http://dx.doi.org/10.62762/TMI.2026.469770

RESEARCH ARTICLE

Context-Aware Large Language Model for Customer Support Chatbots

Rabia Shabbir[1], **Kashif Talpur**[2,*] **and Shakeel Ahmad**[1]

[1] School of Technology and Maritime Industries, Southampton Solent University, Southampton, SO14 0YN, United Kingdom
[2] School of Computing, Engineering, and the Built Environment, University of Roehampton, London, SW15 5PH, United Kingdom

Abstract

Despite their high conversational fluency, large language models (LLMs) tend to produce responses that are either contrived or factually inaccurate, a phenomenon known as hallucination. This constrains their dependability in areas where accuracy is crucial, like customer service. This study leverages a context-aware chatbot built on a Retrieval-Augmented Generation (RAG) pipeline to solve the problem. The system retrieves semantically relevant text from an external knowledge base by integrating neural information retrieval with LLMs to ensure factual correctness and enhanced contextual relevance. These externally retrieved documents are given as a reference during response generation. The LLM-as-a-judge approach is used to evaluate the system by comparing responses to a qualitative performance matrix using GPT-4o. Results show that the RAG-based chatbot increases context precision by a factor of 7.5, decreases hallucinations (measured through Faithfulness) by 73.20% and improves answer relevance by 6.97% when compared to a random retrieval baseline system. This study advances conversational AI by demonstrating how the retrieval method significantly enhances the usefulness and reliability of LLMs for enterprise-level customer service. The results show that the RAG architecture provides a scalable alternative for creating precise, contextually grounded conversational agents, thereby mitigating some of the main drawbacks of LLMs.

Keywords: embedding, large language model, retrieval-augmented generation, LLM-as-a-judge, natural information retrieval.

1 Introduction

The development of the Retrieval-Augmented Generation (RAG) architecture has been motivated by the limits of pre-trained language models in retrieving domain-specific knowledge and real-time information. RAG allows for dynamic access to up-to-date information while answer production is underway by combining the text generation abilities of LLMs with external knowledge retrieval systems. By adding pertinent external data from databases, knowledge bases, and real-time sources, this hybrid approach overcomes the static knowledge

Submitted: 18 April 2026
Revised: 18 May 2026
Accepted: 07 June 2026
Published: 26 August 2026

Vol. 2, **No.** 4, 2026.
10.62762/TMI.2026.469770

***Corresponding author:**
✉ Kashif Talpur
kashif.talpur@roehampton.ac.uk

Citation
Shabbir, R., Talpur, K., & Ahmad, S. (2026). Context-Aware Large Language Model for Customer Support Chatbots. *ICCK Transactions on Machine Intelligence*, 2(4), 190–199.

limitations of pre-trained models. In customer service applications, where reliable automated responses depend on current and domain-specific knowledge, the design has proven very useful.

The main research question, "How can RAG-based architecture enhance the context awareness and response quality of customer-service chatbots?" is addressed with a detailed interpretation and discussion of the obtained results. The Retrieval-Augmented Generation Assessment (RAGAS) framework is used to evaluate model performance. The system evaluation is conducted by comparing the context precision, context recall, answer relevance, and faithfulness metrics of the semantic-based proposed system with a random retrieval baseline system.

The rest of the paper is organised as follows: Section 2 reviews the key research on the application of LLMs in the customer service sector; Section 3 details research methodology; and Section 4 presents analysis and elaborates on a detailed discussion of the research achieved in this research. Finally, this research is concluded in Section 5, which also highlights limitations and future research directions.

2 Related Work

Chatbots can be divided into two main categories: task-oriented and conversational. Task-oriented chatbots manage duties such as taking reservations, setting appointments, or providing customer service [1]. Chat-oriented chatbots engage users in conversation and mimic human-like interaction [3]. However, such systems face systematic challenges, including person-specific errors that degrade response consistency and user experience [2]. It is important to note that both types differ in architecture and assessment techniques. Early conversational AI systems built the framework for human-computer communication. ELIZA [4] used rule-based replies, pattern matching, and keyword recognition to mimic a non-directive psychotherapist. PARRY [5] mimicked paranoid patient behaviour by using belief systems and emotional states to generate suitable responses. These innovations showed how computing methods could mimic psychological states and speech patterns. In the late 2010s and early 2020s, chatbot technology transformed with Large Language Models (LLMs) like Google's BERT and T5, and OpenAI's GPT series. These models introduced new capabilities alongside new risks, including vulnerabilities and privacy concerns that accompany their deployment across sensitive domains [6].

Conversational AI has undergone a significant transformation with the rise of LLMs, particularly in customer service applications, where contextual awareness is crucial. As shown in [7], recent advances in customer service automation using LLMs have demonstrated notable improvements in accuracy, customer satisfaction, and operational efficiency. The growing adoption of LLMs across enterprise settings reflects their expanding reach and transformative impact [8]. LLM-based solutions in customer service show strong potential across implementation approaches and industry sectors. For example, authors in [9] describe an automated customer care system for mobile operators that combines LLMs with a RAG architecture. The researchers in [10] built a customer service chatbot for retail settings using the Llama-7B LLM, illustrating the practical value of LLMs for natural, contextually relevant customer interactions. Additionally, [11] optimized customer feedback summarisation through LLMs and advanced RAG, highlighting how such systems can analyse and synthesize customer feedback to support business intelligence and enhance services. To address the computational demands of context-aware LLM applications, semantic caching has emerged as an optimisation technique. By storing and retrieving semantically similar queries, context-based semantic caching helps reduce redundant computation while maintaining contextual knowledge, as shown by [12]. This approach targets the practical challenge of implementing cost-effective LLM systems in resource-constrained enterprise environments.

Recent advances in the field show that scaling pre-trained language models leads to more complex conversational abilities. GPT-4 [13] stands out due to improved reasoning and a deeper understanding of conversation context. These large-scale models can address a range of consumer inquiries without extensive domain-specific training. As a result, they are especially useful in customer service and support systems. Their influence extends beyond technical uses. As documented in [14], hallucinations in language model outputs—where generated text is unfaithful to the source or factually incorrect—pose a fundamental challenge to deploying LLMs in reliability-sensitive applications such as customer service. However, using pre-trained LLMs in specialised fields, such as customer service, reveals drawbacks that call for better designs. Although these models possess broad language skills, relying solely

on pre-training may result in outdated information, limited domain expertise, and potential contradictions. This gap is evident even in specialised applications such as emotional support chatbots, where LLMs require additional fine-tuning and domain-specific data augmentation to deliver appropriate and contextually grounded responses [15]. Researchers in [16] describe frameworks like LangChain, which address implementation issues by enabling the rapid development of LLM applications. Yet, the core limits of static pre-trained knowledge remain. These problems have led to the development of new methods that enhance pre-trained models with dynamic knowledge retrieval, making conversational systems more resilient and context-aware.

By combining LLMs with external knowledge retrieval, RAG changes how we approach knowledge-intensive NLP applications. Introduced by [17], this method enhances NLP tasks by retrieving and utilising data from external sources while generating text. This hybrid approach addresses the limits of LLMs, which cannot access real-time or domain-specific data not present in their original training. New frameworks have made it simpler to incorporate retrieval into language models, thereby facilitating easier RAG implementation. The authors in [18] explored automated customer support using LangChain, a tool for building custom open-source GPT chatbots for businesses with RAG. Similarly, the work in [16] provided a step-by-step guide for creating LLM applications with LangChain, enabling users to build effective RAG-based LLMs. These tools have broadened RAG's reach, enabling businesses to build complex AI applications without requiring deep technical knowledge of these methods.

Recent studies and benchmarking have focused on RAG systems, examining their capabilities, limitations, efficacy, and performance. For instance, [19] provided comprehensive benchmarking of LLMs in retrieval-augmented generation contexts. Their current work offers key insights into how different model configurations and retrieval tactics affect system performance, and it established essential baseline metrics and assessment frameworks now widely used to compare RAG implementations across domains and use cases.

The development of AI systems, especially in conversational AI and business intelligence, can be viewed through the evolution of RAG architecture. Earlier work such as [3] examined the integration of big data and business intelligence features into chatbots, anticipating many concepts now central to modern RAG implementations. A critical enabler of RAG system evaluation is the RAGAS framework [20], which provides automated, reference-free metrics—including context precision, faithfulness, and answer relevancy—that make it possible to assess both the retrieval and generation components of a RAG pipeline without relying on ground truth human annotations.

Evaluating chatbots and LLMs is a crucial first step in understanding their potential, constraints, and appropriate use cases. The evaluation process typically employs various methods and metrics to assess different aspects of a model's performance, including linguistic skills, knowledge retention, and reasoning, as well as human-centred measures such as relevance, humanness, and empathy. Table 1 provides a comprehensive overview of procedures and considerations used in previous research for both automatic and human evaluation of LLMs.

Table 1. Metrics used in Human and Automatic Evaluation.

Evaluation Metric	Existing Studies
BLEU score	[7, 21]
Mean Absolute Error (MAE), and Mean Squared Error (MSE)	[22]
BERT score, ROUGE, RAGAS	[11, 23]
ROUGE-L	[24, 25]
Embedding Average, Greedy Matching, Vector Extrema	[26]
Identification, Comforting, Suggestion, Overall	[26]
Sensibleness, SSA	[27]
ACUTE-Eval, Humanness	[28]
Engagingness	[29]
Quantity, Relation, Manner	[30]

Researchers use both standard NLP metrics and domain-specific frameworks to assess RAG models. The work in [21] established a baseline in non-English environments by evaluating Korean text quality in HyperCLOVA using BLEU scores. Reference [24] assessed patient-centred medical dialogue creation in PlugMed with Rouge-L, while [25] applied Rouge-L and other metrics to evaluate RAG model domain adaptation for open-domain question answering, demonstrating the importance of specialised retrieval for domain-specific applications.

The author in [26] used Greedy Matching, Vector Extrema, and Embedding Average metrics for their emotional support dialogue systems. These metrics measure semantic similarity using 13 vector representations instead of just surface-level text matching. This offers a more sophisticated assessment of response generation. In both papers, the authors [11, 23] combined BERT Score, ROUGE, and RAGAS metrics for recent work on customer feedback summarisation. BERT Score provides contextualised, embedding-based evaluation. ROUGE measures lexical overlap. RAGAS specifically evaluates retrieval-augmented generation by assessing both retrieval accuracy and generation quality.

Some studies provide a multifaceted assessment framework by [22] that evaluates both statistical performance and semantic quality in RAG implementations. These studies use statistical metrics like Mean Absolute Error (MAE) and Mean Squared Error (MSE) to quantify prediction accuracy in numerical evaluation scenarios. However, while MAE and MSE effectively measure regression accuracy, they do not evaluate the semantic aspects important in Natural Language Processing (NLP) applications. This highlights a gap between what statistical metrics assess and the semantic evaluation metrics required for NLP tasks.

The existing research has generally examined response correctness, coherence, relevance, and some aspects of fluency. Studies on emotional support focused on the chatbot's helpfulness, empathy, and calming qualities [26]. In this paper, we use the Sensibleness and Specificity Average (SSA) [27], which averages specificity (how specific the response is to the context) and sensibleness (how logical the response is). Unlike Perplexity, which measures the likelihood of a given response, SSA directly assesses qualitative traits that reflect human judgment. Additional measures, such as humanness [28], personal manner [30], and engagingness [29] are used to evaluate chatbot behaviour. Most human evaluation research involved surveys with targeted, scale-rated questions, relying on subjective assessment, which differs from automated metrics like SSA and Perplexity.

In conclusion, both retrieval and generative components must be assessed for effective end-to-end RAG evaluation. Thus, RAGAS is the most suitable tool for this purpose.

3 Methodology

The primary objective of this paper is to enhance the factual accuracy of LLM-based customer service chatbots by providing them with relevant information tailored to a specific query. A systematic research framework is adopted to achieve this goal, as we develop a functional system which is evaluated based on predefined criteria. We set up the computing environment using Google Colab notebook with a T4 GPU and the Llama 3-8B LLM orchestrated via the Ollama interface. A domain-specific dataset, the Bitext Customer Support Dataset from Hugging Face, was selected for the RAG system's knowledge base and evaluation. We performed an extensive exploratory data analysis to understand data quality and inform preprocessing. To address LLM token limits, the preprocessing pipeline included chunking, dividing large texts into smaller segments to prevent data loss or model failure.

In the next phase, we built the core RAG pipeline: implemented an embedding model to encode data chunks as dense vector embeddings, built a knowledge base on Chroma DB, and integrated Llama 3 LLM to generate grounded responses. The first step in RAG was transforming pre-processed text into numerical embeddings. This is done with a neural network embedding model, which maps text into high-dimensional embeddings. These embeddings reflect semantic similarity as proximity in the embedding space [31, 32]. The resulting embeddings enable semantic search, outperforming keyword-based search. The Nomic Embed model is a state-of-the-art open-source model that produces high-quality embeddings. The workflow involved iterating over the data chunks from preprocessing and sending them to Ollama to generate embeddings. These embeddings were then stored in Chroma DB for later use.

We evaluated the model's performance using RAGAS, which employs the LLM-as-a-Judge methodology. Here, a separate, stronger LLM assesses the relationship among the query, generated answer, reference answer and retrieved context, from the test set. The judge LLM receives targeted questions, such as whether a chunk is relevant to the query or if it supports a given claim, and gives discrete judgments (e.g., relevant/irrelevant, supported/unsupported). Figure 1 illustrates the RAGAS Framework.

We employed a quantitative research design to objectively measure system performance using

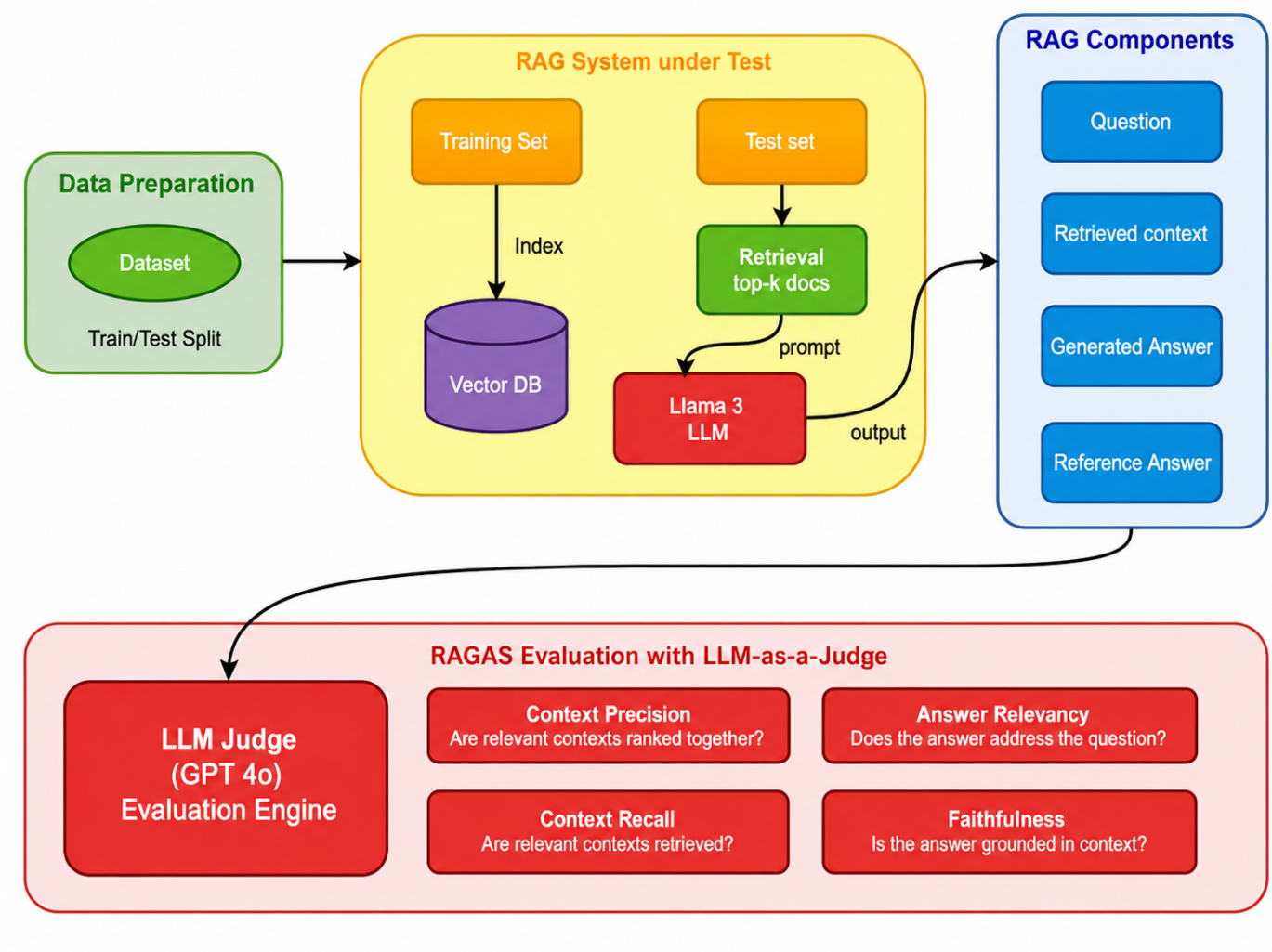

Figure 1. An illustration of RAGAS framework utilised.

statistical metrics. This approach ensures both validity and reliability. We selected the chunk size, the number of retrieved documents (top-k retrieval value), the choice of embedding model, and the choice of LLM as independent variables. The dependent variables are the RAG system's performance metrics: context precision, context recall, answer relevancy, and faithfulness. Each metric ranges from 0 to 1. We sourced data from the Bitext Customer Service Dataset in the Hugging Face repository, split the data into training and test sets using a stratified method to preserve categorical distributions. The training set built the knowledge base for the RAG system and the test set evaluated system performance with the RAGAS framework. The random baseline system is implemented by randomly selecting k documents from the knowledge base without any semantic similarity consideration, serving as a naive retrieval strategy for comparison. RAGAS employed the LLM-as-a-Judge paradigm to automate evaluation, balancing objectivity with scalability. A separate, more capable LLM (GPT-4o, OpenAI, 2024) [13] provided qualitative judgments on RAG system outputs. These judgments inform the calculation of quantitative metrics such as context precision and recall. Finally, we conducted statistical analyses of these metrics to answer the research question.

Figure 2 shows that our RAG pipeline consists of three interconnected stages: indexing, retrieval, and generation. Whenever new documents are incorporated into the knowledge base, the indexing stage initiates. This process transforms the varied incoming documents into smaller, meaningfully grouped chunks through chunking, allowing for optimal storage and future access. Each chunk is represented as an embedding by applying an embedding model, capturing its main semantic attributes. These embeddings are loaded into a scalable vector database, where they are organised for rapid similarity searches and retrievals. When a user submits a query, the retrieval stage begins. The system

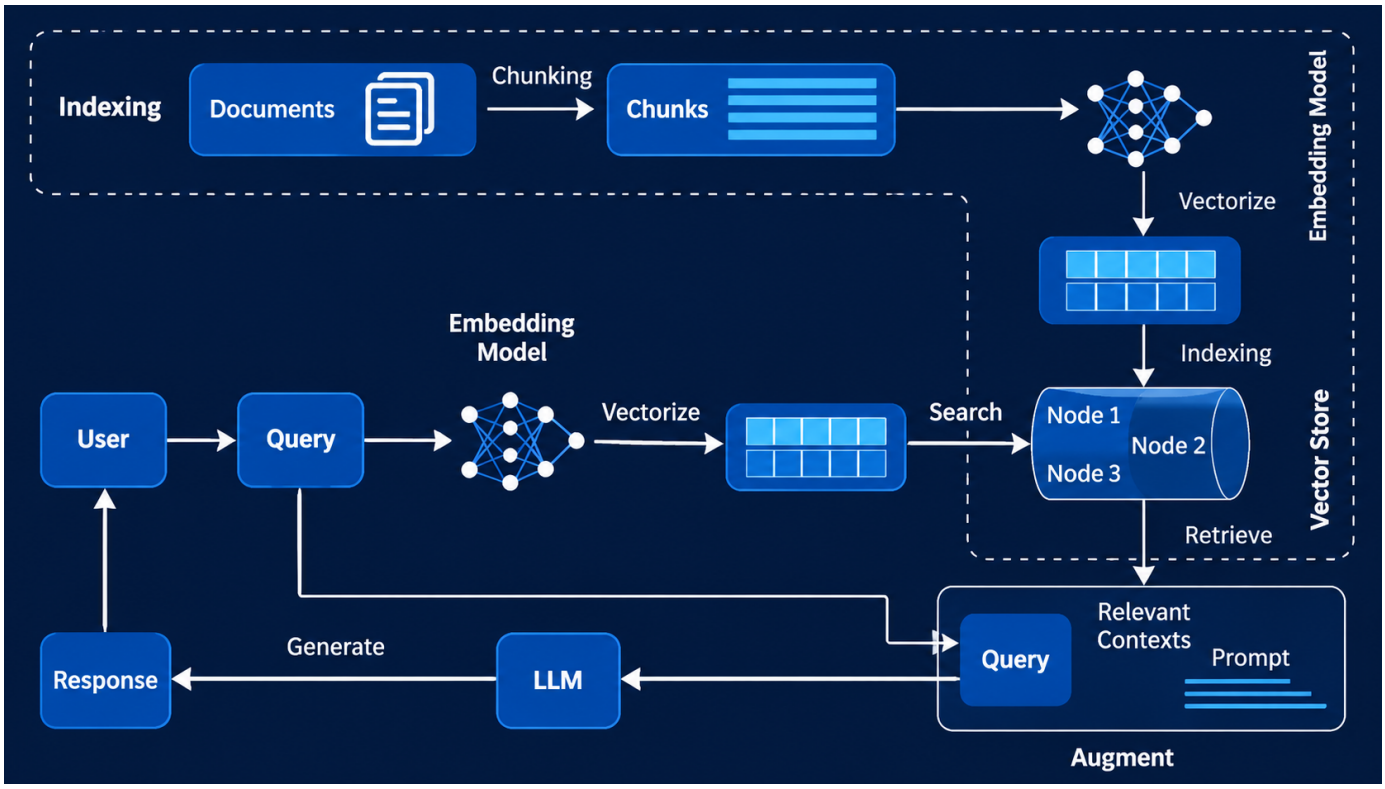

Figure 2. An illustration of RAG pipelines Architecture.

converts the user's query into an embedding using the identical model and methodology as in the indexing phase, guaranteeing uniformity. It then conducts a search for similar embeddings in the vector database and selects the most pertinent chunks. These selected chunks provide the necessary context for the RAG system to address and anchor its answer to the query.

Once relevant data has been identified in the vector database, the generation stage proceeds. At this point, the retrieved contextual chunks are combined with the original user query and passed along to the language model for crafting a response. This method provides the language model with access to specific and relevant information, resulting in more precise and accurate answers than generating responses without aid. The process concludes when the RAG system returns its generated reply to the user.

4 Experimental Results and Analysis

The findings strongly support the claim that response quality and context-awareness are notably enhanced by the information retrieval approach used in RAG systems. As compared to a random baseline system, the proposed vector store retriever reveals measurable improvements in all assessed metrics, with clear advances in context precision, recall, and faithfulness. Table 2 presents a comparison of the performance of the random baseline with the vector store retriever using the metrics context precision, context recall, faithfulness, and answer relevancy.

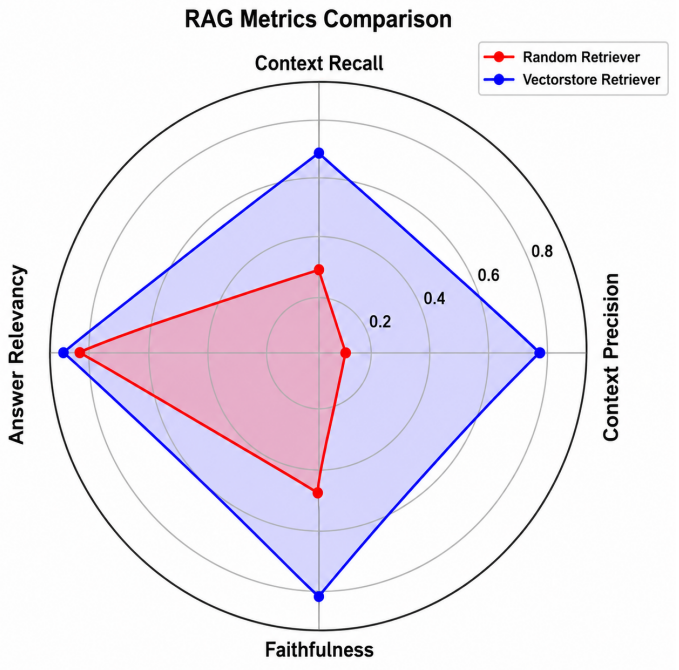

Figure 3. Random Baseline vs Vector store.

Figure 3 shows a slight but meaningful improvement in answer relevancy and a substantial improvement in other performance metrics. The system's response

Table 2. Performance metrics.

Metric	Random Baseline	Vector Store Retriever	Improvement
Context Precision	0.1	0.75	7.5 times
Context Recall	0.31	0.668	2.15 times
Faithfulness	0.4416	0.7647	+73.20 %
Answer Relevancy	0.8275	0.8852	+6.97 %

quality is judged using faithfulness and answer relevance. The baseline faithfulness score of 0.44 means that only 44% of the initial responses were accurate, while 56% contained factual inaccuracies and hallucinations. After introducing the vector store retriever, the faithfulness metric increased by 73.20%, demonstrating a significant reduction in factual errors and hallucinations in the generated responses.

The results indicate that answer relevance, as measured by the relevance score, improved by 6.97%, increasing from an initial score of 0.8275 to 0.8852. This suggests that using RAG led to slightly more accurate and focused answers that better matched user queries. While this improvement is small compared to other metrics, it still represents a meaningful improvement, especially considering that the baseline was already performing quite well due to the underlying model's capabilities. This finding addresses the research question and demonstrates that relevant context enables the RAG system to provide higher-quality responses.

The baseline model's context precision is 0.1, meaning that only 10% of the retrieved documents are actually relevant to the query, while 90% are irrelevant. In comparison, context precision increases to 0.75 (a 7.5-fold improvement) when the vector store retriever is used, meaning that three-quarters of the retrieved documents are relevant to the query. This substantial performance gain confirms that the proposed system enhances contextual awareness and directly answers the research question.

Context recall is also improved by 2.15 times to 0.668, which shows the system can successfully recall two-thirds of relevant documents during the phase of response generation. The results show that further improvement is needed, as approximately one-third of the relevant documents remain unretrieved by the system.

5 Conclusion

While LLMs have improved conversational skills compared to rule-based chatbots, they also introduce new challenges, such as hallucinations that produce factually incorrect responses and issues with losing context due to inadequate integration with external knowledge sources beyond their training data. To address these problems, we developed a RAG-based application that uses Neural Information Retrieval (Neural IR) to supply relevant context to LLMs during runtime. To create a knowledge base, raw documents are converted into embeddings that capture the semantic meaning of the data and allow for similarity-based comparison to find documents relevant to a query. These documents are then used as references for the LLM to generate responses containing accurate and pertinent information. The system is tested using the LLM-as-a-Judge method, which employs a separate LLM to fairly evaluate the responses and assess them through quantitative metrics. The results from performance testing clearly demonstrate that knowledge augmentation enhances the contextual understanding and accuracy of an LLM.

This paper contributes to AI research by demonstrating the effectiveness of RAG in addressing hallucinations and the retrieval of relevant information. The substantial improvement in context precision confirms the utility of vector embeddings in efficient Neural IR. As a result of the proposed methodology, faithfulness was considerably improved, indicating the ability of the RAG systems to reduce factual errors.

The current research maintains some limitations and further enhancements are required, as the improvement in answer relevancy at the retrieval stage is insufficient. The project evaluation was conducted using a domain-specific dataset, which may reduce the generalizability of the RAG system. The model performed well compared to the baseline model, as faithfulness and context recall improved; still, the model gave the incorrect answer occasionally. The

model was evaluated using an automatic evaluation matrix without human intervention because of the limited time.

Future research could explore the use of advanced retrieval methods, such as knowledge graphs and hybrid search with reranking, to enhance performance. Further improvements may be achieved through sophisticated prompt engineering and instruction tuning techniques. Additionally, integrating automated testing with a structured human evaluation framework presents a valuable direction for more robust and reliable assessment.

Data Availability Statement

The dataset used in this study is the Bitext Customer Support LLM Chatbot Training Dataset, an open-access resource publicly available through the Hugging Face repository. The dataset can be accessed at: https://huggingface.co/datasets/bitext/Bitext-customer-support-llm-chatbot-training-dataset. No additional data were generated or analysed during this study beyond those derived from the aforementioned dataset.

Funding

This work was supported by the £10K OfS AI Scholarship granted to Rabia Shabbir to do MSc Applied AI and Data Science at Southampton Solent University from Sep 2024 to Mar 2026.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Li, C. H., Chen, K., & Chang, Y. J. (2019, October). When there is no progress with a task-oriented chatbot: A conversation analysis. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1-6). [CrossRef]

[2] Mitsuda, K., Higashinaka, R., Li, T., & Yoshida, S. (2022, May). Investigating person-specific errors in chat-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 464-469). [CrossRef]

[3] Reshmi, S., & Balakrishnan, K. (2018). Empowering chatbots with business intelligence by big data integration. *International Journal of Advanced Research in Computer Science, 9*, 627-631. [CrossRef]

[4] Weizenbaum, J. (1983). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM, 26*(1), 23-28. [CrossRef]

[5] Colby, K. M., Weber, S., & Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence, 2*(1), 1-25. [CrossRef]

[6] Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing, 4*(2), 100211. [CrossRef]

[7] Shareef, F. (2024, October). Enhancing conversational AI with LLMs for customer support automation. In *2024 2nd international conference on self sustainable artificial intelligence systems (ICSSAS)* (pp. 239-244). IEEE. [CrossRef]

[8] Gokul, A. (2023). LLMs and AI: Understanding its reach and impact. *Preprints*. [CrossRef]

[9] Lovtsov, V. A., & Skvortsova, M. A. (2025, April). Automated mobile operator customer service using large language models combined with RAG system. In *2025 7th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE)* (pp. 1-6). IEEE. [CrossRef]

[10] Tsai, H. C., Jhang, J. W., & Wang, J. F. (2024, December). Constructing a Shopping Mall Customer Service Center Robot Based on the LLAMA-7B Language Model. In *2024 International Conference on Orange Technology (ICOT)* (pp. 1-4). IEEE. [CrossRef]

[11] Praneeth, B., Nattem, E. C., Jetti, K., Kavyashree, B. K., Rakshitha, D., Kumar, P. R., & Sreelakshmi, K. (2025). Optimization of customer feedback summarization using large language models (llm) and advanced retrieval-augmented generation. *IEEE Access, 13*, 124319-124332. [CrossRef]

[12] Mohandoss, R. (2024, June). Context-based semantic caching for llm applications. In *2024 IEEE Conference on Artificial Intelligence (CAI)* (pp. 371-376). IEEE. [CrossRef]

[13] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. [CrossRef]

[14] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM computing surveys, 55*(12), 1-38. [CrossRef]

[15] Zheng, Z., Liao, L., Deng, Y., & Nie, L. (2023). Building

emotional support chatbots in the era of llms. *arXiv preprint arXiv:2308.11584*. [CrossRef]

[16] Topsakal, O., & Akinci, T. C. (2023, July). Creating large language model applications utilizing langchain: A primer on developing llm apps fast. In *International conference on applied engineering and natural sciences* (Vol. 1, No. 1, pp. 1050-1056). [CrossRef]

[17] 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.

[18] Pandya, K., & Holia, M. (2023). Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations. *arXiv preprint arXiv:2310.05421*. [CrossRef]

[19] Chen, J., Lin, H., Han, X., & Sun, L. (2024, March). Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 38, No. 16, pp. 17754-17762). [CrossRef]

[20] Es, S., James, J., Anke, L. E., & Schockaert, S. (2024, March). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations* (pp. 150-158). [CrossRef]

[21] Kim, B., Kim, H., Lee, S. W., Lee, G., Kwak, D., Hyeon, J. D., ... & Sung, N. (2021, November). What changes can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pretrained transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3405-3424). [CrossRef]

[22] Zhang, C., Datla, V. V., Shrivastava, A., Samuel, A., Huang, Z., Kumar, A., & Liu, D. (2025, January). An automatic method to estimate correctness of RAG. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track* (pp. 603-611). https://aclanthology.org/2025.coling-industry.52/

[23] Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024, August). Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data* (pp. 102-120). Singapore: Springer Nature Singapore. [CrossRef]

[24] Dou, C., Jin, Z., Jiao, W., Zhao, H., Zhao, Y., & Tao, Z. (2023, December). Plugmed: Improving specificity in patient-centered medical dialogue generation using in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 5050-5066). [CrossRef]

[25] Siriwardhana, S., Weerasekera, R., Wen, E., Kaluarachchi, T., Rana, R., & Nanayakkara, S. (2023). Improving the domain adaptation of retrieval augmented generation (RAG) models for open domain question answering. *Transactions of the Association for Computational Linguistics, 11*, 1-17. [CrossRef]

[26] Liu, S., Zheng, C., Demasi, O., Sabour, S., Li, Y., Yu, Z., ... & Huang, M. (2021, August). Towards emotional support dialog systems. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 3469-3483). [CrossRef]

[27] Adiwardana, D., Luong, M. T., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., ... & Le, Q. V. (2020). Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*. [CrossRef]

[28] Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., ... & Weston, J. (2021, April). Recipes for building an open-domain chatbot. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume* (pp. 300-325). [CrossRef]

[29] Kasahara, T., Kawahara, D., Tung, N., Li, S., Shinzato, K., & Sato, T. (2022, July). Building a personalized dialogue system with prompt-tuning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies: student research workshop* (pp. 96-105). [CrossRef]

[30] Ngai, E. W., Lee, M. C., Luo, M., Chan, P. S., & Liang, T. (2021). An intelligent knowledge-based chatbot for customer service. *Electronic Commerce Research and Applications, 50*, 101098. [CrossRef]

[31] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*. [CrossRef]

[32] Reimers, N., & Gurevych, I. (2019, November). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3982-3992). [CrossRef]

Rabia Shabbir earned her Master's degree in Applied Artificial Intelligence and Data Science from Southampton Solent University, Southampton, United Kingdom. Rabia earned her Master of Computer Science degree from the Riphah International University. Rabia's areas of interest are large language models (LLMs), biomedical data analytics, speech recognition, object detection, and computer vision. She is actively engaged in research exploring the application of AI in real-world problem-solving, with a particular focus on interdisciplinary domains. Her work also emphasises the ethical and responsible use of artificial intelligence technologies in diverse societal contexts. (Email: Rabiabaig1997@gmail.com)

Kashif Talpur (Professional member ACM, BCS) received the B.S. degree in information technology from the Mehran University of Engineering and Technology, Pakistan. Kashif has extensive industry experience as Software Engineer in Pakistan. Kashif earned his M.S. and PhD. degrees in information technology from the Universiti Tun Hussein Onn Malaysia, Malaysia, in 2015 and 2018, respectively. Kashif also served at the University of Electronic Science and Technology of China, China, as Postdoctoral Researcher. His major research interests include machine learning, optimization, fuzzy neural networks, and metaheuristic algorithms. (Email: kashif.talpur@roehampton.ac.uk)

Shakeel Ahmad (Senior Member IEEE) is an Associate Professor of Computing at Southampton Solent University. Shakeel holds a PhD (Dr.-Ing.) from the University of Konstanz, Germany, and his research interests include Artificial Intelligence, Data Science, Multimedia Communication, and Computer Networks. Shakeel has been actively involved in projects funded by OfS (Office for Students) UK focusing Artificial Intelligence and Data Skills education. Shakeel has extensive experience in designing and leading Applied AI and Data Science curricula for both higher education. (Email: shakeel.ahmad@solent.ac.uk)