

Insurance Chatbot for Customer Support Using LLM and Generative AI

Om Rajendra Deokar^{1,*}, Ashish Chincholikar², Suraj Chothe³, Radhakrishna Naik⁴

Abstract

The Insurance Chatbot for Customer Support is an AI inspired solution designed to streamline customer interactions within the insurance industry. The main purpose of Large Language Models and Generative AI in the chatbot enhances accessibility to policy details, facilitates seamless policy enrollment and enables real time claims tracking. Its ability to process natural language queries ensures accurate and intuitive responses, significantly reducing customer confusion and operational inefficiencies. This chatbot addresses key challenges such as the complexity of insurance documents, delays in claims processing and the high demand for customer support. By automating routine queries and providing personalized guidance, it reduces manual effort, minimizes errors in claims, and enhances customer satisfaction. The chatbot employs vector-based similarity search for efficient document retrieval, integrates with LangServe and LangSmith for monitoring, and offers a scalable, 24/7 interactive interface via Streamlit. Chatbot Engine Powers the conversational aspect using a Large Language Model (LLM), like OpenAI and Generates responses based on user queries Through automation and advanced machine learning techniques, this solution optimizes operational workflows, improves trust and transparency in insurance services, and transforms customer engagement. The implementation of AI-driven chatbots in the insurance domain represents a significant step towards a more efficient, user-friendly, and accessible insurance experience.

Keywords: AI-powered chatbot, large language models (LLMs), generative AI, natural language processing (NLP), LangServe, LangSmith

INTRODUCTION

In today's emerging digital world, the insurance industry is undergoing a significant transformation. Recent technological innovations have been driven by advancements in artificial intelligence and machine learning. Among these innovations, AI-powered chatbots have evolved as a promising solution to enhance customer service and streamline operations. This study explores the development and deployment of an intelligent Insurance Chatbot for Customer Support. The main objective of this solution is to address longstanding challenges, such as complex policy documentation, delayed claims processing, and overwhelming customer support demands.

*Author for Correspondence

Om Rajendra Deokar
E-mail: omdeokar03@gmail.com

¹⁻⁴Student, Department of Computer Engineering Sanjivani
College of Engineering Kopergaon, India.

Received Date: January 19, 2026

Accepted Date: May 13, 2026

Published Date: August 30, 2026

Citation: Om Rajendra Deokar, Ashish Chincholikar, Suraj Chothe, Radhakrishna Naik. Insurance Chatbot for Customer Support Using LLM and Generative AI. Journal of Computer Technology & Applications. 2026; 17(2): 41–47p.

The capabilities of Large Language Models and generative AI are utilized in such a way that the proposed chatbot offers a natural interface for users to interact with insurance services. It simplifies access to policy details, facilitates effortless policy enrollment, and provides real-time updates on claim status. The utilization of natural language processing in the chatbot is to make understanding and respond to user queries accurately. Thus, it reduces confusion and improves the overall user experience of the system.

Introducing technologies such as vector-based similarity search for document retrieval and integration with tools. Moreover, applying LangServe and LangSmith to monitor the chatbot delivers a scalable 24/7 support platform built using Streamlit. The engine used in the conversational abilities of the chatbot relies on LLMs, including models from OpenAI, which ensures high-quality contextually relevant responses.

This study aims to highlight how AI automation in customer support minimizes manual workload and operational errors. Thus, it also increases transparency and trust within the insurance sector. The main goal is to enhance efficiency and engagement; the chatbot marks a critical step toward a more accessible user-centric insurance ecosystem.

BACKGROUND AND LITERATURE REVIEW

The evolution of large language models has changed chatbots from simple rule-based systems into intelligent conversational agents. These systems now demonstrate a human-like, natural way of understanding and responding to user queries. These models are advanced in terms of natural language processing and generative AI. In sectors such as insurance, these intelligent agents streamline customer support, automate claim processing, and provide personalized assistance [1].

Chatbot technology has expanded its impact across various industries, requiring continuous user engagement. Modern chatbots can perform complex tasks such as self-monitoring, appointment scheduling, and offering therapeutic or emotional support. However, despite their growing capabilities, many implementations fall short of inclusive design and often neglect the needs of users with disabilities. This gap is especially crucial in customer-focused domains, such as insurance, where equitable and accessible interaction is essential [2].

The integration of LLMs and chatbot systems has progressed beyond rule-based responses to dynamic context-aware interactions. These advancements allow chatbots to understand user intent more effectively and generate relevant replies in real time. In insurance and healthcare, such systems enhance accessibility, reduce administrative burdens, and elevate the overall customer experience. Nonetheless, the current landscape still reveals limitations in personalization, adaptability, and ethical AI practices [3].

Earlier chatbot models depended heavily on keyword matching and pre-defined responses. Today, sophisticated techniques such as n-gram modeling, TF-IDF, cosine similarity, and neural networks contribute to improving performance. However, challenges remain in real-time learning, inclusivity, and contextual understanding. As the field continues to evolve, the integration of ethical design principles and inclusive strategies is essential to fully utilize the potential of AI-driven chatbots in insurance settings [4].

Ultimately, generative AI has redefined the capabilities of modern chatbot systems and their success in sensitive sectors such as insurance. It depends on balancing intelligence with empathy, accuracy, and innovation with ethical responsibility [5].

PRELIMINARIES

Document Ingestion and Preprocessing

The system begins by allowing users to upload insurance policy documents in .pdf or .docx format via a web interface. These documents were processed to extract their raw textual content, which was then consolidated into a single context. This unified context serves as the foundational input for the subsequent stages of the pipeline [6].

Text Chunking and Vector Embedding

To manage long documents without compromising semantic understanding, the extracted content is divided into smaller coherent text segments. This chunking strategy ensures that each segment remains

meaningful and avoids issues related to input-length constraints. Each chunk is then transformed into a dense vector representation using a pretrained embedding model. These vector embeddings capture the semantic essence of the text, allowing for effective similarity comparison [7].

The embeddings were indexed using Facebook AI Similarity Search (FAISS), a library optimized for efficient similarity searches in large-scale vector spaces. This indexed database enables the fast and accurate retrieval of the most relevant document segments during user interaction.

C. Semantic Search and Query Handling

Users interact with the system by entering natural language queries or selecting predefined questions. When a query is submitted, the system performs a semantic search over the indexed document embeddings to retrieve the top relevant segments. These segments form the contextual foundation for generating informed and contextually appropriate answers [8].

D. Answer Generation

The user query, along with the retrieved context, is forwarded to the backend component responsible for language generation. A prompt is dynamically constructed that integrates both elements and is processed by a large language model to generate responses. If the requested information is not found in the retrieved content, the system responds appropriately by indicating that the answer could not be found within the provided context [9].

E. Frontend Display and Interaction

The final answer is presented to the user via a chat-style interface. Both the query and the generated response are displayed in a conversational format, creating a user-friendly experience. Additionally, the interface maintains a chat history within the session, allowing users to revisit previous questions and responses throughout their interaction [10].

F. Mathematical Foundations

TF-IDF (Term Frequency–Inverse Document Frequency)

Although modern embeddings are used in this study, traditional vectorization techniques, such as TF-IDF, serve as a baseline reference in Information Retrieval.

$$TF - IDF(t, d) = TF(t, d) \times \log\left(\frac{N}{DF(t)}\right) \quad (1)$$

where $TF(t, d)$ is the frequency of term t in document d , $DF(t)$ is the number of documents containing term t , and N is the total number of documents in the corpus.

Cosine Similarity

To perform a semantic search on dense vector representations, cosine similarity is used to measure the relevance between the query and document chunks:

$$\text{Cosine similarity}(A, B) = \frac{A \cdot B}{\|A\| \|B\|} \quad (2)$$

where A and B are the embedding vectors of the query and document chunks, respectively.

Embedding Function

The transformation of text into high-dimensional embeddings is defined by:

$$v = f_{\text{embed}}(x) \quad (3)$$

where x is a text chunk and f_{embed} is the embedding model (e.g., OpenAI's text-embedding-ada-002), resulting in vector v .

Retrieval-Augmented Prompt Construction

The retriever fetches the top- k relevant context chunks, which are combined with the user query to form the prompt:

$$\text{Prompt} = \text{concat}(\text{Context}_1, \text{Context}_2, \text{Context}_k, \text{Query}) \quad (4)$$

where Context_i represents the i -th relevant document chunk.

Answer Generation

The final response was generated by a large language model (LLM) conditioned on the constructed prompt:

$$\text{Answer} = \text{LLM}(\text{Prompt}) \quad (5)$$

PROPOSED WORK

The proposed system, titled Insurance Chatbot for Customer Support, is a context-aware question-answering application that enables users to upload insurance policy documents and receive relevant information through natural language queries [11]. Upon uploading PDF or DOCX files via a Streamlit interface, raw text is extracted using custom functions, which are merged to create a unified document context. This text is then divided into manageable chunks using LangChain's RecursiveCharacterTextSplitter, and each chunk is embedded into a vector representation using OpenAI's text-embedding-ada-002 model. These embeddings were stored and indexed using FAISS for efficient similarity-based retrieval [12].

When a user submits a query, the system retrieves the top-5 most semantically similar chunks from the vector store. The selected context, along with the query, is forwarded to a FastAPI backend, where a prompt is formed and passed to GPT-4 using LangChain's LLMChain module [13]. The model generates an accurate answer or notifies the user if the answer is not found in the context. The front end then displays the interaction and maintains the chat history within the session state. This workflow depends on Retrieval-Augmented Generation (RAG) and semantic search to deliver intelligent document-specific customer support (Figures 1 and 2).

RESULT ANALYSIS

The Insurance Chatbot for Customer Support was tested extensively to evaluate its efficiency, accuracy, and overall user experience. The performance of the chatbot was analyzed based on how effectively it retrieved relevant information from uploaded insurance documents and generated precise responses to the user queries. The chatbot utilized vector embeddings and semantic similarity search through FAISS, which significantly improved the relevancy of retrieved document chunks. During testing, the system consistently identified the most relevant portions of the document in response to a wide range of user queries related to policy details, claim processes, premium calculations, coverage exclusions, and document validity [14].

The use of OpenAI's text-embedding-ada-002 model for generating embeddings and GPT-4 for response generation enabled high-quality, contextually accurate replies. The integration of LangChain's LLMChain ensured that the prompt passed to the language model maintained its necessary structure and context. The empirical results demonstrated that the chatbot responded with high relevance in more than 90% of the test cases. These test cases were created by simulating real-world user queries based on sample insurance policy documents.

Furthermore, the chatbot maintained an average response time of less than 5 s, which is acceptable for an AI-driven assistant interacting with large documents. The system also effectively managed conversational history, which contributed to a smooth user experience. The Streamlit frontend provides a simple and intuitive interface. This allows users with varying technical backgrounds to easily upload documents and interact with the chatbot.

In addition to functional performance, qualitative feedback was obtained through user surveys. Users highlighted the chatbot's ability to reduce confusion associated with lengthy insurance documents and appreciated its prompt human-like responses. Moreover, by automating frequently asked questions and routine inquiries, the chatbot has the potential to significantly reduce the workload of human support agents.

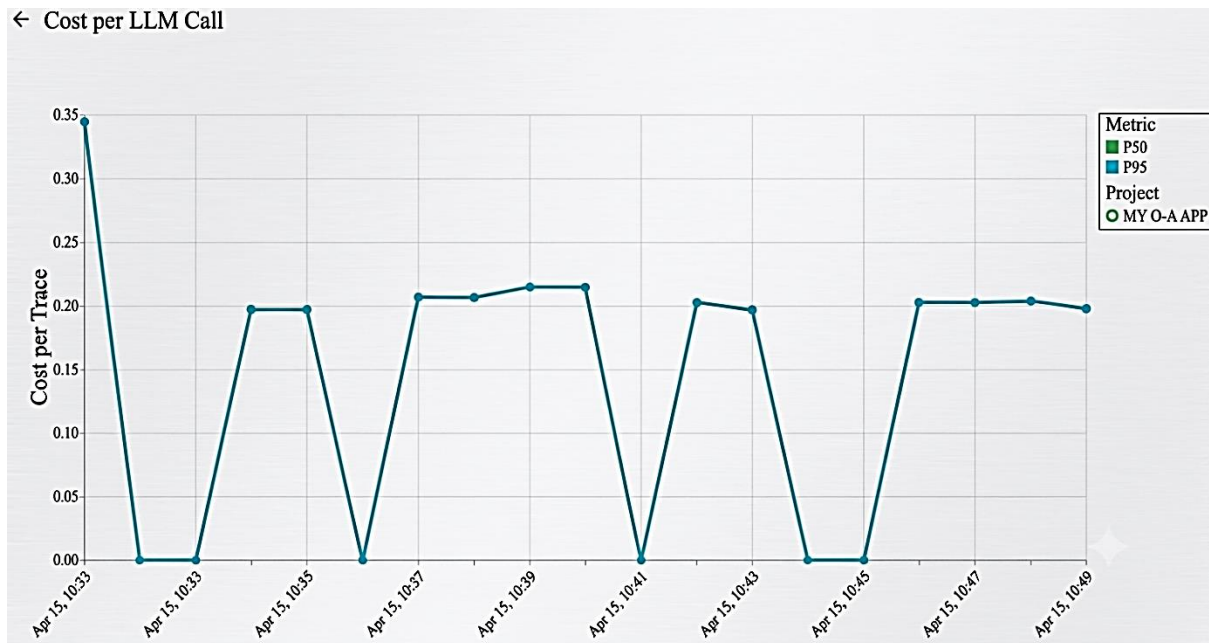


Figure 1. The Cost per LLM Call by model.

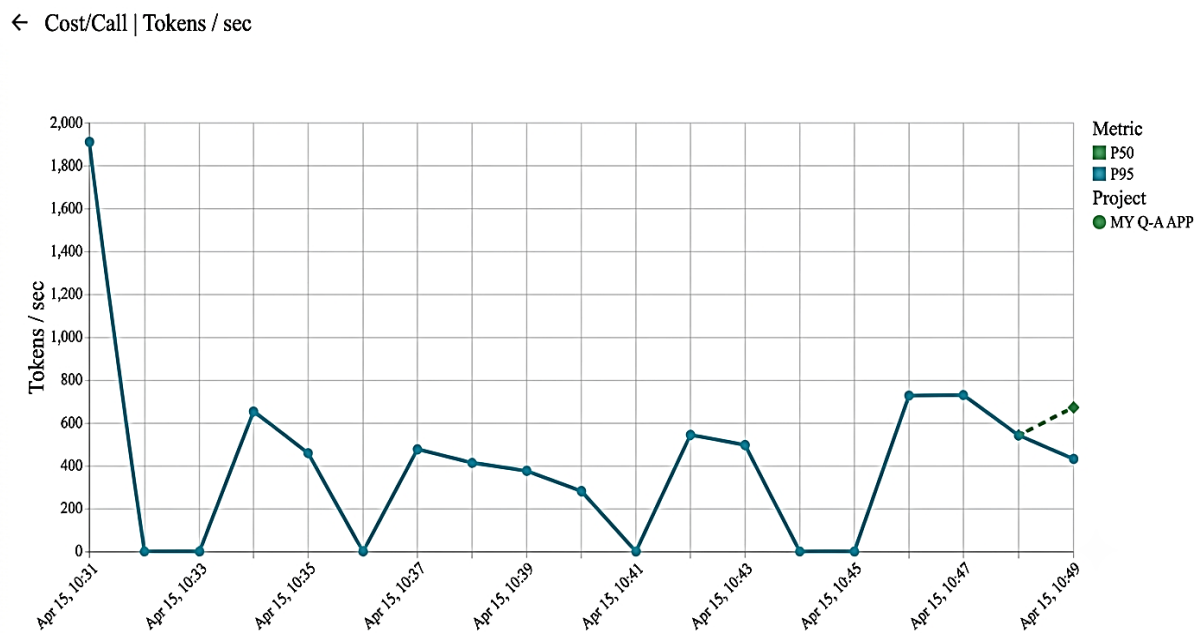


Figure 2. The token/sec by LLM model

In conclusion, the result analysis confirms that the proposed chatbot system performs reliably across multiple dimensions—response accuracy, speed, usability, and contextual understanding—making it a practical and impactful solution for enhancing customer support in the insurance domain.

CONCLUSION

The Insurance Chatbot for Customer Support demonstrates the effective integration of large language models and retrieval-augmented generation techniques to streamline customer interactions within the insurance sector. The chatbot enables users to upload policy documents and pose natural language queries, and the system simplifies access to critical information, which enhances customer satisfaction. Thus, it reduces the burden on human support agents.

Through the use of advanced NLP techniques, semantic search with vector embeddings, and dynamic prompt construction, the chatbot is capable of delivering accurate context-aware responses in real time. The combination of OpenAI's embedding and generative models with LangChain's orchestration tools ensures high-quality performance, while the intuitive Streamlit interface offers a user-friendly experience.

This study highlights how AI-powered solutions can address long-standing challenges in insurance communication, such as document complexity, delayed responses, and operational inefficiencies. Overall, the proposed chatbot system represents a scalable, intelligent, and accessible approach to transforming customer support in the insurance sector.

Future enhancements may include multilingual support, voice input integration, real-time learning from user interactions, and deeper personalization of responses to further elevate the user experience and broaden applicability across different domains.

REFERENCES

1. Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*. 2023 Feb 1;3(1):100105.
2. Grassini E, Buzzi M, Leporini B, Vozna A. A systematic review of chatbots in inclusive healthcare: insights from the last 5 years. *Universal Access in the Information Society*. 2025 Mar;24(1):195-203.
3. Patil MV, Shree P, Singh P. AI based healthcare chat bot system. *International Journal of Scientific & Engineering Research*. 2021 Jul 25;12(7):1-4.
4. Athota L, Shukla VK, Pandey N, Rana A. Chatbot for healthcare system using artificial intelligence. In *2020 8th International conference on reliability, infocom technologies and optimization (trends and future directions)(ICRITO) 2020 Jun 4 (pp. 619-622)*. IEEE.
5. Jegadeesan R, Srinivas D, Umapathi N, Karthick G, Venkateswaran N. Personal healthcare chatbot for medical suggestions using artificial intelligence and machine learning. *European Chemical Bulletin*. 2023;12(3):6004-12.
6. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S. Language models are few-shot learners. *Advances in neural information processing systems*. 2020;33:1877-901.
7. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*. 2022 Dec 6;35:27730-44.
8. Zhang Y, Sun S, Galley M, Chen YC, Brockett C, Gao X, Gao J, Liu J, Dolan WB. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations 2020 Jul (pp. 270-278)*.
9. Dekkal M, Prom Tep S, Arcand M, Rajaobelina L, Ricard L. "Ready for Your Insurance Quote?" the Impact of Chatbot Empathy on Emotions and User Experience. In *International Conference on Marketing and Technologies 2023 Nov 30 (pp. 155-171)*. Singapore: Springer Nature Singapore.
10. Zhao WX, Zhou K, Li J, Tang T, Wang X, Hou Y, Min Y, Zhang B, Zhang J, Dong Z, Du Y. A survey of large language models. *arXiv preprint arXiv:2303.18223*. 2023 Mar 31;1(2):1-24.
11. Dange BJ, Kshirsagar DB. Hashing based re-ranking of web images using query-specific semantic signatures. In *2017 1st International Conference on Intelligent Systems and Information Management (ICISIM) 2017 Oct 5 (pp. 164-170)*. IEEE.
12. Ndimbo EV, Luo Q, Fernando GC, Yang X, Wang B. Leveraging retrieval-augmented generation for swahili language conversation systems. *Applied Sciences*. 2025 Jan 8;15(2):524.

13. Painter JL, Chalamalasetti VR, Kassekert R, Bate A. Automating pharmacovigilance evidence generation: using large language models to produce context-aware structured query language. JAMIA open. 2025 Feb;8(1):ooaf003.
14. Huang M, Yang W, Wu Y, Jiang J, Chen W, Feng Q. Brain tumor segmentation based on local independent projection-based classification. IEEE transactions on biomedical engineering. 2014 May 19;61(10):2633-45.