Evaluating User Experience of a Retrieval-Augmented Generation-Based Customer Service Chatbot

Syuan-Yu Chen and Chang-Yi Kao

Department of Computer Science & Information Management, Soochow University, Taiwan
Email: 104102042g@gmail.com (S.Y.C.), edenkao@scu.edu.tw (C.Y.K.)
Manuscript received December 15, 2025; accepted January 29, 2026; published May 22, 2026.

***Abstract*—As digital services become increasingly widespread, online real-time customer support has become a key way for consumers to seek help. A good service experience can boost customer satisfaction and loyalty, particularly in the financial sector. However, traditional keyword-based chatbots often fail to capture users' intent, limiting the scenarios in which they can be used. This study introduces a system based on Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs), aiming to better understand user queries and provide more helpful responses. Through interviews and sentiment analysis, we explored user reactions and what influenced their emotional responses. Findings revealed that traditional bots often triggered frustration and confusion due to rigid responses and poor understanding. In contrast, the RAG-based system demonstrated stronger natural language handling and was perceived as more empathetic. Some users even reported feeling emotionally supported. However, slow response times and difficulty in handling complex queries remained challenges. Based on these insights, we suggest improving response speed, adding voice interaction, and enhancing response reliability to guide future development of intelligent customer support tools.**

***Keywords*—retrieval-augmented generation, intelligent customer service, chatbot, sentiment analysis**

I. Introduction

In the digital era, consumers can easily compare products and services, leading to a growing demand for real-time and personalized support. Online live chat has become a major communication channel, significantly influencing trust, satisfaction, repurchase behavior, and brand image. However, most existing chatbots still rely on keyword-based rules, which makes it difficult to accurately interpret users' intent. This often results in irrelevant or incorrect responses and high maintenance costs.

Large Language Models (LLMs), such as OpenAI's GPT series, demonstrate strong capabilities in natural language understanding and generation. Yet, their application in domain-specific contexts faces challenges such as hallucinations and high training costs, which reduce their suitability for knowledge-intensive tasks. Retrieval-Augmented Generation (RAG) addresses these issues by combining external knowledge bases with generative models. Through semantic retrieval, RAG enhances response accuracy and reduces hallucinations.

This study applies RAG and LLMs to a company's customer service chatbot, aiming to improve semantic understanding, minimize hallucinations, and deliver precise real-time responses. A mixed-methods approach was adopted, combining qualitative semi-structured interviews with quantitative sentiment analysis, to compare user experiences between the rule-based system and the RAG-enhanced system. The core objectives are:

1) To develop a semantically aware RAG - LLM chatbot.
2) To analyze user experience and identify factors influencing satisfaction.
3) To validate improvements in response accuracy, relevance, and emotional interaction.

By addressing the limitations of domain-specific chatbots, this study provides practical insights into integrating RAG and LLMs into intelligent customer service, offering valuable implications for how RAG technology shapes and improves the user experience within the financial industry.

II. Literature Review

A. Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG), introduced by Lewis *et al*. (2020), addresses key limitations of Large Language Models (LLMs) in domain-specific applications, particularly in knowledge freshness and expertise. RAG integrates a retrieval module with a generation module: relevant documents are first retrieved from an external knowledge database based on user input, then used as context for generating more accurate and reliable responses.

1) Naive RAG

Naive RAG, the earliest form of RAG (Gao, 2023), follows a basic retrieve-augment-generate pipeline. It typically relies on simple semantic similarity for document selection, making it suitable for small datasets but less effective for complex queries or large-scale knowledge bases. Common limitations include low retrieval accuracy and recall, difficulty in integrating retrieved content, and hallucinations or irrelevant outputs during generation. Despite being cost-effective and outperforming vanilla LLMs, Naive RAG struggles with scalability and precision, these limitations highlight the need for more refined approaches.

2) Advanced RAG

Advanced RAG improves performance through multiple stages. In data preparation, techniques like sliding windows and recursive splitting preserve semantic context while enhancing granularity. Metadata enrichment and structural indexing further refine retrieval precision. Fine-tuning embedding models helps vector representations align better for domain-specific purpose.

During query processing, methods such as query rewriting, expansion, and routing enhance clarity and relevance. These techniques address ambiguity and improve retrieval outcomes, especially in specialized contexts. In the augmentation and generation stages, reranking, context compression, and hybrid augmentation reduce noise and

improve coherence. Modular RAG allows iterative refinement, enabling dynamic feedback loops between retrieval and generation. Recursive retrieval strategies like IRCoT and ToC further support complex reasoning by progressively refining queries and focusing on relevant information.

RAG has demonstrated strong potential in domains such as finance and customer service, where timely and accurate information is critical (Izacard *et al*., 2021). However, challenges remain, including dependency on up-to-date knowledge bases and high computational demands when scaling to large datasets (Gao *et al*., 2023).

B. Applications of RAG and LLM in the Insurance and Customer Service Sectors

Traditional insurance products are often difficult for consumers to understand due to lengthy contracts, technical jargon, and long coverage periods. The sales process remains highly agent-dependent, contrasting with the customer-driven models of banking and e-commerce. However, younger generations (X, Y, Z) of digital natives prefer self-directed exploration of financial services, creating opportunities for AI-driven customer support to reshape insurance sales.

In customer service, intelligent systems have become essential for improving efficiency and user experience, and even reducing operational costs (Adam, 2021). Keyword-based chatbots are limited, particularly when handling complex or multi-intent queries. RAG addresses this by combining knowledge retrieval with generative models, enabling accurate, context-aware responses that improve customer satisfaction (Mero, 2018). Integration with knowledge graphs, as demonstrated by Xu *et al*. (2024) in LinkedIn's support system, has shown measurable gains in efficiency.

Beyond insurance, RAG and LLM applications extend to healthcare, e-commerce, and multilingual support, where they enhance recommendation accuracy and generate empathetic responses.

Despite these advantages, some challenges still remain. First, database quality and timeliness directly affect retrieval accuracy, especially in finance where rules or product updates happen often. Outdated knowledge bases may lead to inaccurate answers, undermining trust (Bender, 2021). Second, LLMs are computationally expensive, with high energy consumption and carbon emissions (Strubell, 2019). This raises concerns for enterprises, particularly in the financial sector, where Environmental, Social, and Governance (ESG) requirements are increasingly important. Research into techniques such as model distillation, distributed computing, and efficient hardware design (Lan, 2019) is crucial to address these issues.

Another problem is language variety. While LLMs work well in high-resource languages, they struggle with complicated meanings and culture-specific expressions in under-resourced languages (Joshi, 2020). This is critical in finance, where metaphorical or culture-specific language may shape customer questions. In addition, transparency and explainability are urgent issues: although LLMs produce fluent answers, how they make decisions is not clear, creating risks in areas like finance (Gao *et al*., 2023). Making RAG systems which are easy to understand and show sources clearly is therefore necessary.

In conclusion, although RAG and LLM technologies face multiple challenges, they offer new opportunities for improving insurance and customer service. Combining them not only allows the system to handle simple Q&A, but also supports more interactive, decision-making, and empathetic tasks, showing strong potential to enhance financial services through greater efficiency, personalization, and trust.

C. Sentiment Analysis

1) Overview and applications

Sentiment analysis, or opinion mining, is a core NLP technique used to classify emotional polarity in text—positive, negative, or neutral. It is widely applied to online reviews, social media, and interviews, offering insights into user attitudes and public opinion. Sentiment analysis originate from early public opinion research, and its relevance surged with the rise of online content in the 2000s, leading to rapid academic growth (Mäntylä *et al*., 2016).

Technically, sentiment analysis methods include lexicon-based and machine learning approaches. Early foundational work by Pang, Lee, and Vaithyanathan (2002) demonstrated that machine learning models such as Naïve Bayes and Support Vector Machines outperform lexicon-based and intuition-driven methods in binary classification tasks (positive/negative). Nevertheless, lexicon-based techniques retain advantages in interpretability and semantic consistency, particularly in non-English contexts.

For instance, Ku and Chen (2007) constructed a Chinese sentiment lexicon tailored to news and blogs by leveraging the General Inquirer (GI) and Chinese Network Sentiment Dictionary (CNSD) as seed sources, followed by manual refinement. Their model incorporated strategies such as negation handling, opinion holder weighting, and sentiment aggregation at the sentence and document levels. Results showed higher accuracy in blog corpora than in news data, since blog opinions are explicit and news are typically more objective. Similarly, Chen (2010) applied a lexicon-based approach to Chinese movie reviews with effective results, highlighting the continued value of lexicon methods in Chinese sentiment analysis.

Recently, deep learning has advanced fine-grained sentiment analysis, moving beyond polarity classification to detect emotions such as joy, anger, sadness, and surprise. Models such as CNN, LSTM, and BERT have proven effective in capturing refined sentiment and intensity (Tan, 2023). Reinforcement learning techniques are also being explored to address challenges such as sarcasm and ambiguity. These developments not only improve analytical accuracy but also enrich the contextual understanding of emotions in text.

2) Interviews as sentiment analysis data

Interviews are a key source of qualitative data for sentiment analysis, offering rich insights into participant experiences. Structured interviews follow a fixed set of questions, making the data more consistent. Unstructured interviews are open and flexible, with little or no pre-planned structure. Unstructured interviews can be time-consuming and harder to manage but is helpful when exploring new or

complex topics. Semi-structured interviews offer a balance between structure and flexibility. They include key questions to guide the conversation but also allow follow-up questions based on the participant's responses. This format helps gather richer data and often reveals insights that researchers may not have expected (Gill *et al*., 2008).

Interviews are used to explore people's attitudes, behaviors, and social contexts in many areas, such as healthcare, education, psychology, and sociology. As a qualitative tool, interviews provide valuable emotional and contextual information that complements quantitative data and helps researchers better understand complex issues.

III. Materials and Methods

A. System Implementation

1) Data preparation and system development

Relevant data was collected using web scraping techniques. The Advanced RAG framework with Hypothetical Document Embeddings (HyDE) was applied, where FAQ questions were embedded together with generated hypothetical documents and saved into a vector database. The system further incorporated Router Chains, Chain-of-Thought (CoT) reasoning, and Query Rewriting techniques.

When a user submits a query, the system classifies the query, directs it through an appropriate process, retrieves relevant content from the vector database, summarizes and refines the response, and then delivers it back to the user.

B. Interviews

1) Participant selection

This study employed a combination of convenience and random sampling. Participants were recruited voluntarily without restrictions on prior familiarity with intelligent customer service systems. This approach enabled efficient recruitment under limited resources, while ensuring practicality and flexibility for an exploratory study (Etikan *et al*., 2016; Marshall, 1996).

2) Interview method

Semi-structured interviews were adopted as the primary data collection method. Compared with structured interviews, this approach allows for deeper exploration of participants' emotions and experiences. Compared with unstructured interviews, it ensures guidance and efficiency while avoiding excessive variability in responses.

3) Interview procedure

Each interview lasted about 60 to 80 min and included the following sections:

- User background: Understanding the participant's age, occupation, and how often they use chatbots, along with their main purpose for using them.
- Initial impressions: capturing the participant's first reactions and emotional responses to the system.
- Emotional responses: Exploring emotional changes during use, with specific examples and feedback.
- System rating: Asking participants to rate the system, providing useful quantitative data for later analysis.
- Suggestions for improvement: Collecting ideas and expectations for system enhancement.

All interviews were recorded and transcribed for sentiment analysis. After transcription, the text was cleaned and refined to improve clarity and make it suitable for analysis.

C. Sentiment Analysis

The quantitative sentiment analysis process included the following steps:

1) Initialization of variables
   - sentiment_score: overall sentiment score, initialized as 0.
   - words: total number of words in the transcript.
   - dictwords: number of words matched with the sentiment dictionary.
   - negation_amount: count of negation words in the sentence.
   - emotional_degree: weighted intensity of the nearest degree adverb (default = 1).
   - positive_sentence and negative_sentence: counts of sentences classified as positive or negative.
2) Sentence-level processing
   - Iterate through all words in each sentence.
   - If a word is a negation, increment negation_amount.
   - If a word is a degree adverb (e.g., "very," "slightly"), adjust emotional_degree accordingly (e.g., very = 1.5, slightly = 0.5).
   - If a word is an emotional word, determine polarity (positive or negative).
     - If negation_amount is odd, polarity is reversed. If even, polarity remains.
     - Contribution score = polarity (±1) × emotional_degree. Add to sentiment_score.
   - At the end of each sentence:
     - If the sentiment score > 0, increment positive_sentence.
     - Otherwise, increment negative_sentence.
3) Score adjustment and normalization

Since transcripts varied in length, raw sentiment scores were adjusted (Chen, 2019):

$$\text{adjusted\_sentiment\_score} = \frac{sentiment\_score \times dictwords}{words}$$

To ensure comparability, adjusted scores were normalized to the range −1 to +1:

$$\text{normalize\_to\_one} = \frac{\text{adjusted\_sentiment\_score}}{\sqrt{(\text{adjusted\_sentiment\_score})^2 + 2}}$$

4) Overall sentiment classification
   - If normalize_to_one > 0 and positive_sentence > negative_sentence: overall sentiment is positive.
   - If normalize_to_one < 0 and negative_sentence > positive_sentence: overall sentiment is negative.
   - Otherwise, the sentiment is considered ambiguous or mixed and requires further qualitative interpretation.

IV. Result

A. Data Results

Key performance indicators before and after optimization are shown below:

Table 1. Data results

Metric	Before (avg.)	After (avg.)	Change (After − Before)
Adjusted Sentiment Score	−1.4450	3.3344	+4.7794
Normalized Score	−0.6684	0.7302	+1.3986
Positive Sentences	12.2	23.6	+11.4
Negative Sentences	20.4	8.6	−11.8
Sentiment Dictionary Matches	63.6	61.8	−1.8
User Ratings	3/10	7.4/10	+4.4

The data reveal significant improvements in sentiment metrics after system optimization. The overall sentiment shifted noticeably toward the positive. Normalized scores showed a clear improvement, transitioning from negative to positive values. This change was reflected in the increased number of positive sentences and a corresponding decrease in negative sentences. The optimized system showed better results in different ways of measuring sentiment, such as polarity distribution and normalized scores.

B. Qualitative Evaluation

To gain a more comprehensive understanding of how system optimization influences user experience, this study not only analyzed quantitative data but also qualitative one. Feedback regarding both traditional keyword-based customer service chatbots and the optimized Retrieval-Augmented Generation (RAG) system was collected and examined.

1) User backgrounds

Interview results revealed that participants' prior experiences with customer service chatbots were primarily routine inquiries or operational guidance, typically involving either telephone-based voice menus or pre-set options embedded in messaging apps. For example, Participant 5 frequently used a voice-based service for credit card payments, where tasks could be completed simply by pressing numerical options (e.g., "Press 1"). Other participants often engaged with hotel services or event-related queries through fixed menu buttons such as "Check-in," "Check-out," or "Wi-Fi password." In general, prior experiences of chatbot usage were described as "simple operations, fixed contexts, and almost no need for text-based interaction."

2) Interview findings

Regarding the keyword-based chatbot used in this study, participants consistently reported dissatisfaction and frustration. They noted the system's limited ability to interpret queries, often producing responses like "Sorry, I don't understand," and failing to provide answers when identical questions were rephrased. One participant stated, "If I phrase the same question in a more casual way, it no longer understands." Another said, "I had to use nearly identical wording, otherwise it wouldn't recognize what I meant." The other said, "It kept asking me to repeat my question though it was already explicit enough."

Such experiences led to perceptions of "ineffective communication" and triggered negative emotions, ranging from frustration and confusion to outright anger. One participant described their emotional experience as "going from feeling excited, to getting frustrated, and finally feeling let down," while another gave a zero out of ten rating, remarking, "It never understood my questions, just kept saying it didn't understand."

In contrast, the optimized RAG-based system received mostly positive feedback. Users reported that the chatbot successfully understood their queries, and they also explicitly expressed satisfaction with the responses. Beyond factual accuracy, participants highlighted the system's emotional support. For example, one participant described feeling "comforted" during a health-related insurance inquiry, saying, "It felt like it truly empathized with me." Another participant recalled, "When I asked about borrowing against my insurance policy, it understood me immediately—I was thrilled!" These comments emphasized the improvements in empathetic tone and tolerance for semantic variation, which reduced emotional stress.

Participants also noted that the RAG system displayed a higher tolerance for long, complex, or incomplete queries, as well as for typos and irrelevant content. For instance, one participant explained, "I mainly wanted to update the address listed in my insurance policy, but when I said I was moving house, it still understood." Another user asked the chatbot "I was diagnosed with Norovirus. Can I file a claim?" The system correctly interpreted Norovirus as a case of food poisoning, which surprised the user.

However, all participants pointed out a critical drawback: the slow response speed. Since the RAG system involved local computation and multiple calls to LLMs or other models—sometimes up to four per query—the waiting time often stretched to two or three minutes. As participants described: "Every question felt laggy," "The delay made the experience feel less like talking to a real person," and "The slowness was frustrating." In addition, some participants observed occasional semantic misinterpretations, particularly when questions contained multiple intentions or implicit meanings. For example, one participant reported, "I asked about kidney stones and expected both coverage details and claim procedures, but it only addressed claims." Furthermore, an older participant suggested that voice input would make the system more convenient, showing that the design should work well for people of all ages.

Overall, the RAG system improved semantic understanding and emotional resonance, especially in natural language processing and empathetic responses. Remaining challenges included response time, complex query handling, and multimodal interaction.

V. Conclusion and Recommendation

A. Conclusion

This study used a mixed-methods approach, combining quantitative sentiment scores with qualitative analysis, to compare traditional keyword-based chatbots with an optimized RAG system. We focused on three areas: semantic understanding, response quality, and user emotions.

Our analysis first reveals a clear gap between the two

systems. Traditional systems often fail to understand natural language, leading to communication errors and making users feel frustrated, confused, or even angry. Our sentiment dictionary method successfully quantified this frustration. In contrast, the RAG system showed a significant improvement. It can handle incomplete sentences and understand the user's hidden intent. Most participants felt the RAG system truly understood them, some even describing the experience as "comforting" or "empathetic." This proves that RAG not only improves efficiency but also builds trust.

However, despite its success, the RAG system still has room for improvement. Users reported issues like slow response speeds and difficulty handling multi-intent questions. Specifically, we identified four main risks in LLM-generated responses:

1) Misunderstanding complex or hidden messages.
2) Failing to retrieve all the necessary information.
3) Retrieving the right information but failing to include it in the final answer.
4) Hallucinations, where the AI makes up fake information.

For high-risk fields, these limits mean we must build strict verification tools to ensure the accuracy and completeness.

Finally, this study contributes not only to practice but also to research methodology. Traditionally, interview data analysis relies on purely qualitative methods, such as grounded theory. This research adopted a mixed-methods approach. We simultaneously conducted qualitative analysis and calculated quantitative sentiment scores from the same interview transcripts. This effectively fills the gap in traditional research by allowing us to quantify the intensity of emotions that qualitative analysis alone might miss.

B. Recommendation

To address the challenges mentioned above, we offer the following recommendations:

1) Practical implementation

1) Speed Up the System: Future developers should optimize model invocation to avoid unnecessary computing and makes the chatbot respond faster.
2) Add Multi-modal Features: For elderly users or those who struggle with typing, adding voice input, voice feedback would make the system much easier to use.
3) Build Trustworthy Outputs: Providing source links and confidence scores with each answer can help users trust the information and reduce the impact of hallucinations.
4) Handle Multiple Intentions: Use "semantic decomposition" or "multi-turn reasoning" to break down the user's goals and give more focused answers.

2) Theoretical and future research

1) Expand Sample Sizes: Future research should involve more participants from different age groups and digital literacy levels. This will help confirm if our findings apply to everyone in different situations.
2) Balance Efficiency and Privacy: Companies must find the "sweet spot" between speed, accuracy, and data privacy. For instance, cloud systems are fast but risky for privacy, while private servers are secure but slower. Finding this balance is a key topic for future AI research.
3) Apply Sentiment Tools to Other Fields: We recommend testing our sentiment analysis method, which uses interview transcripts, is reliable and broadly useful in other text-heavy research areas.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

SYC designed the system, conducted the experiments, and wrote the manuscript; Prof. CYK supervised the research and provided critical revisions; All authors had approved the final version.

REFERENCES

Adam, M., Wessel, M., & Benlian, A. 2021. AI-based chatbots in customer service and their effects on user compliance. ***Electronic Markets,*** 31(2): 427–445.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big?. ***Proceedings of the 2021 ACM conference on fairness, accountability, and transparency***: 610–623.

Chen, J., Becken, S., & Stantic, B. 2019. Lexicon Based Chinese Language Sentiment Analysis Methods. ***Computer Science and Information Systems,*** 16(2): 639–655.

Chen, L. 2010. ***A study on automatic classification of Chinese sentiment semantics.*** Master's thesis. Graduate Institute of Computer Science and Information Engineering, National Taiwan Normal University., 1–39.

Etikan, I., Musa, S. A., & Alkassim, R. S. 2016. Comparison of convenience sampling and purposive sampling. ***American Journal of Theoretical and Applied Statistics,*** 5(1): 1–4.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2(1).

Gill, P., Stewart, K., Treasure, E., & Chadwick, B. 2008. Methods of data collection in qualitative research: interviews and focus groups. ***British Dental Journal,*** 204(6): 291–295.

Izacard, G., & Grave, E. 2020. Leveraging passage retrieval with generative models for open domain question answering. arXiv preprint, arXiv:2007.01282.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. arXiv preprint, arXiv:2004.09095.

Ku, L. W., Liang, Y. T., & Chen, H. H. 2006. Opinion extraction, summarization and tracking in news and blog corpora. ***AAAI spring symposium: Computational approaches to analyzing weblogs***, vol. 100107: 1–167.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. arXiv preprint, arXiv:1909.11942.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. ***Advances in neural information processing systems,*** 33: 9459–9474.

Mäntylä, M. V., Graziotin, D., & Kuutila, M. 2018. The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. ***Computer Science Review,*** 27: 16–32.

Marshall, M. N. 1996. Sampling for qualitative research. ***Family Practice,*** 13(6): 522–526.

Mero, J. 2018. The effects of two-way communication and chat service usage on consumer attitudes in the e-commerce retailing sector. ***Electronic Markets,*** 28: 205–217.

Pang, B., Lee, L., & Vaithyanathan, S. 2002. Thumbs up? Sentiment classification using machine learning techniques. arXiv preprint,

cs/0205070.

Strubell, E., Ganesh, A., & McCallum, A. 2020. Energy and policy considerations for modern deep learning research. ***Proceedings of the AAAI conference on artificial intelligence*** (Vol. 34(09): 13693–13696.

Tan, K. L., Lee, C. P., & Lim, K. M. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. ***Applied Sciences,*** 13(7): 4550.

Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. 2024. Retrieval-augmented generation with knowledge graphs for customer service question answering. ***Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval:*** 2905–2909.