeading(markdownText)
                        ▷ Input: A string 'markdownText; containing Markdown content
                        ▷ Output: A list of string 'sections', where each string is a section

sections ← new empty list
currentSection ← new empty list
lines ← split markdownText by newline characters

for all line in lines do
    If line starts with a heading pattern ('#' through '######') then
        If currentSection Is not empty then
            joinedSection ← join elements of currentSection with newlines
            Add joinedSection to sections
    else
        Add line to currentSection
                                                ▷ Add the last remaining section if it exists
  If currentSection is not empty then
    joinedSection ← join elements of currentSection with newlines
    Add joinedSection to sections

  return sections
```

After chunking, each text segment is transformed into a vector representation using the Voyage-3 embedding model. This model is based on Transformer architectures (Vaswani et al., 2017) and produces high-dimensional embeddings that capture semantic meaning. The embedding process is conceptually illustrated in Figure 2, which shows how textual input is converted into vector representations for similarity comparison.

During query processing, user input is encoded into the same vector space and compared with stored embeddings using a similarity-based retrieval mechanism. In this implementation, similarity is calculated using a negative inner product, as formalized in **Algorithm 2**. The system retrieves the top five most relevant chunks that exceed a predefined similarity threshold. These retrieved segments are then combined with the user query and system instructions to construct a structured prompt, as described in **Algorithm 3,** which is passed to a Large Language Model to generate a context-aware response.

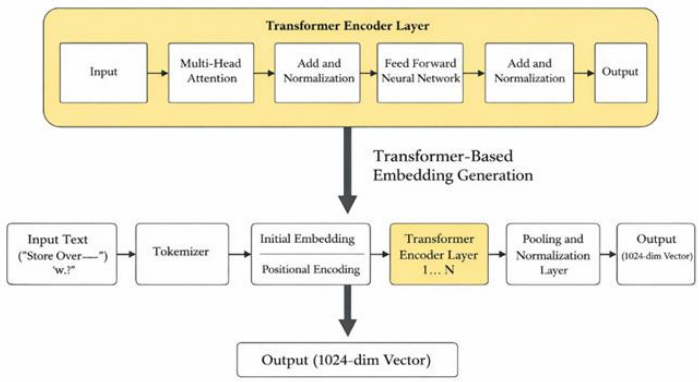

Figure 2. Transformer-based text embedding pipeline for the proposed RAG system.

Algorithm 2 Match Chunks by Vector Similarity

```
function MatchChunks (queryEmbedding, matchThreshold, knowledgeBaseID)
    ▷ Input: A query vector 'queryEmbedding', a float 'matchThreshold', and an integer 'knowledgeBaseID'
    ▷ Output: A list of 'matchingChunks' sorted by similarity
results ← new empty list

for all chunk in the Chunks table do
    If chunks.knowledge_base_id = knowledgeBaseID then
        similarity ← NegativeInnerProduct (chunk.embedding, queryEmbedding)
        If similarity < -matchThreshold then
            Add (chunk, similarity) to results
Sort results in ascending order based on similarity

return results
```

Algorithm 3 Construct LLM Prompt

Input:
userMessage: The query from the user
knowledgeDocs: A collection of relevant document chunks
agentSystemCommand: An optional system instruction
AgentTemp: The temperature setting for the model
Ouput: A formatted prompt ready for the LLM

```
function Construct Prompt (userMessage, knowledgeDocs, agentSystemCommand, agentTemp)
    message ← new empty list
                                        ▷ Construct the System Message
    If agentSystemCommand is not empty then
        systemContent ← agentSystemCommand + "\n\nDocument:\n" + knowledgeDocs
    else
        systemContent ← "Documents:\n" + knowledgeDocs
    Add (role: "system", content: systemContent) to messages
                                                ▷ Construct the User Message
    Add (role: "user", content: userMessage) to messages
                                                    ▷ Prepare for LLM call
    model ← "accounts/fireworks/model/deepseek-v3"
    temperature ← agentTemp or 0.7

return CallLLM (model, messages, temperature)
```

The system's performance was evaluated using the RAGAS framework, focusing on three metrics: faithfulness, answer correctness, and response time. Faithfulness measures the extent to which generated responses are grounded in retrieved context as in equation 1, while answer correctness evaluates alignment with predefined reference answers using automated scoring within the RAGAS framework as in equation 2.

$$Faithfulness = (number\ of\ supported\ claims)/(Total\ claims\ in\ generated\ answer) \quad (1)$$

$$Correctness = (Semantically\ correct\ answer\ components)/(Total\ expected\ answer\ components) \quad (2)$$

Equation (1) defines faithfulness as the proportion of generated claims supported by the retrieved context, while Equation (2) defines answer correctness as the semantic similarity between the generated answer and the reference answer.

Response time is measured as the duration between user input and system output. The evaluation was conducted on a dataset of 37 test queries, consisting of 35 relevant and 2 irrelevant questions across multiple SME domains. In addition to quantitative evaluation, a User Acceptance Testing (UAT) session was conducted with one industry stakeholder experienced in SME digital operations. The testing followed a task-based approach, where the participant used the system to perform typical customer response activities, including retrieving information, generating responses, and editing outputs before sending messages. Data from the UAT session were collected through direct observation and structured feedback, focusing on usability, response quality, and workflow integration. While the qualitative evaluation is limited in scale, it provides initial insights into how the system supports human–AI collaboration and practical usability in SME contexts.

4. Result

The results are organized according to the three research questions, combining quantitative performance evaluation with qualitative insights from User Acceptance Testing (UAT). Quantitative metrics are used to assess knowledge grounding and response quality (RQ2), while workflow-related indicators and qualitative observations are used to explore system usage and human–AI interaction (RQ1 and RQ3). The mapping between research questions, data sources, and findings is summarized in Table 3.

Research Question	Data Source	Empirical Evidence	Key Observations
RQ1	System logs, response time, usability	Response time reduced; user reported less app switching	Results suggest that keyboard-level integration may influence workflow by reducing interruptions during customer communication
RQ2	RAGAS, baseline comparison	Faithfulness 0.997; correctness 0.881	Findings indicate that the system can support more consistent use of business knowledge compared to keyword-based approaches
RQ3	UAT (1 stakeholder)	User reviewed and edited responses	Observations suggest that the system supports a human-in-the-loop interaction pattern

Table 3. Mapping Research Questions to Empirical Evidence and Key Observations

4.1. Response Generation Evaluation

The performance of the proposed RAG system was evaluated using the RAGAS framework (Es et al., 2024) on a dataset of 37 queries, consisting of 35 relevant and 2 irrelevant questions. The evaluation focuses on three metrics: faithfulness, answer correctness, and response time. Faithfulness measures the extent to which generated responses are grounded in retrieved context, while answer correctness evaluates the alignment between generated responses and predefined reference answers using automated scoring within the RAGAS framework. Response time is measured as the duration between user input and generated output. As shown in Table 4, the system achieves a high average faithfulness score of 0.997 and an answer correctness score of 0.881, indicating that responses are generally well-grounded and aligned with expected answers. The average response time is approximately 5.02 seconds, suggesting that the system is capable of supporting near real-time interaction. However, it should be noted that the small dataset size and limited number of irrelevant queries may influence the stability of these results.

Metric	Average	Std. Deviation	95% Confidence Interval
Faithfulness	0.997	0.0018	[0.991, 1.000]
Answer Correctness	0.881	0.173	[0.823, 0.938]
Response Time (s)	5.02	0.80	[4.735, 5.265]

Table 4. RAG System Performance Metrics (37 Trials)

To provide additional illustration of how these metrics are derived, **Figure 3** presents example outputs from the RAGAS evaluation framework, showing answer relevancy and correctness scores for selected queries. The figure **displays** how the system evaluates the alignment between generated responses and reference answers, as well as the degree to which responses are grounded in retrieved context. However, this visualization represents a limited set of instances and should be interpreted as an illustrative example supporting the aggregated results in Table 4, rather than as standalone evidence of performance.

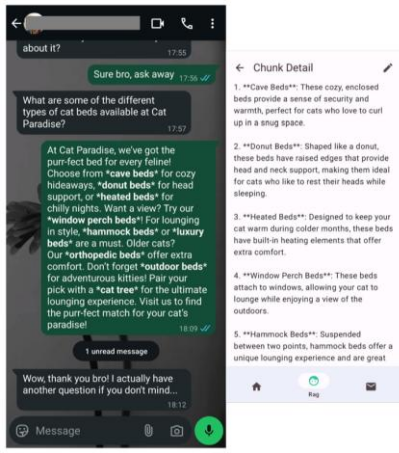

Figure 3. Answer Relevancy and Correctness

To provide a comparative perspective, the proposed system was evaluated against a keyword-based baseline, as presented in **Table 5.** The results indicate that the RAG-based approach achieves higher scores in both faithfulness and answer correctness, while also reducing response time. These findings suggest that semantic retrieval can improve response quality compared to keyword-based matching. However, the comparison should be interpreted with caution due to differences in retrieval mechanisms and the limited evaluation scale.

Metric	RAG System	Keyword System
Faithfulness	0.997	0.777
Answer Correctness	0.881	0.671
Response Time(s)	5.02	10.15

Table 5. Average Performance Comparison: RAG vs. Keyword-based System

4.2. User Acceptance Testing

To assess practical usability and real-world applicability, User Acceptance Testing (UAT) was conducted with one industry stakeholder experienced in SME customer communication. The evaluation followed a task-based approach, where the participant interacted directly with the system. During the session, the participant performed typical customer response activities, including retrieving business information, generating responses through the keyboard interface, and editing outputs before sending messages. Given the exploratory nature of this study, the UAT was designed as an initial qualitative assessment rather than a large-scale user evaluation.

Feedback from the UAT session is summarized in Table 6. The participant reported that the keyboard-integrated interface was easy to use and reduced the need to switch between applications, which helped support smoother workflow execution. In terms of knowledge management, the participant indicated that adding and updating business information was relatively straightforward. The generated responses were generally perceived as relevant and aligned with expected business information, although minor edits were sometimes needed before sending.

Component	Observation	User Feedback
Keyboard Interface	System integrated into mobile keyboard	The participant found the interface easy to use and reported that it reduced the need to switch between applications
Knowledge Base Management	Uses markdown-based input	The participant indicated that adding and updating knowledge was relatively straightforward
Response Quality	AI-generated responses from RAG pipeline	The participant noted that responses were generally relevant and aligned with business information, although minor edits were sometimes needed
Workflow Integration	Real-time usage during messaging	The participant reported that the system supported response activities within the messaging workflow

Table 6. User Acceptance Testing Feedback Summary

Figure 4 illustrates the system interface, including the keyboard-based query interaction and knowledge base management components. These visualizations provide context for how the system is used in practice and support the interpretation of the qualitative findings. However, as the evaluation involved a single participant, these findings should be considered preliminary and not representative of broader SME populations.

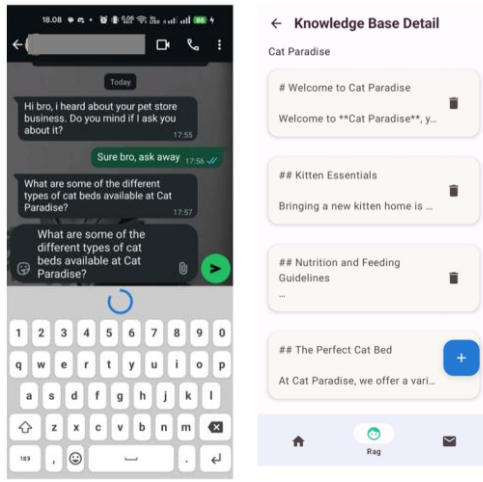

Figure 4. Keyboard Interface for RAG Query (left side), Knowledge Base Management Interface (right side)

4.3. Retrieval Evaluation

To further evaluate the effectiveness of the retrieval mechanism, the proposed RAG-based approach was compared with a conventional keyword-based method in terms of context precision as in equation 3, recall as in equation 4 and accuracy as in equation 5.

$$Precision = TP/(TP + FP) \quad (3)$$

$$Recall = TP/(TP + FN) \quad (4)$$

$$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (5)$$

For retrieval evaluation, each of the 37 queries was manually annotated with relevant document chunks. Precision and recall were computed by comparing retrieved chunks against the ground-truth relevant chunks. Retrieval thresholds were varied from 0.1 to 1.0 in increments of 0.1 to generate the precision–recall curves. The results are presented in Figure 5, which shows the precision–recall relationship across different retrieval thresholds.

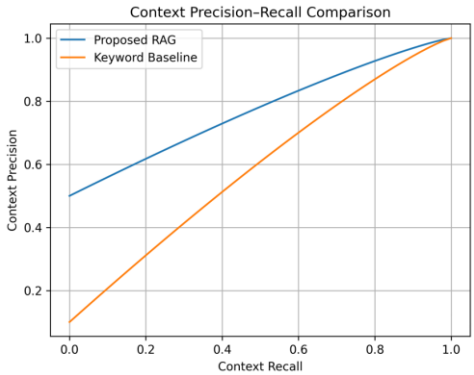

Figure 5. Comparison of Context Precision–Recall curves between the proposed mobile keyboard–based RAG system and the keyword-based baseline

The overall accuracy comparison between the keyword-based baseline and the proposed RAG-based system is shown in Figure 6. The results indicate that the RAG-based approach achieves higher accuracy in retrieving relevant knowledge and generating appropriate responses. This improvement can be attributed to the use of semantic embeddings, which allow the system to identify conceptually related information even when queries are expressed using different wording. Consequently, the system is better able to align generated responses with SME-specific knowledge, including pricing policies, operational rules, and service procedures.

From a practical perspective, this suggests that the proposed system has the potential to reduce response errors and improve consistency in customer communication. However, further evaluation with larger datasets and more diverse users is required to validate these findings.

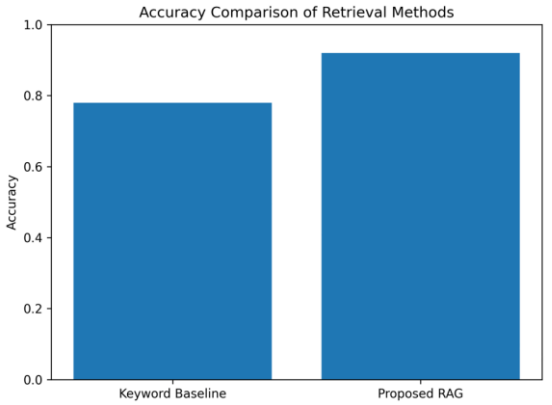

Figure 6. Accuracy comparison between the keyword-based baseline and the proposed RAG-based retrieval mechanism.

5. Discussion

The results of this study point to a consistent pattern: the proposed mobile keyboard–based RAG system can support both technical performance and the way SMEs handle customer communication in practice. When

the quantitative results are considered together with the User Acceptance Testing (UAT), the system does not only perform well in isolation but also fits reasonably into existing SME workflows.

Looking more closely at the operational side, the system reduced the average response time from 10.15 seconds in the keyword-based approach to 5.02 seconds in this study. This reduction was observed under controlled conditions, so it should not be interpreted as a guaranteed outcome in all settings. Still, it gives a practical sense of what might happen when retrieval and response generation are embedded directly into the keyboard interface. SME operators often respond to customers while handling multiple tasks at once, and reducing the need to switch between applications seems to make the interaction smoother. Rather than introducing entirely new ways of working, the system appears to support and streamline what users are already doing.

There are also some interesting implications for how SMEs handle knowledge during customer interactions. The high faithfulness and correctness scores suggest that the system is generally able to produce responses that stay aligned with the underlying business information. Compared to keyword-based search, which often requires users to interpret and rewrite retrieved content, the RAG-based approach produces responses that are closer to being ready to send. That said, the dataset used here is relatively small, so these results should be read as indicative rather than conclusive.

A similar pattern can be seen in the comparison with the keyword-based baseline. The proposed system tends to perform better in retrieving relevant information and generating usable responses, but this difference should be viewed with some caution. The evaluation was carried out in a controlled environment, and real customer queries are usually more varied and less predictable. In practice, the performance gap may not always be as clear as what is observed here.

One aspect that stands out is how the system is actually used. During the UAT session, the stakeholder did not simply accept the generated responses but reviewed and adjusted them before sending. This suggests that the system works more as a support tool than as a replacement for human input. The interaction feels closer to collaboration: the AI drafts a response, and the user refines it. In SME settings, where communication style and relationships with customers matter, this balance seems important.

From an Information Systems perspective, this points to the importance of how AI is embedded into everyday work. Much of the existing literature focuses on standalone tools or chatbot interfaces. Here, the system is integrated at the keyboard level, placing it directly within the flow of communication. This reduces the gap between accessing information and responding to customers, which may explain why the system feels usable even in a relatively simple setup. For SMEs, where resources are often limited, this kind of lightweight integration may be more practical than adopting more complex systems.The study suggests that a keyboard-integrated RAG system can support more efficient and consistent customer communication in SMEs, at least within the context examined here. At the same time, the scope of the evaluation remains limited. The dataset consists of a relatively small number of queries, and the qualitative insights are based on a single stakeholder. For that reason, the findings should be treated as exploratory. Further work involving more participants, broader datasets, and longer periods of use would help clarify how the system performs in real-world SME environments.

6. Conclusion

This study explores how a mobile keyboard–based Retrieval-Augmented Generation system can support customer communication in Small and Medium Enterprises from an Information Systems perspective. Instead of focusing only on technical performance, the study looks at how such a system fits into everyday work, particularly how SME operators access and use business knowledge while interacting with customers. The results suggest that embedding AI directly into a keyboard interface can make a noticeable difference in how customer responses are handled. Rather than acting as a separate tool, the system becomes part of the communication process itself. This seems to help reduce the effort required to retrieve information and draft responses, especially in situations where users need to respond quickly while managing other tasks. Another point that emerges from the study is how knowledge is used in practice. By combining semantic retrieval with generative models, the system is able to produce responses that are generally aligned with existing business information. This can be particularly useful in SME contexts, where knowledge is often scattered across documents and not formally structured. At the same time, the system does not replace human input. Users still review and adjust responses, which highlights a collaborative pattern where AI supports the task rather than fully automating it. These observations contribute to Information Systems research by showing that the value of AI does not only depend on model performance, but also on how and where the system is integrated. In this case, placing the functionality at the keyboard level reduces the gap between accessing knowledge and acting on it. For SMEs, this kind of lightweight integration may be more practical than adopting more complex

systems that require significant changes to existing workflows. There are, however, several limitations that need to be acknowledged. The evaluation is based on a relatively small number of queries and involves SMEs located in Cikarang and Bekasi, Indonesia. In addition, the qualitative insights come from a single stakeholder and reflect short-term system use. These factors mean that the findings should be interpreted with caution and cannot be generalized without further validation. Future research could explore how similar systems perform across different industries and regions, as well as over longer periods of use. It would also be useful to examine how such tools affect broader outcomes, such as customer satisfaction, response consistency, or sales performance. Looking at how these systems integrate with other organizational processes, such as training or onboarding, may also provide a deeper understanding of their role within SME information systems.

References

Aurora, C., & Tuga Mauritsius. (2025). Creation of RAG Chatbot in Answering Queries Related to Banking Terms Using Microsoft Azure. *International Journal of Cyber and IT Service Management*, *5*(2), 144–155. https://doi.org/10.34306/ijcitsm.v5i2.209

Benita, J., Tej, K. V. C., Kumar, E. V., Subbarao, G. V., & Venkatesh, CH. (2024). Implementation of Retrieval-Augmented Generation (RAG) in Chatbot Systems for Enhanced Real-Time Customer Support in E-Commerce. *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pp.1381–1388. https://doi.org/10.1109/icacrs62842.2024.10841586

Bostrom, R. P., & Heinen, J. S. (1977). MIS Problems and Failures: A Socio-Technical Perspective. Part I: The Causes. *MIS Quarterly*, 1(3), 17–32. https://doi.org/10.2307/248710

Campi, R., Giudici, M., Pinciroli Vago, N. O., Brambilla, M., & Fraternali, P. (2025). Enhancing Human-AI Collaboration through a Conversational Agent for Energy Efficiency. *Proceedings of the AAAI Symposium Series*, vol. *5,* no. 1, pp. 52–55. https://doi.org/10.1609/aaaiss.v5i1.35554

Chen, C.-C., & Chang, C.-C. (2025). AI-enhanced reference services in special libraries: a case study of the Hakka Literary Museum. *The Electronic Library*, vol. 43, no. 5, pp. 715–732. https://doi.org/10.1108/el-05-2025-0168

Chien, C.-Y., Wang, H.-C., Wang, S.-M., & Wu, K.-C. (2025). A Novel Intermediator of Food Experience: Applying Retrieval Augmented Generation Technology for Emotional Menu Design. *Proceeding of IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, pp. 87–88. https://doi.org/10.1109/icce-taiwan66881.2025.11208075

Choudhary, N., Goyal, P., Devashish Siwatch, Atharva Chandak, Mahajan, H., Khurana, V., & Kumar, Y. (2025). AdQuestA: Knowledge-Guided Visual Question Answer Framework for Advertisements. *Proceeding of IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5812–5821. https://doi.org/10.1109/wacv61041.2025.00567

Church, K. W., Sun, J., Yue, R., Vickers, P., Saba, W., & Chandrasekar, R. (2024). Emerging trends: a gentle introduction to RAG. *Natural Language Engineering,* vol. 30, no. 4, pp. 870–881. https://doi.org/10.1017/s1351324924000044

Coen, E., Del Fiol, G., Kaphingst, K. A., Borsato, E., Shannon, J., Smith, H., Masino, A., & Allen, C. G. (2025). Chatbot for the Return of Positive Genetic Screening Results for Hereditary Cancer Syndromes: Prompt Engineering Project. *JMIR Cancer*, vol. 11, e65848. https://doi.org/10.2196/65848

Cook, S. D. N., & Brown, J. S. (1999). Bridging Epistemologies: The Generative Dance Between Organizational Knowledge and Organizational Knowing. *Organization Science*, vol. 10, no. 4, pp. 381–400. https://doi.org/10.1287/orsc.10.4.381

Decker, B., Grzemski, S., & Hengstebeck, I. (2025). Implementation of an AI Support Chatbot Based on Microsoft Azure OpenAI with Special Consideration of Quality. *Proceedings of EUNIS 2025 Annual Congress in Belfast*, vol. 107, pp. 237–226. https://doi.org/10.29007/l6ss

Es, S., James, J., Anke, L. E., & Schockaert, S. (2024). RAGAs: Automated Evaluation of Retrieval Augmented Generation. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, 150–158. https://doi.org/10.18653/v1/2024.eacl-demo.16

Mehul, Kanagavalli, V. R., K R, S., P N, G., M P, S., U, S., R, G., S, G., & R, N. (2025). Gen AI Driven FAQ Chatbot Using Advanced RAG Architecture for Querying Annual Reports. *2025 International Conference on Computing and Communication Technologies (ICCCT)*, 1–6. https://doi.org/10.1109/iccct63501.2025.11020025

Mumford, E. (2006). The story of socio-technical design: reflections on its successes, failures and potential. *Information Systems Journal*, vol. 16, no. 4, pp. 317–342. https://doi.org/10.1111/j.1365-2575.2006.00221.x

Lee, H.-C., Hung, K., Man, G. M.-T., Ho, R., & Leung, M. (2024). Development of an RAG-Based LLM Chatbot for Enhancing Technical Support Service. *TENCON 2024 - 2024 IEEE Region 10 Conference (TENCON)*, pp. 1080–1083. https://doi.org/10.1109/tencon61640.2024.10902801

Li, Y., Zhao, H., Lin, J., Zhou, Z., & Zhang, Y. (2025). Research on Personalized Insurance Intelligent Customer Service Dialogue Generation Based on Large Language Models. *Proceedings of the 2025 2nd International Conference on Generative Artificial Intelligence and Information Security*, pp. 178–183. https://doi.org/10.1145/3728725.3728753

Lukman Arif Sanjani, Sarno, R., Sungkono, K. R., Agus Tri Haryono, Abdullah Faqih Septiyanto, & Dwi Sunaryono. (2025). Performance Analysis of LLM Models with RAG and Fine-Tuning T5 for Chatbot Optimization in Call Centers. *Proceeding of International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*, pp. 152–157. https://doi.org/10.1109/icocseti63724.2025.11018908

Lystbæk, M. S., & Ulrik Sahl Lystbæk. (2025). A Norm-Chatbot: Local LLM Vision with Vision-Based RAG for Complex Production Documents and Task-Specific Responses. *Communications in Computer and Information Science*, pp. 17–30. https://doi.org/10.1007/978-3-031-96199-1_2

Nguyen, H. T., NGO, Q.-D., & DANG, Q.-D. (2024). An Intent-Filtered Retrieval-Augmented Generation Chatbot for University Admissions in Vietnam. *Proceedings of International Conference on Advanced Technologies for Communications (ATC)*, pp. 928–933. https://doi.org/10.1109/atc63255.2024.10908188

Orlikowski, W. J., & Yates, J. (2002). It's About Time: Temporal Structuring in Organizations. *Organization Science*, vol. 13, no. 6, pp. 684–700. https://doi.org/10.1287/orsc.13.6.684.501

Packowski, S., Halilovic, I., Schlotfeldt, J., & Smith, T. (2024). Optimizing and Evaluating Enterprise Retrieval-Augmented Generation (RAG): A Content Design Perspective. *Proceedings of the 2024 8th International Conference on Advances in Artificial Intelligence*, pp. 162–167. https://doi.org/10.1145/3704137.3704181

Patel, N., Mouratidis, H., & Zhi, K. N. K. (2025). LLM-Based Automated Hallucination Detection in Multilingual Customer Service RAG Applications. *IFIP Advances in Information and Communication Technology*, pp. 360–373. https://doi.org/10.1007/978-3-031-96235-6_26

Praneeth, B., Mohana, Nattem, E. C., Jetti, K., Kavyashree, B. K., Rakshitha, D., Ramakanth Kumar, P., & Sreelakshmi, K. (2025). Optimization of Customer Feedback Summarization Using Large Language Models (LLM) and Advanced Retrieval-Augmented Generation. *IEEE Access*, vol. 13, pp. 124319–124332. https://doi.org/10.1109/access.2025.3588337

Qi, Y., Gu, S., Ma, W., Wang, N., & Wang, Q. (2025). DeepSem: Multilayer Semantic Image Retrieval for High-Accuracy Multimodal RAG in Industrial Applications. *2025 IEEE 49th Annual Computers, Software, and Applications Conference (COMPSAC)*, pp. 1388–1393. https://doi.org/10.1109/compsac65507.2025.00174

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *The Academy of Management Review*, vol 46, no. 1, pp. 192–210. https://doi.org/10.5465/amr.2018.0072

Rao, D., Zhuang, J., & Jiang, Z. (2025). Leveraging dynamic few-shot prompting and ensemble method for task-oriented dialogue with subjective knowledge. *Information Processing & Management*, vol. 63, no. 2, pp. 104317–104317. https://doi.org/10.1016/j.ipm.2025.104317

Sandakelum, D., Rajapakse, C., & Jayalath, N. (2025). Real-Time Knowledge Retrieval for Banking Chatbots: A RAG-Based Approach to Employee Assistance. *Proceeding of International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pp. 1–6. https://doi.org/10.1109/scse65633.2025.11030999

Shan, R., & Shan, T. (2024). Retrieval-Augmented Generation Architecture Framework: Harnessing the Power of RAG. *Lecture Notes in Computer Science*, pp. 88–104. https://doi.org/10.1007/978-3-031-77954-1_6

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Three Fresh Ideas. *AIS Transactions on Human-Computer Interaction*, vo. 12, no. 3, pp. 109–124. https://doi.org/10.17705/1thci.00131

Singh, R. K., Singh, K., Kunde, S., Mishra, M., Singhal, R., & Nambiar, M. (2025). RAGs to Riches: Cost-efficient Complex Query Orchestration. *Proceeding of ACM/SPEC International Conference on Performance Engineering*, pp. 114–120. https://doi.org/10.1145/3680256.3721327

Thiyagarajan, K. (2025). Multimodal RAG for Enhanced Information Retrieval and Generation in Retail. *2025 International Conference on Visual Analytics and Data Visualization (ICVADV)*, pp. 102–106. https://doi.org/10.1109/icvadv63329.2025.10961713

Varga, I., & Yamashita, Y. (2025). Inquiry Assistant Using LLM-Generated Knowledge Graphs. *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 4319–4323. https://doi.org/10.1145/3726302.3731956

Vaswani, A., et al. (2017) Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 6000-6010.

Wang, J., & Shalaby, A. (2025). TransitTalk: Large language model-based digital assistants for enhancing transit customer experience and staff performance. *Journal of Intelligent Transportation Systems*, pp 1–18. https://doi.org/10.1080/15472450.2025.2576883

Wulf, J., & Meierhofer, J. (2025). The Impact of Large Language Models on Task Automation in Manufacturing Services. *Procedia CIRP*, *134*, pp. 1089–1094. https://doi.org/10.1016/j.procir.2025.03.071

Yang, Y.-T., Jiang, J.-Y., Lin, Y.-T., & Chang, C.-Y. (2025). Enhancing Retrieval-Augmented Generation with Knowledge Graph-Based Soft-Labeling and Triplet Similarity SBERT. *Proceeding of IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, pp. 487–488. https://doi.org/10.1109/icce-taiwan66881.2025.11208012

Yildirim, S., & Ruya Samli. (2025). A RAG-Based Automotive Sector AI Assistant for Enhanced Information Retrieval. *The Eurasia Proceedings of Science Technology Engineering and Mathematics*, vol. 33, pp. 20–27. https://doi.org/10.55549/epstem.1729714