Real-Time Vehicle Query Resolution Using Retrieval-Augmented Generation and Large Language Models

[1]Kanika Dogra

[1]Student
[1]Computer Science and Engineering,
[1]Parul Universiy, Vadodra, India
[1]kanika2005dogra@gmail.com

***Abstract*—This research explores the development of a next-generation RAG-based chatbot designed to handle vehicle-related queries with improved accuracy and efficiency. By integrating advanced NLP models, vector-based retrieval using Qdrant, and structured data management through PostgreSQL, the chatbot delivers real-time, context-aware responses. The system leverages Flask for seamless API integration, ensuring quick and reliable communication. The study demonstrates how combining retrieval and generation techniques enhances user satisfaction and streamlines customer support in the automotive sector**

***Index Terms*—RAG-based chatbot, AI-driven customer support, vehicle query resolution, Retrieval-Augmented Generation, natural language processing (NLP), intelligent virtual assistant.**

I. Introduction

The rapid advancement of artificial intelligence has revolutionized customer support systems across various industries, including the automotive sector. Traditional vehicle query resolution mechanisms, such as call centers and static FAQs, often fall short in delivering timely, personalized, and context-aware responses. These conventional methods suffer from limitations such as slow response times, lack of scalability, and difficulty in handling complex or multi-turn queries.

To address these challenges, Retrieval-Augmented Generation (RAG)-based chatbots have emerged as a powerful solution, combining the strengths of retrieval-based and generative AI models. By leveraging large language models (LLMs) and vector-based knowledge retrieval, these systems ensure more accurate, context-driven interactions. The proposed system integrates LLaMA and Cohere embeddings for enhanced natural language understanding, Qdrant DB for efficient vector similarity searches, and PostgreSQL for structured data management. This research delves into the architecture, implementation, and impact of an AI-powered chatbot designed to handle vehicle-related queries in real-time, providing a seamless and intelligent support experience for users.

The paper will explore the advantages of using RAG-based chatbots in vehicle service systems, compare them with traditional support models, and discuss their scalability and adaptability for future improvements. Through empirical analysis and case studies, we demonstrate the system's potential in revolutionizing the way automotive service providers manage customer interactions.

II. Background and Motivation for the Study

The increasing complexity of vehicle management and customer support in the automotive industry has necessitated the development of intelligent, automated solutions. Traditional support systems, such as call centers, email-based queries, and static FAQs, often fail to provide real-time, personalized assistance, leading to customer dissatisfaction and inefficiencies in service management. Additionally, as the demand for seamless, 24/7 support grows, businesses struggle to scale their support infrastructure without significantly increasing operational costs.

The advent of Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG) techniques presents a transformative opportunity for improving vehicle query resolution. By integrating AI-driven chatbots with vector-based retrieval systems, companies can ensure faster response times, higher accuracy, and enhanced user experience. The proposed chatbot leverages advanced NLP models, including LLaMA and Cohere embeddings, combined with Qdrant DB for vector-based search and PostgreSQL for structured data management. This hybrid approach enhances contextual understanding, enabling the chatbot to handle complex queries more effectively than conventional rule-based or retrieval-only systems.

The motivation behind this study stems from the need to bridge the gap between static, pre-defined FAQ systems and human-like conversational AI in the automotive sector. By analyzing and evaluating the effectiveness of a RAG-based chatbot for vehicle query resolution, this research aims to demonstrate its potential in optimizing customer interactions, improving service efficiency, and setting a new standard for AI-driven automotive support systems.

III. Objectives and Scope of the Research

The primary objective of this research is to design and implement an AI-driven chatbot that enhances vehicle query resolution by integrating advanced natural language processing and retrieval-augmented generation (RAG) techniques. The chatbot is developed to automate and optimize real-time customer interactions, improving efficiency and accuracy in the automotive support domain.

The key functionalities of the proposed system include:

- Context-Aware Query Resolution: The chatbot utilizes advanced NLP models, including LLaMA and Cohere embeddings, to provide accurate, context-aware responses to vehicle-related queries.
- Real-Time Information Retrieval: By leveraging Qdrant DB for vector-based search and PostgreSQL for structured data management, the chatbot ensures quick and relevant data retrieval.

- Seamless API Integration: A Flask-based REST API is implemented to facilitate smooth communication between the chatbot, databases, and external systems, ensuring efficient data flow.
- Scalability and Performance Optimization: The system is designed to handle increasing user demand while maintaining low latency and high response accuracy.

The scope of this research includes the development, deployment, and evaluation of the chatbot in a real-world automotive service setting. The study focuses on assessing improvements in response time, accuracy, and user satisfaction compared to traditional support methods. The research also explores the adaptability of the system to other domains that require AI-driven customer support solutions.

IV. Existing RAG based chatbot

Traditional chatbots in the automotive sector face several limitations:

- Limited Context Awareness: Many rely on rule-based logic, struggling with complex or nuanced queries.
- Lack of Real-Time Data Retrieval: Static FAQs limit their ability to provide dynamic, up-to-date responses.
- Poor Database Integration: Inefficient handling of structured (PostgreSQL) and vector (Qdrant DB) data reduces accuracy.
- Scalability Issues: Many struggle with high query volumes, leading to slow responses and system inefficiencies.
- Security Concerns: Weak encryption and insecure APIs put sensitive user and vehicle data at risk.
- Lack of Personalization: Without adaptive learning, these chatbots fail to improve responses based on user interactions.

These shortcomings highlight the need for an advanced RAG-based chatbot for real-time, intelligent query resolution in vehicle management.

V. Proposed Chatbot System

The proposed RAG-based chatbot enhances vehicle query resolution by integrating advanced AI, real-time data retrieval, and secure system architecture:

- Intelligent Query Processing: Combines retrieval (Qdrant DB) and generative models (LLaMA, Cohere) for context-aware responses.
- Real-Time Data Access: Ensures up-to-date information by integrating structured (PostgreSQL) and unstructured data sources.
- Enhanced User Experience: Provides a seamless, interactive web interface for accurate and instant support.

Secure & Scalable Infrastructure: Implements encryption and API security while leveraging scalable cloud-based deployment.

- Automated Learning & Optimization: Continuously improves responses through user feedback and adaptive AI models.

This system overcomes traditional chatbot limitations, delivering fast, intelligent, and personalized support for vehicle-related queries.

VI. Comparative Analysis of Research in AI-Based Chatbot Systems

Table 1 Comparative analysis

Title	Year	Benefits	Limitations
A Chatbot for Conflict Detection and Resolution	2019	Analyzes customer preferences for chatbots in non-complex, non-urgent queries; offers managerial insights.	Does not address complex or urgent customer queries effectively.
Enhancing Academic Administration with AI: Case Studies and Best Practices	2021	Explores AI integration in academic administration; presents case studies and best practices.	Challenges in AI integration, data privacy concerns, and lack of standardized implementation methods.
EQRbot: A Chatbot Delivering EQR Argument-Based Explanations	2023	Develops a chatbot that provides argument-based explanations to enhance user understanding.	Complexity of argumentation may not be suitable for all user interactions.
Analysis of Language-Model-Powered Chatbots for Query Resolution in PDF-Based Automotive Manuals	2023	Evaluates LLM-based chatbots for resolving queries in automotive manuals; provides insights into AI strategies.	Focuses on PDF-based manuals; may not apply to other document formats.
When Chatbots Struggle with Customer Queries: Can Humans Help Them?	2024	Analyzes customer preferences for chatbots in non-complex, non-urgent queries; offers managerial insights.	Does not address complex or urgent customer queries effectively.

VII. Overview of Document

This document outlines a research framework for developing an advanced chatbot system for real-time vehicle query resolution. It examines existing chatbot solutions, identifying key limitations such as lack of contextual awareness, inefficient data retrieval, and scalability issues. The proposed system leverages a RAG-based architecture, integrating NLP models, vector databases, and efficient data management techniques to enhance accuracy and response time. Additionally, the document includes a comprehensive literature review, comparing traditional rule-based chatbots with modern AI-driven approaches to highlight advancements in intelligent query handling.

VIII. Product Prospectives

Table 2 Product Prospective

Feature	Description
Vehicle Query Resolution	Provides real-time answers to user queries related to vehicle troubleshooting, maintenance schedules, and specifications using a RAG-based chatbot.
Service & Maintenance Details	Offers users access to service history, upcoming maintenance schedules, and personalized reminders for vehicle servicing.
Spare Parts & Availability	Helps users check the availability of spare parts, pricing, and compatibility with their vehicle model by integrating with a parts database..
Roadside Assistance & Emergency Support	Enables users to request immediate roadside assistance for breakdowns, towing services, and emergency contacts.
Personalized Recommendations	Uses AI-driven insights to recommend vehicle upgrades, maintenance tips, and cost-saving strategies based on user behavior and driving patterns.

IX. Literature Review

1. Existing Chatbot Systems and Their Limitations:

Traditional chatbot systems often rely on rule-based or retrieval-based methods, limiting their ability to handle complex or dynamic queries. Many chatbots lack contextual awareness, leading to repetitive or inaccurate responses. Additionally, they struggle with understanding domain-specific terminology, resulting in poor user satisfaction. Legacy chatbot architectures also have limited integration with real-time data sources, making them less effective in providing up-to-date information.

2. Emerging Technologies in Conversational AI:

Advancements in AI, particularly Large Language Models (LLMs) and Retrieval-Augmented Generation (RAG), have improved chatbot capabilities. These technologies enhance contextual understanding, allowing chatbots to generate more accurate and dynamic responses. Vector databases like Qdrant enable efficient information retrieval, while cloud-based deployment ensures scalability. Additionally, reinforcement learning from human feedback (RLHF) is being used to refine chatbot interactions based on user preferences.

3. Comparison of Traditional and AI-Powered Chatbots:

Modern AI-driven chatbots outperform traditional rule-based systems by offering natural language understanding, context retention, and real-time learning. While traditional chatbots follow pre-defined scripts and struggle with unstructured queries, AI-powered chatbots leverage retrieval and generation techniques for more accurate responses. However, challenges such as model bias, computational costs, and data privacy concerns remain key areas of research.

X. Acknowledgment

I would like to express my sincere gratitude to my college mentor, Ms. Meenakshi Prajapati, for her valuable guidance and support throughout this project. I am also thankful to my internship mentor, Mr. Jitesh Jani, for his insightful feedback and encouragement, which have been instrumental in the successful completion of this research. Their expertise and continuous support have greatly contributed to my learning and growth.

References

[1] Bender, E. M. (2024). 'Chatbots are like parrots, they repeat without understanding'. Le Monde.
[2] Jones, C. R., & Bergen, B. K. (2025). 'Does GPT-4 Pass the Turing Test?'. Proceedings of the NeurIPS AI Conference.
[3] DeepMind. (2022). 'Sparrow: Towards Safer Dialogue Agents'. DeepMind Blog.
[4] Vincent, J. (2022). 'DeepMind says its new AI coding engine is as good as an average human programmer'. The Verge.
[5] Gershgorn, D. (2017). 'Google's voice-generating AI is now indistinguishable from humans'. Quartz.