International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

SMART RAG VEHICLE ASSISTANT

Mrs. Kamalaveni V, Aarthi S R, Logupriya A, Divya J

Department of Artificial Intelligence and Data Science
Sri Shakthi Institute of Engineering and Technology

ABSTRACT:

This paper presents a Smart RAG Vehicle Assistant based on Retrieval-Augmented Generation (RAG) to provide intelligent and real-time support for drivers. The system is designed by uploading vehicle manuals into a knowledge base, enabling accurate and domain-specific information retrieval. Users can interact with the system through both text and voice input. When a query is given, the system retrieves relevant content from the stored manual using a vector database and processes it using the Gemini Large Language Model to generate a human-like response. The response is delivered to the user in both text and voice formats using a text-to-speech module, while speech recognition is used for voice-based input. The system also incorporates emergency response handling for critical situations. This approach improves response accuracy, enhances user interaction, and reduces driver distraction. The proposed system demonstrates efficient performance and provides a reliable solution for intelligent vehicle assistance.

Key Words: Smart Assistant, RAG, Gemini LLM, Vector Database, Voice Interaction, Vehicle System.

1.INTRODUCTION

Artificial Intelligence has significantly transformed modern systems by enabling intelligent and interactive solutions, especially in the automotive domain where smart assistants are used to enhance driver experience and safety. However, traditional vehicle assistants rely on predefined responses and limited datasets, making them less effective for handling complex and context-specific queries. To overcome this limitation, this paper proposes a Smart Vehicle Assistant based on Retrieval-Augmented Generation (RAG), where vehicle manuals are uploaded and converted into a structured knowledge base for efficient retrieval. When a user asks a question through voice or text, the system retrieves relevant information from the stored manual and processes it using the Gemini Large Language Model to generate a clear, accurate, and human-like response.

The assistant supports both voice and text interaction, allowing users to communicate naturally while driving and access information such as vehicle usage instructions, troubleshooting steps, and safety guidelines without manual searching. The integration of retrieval mechanisms with a generative model ensures context-aware responses, reduces driver distraction, and improves accessibility to information, thereby enhancing overall driving safety and user experience.

2. BODY OF PAPER

The Smart Vehicle Assistant is developed using a Retrieval-Augmented Generation (RAG) approach combined with a Large Language Model (LLM) to provide intelligent and context-aware responses. The system begins by creating a knowledge base from vehicle manuals, which are uploaded in Portable Document Format (PDF) and processed for efficient retrieval. The documents are divided into smaller segments and converted into embeddings, which are stored in a vector database to enable fast and accurate similarity search.

When a user provides a query through either voice or text, the system processes the input and forwards it to the retrieval module. For voice input, a speech recognition mechanism converts the spoken query into text before further processing. The retrieval module searches the vector database and extracts the most relevant content corresponding to the user query. This ensures that the response is based on actual vehicle documentation rather than predefined or generic answers.

The retrieved content is then provided as context to the generative model, where the Gemini Large Language Model (LLM) generates a clear and human-like response. The generated output is presented to the user in text format and is also converted into speech using a text-to-speech mechanism, enabling voice output. This dual interaction mode improves usability and allows users to interact with the system naturally while driving.

The system also incorporates basic safety handling by identifying critical queries and providing immediate responses when necessary. The use of a vector database improves retrieval efficiency, while the generative model enhances the quality and readability of responses. Compared to traditional assistants, the proposed system offers improved accuracy, better context understanding, and enhanced user interaction by combining document-based retrieval with intelligent response generation. This approach demonstrates the effectiveness of integrating retrieval-based techniques with generative models for real-world applications.

Table -1: System Comparison

Feature	Traditional Assistant	Smart RAG Assistant
Data Source	Static Data	Vehicle Manuals
Interaction Mode	Text Only	Voice and Text
Response Type	Predefined	Context-Aware
Accuracy	Moderate	High
Information Source	Limited	Dynamic Retrieval

2.1 SYSTEM ARCHITECTURE

The system architecture is designed based on a Retrieval-Augmented Generation approach. The process begins with uploading vehicle manuals, which are converted into smaller segments and stored in a vector database using embeddings. When a user provides a query, the retrieval module searches the database to find relevant information. This information is then passed to the Gemini Large Language Model, which generates a context-aware response. The final output is delivered to the

user in both text and voice formats, enabling effective interaction.

The architecture ensures efficient handling of large volumes of data by using a vector-based storage mechanism, which allows fast similarity search. The integration of retrieval and generation components enables the system to produce accurate and meaningful responses. This design improves scalability and makes the system adaptable to different vehicle models and document types without significant changes.

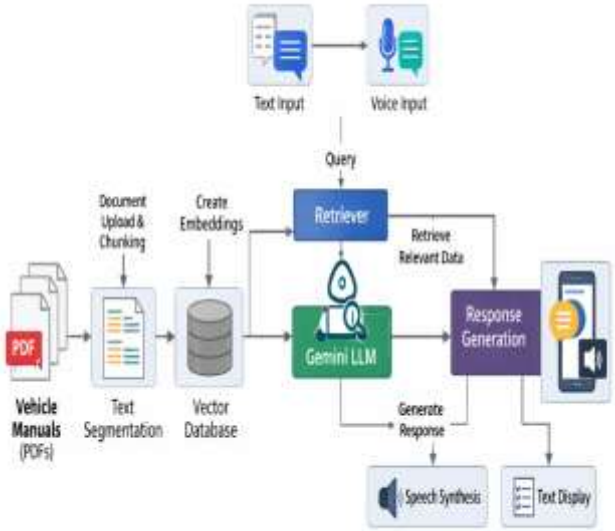

Fig -1: System Architecture

2.2 WORKING OF THE SYSTEM

The system operates by first converting vehicle manuals into a searchable format and storing them in a vector database. When a user submits a query through voice or text, the system processes the input and retrieves the most relevant information from the stored data. The retrieved content is then given to the Gemini Large Language Model, which generates a meaningful and human-like response. The response is displayed on the interface and also converted into speech for voice output, ensuring a smooth user experience.

The working process is designed to be efficient and user-friendly, allowing quick response generation with minimal delay. The retrieval mechanism ensures that only relevant information is used, which improves accuracy. The integration of generative models enhances the quality of responses, making them more understandable and conversational for the user. The working process is optimized to ensure fast and reliable response generation with minimal delay. By combining efficient retrieval with generative capabilities, the system ensures that only relevant and accurate information is used for answering user queries. This reduces unnecessary processing and improves overall system performance. The integration of real-time voice and text interaction further enhances usability, making the system practical for everyday driving scenarios and ensuring a smooth and responsive user experience.

2.3 VOICE INTERACTION MODULE

The voice interaction module enables users to communicate with the system using speech. A speech recognition component captures the user's voice input and converts it into text for processing. After the response is generated by the system, a text-to-speech module converts the output text into audio. This allows users to receive responses in voice format, making the system more accessible and convenient, especially while driving.

This module improves user experience by enabling hands-free interaction, which is essential for driver safety. It reduces the need for manual input and allows users to focus on driving while receiving assistance. The combination of speech recognition and voice output enhances accessibility and ensures that the system can be used effectively in real-time driving conditions.

The voice interaction module is optimized to provide accurate and responsive communication between the user and the system. By integrating efficient speech recognition with real-time text-to-speech conversion, the system ensures seamless interaction without noticeable delay. This enhances the practicality of the assistant in real-world driving conditions, where quick and reliable responses are essential. The module contributes significantly to reducing driver distraction and improving overall safety, making the system more efficient and user-friendly.

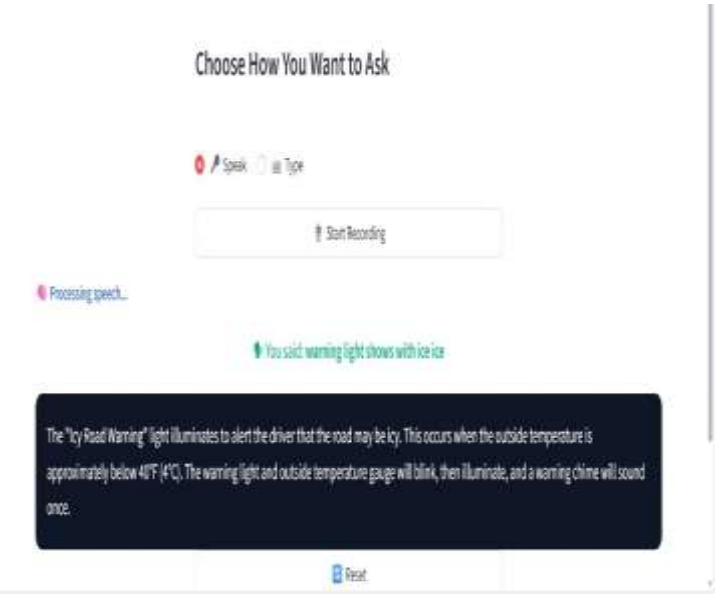

Fig -2: UI Page with output

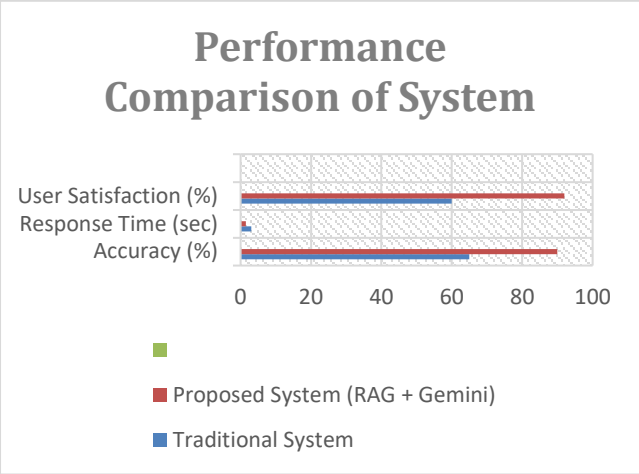

Fig-3: Performance chart

3. CONCLUSIONS

The Smart RAG-based Vehicle Assistant presents an effective approach for enhancing driver support by integrating retrieval-based techniques with generative AI models. The system enables users to interact through both voice and text, providing accurate and context-aware responses by retrieving information directly from vehicle manuals. The use of the Gemini Large Language Model improves the quality of responses by generating human-like and meaningful outputs.

The proposed system reduces the effort required to manually search for information and minimizes driver distraction by delivering quick and reliable assistance. The integration of voice input and output further enhances usability, making the system suitable for real-time driving scenarios. Overall, the system demonstrates improved performance in terms of accuracy, response time, and user interaction compared to traditional assistants.

This approach highlights the potential of combining vector-based retrieval with advanced language models for practical applications. The system can be further enhanced by integrating real-time data sources and extending support to multiple vehicle models, making it a scalable and efficient solution for intelligent vehicle assistance.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the institution and faculty members for their continuous guidance and support throughout the development of this project. The authors also thank their peers for their valuable suggestions and encouragement, which helped in successfully completing this work.

REFERENCES

1.Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS), 33, 9459–9474.

2.Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS), 30, 5998–6008.

3.Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS), 33, 1877–1901.

4.Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT 2019, 4171–4186.