IEEE Open Journal of
Vehicular Technology

Received 3 August 2024; accepted 19 August 2024. Date of publication 21 August 2024;
date of current version 10 September 2024. The review of this article was coordinated by Editor Darren Galpin.

Digital Object Identifier 10.1109/OJVT.2024.3447449

IDAS: Intelligent Driving Assistance System Using RAG

LUIS-BERNARDO HERNANDEZ-SALINAS[1], JUAN TERVEN[1] (Member, IEEE), E. A. CHAVEZ-URBIOLA[1] (Member, IEEE), DIANA-MARGARITA CÓRDOVA-ESPARZA[2] (Member, IEEE), JULIO-ALEJANDRO ROMERO-GONZÁLEZ[2] (Member, IEEE), AMADEO ARGUELLES[3] (Member, IEEE), AND ILSE CERVANTES[1] (Senior Member, IEEE)

[1]Instituto Politecnico Nacional, CICATA, Querétaro 76090, Mexico
[2]Facultad de Informatica, Universidad Autónoma de Queretaro, Queretaro 76230, Mexico
[3]Instituto Politecnico Nacional, Centro de Investigacion en Computacion, Mexico City 07738, Mexico

CORRESPONDING AUTHOR: JUAN TERVEN (e-mail: jrtervens@ipn.mx).

This work was supported in part by the Secretaría de Investigación y Posgrado-IPN, Grant PRORED “SIAC: Sistema Inteligente de Asistencia a la Conduccion”.

ABSTRACT In the fast-growing automotive technology sector, it has become increasingly clear that there is a need for cars with smarter and more interactive systems. This article presents the Intelligent Driving Assistance System (IDAS), an artificial intelligence system that enables the driver to use voice commands to access various features of a car. The primary component of IDAS is a Large Language Model (LLM), which, through retrieval augmented generation (RAG), can efficiently read and understand the car manual for immediate context-based aid. In addition, this system incorporates speech recognition and speech synthesis capabilities, it can understand commands given in multiple languages, improving user experiences among diverse driver communities. Our results show a minimum response time of one second for the pipeline using GPT-4o-mini and Mistral Nemo.

INDEX TERMS Artificial intelligence, human-computer interaction, intelligent agents, large language models, retrieval augmented generation (RAG).

I. INTRODUCTION

Intelligent agents in vehicle technology are systems that can make decisions and perform tasks without human intervention [1]. These are programmed to improve safety [2], [3], efficiency [4], and convenience in transportation [5]. The idea behind using intelligent agents in vehicles has been growing rapidly with the help of artificial intelligence (AI), machine learning (ML), and sensor technologies, among others. Research in artificial intelligence and robotics paved the way for the birth of intelligent vehicle agents. Initially, their focus was on creating basic autonomous functions such as cruise control [6] and anti-lock braking systems [7]. As time passed, these became more advanced, which led to the development of advanced driver assistance systems (ADAS) [8] and the quest to achieve the production of fully autonomous vehicles [9], [10].

Key advancements in electronics and sensor technology have allowed modern vehicles to be equipped with a variety of sensors, including cameras, radar, and LiDAR, which provide real-time data about the vehicle’s surroundings. At the same time, new machine learning algorithms allow vehicles to learn from data and improve their performance over time. This includes object detection [11], path planning [12], and decision making [13]. In addition to this, the evolution of connectivity technologies has allowed vehicles to connect to the Internet and other devices, allowing real-time communication and data exchange. This connectivity supports features such as traffic updates [14], remote diagnostics [15], and vehicle-to-vehicle (V2V) communication [16]. Finally, the exponential increase in computing power has enabled the processing of large amounts of data in real-time, which is crucial for the operation of intelligent agents. These advancements have opened up a world of possibilities, allowing the development of intelligent agents that can perform a broad spectrum of functions, from essential driver assistance to the quest for autonomous driving.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License. For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

In the field of conversational agents for vehicle technology, there are several gaps and limitations with which researchers and developers are currently grappling. For example, previous conversational agents often struggle with maintaining context during long interactions [17]. This can lead to misunderstandings or irrelevant responses, which is particularly problematic in a driving environment where fast and accurate communication is essential. Vehicles present a unique challenge as they require real-time multimodal interaction capabilities, including voice, touch, and possibly even gesture recognition. Integrating these different modes of communication seamlessly is still an ongoing challenge. Although some progress has been made, current systems often cannot adapt to individual user preferences and behaviors effectively [18]. Personalization can also improve the user experience and ensure that the conversational agent can provide relevant and timely assistance [19], [20]. Also, ensuring that conversational agents do not distract the driver and can operate reliably under various conditions (e.g., noisy environments, different accents) is a major challenge.

Despite these challenges, the possibilities of large language models (LLMs), such as GPT [21], in the automotive industry are promising. LLMs can significantly improve the natural language understanding capabilities, reducing misunderstandings and improving the overall user experience. With their ability to learn behaviors on the fly [22], LLMs can offer more advanced personalization, adapting their function to user preferences and diverse scenarios, including technologies that contribute to tailoring driving to user preferences and driving habits [23]. In addition, LLMs can be integrated with other AI technologies to support multimodal interactions, combining voice, touch, and gesture recognition to create a seamless user experience [24].

Using LLMs with Retrieval-Augmented Generation (RAG) [25] can considerably aid the automotive industry by providing accurate and contextual information about vehicles. It works by leveraging both a retrieval mechanism and a generative model, enabling it to access and utilize vast amounts of relevant data. This system reduces the likelihood of hallucinations, a common issue in which AI generates incorrect or non-sensical information [26]. In the automotive industry, RAG can help by providing precise vehicle specifications from authoritative sources, improving customer support through accurate maintenance procedures and vehicle-related information, improving vehicle diagnostics with data-driven insights based on historical maintenance records and manufacturer data.

In this paper, we introduce IDAS, an intelligent driving assistant that can communicate with car users and answer questions about the vehicle. IDAS incorporates detailed knowledge of the car that it accesses using RAG from the car user manual. Fig. 1 illustrates the overall system, where a car user, such as the driver, can ask a question using voice commands. The system then performs speech transcription, and via a conversational RAG pipeline, it can access specific information about the vehicle, such as the car's manual, to provide accurate responses. The responses are converted to audio and played to the user. The code for this project and a video demo can be found at https://github.com/jrterven/IDAS.

A. RELATED WORK

The development of intelligent systems in the automotive industry has greatly improved the safety and comfort of road travel. This section reviews related work on intelligent driving assistants, focusing on the integration of artificial intelligence (AI) and machine learning (ML) technologies to improve driver assistance systems.

Vochin et al. [27] discussed enhancing navigation systems by collecting traffic data from integrated car sensors and security systems to alert other users. Their work covered system concepts, architecture, preliminary design, performance evaluation, and task organization for a cost-effective embedded implementation using a real-time kernel. Complementing this research, Teran et al. [3] presented an intelligent driving assistant that takes advantage of vehicle telemetry and road accident risk map analysis to alert drivers about potential dangers, reducing the likelihood of traffic accidents through real-time audiovisual alerts. An intelligent assistance agent based on fuzzy reasoning supported drivers by integrating telemetry data, environmental conditions, and safety principles, with real-car tests demonstrating its effectiveness.

Dinesh et al. [2] proposed a virtual assistant that provides users with fundamental chores and information in a natural language. However, it could not perform complex tasks. Addressing this gap, Mathis et al. [28] focused on creating more adaptive and interactive voice assistants in vehicles. As vehicles became more automated, users shifted from being drivers to passengers. This transition brought about new opportunities for interactions with assistants. Their study delved into proactive use cases involving experts and users and, based on driving simulator results, found that use cases related to work and well-being were well received. This emphasized the importance of adaptability and non-intrusive communication.

Naz et al. [29] and Jishma et al. [30] both examined AI-powered machines and vehicle automation. Naz et al. detailed the use of onboard sensors such as laser, radar, lidar, GPS, and vehicular communication networks to gather environmental data, aiding in path planning and control techniques for autonomous driving. They presented a comprehensive review of state-of-the-art techniques that highlights the role of AI in-vehicle systems. Similarly, Jishma et al. explored various automation techniques focusing on cognitive cars that use machine learning algorithms. Their study involved lane detection, obstacle detection, road sign detection, and traffic light detection, suggesting that cognitive cars will become an asset for future generations due to their intelligent features and increased safety aspects.

On the other hand, Lee et al. [31] and Tanaka et al. [32] focused on intelligent driver assistance systems to improve

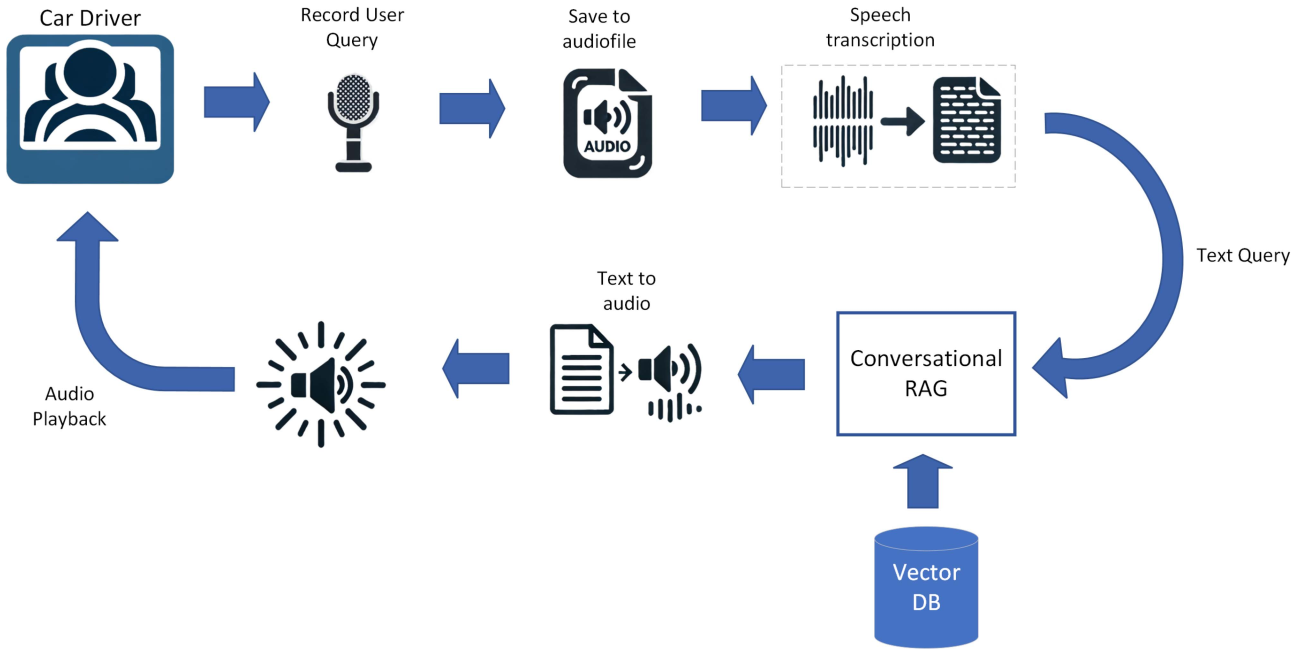


FIGURE 1. Intelligent driving assistance system (IDAS): The agent receives an audio command from the user, records and processes the audio, and converts it to text. It then retrieves context from the car manual via retrieval augmented generation and generates a text response using a large language model. Finally, it converts the text back to an audio response.

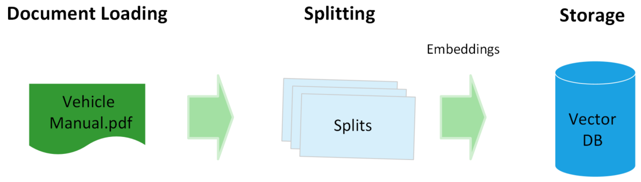


FIGURE 2. Generation of a vector database: The car user manual is loaded and split into multiple chunks. Each chunk of text is converted into a vector embedding and stored in a vector database.

situational awareness and address distracted driving. Lee et al. developed a system that reminds drivers to turn on headlights or wipers at night or on rainy days and alerts them when nearby vehicles are too close, using a lightweight deep learning model and a distance estimation method. Tanaka et al. proposed an enhanced system using YOLOv5m for object detection to detect distracted driving behaviors by observing hand movements, with high accuracy in detecting distractions.

The integration of emotional and situational awareness into driver assistance systems was explored by Lee Heung-Gu et al. [33] and Mahima & Ginige [34]. Lee Heung-Gu introduced a human-machine interface designed to enhance ADAS by incorporating driver emotional and situational awareness, employing a 1D convolutional neural network model using multimodal biosignals to control semiautonomous vehicles. Their system classifies driving scenarios and regulates vehicle speed accordingly, with experiments showing rapid response times and accurate recognition of scenarios and emotions. In a similar vein, Mahima & Ginige [34] focused on developing a new vehicle assistant that detects and responds to driver emotions, with the aim of reducing accidents caused by driver distraction.

Finally, Hsieh et al. [35] introduced an architecture for adaptive autonomous driving assistance that improves safety by understanding the trust of drivers in automation. Using sensory fusion models, their system provided voice reminders based on predicted driving status, suggesting that quantified trust can lead to safer and more appropriate use of ADAS.

Our work distinguishes itself by integrating Retrieval-Augmented Generation (RAG) into an intelligent driving assistant to enhance its ability to respond to user queries by referencing car manuals in real-time. Unlike previous systems that focus primarily on predefined tasks, situational recognition, or emotional awareness, our approach leverages the vast information contained within car manuals to provide precise and contextually relevant responses to user inquiries. This integration of RAG enables more dynamic assistance, addresses a wide range of user needs, and potentially improves overall driving safety and user satisfaction.

II. METHODS

To build the intelligent driving assistant system, we followed these steps:

1) Build a conversational agent using speech recognition, GPT-based intelligence, and text-to-speech to generate audio responses.

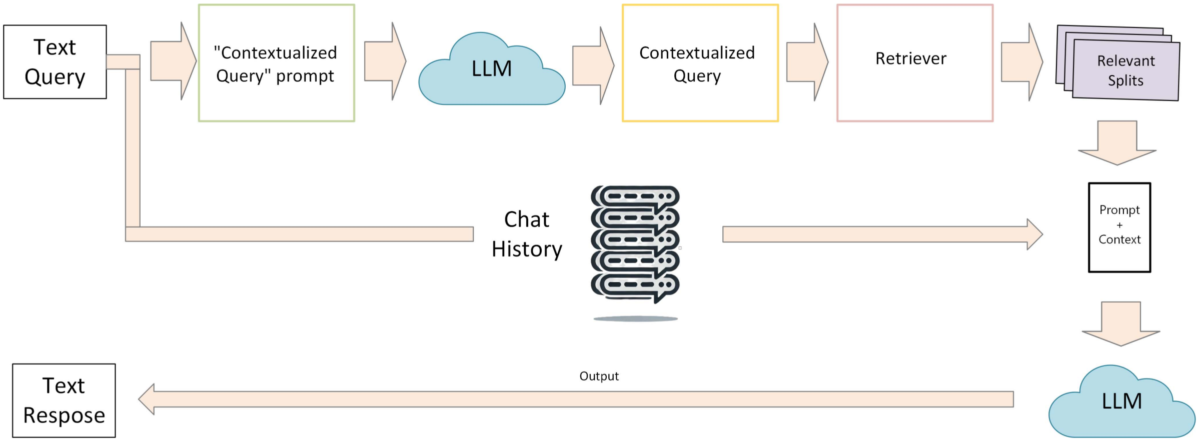


FIGURE 3. **The retrieval-augmented generation (RAG) pipeline: A user's query is combined with past chat history and passed through an LLM to generate a *contextualized query*, then this query is converted into an embedding, which is then used to search a vector database for relevant information. The retrieved information serves as context for a language model to generate an appropriate response.**

2) Develop an RAG pipeline to provide the agent with car-specific knowledge based on the car user's manual.

In the following subsections, we will delve into the details of these two component steps.

A. CONVERSATIONAL AGENT

We developed a conversational agent to provide real-time assistance to drivers. Fig. 1 shows a block diagram of the agent.

We developed the system in Python, integrating two external APIs. We used OpenAI Whisper [36] for speech-to-text conversion, GPT-4o [21] for natural language processing, and ElevenLabs [37] for text-to-speech conversion and audio streaming.

We composed a system prompt to define the agent's role and behavior. This prompt, as shown below, established the agent's tone and personality and included instructions for responding to vehicle-related questions. If the question is not about the vehicle, the system will respond from its comprehensive knowledge acquired during training.

> *You are a helpful assistant called IDAS who is always willing to answer the questions that are asked to you. If the question is about the vehicle, use the provided car manual information to answer the question at the end. If you don't know the answer even with the provided context, say, "I am sorry, I do not know the answer." Don't try to make up an answer. Respond in the most attentive way possible. Use a maximum of three sentences. Keep the answer as concise as possible.*
>
> *Context: {context}*
> *Question: {question}*
> *Helpful Answer:*

When the user speaks, the system records the audio, transcribes it using OpenAI's Whisper, and generates a response with a large language model. This response is subsequently converted to speech via the ElevenLabs API and played back to the user.

B. RAG PIPELINE FOR SPECIFIC CAR KNOWLEDGE

Retrieval Augmented Generation (RAG) is a technique that combines the strengths of information retrieval and natural language generation to improve the quality and relevance of generated text. It was introduced in 2020 by Lewis et al. at Facebook Artificial Intelligence Research (FAIR) [25]. The main idea is to enhance the generation process by retrieving relevant documents or pieces of information from a large corpus that cannot fit inside the working memory of a large language model. These relevant documents can then be used as context information for the LLM to inform and guide the generation of the final output.

RAG models can produce more accurate, informative, and contextually relevant responses compared to traditional generation models that rely solely on the input query without additional context.

The RAG process typically involves three main components:

1) *Vector database generation:* In this component, we create a database of the documents that later can be used to search for relevant information efficiently.
2) *Retriever:* This component is responsible for searching and retrieving relevant documents or passages from a vector database based on the input query.
3) *Generator:* Once the relevant information is retrieved, the generator component uses this information to produce a coherent and contextually appropriate response. The generator is a language model that can leverage the retrieved documents to generate high-quality text.

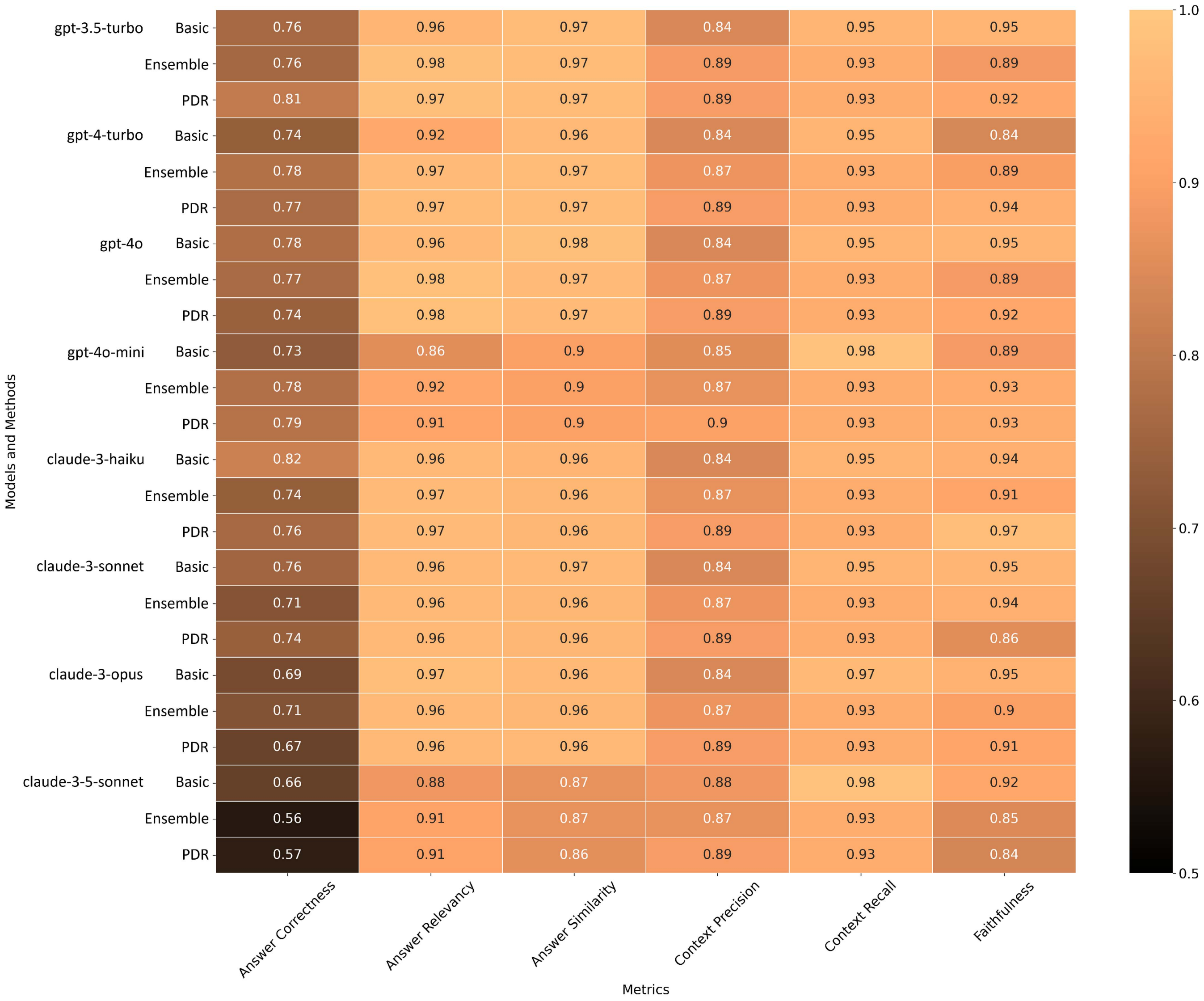


FIGURE 4. **Performance of the RAG pipeline using proprietary models evaluated on the automatically generated questions. We report the results of six metrics: answer correctness, answer relevancy, answer similarity, context precision, context recall, and faithfulness.**

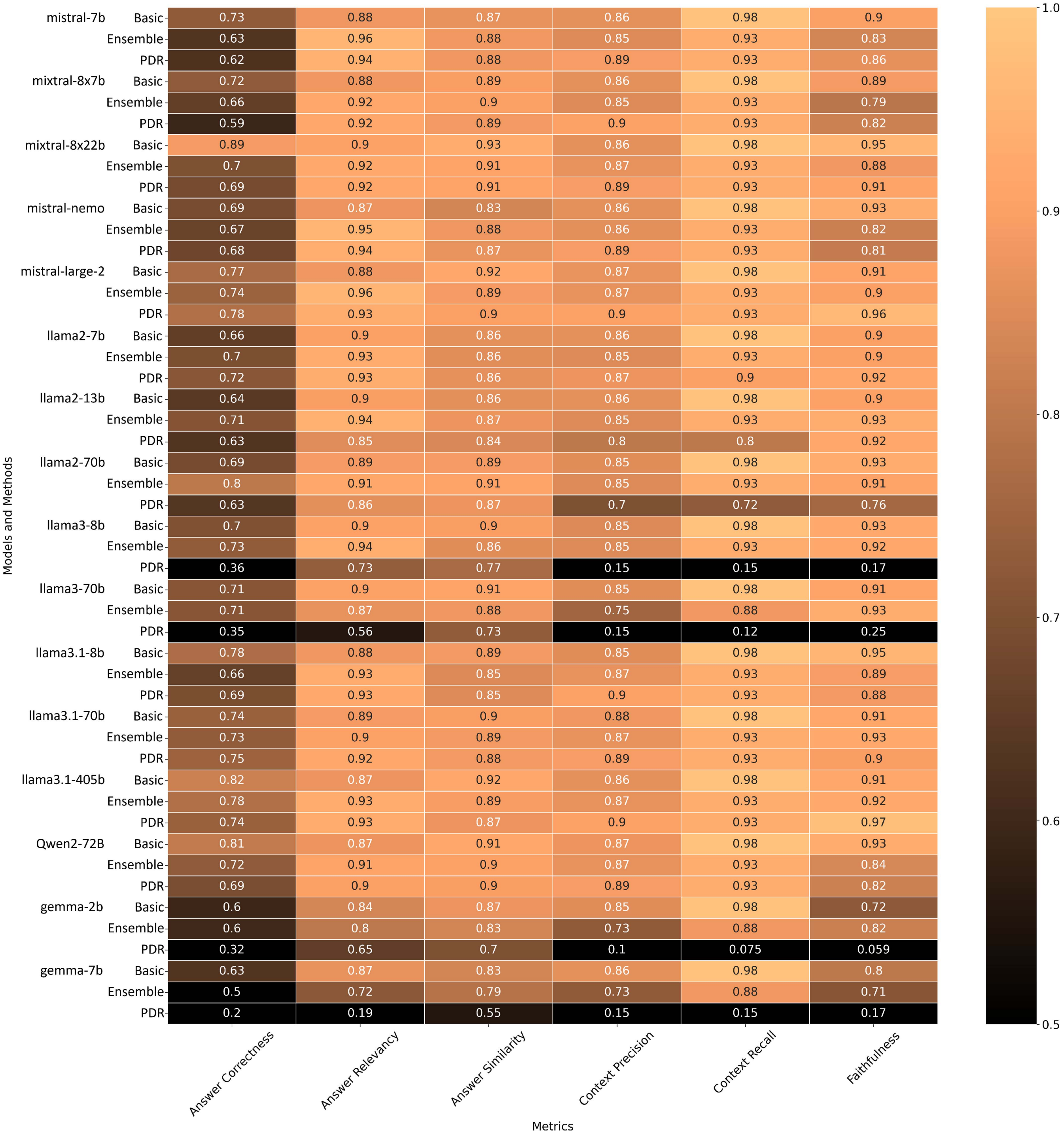
1) VECTOR DATABASE GENERATION

Every RAG pipeline starts with the generation of a vector database. In this step, which is usually performed for every new data corpus, the documents, in our case, the car user's manual, are stored in a vector database. Fig. 2 shows the process.

The documents are split into smaller chunks that, during retrieval, provide context information to the language model that can fit into its memory. To split the document, we use a recursive character text splitter from Langchain [40]. This splitter is defined by a series of characters. It attempts to divide based on these characters in sequence until the resulting parts are sufficiently small. The default series is `[”\n\n”, “\n”, “ ”, ””]`. This means that it tries to maintain the integrity of paragraphs, sentences, and words as much as possible, as these are generally the most closely related pieces of text in terms of meaning [41].

We opted for a chunk size of 2000 characters and an overlap of 150 characters, resulting in 243 splits for our test car manual. These splits are then transformed into embeddings using the OpenAI text embedding model *text-embedding-3-large*, which has a dimension of 3072. An embedding is essentially a set of floating-point numbers that characterize a piece of text. These vectors of numbers live in Euclidean space and are used to search for similar concepts based on their distances. Short distances indicate strong relatedness, while long distances indicate weak relatedness [42].

Vector embeddings are stored along with the corresponding text chunk in a vector database. For this, we use Chroma [43], an open source vector database.



Models and Methods		Answer Correctness	Answer Relevancy	Answer Similarity	Context Precision	Context Recall	Faithfulness
mistral-7b	Basic	0.73	0.88	0.87	0.86	0.98	0.9
	Ensemble	0.63	0.96	0.88	0.85	0.93	0.83
	PDR	0.62	0.94	0.88	0.89	0.93	0.86
mixtral-8x7b	Basic	0.72	0.88	0.89	0.86	0.98	0.89
	Ensemble	0.66	0.92	0.9	0.85	0.93	0.79
	PDR	0.59	0.92	0.89	0.9	0.93	0.82
mixtral-8x22b	Basic	0.89	0.9	0.93	0.86	0.98	0.95
	Ensemble	0.7	0.92	0.91	0.87	0.93	0.88
	PDR	0.69	0.92	0.91	0.89	0.93	0.91
mistral-nemo	Basic	0.69	0.87	0.83	0.86	0.98	0.93
	Ensemble	0.67	0.95	0.88	0.86	0.93	0.82
	PDR	0.68	0.94	0.87	0.89	0.93	0.81
mistral-large-2	Basic	0.77	0.88	0.92	0.87	0.98	0.91
	Ensemble	0.74	0.96	0.89	0.87	0.93	0.9
	PDR	0.78	0.93	0.9	0.9	0.93	0.96
llama2-7b	Basic	0.66	0.9	0.86	0.86	0.98	0.9
	Ensemble	0.7	0.93	0.86	0.85	0.93	0.9
	PDR	0.72	0.93	0.86	0.87	0.9	0.92
llama2-13b	Basic	0.64	0.9	0.86	0.86	0.98	0.9
	Ensemble	0.71	0.94	0.87	0.85	0.93	0.93
	PDR	0.63	0.85	0.84	0.8	0.8	0.92
llama2-70b	Basic	0.69	0.89	0.89	0.85	0.98	0.93
	Ensemble	0.8	0.91	0.91	0.85	0.93	0.91
	PDR	0.63	0.86	0.87	0.7	0.72	0.76
llama3-8b	Basic	0.7	0.9	0.9	0.85	0.98	0.93
	Ensemble	0.73	0.94	0.86	0.85	0.93	0.92
	PDR	0.36	0.73	0.77	0.15	0.15	0.17
llama3-70b	Basic	0.71	0.9	0.91	0.85	0.98	0.91
	Ensemble	0.71	0.87	0.88	0.75	0.88	0.93
	PDR	0.35	0.56	0.73	0.15	0.12	0.25
llama3.1-8b	Basic	0.78	0.88	0.89	0.85	0.98	0.95
	Ensemble	0.66	0.93	0.85	0.87	0.93	0.89
	PDR	0.69	0.93	0.85	0.9	0.93	0.88
llama3.1-70b	Basic	0.74	0.89	0.9	0.88	0.98	0.91
	Ensemble	0.73	0.9	0.89	0.87	0.93	0.93
	PDR	0.75	0.92	0.88	0.89	0.93	0.9
llama3.1-405b	Basic	0.82	0.87	0.92	0.86	0.98	0.91
	Ensemble	0.78	0.93	0.89	0.87	0.93	0.92
	PDR	0.74	0.93	0.87	0.9	0.93	0.97
Qwen2-72B	Basic	0.81	0.87	0.91	0.87	0.98	0.93
	Ensemble	0.72	0.91	0.9	0.87	0.93	0.84
	PDR	0.69	0.9	0.9	0.89	0.93	0.82
gemma-2b	Basic	0.6	0.84	0.87	0.85	0.98	0.72
	Ensemble	0.6	0.8	0.83	0.73	0.88	0.82
	PDR	0.32	0.65	0.7	0.1	0.075	0.059
gemma-7b	Basic	0.63	0.87	0.83	0.86	0.98	0.8
	Ensemble	0.5	0.72	0.79	0.73	0.88	0.71
	PDR	0.2	0.19	0.55	0.15	0.15	0.17

FIGURE 5. Performance of the RAG pipeline using open source models evaluated on the automatically generated questions. We report the results of six metrics: answer correctness, answer relevancy, answer similarity, context precision, context recall, and faithfulness.

2) INFORMATION RETRIEVAL

Now that we have created the vector database, we can retrieve the information from it. Fig. 3 shows the conversational RAG pipeline, which includes query embedding, retrieving relevant context search, and combining user query with the context and past conversation history to produce and retrieve an augmented generated response.

For the conversational RAG pipeline, we use the Langchain library [40], where we provide the system prompt shown before, along with the language model and the vector database.

In the following, we show the model response generated by our system for the query *"What is the recommended fuel for this vehicle?"*

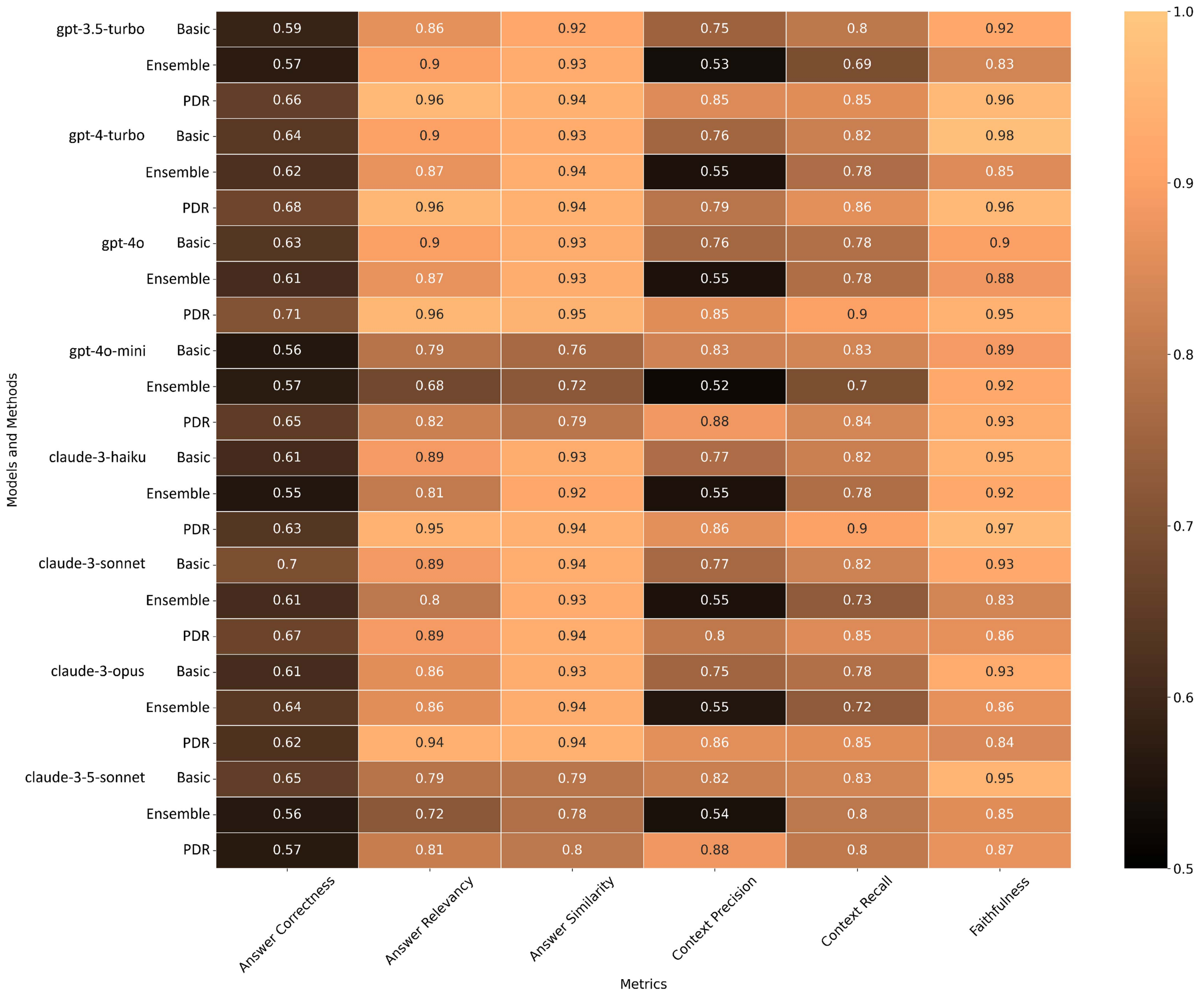


FIGURE 6. **Performance of the RAG pipeline using proprietary models evaluated on the manually crafted questions. We present the results of six metrics: answer correctness, answer relevancy, answer similarity, context precision, context recall, and faithfulness.**

According to the car manual, the recommended fuel for this vehicle is 87 or higher Octane rating fuel. The manual states: "Fuel quality and additives contained in fuel have a significant effect on power output, driveability, and life of the engine. Fuel with too low an Octane number can cause engine knock. Use of 87 or higher Octane rating fuel is recommended." The manual also cautions that "The use of gasoline with an Octane rating lower than 87 may cause engine to knock. Continued use may cause engine damage and manufacturer warranty stands void."

a) Parent Document Retriever (PDR): Conflicting desires often arise when dividing documents for retrieval. On the one hand, we may want to keep the documents small to ensure that their embeddings accurately reflect their meaning. If the documents are too long, the embeddings can lose their meaning. On the other hand, we want the documents long enough to preserve the context of each piece [44].

The Parent Document Retriever (PDR) method achieves this balance by splitting and storing small data chunks. During retrieval, it first fetches the small chunks, then looks up the parent IDs for those chunks, and returns the larger documents. Algorithm 1 describes the steps taken to implement this retrieval method. We use *ParentDocumentRetriever* from Langchain [45], taking two different splitters, one with a 2000-character chunk size and the other with a 400-character chunk size.

b) Ensemble retriever: The EnsembleRetriever accepts a set of retrievers as an input and combines their results, then reorders the results using the Reciprocal Rank Fusion algorithm [46]. Algorithm 2 describes the algorithm followed to perform the ensemble retrieval.

Algorithm 1: Parent Document Retriever Algorithm.

```
Procedure RetrieveParentDocuments
1: Obtain User Query
2: Retrieve child documents using Dense Vector Retrieval
3: for each child document do
    4: Merge the child documents based on their parents
    5: if child documents have the same parents then
        6: Merge them
    7: end if
8: end for
9: Replace the child documents with their respective parent documents from an in-memory-store
10: Use the parent documents to augment generation
End Procedure
```

Algorithm 2: Ensemble Retrieval Algorithm.

```
Procedure EnsembleRetrieval
1: Obtain User Question
2: Hit the Retriever Pair
3: Retrieve Documents with BM25 Sparse Vector Retrieval
4: Retrieve Documents with Dense Vector Retrieval Method
5: Collect and fuse the retrieved documents based on their weighting using the Reciprocal Rank Fusion algorithm into a single ranked list
6: Use those documents to augment our generation
End Procedure
```

C. METRICS TO EVALUATE RAG PIPELINE

Assessing RAG architectures presents challenges due to several factors: the retrieval system's effectiveness in locating relevant context passages, the language model's ability to accurately utilize these passages, and the overall quality of the generated output. To address these multiple factors, we used Ragas [47], a framework designed to evaluate RAG pipelines. In evaluating our system's performance, we considered the following metrics: answer correctness, relevancy, semantic similarity, context precision, context recall, and faithfulness, which are explained below.

a) Answer semantic similarity: Answer Semantic Similarity evaluates how closely the generated answer aligns with the ground truth in meaning. This metric uses a cross-encoder model to calculate the semantic similarity score, ranging from 0 to 1, where higher scores indicate better alignment [47], [48].

$$\text{Semantic Similarity} = \cos(E_g, E_t), \tag{1}$$

where E_g is the embedding of the generated answer, and E_t is the embedding of the ground truth.

b) Answer correctness: Answer Correctness assesses how well a generated answer matches the ground truth, considering factual correctness and semantic similarity. The metric uses an F1 Score to balance precision and recall, followed by a weighted average with semantic similarity [47], [49].

$$\text{Answer Correctness} = \alpha \times \text{F1} + (1 - \alpha) \times \text{Semantic Similarity} \tag{2}$$

$$\text{F1 Score} = \frac{2 \times |\text{TP}|}{2 \times |\text{TP}| + |\text{FP}| + |\text{FN}|} \tag{3}$$

Where true positives (TP) are the elements in the generated answer that match the correct response. False positives (FP) are the incorrect elements in the generated answer not present in the ground truth. False negatives (FN) are the correct elements in the ground truth missing from the generated answer. The Semantic Similarity described previously is the cosine similarity between embeddings of generated and ground truth answers, and α is a weight factor balancing F1 and Semantic Similarity.

c) Answer relevance: Answer relevance measures how well a generated answer addresses the given question. This metric penalizes incomplete or redundant responses by computing the mean cosine similarity between the original question and multiple questions generated from the answer [47].

$$\text{Answer Relevance} = \frac{1}{N} \sum_{i=1}^{N} \cos(E_{g_i}, E_o), \tag{4}$$

where E_{g_i} is the embedding of the generated question i, E_o is the embedding of the original question, and N is the number of generated questions.

A response is considered relevant when it directly and properly addresses the original question. It is important to note that answer relevance does not consider accuracy but penalizes instances where the answer is incomplete or contains duplicate information. To calculate this score, the LLM is tasked with generating a suitable question for the answer provided multiple times, and then the average cosine similarity between these generated questions and the original question is determined. The fundamental idea is that if the provided answer effectively responds to the initial question, the LLM should be able to generate questions from the answer that align with the original question [50].

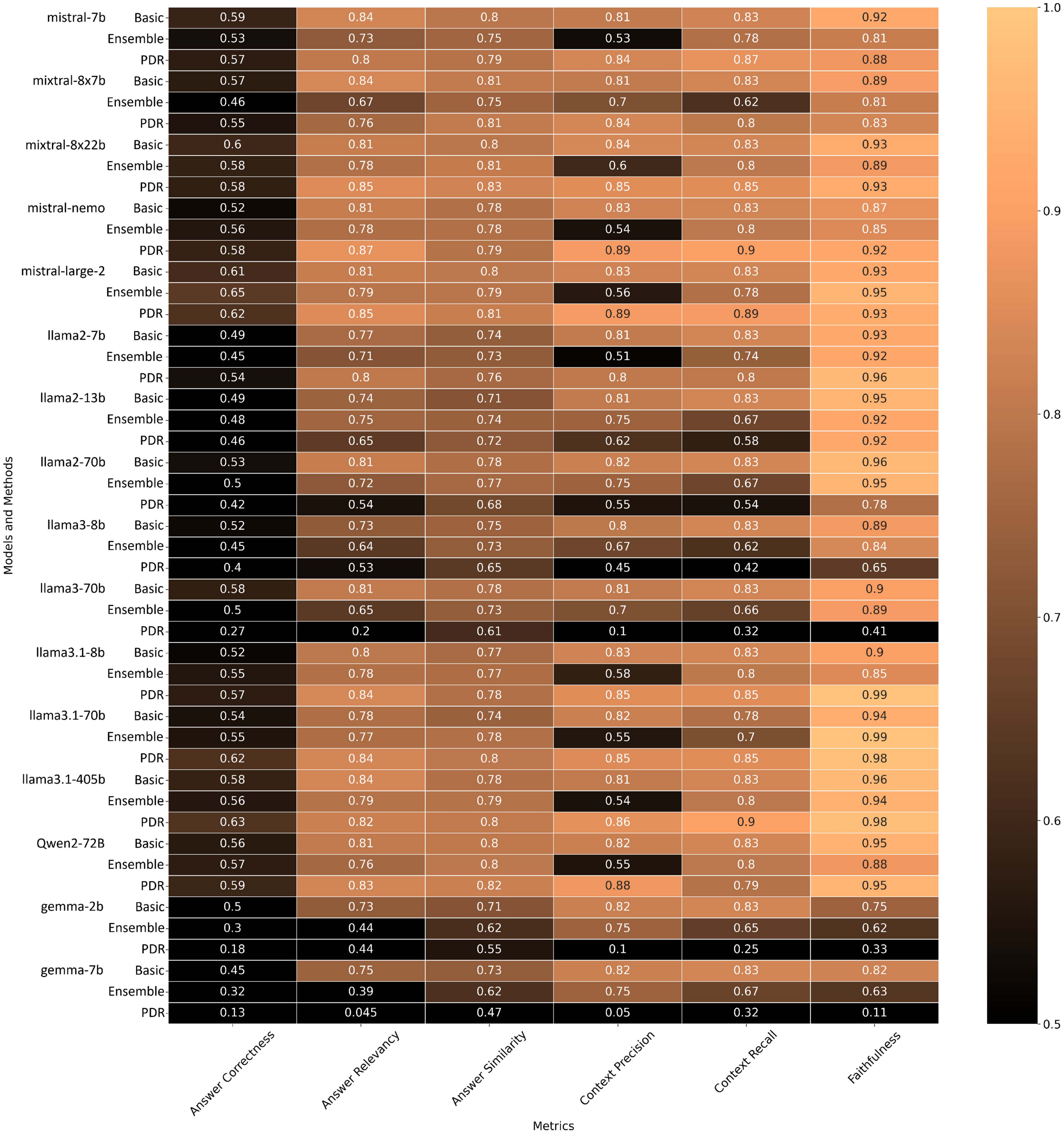
d) Context precision: Context Precision (CP) evaluates if the most relevant items in the context are ranked higher. It uses the question, ground truth, and context to calculate a score between 0 and 1, with higher scores indicating better precision [47], [51].

$$\text{CP@K} = \frac{\sum_{k=1}^{K} (\text{Precision at } k \times v_k)}{\text{Total relevant items in top } K} \tag{5}$$

$$\text{Precision at } k = \frac{\text{TP at } k}{\text{TP at } k + \text{FP at } k} \tag{6}$$

Where v_k is 1 if the item at rank k is relevant, otherwise 0.

e) Context recall: Context recall evaluates how well the retrieved context aligns with the ground truth answer. It is



Models and Methods		Answer Correctness	Answer Relevancy	Answer Similarity	Context Precision	Context Recall	Faithfulness
mistral-7b	Basic	0.59	0.84	0.8	0.81	0.83	0.92
	Ensemble	0.53	0.73	0.75	0.53	0.78	0.81
	PDR	0.57	0.8	0.79	0.84	0.87	0.88
mixtral-8x7b	Basic	0.57	0.84	0.81	0.81	0.83	0.89
	Ensemble	0.46	0.67	0.75	0.7	0.62	0.81
	PDR	0.55	0.76	0.81	0.84	0.8	0.83
mixtral-8x22b	Basic	0.6	0.81	0.8	0.84	0.83	0.93
	Ensemble	0.58	0.78	0.81	0.6	0.8	0.89
	PDR	0.58	0.85	0.83	0.85	0.85	0.93
mistral-nemo	Basic	0.52	0.81	0.78	0.83	0.83	0.87
	Ensemble	0.56	0.78	0.78	0.54	0.8	0.85
	PDR	0.58	0.87	0.79	0.89	0.9	0.92
mistral-large-2	Basic	0.61	0.81	0.8	0.83	0.83	0.93
	Ensemble	0.65	0.79	0.79	0.56	0.78	0.95
	PDR	0.62	0.85	0.81	0.89	0.89	0.93
llama2-7b	Basic	0.49	0.77	0.74	0.81	0.83	0.93
	Ensemble	0.45	0.71	0.73	0.51	0.74	0.92
	PDR	0.54	0.8	0.76	0.8	0.8	0.96
llama2-13b	Basic	0.49	0.74	0.71	0.81	0.83	0.95
	Ensemble	0.48	0.75	0.74	0.75	0.67	0.92
	PDR	0.46	0.65	0.72	0.62	0.58	0.92
llama2-70b	Basic	0.53	0.81	0.78	0.82	0.83	0.96
	Ensemble	0.5	0.72	0.77	0.75	0.67	0.95
	PDR	0.42	0.54	0.68	0.55	0.54	0.78
llama3-8b	Basic	0.52	0.73	0.75	0.8	0.83	0.89
	Ensemble	0.45	0.64	0.73	0.67	0.62	0.84
	PDR	0.4	0.53	0.65	0.45	0.42	0.65
llama3-70b	Basic	0.58	0.81	0.78	0.81	0.83	0.9
	Ensemble	0.5	0.65	0.73	0.7	0.66	0.89
	PDR	0.27	0.2	0.61	0.1	0.32	0.41
llama3.1-8b	Basic	0.52	0.8	0.77	0.83	0.83	0.9
	Ensemble	0.55	0.78	0.77	0.58	0.8	0.85
	PDR	0.57	0.84	0.78	0.85	0.85	0.99
llama3.1-70b	Basic	0.54	0.78	0.74	0.82	0.78	0.94
	Ensemble	0.55	0.77	0.78	0.55	0.7	0.99
	PDR	0.62	0.84	0.8	0.85	0.85	0.98
llama3.1-405b	Basic	0.58	0.84	0.78	0.81	0.83	0.96
	Ensemble	0.56	0.79	0.79	0.54	0.8	0.94
	PDR	0.63	0.82	0.8	0.86	0.9	0.98
Qwen2-72B	Basic	0.56	0.81	0.8	0.82	0.83	0.95
	Ensemble	0.57	0.76	0.8	0.55	0.8	0.88
	PDR	0.59	0.83	0.82	0.88	0.79	0.95
gemma-2b	Basic	0.5	0.73	0.71	0.82	0.83	0.75
	Ensemble	0.3	0.44	0.62	0.75	0.65	0.62
	PDR	0.18	0.44	0.55	0.1	0.25	0.33
gemma-7b	Basic	0.45	0.75	0.73	0.82	0.83	0.82
	Ensemble	0.32	0.39	0.62	0.75	0.67	0.63
	PDR	0.13	0.045	0.47	0.05	0.32	0.11

FIGURE 7. Performance of the RAG pipeline using open source models evaluated on the manually crafted questions. We present the results of six metrics: answer correctness, answer relevancy, answer similarity, context precision, context recall, and faithfulness.

calculated by determining the proportion of sentences in the ground truth that can be attributed to the retrieved context [47], [52].

$$\text{Context Recall} = \frac{|\text{GT}_{\text{sentences attributed}}|}{|\text{GT}_{\text{total sentences}}|}, \tag{7}$$

where $|\text{GT}_{\text{sentences attributed}}|$ is the number of ground truth sentences attributable to the context, and $|\text{GT}_{\text{total sentences}}|$ is the total number of sentences in the ground truth.

f) Faithfulness: Faithfulness measures the factual consistency between a generated answer and the given context. Assesses whether the claims in the answer can be inferred from the context. To compute this, each claim in the answer is

TABLE 1. **Language Models Used to Evaluate Our System**

Model	Source	Released date	Context size	Model size
gpt-3.5-turbo [55]	OpenAI	2022-11-30	16k	-
gpt-4-turbo [21]	OpenAI	2023-03-14	128k	-
gpt-4o [56]	OpenAI	2024-05-13	128k	-
gpt-4o-mini [56]	OpenAI	2024-07-18	128k	-
claude-3-haiku [57]	Anthropic	2024-03-07	200k	-
claude-3-sonnet [57]	Anthropic	2024-02-29	200k	-
claude-3-opus [57]	Anthropic	2024-02-29	200k	-
claude-3.5-sonnet [57]	Anthropic	2024-06-20	200k	-
mistral-7b [59]	Mistral	2023-09-27	8k	7
mixtral-8x7b [60]	Mistral	2024-01-08	8k	56
mixtral-8x22b [61]	Mistral	2024-04-17	64k	176
mistral-nemo [62]	Mistral	2024-07-18	128k	12
mistral-large-2407 [63]	Mistral	2024-07-24	128k	123
llama2-7b-chat [64]	Meta	2023-07-18	4k	7
llama2-13b-chat [64]	Meta	2023-07-18	4k	13
llama2-70b-chat [64]	Meta	2023-07-18	4k	70
llama3-8b [65]	Meta	2024-04-18	8k	8
llama3-70b [65]	Meta	2024-04-18	8k	70
llama3.1-8b [66]	Meta	2024-07-23	128k	8
llama3.1-70b [66]	Meta	2024-07-23	128k	70
llama3.1-405b [66]	Meta	2024-07-23	128k	405
Qwen2-72B [67]	Alibaba	2024-07-15	32k	72
gemma-7b [68]	Google	2024-02-21	8k	7
gemma-2b [68]	Google	2024-02-21	8k	2

The first eight models are closed proprietary models, with undisclosed sizes, from OpenAI and Anthropic. The subsequent 16 models are open source, detailing their source, release dates, context size in tokens, and model sizes in billions of parameters.

cross-checked with the context. The faithfulness score is the ratio of the claims supported by the context to the total claims made [47], [53].

$$\text{Faithfulness score} = \frac{C_s}{C_t}, \tag{8}$$

where C_s is the number of claims supported by the context , and C_t is the total number of claims.

The following section describes the performance of the RAG pipeline in producing the correct responses.

III. RESULTS

In order to assess our system's response, we compared 24 large language models, including eight closed models and 16 open source models. Table 1 provides a list of the models used, their sources, release dates, and sizes. While the sizes of the proprietary models from OpenAI and Anthropic have not been disclosed, they are known to outperform open source models in various benchmarks [54].

To evaluate our RAG pipeline, we generated two sets of 20 questions. The first set consisted of questions was automatically generated using GPT-4o on the data splits used to build the vector database. The second set was created manually by reviewing the car manual and selecting common questions that a user might ask.

A. EVALUATION USING AUTOMATICALLY GENERATED QUESTIONS AND ANSWERS DATASET

To generate automatic questions, we iterated through 20 randomly selected sections of the car manual and used GPT-4o to create the questions, using the prompt shown below.

> *You are a car expert creating a test for car users. For each context, create a question that is specific to the context. Avoid creating generic or general questions. All questions must be in English.*
> *question: a question about the context.*
> *Format the output as JSON with the following keys:*
> *question: {question}*
> *context: {context}*

Then, for each question and context generated, we also automatically generated the answers using GPT-4o with the following prompt.

> *You are a car expert creating a test for car users. For each question and context, create an answer.*
> *answer: an answer about the context.*
> *Format the output as JSON with the following keys:*
> *answer: {answer}*
> *question: {question}*
> *context: {context}*

We use GPT-4o because it is the state-of-the-art model at the time of writing.

For brevity, we display only the initial five samples and provide the full set in the appendix.

1) **What should you do before using the air conditioning system if your vehicle has been parked in direct sunlight?**
   Before using the air conditioning system, open the windows for a few minutes to allow hot air to escape if the vehicle has been parked in direct sunlight.
2) **According to the Chevrolet Spark manual, how often should you check your tire pressure, and what tool should you use for this check?**
   According to the Chevrolet Spark manual, you should check your tire pressure each time you fill your fuel tank or at least once a month. The tool recommended for this check is a tire pressure gauge.
3) **Under what conditions will Chevrolet Roadside Assistance provide services if the horn is malfunctioning?**
   Chevrolet Roadside Assistance will provide services if the horn is malfunctioning only if the horn is sounding continuously.
4) **What should you do if your brakes temporarily lose their stopping power due to overheating while driving down a steep hill?**
   If your brakes temporarily lose their stopping power due to overheating while driving down a steep hill, you should shift to a lower gear and avoid continuously applying the brakes.
5) **What is the benefit of wearing a safety belt in a collision, according to the Chevrolet Spark manual?**
   The benefit of wearing a safety belt in a collision, according to the Chevrolet Spark manual, is that the

TABLE 2. **Propietary Model Responses for the Question *How Many Safety Belts Does This Car Have?* the Incorrect Answers are Highlighted for Easier Visualization**

Model	Retrieval method	Response
gpt-3.5-turbo	Basic	**This car has three-point safety belts for each front seat and each outboard rear seat, totaling four safety belts in total.**
	PDR	**The car mentioned in the context has three-point safety belts for each front seat and each outboard rear seat. So, it has a total of four safety belts.**
	Ensemble	**The car has three-point safety belts for each front seat and each outboard rear seat, totaling to four safety belts in total.**
gpt-4-turbo	Basic	The CHEVROLET vehicle described in the manual is equipped with a total of five safety belts. There are three-point lap-and-shoulder safety belts for each of the two front seats and the two outboard rear seating positions, and a lap safety belt for the center seat in the rear.
	PDR	The Chevrolet vehicle described has a total of five safety belts. Each front seat is equipped with a three-point lap-and-shoulder safety belt, and the rear seat has two outboard seating positions with three-point lap-and-shoulder safety belts, plus a center seat equipped with a lap safety belt.
	Ensemble	The car has five safety belts. Each front seat is equipped with a three-point lap-and-shoulder safety belt, and the rear seat is equipped with two outboard three-point lap-and-shoulder safety belts and a center lap safety belt.
gpt-4o	Basic	The car is equipped with a total of five safety belts: three-point safety belts for each of the two front seats and the two outboard rear seats, and a lap safety belt for the center rear seat.
	PDR	The car is equipped with a total of five safety belts. Each of the two front seats and the two outboard rear seats have three-point lap-and-shoulder safety belts, and the center rear seat has a lap safety belt.
	Ensemble	The car is equipped with a total of five safety belts. There are three-point safety belts for each of the two front seats and the two outboard rear seats, and a lap safety belt for the center rear seat.
gpt-4o-mini	Basic	The car is equipped with a total of five safety belts: two three-point safety belts for the front seats, two three-point safety belts for the outboard rear seating positions, and one lap safety belt for the center rear seat.
	PDR	The car is equipped with a total of five safety belts: two for the front seats (three-point lap-and-shoulder safety belts) and three for the rear seats (two outboard positions with three-point lap-and-shoulder safety belts and one center seat with a lap safety belt).
	Ensemble	The car is equipped with three-point safety belts for each front seat and the two outboard rear seating positions. The center rear seat is equipped with a lap safety belt.
claude-3-haiku	Basic	According to the car manual information provided: The vehicle has the following safety belt configurations: - Each front seat is equipped with a three-point lap-and-shoulder safety belt. - The two outboard rear seating positions are each equipped with three-point lap-and-shoulder safety belts.- The center rear seat is equipped with a lap safety belt. So in total, this vehicle has 5 safety belts - 2 in the front seats and 3 in the rear seats.
	PDR	Based on the information provided in the context, this vehicle has the following safety belt configurations: - Each front seat is equipped with an adjustable seat and seatback with a height-adjustable head restraint and a three-point lap-and-shoulder safety belt. - The rear seat has two outboard seating positions that each feature a three-point lap-and-shoulder safety belt. - The rear seat also has a center seating position that is equipped with a lap safety belt. So in total, this vehicle has 5 safety belts - two three-point belts in the front seats, two three-point belts in the outboard rear seats, and one lap belt in the rear center seat.
	Ensemble	Based on the information provided in the passage, this CHEVROLET vehicle has the following safety belt configuration: - Each front seat is equipped with an adjustable seat and seatback, a height-adjustable head restraint, and a three-point lap-and-shoulder safety belt. - The rear seat has two outboard seating positions, each equipped with a three-point lap-and-shoulder safety belt. - The rear seat also has a center seating position, which is equipped with a lap safety belt. So in total, this CHEVROLET vehicle has 5 safety belts - 2 three-point safety belts for the front seats, 2 three-point safety belts for the outboard rear seats, and 1 lap safety belt for the rear center seat.
claude-3-sonnet	Basic	Based on the information provided in the car manual, this vehicle has the following safety belts: 1. Three-point safety belts for the two front seats and the two outboard rear seating positions. 2. A lap safety belt for the center rear seating position. So in total, there are five safety belts in this car - two three-point belts in the front, two three-point belts for the outboard rear seats, and one lap belt for the center rear seat.
	PDR	Based on the information provided, this Chevrolet vehicle has the following safety belts: - Three-point lap-and-shoulder safety belts for the two front seats - Three-point lap-and-shoulder safety belts for the two outboard rear seating positions - A lap belt for the center rear seating position. So in total, it has 5 safety belts - 4 three-point lap-and-shoulder belts (2 front, 2 rear outboard) and 1 lap belt for the rear center seat.
	Ensemble	Based on the information provided, this vehicle has: - 2 three-point (lap and shoulder) safety belts for the front seats - 2 three-point (lap and shoulder) safety belts for the outboard rear seats - 1 lap belt for the center rear seat. So in total, there are 5 safety belts in this vehicle - 4 three-point belts and 1 lap belt.
claude-3-opus	Basic	According to the car manual, the vehicle is equipped with three-point safety belts for the two front seats and the two outboard rear seating positions. The rear center seat is equipped with a lap safety belt.
	PDR	Based on the information provided in the owner's manual excerpts, the Chevrolet vehicle is equipped with the following safety belts: - Each of the two front seats has a three-point (lap and shoulder) safety belt - The two outboard rear seating positions each have a three-point (lap and shoulder) safety belt - The rear center seating position has a lap safety belt. So in total, this Chevrolet vehicle is equipped with 5 safety belts - 3 three-point belts (2 front, 2 rear outboard) and 1 lap belt (rear center).
	Ensemble	Based on the information provided in the owner's manual excerpt, this Chevrolet vehicle is equipped with the following safety belts: 1. The two front seats each have a three-point (lap and shoulder) safety belt. 2. The two outboard rear seating positions each have a three-point (lap and shoulder) safety belt. 3. The rear center seating position has a lap safety belt. So in total, the vehicle has 5 safety belts - 4 three-point lap/shoulder belts (2 in front, 2 in rear outboard positions) and 1 lap belt for the rear center seat.
claude-3.5-sonnet	Basic	Based on the car manual information provided, this vehicle has the following safety belts: - Two three-point safety belts for the front seats - Two three-point safety belts for the rear outboard seats - One lap safety belt for the rear center seat In total, the car has 5 safety belts - 4 three-point belts and 1 lap belt.
	PDR	Based on the information provided in the context, this CHEVROLET vehicle is equipped with the following safety belts: 1. Two three-point safety belts for the front seats (driver and passenger). 2. Two three-point safety belts for the outboard rear seats (left and right sides). 3. One lap safety belt for the center rear seat. In total, the vehicle has 5 safety belts: - 4 three-point lap-and-shoulder safety belts (2 front, 2 rear outboard) - 1 lap safety belt (rear center)
	Ensemble	Based on the information provided in the context, this CHEVROLET vehicle has the following safety belts: 1. Two three-point safety belts for the front seats (one for each front seat). 2. Two three-point safety belts for the rear outboard seats (one for each outboard rear seat). 3. One lap safety belt for the rear center seat. In total, the vehicle has 5 safety belts: - 4 three-point safety belts (2 front, 2 rear outboard) - 1 lap safety belt (rear center)

We report the responses of the eight proprietary models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

TABLE 3. Propietary Models Answers for the Question *What is the Maximum Recommended RPM for This Car?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
gpt-3.5-turbo	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	The maximum recommended RPM for the car is mentioned as above 4500 RPM. Excessive engine speed can damage the engine, resulting in costly repairs.
	Ensemble	**I don’t have the specific information on the maximum recommended RPMs for your car. It’s best to refer to your vehicle’s manual or contact a Chevrolet retailer for accurate details on the recommended RPM range for your specific vehicle.**
gpt-4-turbo	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	The maximum recommended RPM to operate the engine is 4500 RPM. Operating the engine at higher RPMs can cause damage and result in costly repairs.
	Ensemble	**The provided text does not specify the maximum recommended RPMs to force the car. For accurate information on this topic, please refer to the vehicle’s owner manual or contact a CHEVROLET retailer.**
gpt-4o	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	The maximum recommended RPMs for operating the engine is 4500 RPM. Exceeding this limit can damage the engine and result in costly repairs.
	Ensemble	**The provided context does not specify the maximum recommended RPMs for your vehicle. For this information, it is best to consult your vehicle’s owner’s manual or contact a CHEVROLET retailer directly. They can provide specific details based on your vehicle model and engine type.**
gpt-4o-mini	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	The maximum recommended RPM to operate the engine is 4500 RPM.
	Ensemble	**I don’t know.**
claude-3-haiku	Basic	**According to the car manual, the manual states: ”Do not idle the engine at high rpm for more than 3 minutes. Excessive temperature can damage the exhaust system/catalytic converter.” Therefore, the maximum recommended RPMs to idle the engine is 3 minutes. The manual does not provide a specific maximum RPM value, but it cautions against idling the engine at high RPMs for an extended period of time, as it can damage the exhaust system and catalytic converter.**
	PDR	According to the information provided in the manual, the digital tachometer indicates the engine speed in revolutions per minute (RPM). The manual cautions: ”Neveroperate the engine at higher rpm (above 4500 rpm). Excessive engine speed can damage the engine. This could result in costly re-pairs.” So the maximum recommended RPMs is 4500 rpm. The manual advises not to operate the engine at higher RPMs above that limit, as it can damage the engine and lead to costly repairs.
	Ensemble	**I do not see any information in the provided context about the maximum recommended RPMs to ”force” the car. The context covers topics like assist handles, luggage carriers, starting the engine, using the manual transmission, and maintaining tires and safety belts. It does not contain any specific guidance about pushing the engine to high RPMs. Without more relevant information, I cannot provide a definitive answer to your question. The owner’s manual or guidance from the vehicle manufacturer would be needed to determine the recommended maximum RPMs for safe operation of the vehicle.**
claude-3-sonnet	Basic	**I am sorry, I did not find the answer to the maximum recommended RPMs in the provided car manual. The manual does not seem to specify a maximum RPM limit.**
	PDR	**The owner’s manual does not provide a specific maximum recommended RPM limit for this vehicle. In general, it advises against racing the engine or full-throttle acceleration in low gears, as this can damage the engine. The manual recommends allowing the engine to warm up before driving, avoiding sudden accelerations, and not towing another vehicle during the break-in period to improve performance and longevity. Without more specifics from the manufacturer, I cannot give you a definitive maximum RPM limit to avoid exceeding.**
	Ensemble	**There is no mention of a maximum recommended RPM limit or guidance on ”forcing” the engine RPMs in the provided context from the owner’s manual. Vehicle owner’s manuals typically do not recommend exceeding the normal operating RPM range of the engine, as this can lead to excessive wear or potential engine damage. Forcing abnormally high RPMs is generally not advised.**
claude-3-opus	Basic	**I’m sorry, I did not find any information in the provided car manual about the maximum recommended RPMs to force the car. The manual provides instructions for starting the engine and driving the vehicle safely, but does not mention anything about forcing the car to a certain RPM level.**
	PDR	According to the information provided in the owner’s manual, you should never operate the engine at higher rpm, specifically above 4500 rpm. The manual states: ”Never operate the engine at higher rpm (above 4500 rpm).” It also includes this caution: ”CAUTION Excessive engine speed can damage the engine. This could result in costly repairs.” So the maximum recommended RPMs to run the engine is 4500 rpm. Exceeding this limit could cause engine damage that would be expensive to repair. The manual advises against revving the engine above 4500 rpm.
	Ensemble	**I could not find any information in the provided vehicle owner’s manual excerpts about the maximum recommended RPMs to force or push the engine to. In general, it is not advisable to force or push a vehicle engine to very high RPMs, as this can cause excessive wear and potentially damage the engine. The owner’s manual typically specifies the maximum safe engine speed, but those details were not included in the excerpts provided. I would recommend consulting the full owner’s manual or a Chevrolet dealer for guidance on the maximum safe RPMs for your specific vehicle model.**
claude-3.5-sonnet	Basic	**I apologize, but I did not find any specific information about maximum recommended RPMs in the car manual provided. The manual does advise avoiding full-throttle starts, racing the engine, and prolonged high-speed driving during the break-in period to avoid engine damage and conserve fuel. However, it does not mention a specific RPM limit.**
	PDR	According to the information provided, you should never operate the engine at higher than 4500 RPM. Specifically, the manual states: ”The tachometer indicates engine speed in revolutions per minute (RPM). Never operate the engine at higher rpm (above 4500 rpm).” It also includes a caution that says: ”Excessive engine speed can damage the engine. This could result in costly repairs.” So the maximum recommended RPM is 4500, and exceeding this could potentially damage the engine.
	Ensemble	**I apologize, but I don’t see any specific information in the given context about maximum recommended RPMs for forcing or pushing the car. The provided text contains general information about safety belts, manual transmission operation, towing, and getting unstuck from snow or mud, but does not mention specific RPM limits. Without more context from the vehicle’s manual or specifications, I cannot provide a reliable answer about maximum recommended RPMs. It’s generally best to consult the specific vehicle’s owner’s manual for such technical details, as they can vary between different makes and models.**

We report the responses of the eight proprietary models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble retriever.

TABLE 4. Propietary Models Responses for the Question *What Engine Does This Car Have?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
gpt-3.5-turbo	Basic	**The engine identification number is stamped on the cylinder block under the No. 4 exhaust manifold of the engine.**
	PDR	The car has a 4-cylinder/in-line engine with a single overhead camshaft (SOHC) and 8 valves. The engine displacement is 995cc with a bore x stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3 : 1.
	Ensemble	**The engine in the car is a 1.0 SOHC model.**
gpt-4-turbo	Basic	The car is equipped with a 4-cylinder, in-line, SOHC 8-valve engine with a displacement of 995 cc.
	PDR	**The car has a 1.0 SOHC (Single Overhead Camshaft) engine model.**
	Ensemble	**The car has a 1.0 SOHC (Single Overhead Camshaft) engine model.**
gpt-4o	Basic	The car has a 4-cylinder, in-line, SOHC 8-valve engine with a displacement of 995 cc.
	PDR	The car has a 1.0 SOHC (Single Overhead Camshaft) engine with 4 cylinders in-line and 8 valves. The engine has a displacement of 995 cc.
	Ensemble	**The car has a 1.0 SOHC (Single Overhead Camshaft) engine.**
gpt-4o-mini	Basic	The car is equipped with a 1.0 SOHC 4-cylinder in-line engine with 8 valves. It has a displacement of 995 cc and a compression ratio of 9.3 ± 0.3:1.
	PDR	The car has a 4-cylinder, in-line SOHC engine with a displacement of 995 cc, a bore x stroke of 68.5 x 67.5 mm, and a compression ratio of 9.3 ± 0.3 : 1.
	Ensemble	The car has a 4-cylinder in-line engine with a displacement of 995 cc, a bore x stroke of 68.5 x 67.5 mm, and a compression ratio of 9.3 ± 0.3 : 1. It produces a maximum power of 45.0 kW at 5400 rpm and a maximum torque of 86.8 Nm at 4200 rpm. The fuel system is multi-point injection.
claude-3-haiku	Basic	According to the information provided in the car manual, this vehicle has a 4-cylinder/in-line SOHC 8 valve engine with a displacement of 995 cc.
	PDR	Based on the technical information provided: The engine in this vehicle is a 4-cylinder, in-line SOHC (Single Overhead Cam) engine with 8 valves and a displacement of 995 cc. The key engine specifications are: - Displacement: 995 cc - Bore x Stroke: 68.5 mm x 67.5 mm - Compression Ratio: 9.3 ± 0.3 : 1 - Max. Power: 45.0 kW @ 5400 rpm - Max. Torque: 86.8 Nm @ 4200 rpm So in summary, this vehicle is equipped with a 995 cc, 4-cylinder SOHC engine.
	Ensemble	**The context does not specify the exact engine model for this vehicle. The service and vehicle care section mentions an ”1.0 SOHC model”, which indicates the vehicle has a 1.0 liter single overhead cam (SOHC) engine, but the specific engine type is not provided. Without more details about the make, model and year of the vehicle, I don’t have enough information to definitively state what engine it has.**
claude-3-sonnet	Basic	**Based on the information provided in the car manual, this vehicle has a 1.0 SOHC engine. The engine compartment diagram shows details for the ”1.0 SOHC MODEL”.**
	PDR	Based on the technical information provided, this vehicle has a 4-cylinder in-line SOHC 8 valve engine with a displacement of 995 cc (99.5 x 68.5 mm bore x stroke) and a compression ratio of 9.3 +/- 0.3:1.
	Ensemble	**Unfortunately, the provided information does not explicitly state what specific engine this vehicle has. The context mentions an ”engine identification number” that is stamped on the cylinder block, and that there is a 1.0 SOHC model engine compartment layout shown, but it does not give details on the actual engine specifications like displacement, number of cylinders, etc.**
claude-3-opus	Basic	According to the vehicle specifications provided in the car manual, this car has a 4-cylinder/in-line engine with a single overhead camshaft (SOHC) and 8 valves. The engine displacement is 995 cc with a bore x stroke of 68.5 x 67.5 mm.
	PDR	According to the technical specifications provided, this car has a 4-cylinder inline SOHC 8-valve engine with a displacement of 995 cc. The engine has a bore x stroke of 68.5 x 67.5 mm and a compression ratio of 9.3 ± 0.3 : 1. It produces a maximum power output of 45.0 kW at 5400 rpm and maximum torque of 86.8 Nm at 4200 rpm.
	Ensemble	**Based on the image and information provided, this vehicle has a 1.0L SOHC (Single Overhead Camshaft) engine. The engine compartment diagram is labeled ”1.0 SOHC MODEL” at the top, indicating the engine size and configuration.**
claude-3.5-sonnet	Basic	According to the car manual, this vehicle has a 4-cylinder/in-line engine with SOHC (Single Overhead Camshaft) 8 valve configuration. The engine displacement is 995cc with a bore x stroke of 68.5 x 67.5 mm and a compression ratio of 9.3 ± 0.3 : 1.
	PDR	According to the technical information provided, this car has a 1.0 liter SOHC (Single Overhead Camshaft) engine. Specifically, the engine details are: - 4-cylinder/in-line, SOHC 8 valves, Displacement: 995 cc, Bore x stroke: 68.5 x 67.5 mm, Compression ratio: 9.3 ± 0.3:1, Max Power: 45.0 kW at 5400 rpm, Max Torque: 86.8 Nm at 4200 rpm, Fuel system: Multi-point injection. This information is consistent with a 1.0 liter SOHC engine configuration.
	Ensemble	According to the technical specifications provided, this car has the following engine: 4-cylinder, in-line engine, SOHC (Single Overhead Camshaft) with 8 valves, Displacement: 995 cc, Bore x stroke: 68.5 x 67.5mm, Compression ratio: 9.3 ± 0.3:1, Maximum Power: 45.0 kW at 5400 rpm, Maximum Torque: 86.8 Nm at 4200 rpm, Fuel system: Multi-point injection. The engine appears to be a small 1.0 liter (995 cc) 4-cylinder engine with modest power output, likely designed for fuel efficiency in a compact car.

We report the responses of the eight proprietary models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble retriever.

belted occupant’s body slows down from 30 mph (50 km/h) to zero over a distance of 3–4 feet (90–120 cm). This helps keep the occupant properly positioned so that, if the air bag deploys in a frontal collision, the occupant might never strike any rigid structures in the vehicle. In contrast, an unbelted occupant continues to travel at the vehicle’s pre-crash speed until striking a hard object, resulting in a greater risk of serious injury or death.

Figs. 4 and 5 show the evaluation of our RAG system with the 20 automatically generated questions using the proprietary and open source models, respectively. The proprietary models achieved superior results in all six metrics, with *answer correctness* having the lowest scores. As explained in (2), the correctness of the answer measures how accurately a generated answer aligns with the ground truth. Answer correctness is a weighted average of answer similarity and F1 score. Answer similarity focuses on the overall meaning, while answer correctness also considers exact factual details. Minor inaccuracies or missing specific details can lower the correctness score. Since the answer similarity (as shown in the third column) is high, the element that affects this metric is the F1 score given by the presence of incorrect elements (false positives) or the absence of correct elements (false negatives).

It should be noted that certain open source models did not perform well with the parent document retrieval (PDR) method, as shown in the darker rows of Fig. 5. This is because their context windows are more limited and could not accommodate the extra text provided by this retrieval. One way to solve this issue is to use a smaller split size during retrieval.

B. EVALUATION USING MANUALLY GENERATED QUESTIONS AND ANSWERS DATASET

Given the impressive performance of the RAG pipeline in the automatically generated dataset, we hypothesized that the optimistic results stemmed from the method of selecting the context of the questions. Specifically, the context used to generate the automatic questions was derived from the splits used to create the vector dataset, ensuring that the specific context used for the question existed in the database. For this reason, we decided to manually construct an evaluation dataset by reviewing the car manual and choosing questions that an ordinary user might ask.

The following are the questions and answers that we designed manually to evaluate our system. For the sake of brevity, we are only showcasing four questions that we found challenging for the models to answer correctly, as determined through manual inspection. The complete set of questions is in the appendix.

1) *How many safety belts does this car has?* A total of five. Each front seat is equipped with an adjustable seat and seatback with height adjustable headrest and three point lap and shoulder safety belts. The rear seat is equipped with two outboard seating positions featuring three point lap-and-shoulder safety belts and a center seat equipped with a lap safety belt.
2) *What is the maximum recommended RPM for this car?* Never operate the engine at more than 4500 rpm.
3) *Can I use the phone while driving?* One significant distraction in today's driving environment is the use of hand-held phones while driving. Studies have found that the use of handheld phones while driving increases the risk of being involved in an accident. At least one scientific study found that the use of any hand-held or hands-free phone system increases the risk of a collision by 400%. Using phones, two-way radios or other electronic devices such as computers, organizers, games, video, GPS, and other navigation aids similarly increases the risk of collision. We discourage the use of equipment while driving your vehicle.
4) *What engine does this car have?* This car is equipped with a 4-cylinder in-line engine. It features a single overhead camshaft (SOHC) with eight valves and has a displacement of 995 cc. The bore and stroke dimensions are 68.5 mm by 67.5 mm and have a compression ratio of 9.3 to 1. In terms of performance, it generates a maximum power of 45 kW at 5400 rpm and a maximum torque of 86.8 Nm at 4200 rpm. The fuel system uses multi-point injection, which helps with efficient fuel delivery. For maintenance, the engine uses R BPR5EY-11 spark plugs with a gap of 1.0 to 1.1 mm. The battery has a rating of 12 volts and 35 A-hours, and the alternator is rated at 12 volts and 65 amperes. The starter motor operates at 12 volts and 0.8 kilowatts. In general, this engine is designed for reliability and efficiency, suitable for everyday driving needs.

Figs. 6 and 7 show the performance of the RAG methods using the proprietary and open source models evaluated with the 20 questions designed manually. As anticipated, the results indicate worse performance compared to the automatically selected questions shown in Figs. 4 and 5. This is because the manually designed questions were not chosen from the dataset splits, but rather directly from the manual, reflecting a more realistic scenario. Again, the proprietary models achieved superior results in comparison with their open source counterparts.

The majority of the metrics are lower compared with the results of the automatically generated questions, especially in terms of context precision (4th column) using the ensemble retrieval method. This indicates that the ensemble retrieval approach we tested did not effectively retrieve relevant items for the contexts, as demonstrated in Tables 3 and 4.

To analyze qualitative results, we selected three questions that we found challenging for the RAG pipeline and showed the detailed responses of each model with each retrieval method. The questions selected were *How many safety belts does this car have?*, *What is the maximum recommended RPM for this car?*, and *What engine does this car have?*

To display the responses to these questions, we divided the models into three categories: (1) proprietary models, including GPT and Claude models; (2) small open source models, including Gemma and the smaller versions of Mistral and Llama; and (3) large open source models including bigger versions of Llama and Mixtral models. All models received the same context from the three retrieval methods so that we could compare the ability of the models to interpret the retrieved information.

Tables 2–4 show the responses of the proprietary models for each of the three questions.

Tables 5, 7, and 9 show the responses of the small open source models for each of the threes questions, and Tables 6, 8, and 10 show the responses of the large open source models.

Let us now analyze these results in detail. As shown in Table 2, GPT-3.5 was the only model that had problems in interpreting the context to provide a correct answer for the question *How many safety belts does this car have?*. The rest of the models were able to produce correct responses, although some of the models were more verbose than others. In our opinion, GPT-4o was the model that produced the best responses to this question while keeping the responses concise. This question was hard to respond to correctly because the manual specifies three types of belts for three different types of seats, making the context a bit hard to understand and the required ability of the model to perform calculations.

TABLE 5. Small Open Models Responses for the Question *How Many Safety Belts Does This Car Have?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
mistral-7b	Basic	**This car has a total of 6 safety belts, with three-point lap-and-shoulder safety belts for each front seat and the two outboard rear seating positions. The center rear seat is equipped with a lap safety belt.**
	PDR	**This car has a total of 6 safety belts. Each front seat and each outboard rear seat is equipped with a three-point safety belt system. The two front seats have lap-and-shoulder safety belts, and the two outboard rear seating positions also have lap-and-shoulder safety belts. Additionally, the center rear seat is equipped with a lap safety belt.**
	Ensemble	**This car has a total of 6 safety belts. Each front seat and the two outboard rear seating positions are equipped with three-point safety belts, making a total of 6 safety belts.**
mistral-nemo	Basic	This car has a total of five safety belts: two three-point lap-and-shoulder belts for the front seats, two three-point lap-and-shoulder belts for the rear outboard seats, and one lap belt for the center rear seat.
	PDR	This car has a total of 5 safety belts: 2 for the front seats and 3 for the rear seats (2 outboard and 1 center).
	Ensemble	This car has a total of five safety belts: two for the front seats and three for the rear seats (two outboard seats with lap-and-shoulder belts, and one center seat with a lap belt).
llama2-7b	Basic	**According to the provided car manual, the car is equipped with three-point safety belts in the front seats and three-point safety belts in the outboard rear seating positions, as well as a center seat equipped with a lap safety belt.**
	PDR	According to the provided context, the car has: 3-point safety belts in the two front seats, 3-point safety belts in the two outboard rear seating positions, a lap safety belt in the center rear seat. So, in total, the car has 5 safety belts.
	Ensemble	**According to the provided context, the car has: 2 front seats with adjustable seat and seatback, each equipped with a three-point lap-and-shoulder safety belt, 2 outboard rear seating positions with three-point lap-and-shoulder safety belts 1 center rear seat with a lap safety belt. So, in total, the car has 6 safety belts.**
llama2-13b	Basic	**According to the provided car manual, the car is equipped with three-point safety belts in the front seats and three-point safety belts in the outboard rear seating positions, as well as a center seat equipped with a lap safety belt.**
	PDR	**According to the provided text, each front seat and each outboard rear seat of the CHEVROLET vehicle is equipped with a three-point safety belt system. This means that there are at least 4 safety belts in total (2 front seats and 2 outboard rear seats). However, it's possible that there may be additional safety belts in the vehicle, such as center rear seat belts or additional belts for passengers in the rear seats.**
	Ensemble	**According to the provided context, the car has 3-point safety belts in the two front seats, 3-point safety belts in the two outboard rear seating positions, A lap safety belt in the center rear seat. So, in total, the car has 6 safety belts.**
llama3-8b	Basic	According to the provided car manual, the car has three-point safety belts in the two front seats and the two outboard rear seating positions, and a lap safety belt in the center rear seat.
	PDR	**According to the text, each front seat and each outboard rear seat of the CHEVROLET vehicle is equipped with a three-point safety belt system. This means that the car has a total of 6 safety belts: 2 for the front seats and 2 for each of the outboard rear seats.**
	Ensemble	**According to the provided text, the car has: Two front seats with three-point lap-and-shoulder safety belts, Two outboard rear seating positions with three-point lap-and-shoulder safety belts. A center seat in the rear with a lap safety belt. So, in total, the car has at least 6 safety belts.**
llama3.1-8b	Basic	According to the provided car manual information, this car has three-point safety belts in the two front seats and the two outboard rear seating positions, and a lap safety belt in the center rear seat. Therefore, the car has a total of 6 safety belts.
	PDR	**According to the text, the car has: * 2 front seats with 3-point lap-and-shoulder safety belts * 2 outboard rear seats with 3-point lap-and-shoulder safety belts * 1 center rear seat with a lap safety belt So, in total, the car has 6 safety belts.**
	Ensemble	**According to the text, this car has: * 2 front seats with 3-point lap-and-shoulder safety belts * 2 outboard rear seats with 3-point lap-and-shoulder safety belts * 1 center rear seat with a lap safety belt So, in total, the car has 6 safety belts.**
gemma-2b	Basic	**This car has 3-point safety belts for front and rear occupants.**
	PDR	The car has 3-point safety belts on the front seats and 2-point safety belts on the rear seats.
	Ensemble	The car has 3-point safety belts on the front seats and 2-point safety belts on the rear seats.
gemma-7b	Basic	**The provided text indicates that the Chevrolet vehicle has a total of **six three-point safety belts** for the front and rear seats.**
	PDR	**I am unable to access real-time information or specific car data, and am unable to provide information regarding the safety belt count of a particular car. For the most up-to-date and accurate information, please refer to the official vehicle specifications or the car manufacturer's website.**
	Ensemble	**The provided text does not contain information regarding the exact number of safety belts in the car, so I am unable to answer this question from the given context.**

We report five small open models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

Table 3 shows the detailed responses of the proprietary models for the question *What is the maximum recommended RPM for this car?*. This question proved to be quite challenging for the retriever, as the parent document retriever (PDR) approach was the only one capable of providing the correct context for the models. We hypothesize that this question is challenging for the retriever because the response is mentioned once in the digital tachometer section making it harder to find when asked about the car instead of the tachometer.

Table 4 shows the responses of the proprietary models for the question *What engine does this car have?* This question presented challenges for the retriever to produce the right context. Specifically, the embedded retrieval approach failed to provide accurate responses in all cases. In addition, the question also challenged the models to generate the desired answer. This question is challenging for the retriever because the word engine appears 230 times, but the engine specifications appear once in a table at the lower end of the manual.

Now, moving to open source models, Table 5 shows that all small models had trouble interpreting correctly the context provided by the retrieval and responding with the correct number of safety belts for the question *How many safety belts does this car have?* Table 6 shows that larger open source models

TABLE 6. Large Open Models Responses for the Question *How Many Safety Belts Does This Car Have?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
mixtral-8x7b	Basic	**This car has a total of six safety belts: three-point safety belts for the two front seats and the two outboard rear seating positions, and a lap safety belt for the center rear seat.**
	PDR	Based on the provided context, this Chevrolet vehicle has: Three-point safety belts for each of the two front seats, Three-point safety belts for the two outboard rear seating positions, A lap safety belt for the rear center seat
	Ensemble	Based on the provided context, this Chevrolet vehicle is equipped with Three-point safety belts for the two front seats, Three-point safety belts for the two outboard rear seating positions, A lap safety belt for the rear center seat
mixtral-8x22b	Basic	This car has a total of five safety belts. There are two three-point safety belts in the front seats and two three-point safety belts in the outboard rear seating positions. Additionally, there is a lap safety belt in the center seat of the rear.
	PDR	The car has a total of five safety belts. There are two front seats equipped with three-point lap-and-shoulder safety belts, and the rear seat has two outboard seating positions also featuring three-point lap-and-shoulder safety belts, along with a center seat equipped with a lap safety belt.
	Ensemble	**The car has a total of six safety belts. There are two front seats equipped with three-point lap-and-shoulder safety belts, and the rear seat has two outboard seating positions also featuring three-point lap-and-shoulder safety belts, along with a center seat equipped with a lap safety belt.**
mistral-large	Basic	The car has a total of five safety belts. There are three-point lap-and-shoulder safety belts for both front seats and the two outboard rear seating positions, and a lap safety belt for the center rear seat.
	PDR	Based on the provided context, the car has the following safety belts: - Two front seats with three-point lap-and-shoulder safety belts. - Two outboard rear seating positions with three-point lap-and-shoulder safety belts. - One center rear seat with a lap safety belt. So, in total, the car has 5 safety belts.
	Ensemble	Based on the provided context, the car has the following safety belts: - Two front seats equipped with three-point lap-and-shoulder safety belts. - Two outboard rear seating positions equipped with three-point lap-and-shoulder safety belts. - One center rear seat equipped with a lap safety belt. So, in total, the car has 5 safety belts.
llama2-70b	Basic	According to the car manual, the car has three-point safety belts in the two front seats and the two outboard rear seating positions, and a lap safety belt in the center rear seat.
	PDR	**According to the text, each front seat and each outboard rear seat of the CHEVROLET vehicle is equipped with a three-point safety belt system. This implies that there are at least 4 safety belts in the vehicle, one for each front seat and two for the outboard rear seats.**
	Ensemble	**The car has a total of six safety belts. There are two front seats equipped with three-point lap-and-shoulder safety belts, and the rear seat has two outboard seating positions also featuring three-point lap-and-shoulder safety belts, along with a center seat equipped with a lap safety belt.**
llama3-70b	Basic	This car has a total of five safety belts. There are two three-point safety belts in the front seats and two three-point safety belts in the outboard rear seating positions. Additionally, there is a lap safety belt in the center seat of the rear.
	PDR	The car has a total of five safety belts. There are two front seats equipped with three-point lap-and-shoulder safety belts, and the rear seat has two outboard seating positions also featuring three-point lap-and-shoulder safety belts, along with a center seat equipped with a lap safety belt.
	Ensemble	According to the text, the car has 2 three-point safety belts for the front seats, 2 three-point safety belts for the outboard rear seating positions, 1 lap safety belt for the center rear seat. So, in total, the car has 5 safety belts.
llama3.1-70b	Basic	According to the car manual, the car has a total of 5 safety belts: 2 three-point safety belts for the front seats, 2 three-point safety belts for the outboard rear seating positions, and 1 lap safety belt for the center rear seat.
	PDR	According to the text, the car has: * 2 three-point safety belts for the front seats * 2 three-point safety belts for the outboard rear seating positions * 1 lap safety belt for the center rear seat So, in total, the car has 5 safety belts.
	Ensemble	According to the text, the car has: * 2 three-point safety belts in the front seats * 2 three-point safety belts in the outboard rear seating positions * 1 lap safety belt in the center rear seat So, in total, the car has 5 safety belts.
llama3.1-405b	Basic	This car has a total of 5 safety belts: 2 three-point safety belts for the front seats, 2 three-point safety belts for the outboard rear seats, and 1 lap safety belt for the center rear seat.
	PDR	This car has a total of 5 safety belts: * 2 three-point safety belts for the front seats * 2 three-point safety belts for the outboard rear seats * 1 lap safety belt for the center rear seat
	Ensemble	The car has a total of 5 safety belts: * 2 three-point lap-and-shoulder safety belts for the front seats * 2 three-point lap-and-shoulder safety belts for the outboard rear seats * 1 lap safety belt for the center rear seat
Qwen2-72B	Basic	The car has a total of five safety belts: two for the front seats, two for the outboard rear seating positions, and one center rear lap belt.
	PDR	The car is equipped with safety belts for each front seat and each outboard rear seat. Specifically, there are three-point safety belts for the two front seats and the two outboard rear seating positions. The center rear seat is equipped with a lap safety belt.
	Ensemble	The car is equipped with several safety belts. Specifically, each front seat has an adjustable seat and seatback with a height-adjustable head restraint and a three-point lap-and-shoulder safety belt. The rear seat has two outboard seating positions featuring three-point lap-and-shoulder safety belts and a center seat equipped with a lap safety belt. So, in total, there are at least five safety belts: two in the front and three in the rear (two three-point belts and one lap

We report five large open models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

were able to correctly respond to these questions, demonstrating their higher ability to understand text and perform calculations.

For the second question *What is the maximum recommended RPM for this car?* where even large proprietary models had a hard time responding correctly, Table 7 shows that small open source models were, in most cases, unable to produce a good response from the retrieval context. Table 8 shows that only Mixtral models with Parent Document Retrieval (PDR) were able to produce the correct response.

Finally, Table 9 shows that small models in most cases were unable to produce the correct response for the question *What engine does this car have?* sometimes even stating that the context does not provide any information about the engine. On the other hand, Table 10 shows that most large open source models were able to produce the correct response.

In terms of inference times, Fig. 8 illustrates the average response times for each language model within the entire Retrieval-Augmented Generation (RAG) pipeline. The results represent the average time, calculated over ten runs, required

TABLE 7. Small Open Models Responses for the Question *What is the Maximum Recommended RPM for This Car?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
mistral-7b	Basic	**The car manual does not provide information on forcing the car or its maximum recommended RPMs for such an action. It's important to note that forcing the car can potentially cause damage to the engine and other components. Always drive safely and follow the recommended driving procedures in the manual.**
	PDR	The provided context does not indicate any information about forcing or racing the car. However, it does mention a maximum recommended RPM of 4500 rpm for the engine, which should not be exceeded to prevent damage to the engine. If you are asking about the maximum RPM for safe and efficient driving, it would be best to consult a professional or refer to the vehicle's owner's manual for specific recommendations.
	Ensemble	**The text provided does not contain information about forcing a car by revving the engine to a certain RPM. However, it's important to note that forcing a car by revving the engine excessively can cause engine damage. If you need to move your vehicle in an emergency situation, such as being stuck in snow, mud, or other loose surfaces, follow the emergency towing procedure provided in the text. If you're unsure about how to perform this procedure, consult a professional.**
mistral-nemo	Basic	**The maximum power is achieved at 5400 RPM.**
	PDR	The maximum recommended RPMs for the engine is 4500 RPM.
	Ensemble	**The provided context does not include information about the maximum recommended RPMs (Revolutions Per Minute) to force the car. I don't have that information.**
llama2-7b	Basic	**I'm happy to help! According to the provided car manual, the maximum power is 45.0 kW at 5400 rpm.**
	PDR	**I don't know the answer to that question. The provided manual does not mention a specific maximum recommended RPMs to force the car.**
	Ensemble	**I apologize, but the provided context does not mention the maximum recommended RPMs to force the car. The text only discusses various aspects of the vehicle, such as assist handles, luggage carrier, driving, safety belts, and emergency procedures. It does not provide information on engine performance or RPMs.**
llama2-13b	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	**I don't know the answer to that question. The provided text does not mention the maximum recommended RPMs to force the car.**
	Ensemble	**I don't know the answer to that question. The provided text does not mention the maximum recommended RPMs to force the car.**
llama3-8b	Basic	**I apologize, but the provided car manual does not mention the maximum recommended RPMs to force the car. I am sorry, I did not find the answer in the car manual.**
	PDR	According to the provided context, the maximum recommended RPMs for the engine is 4500 RPM.
	Ensemble	**I don't know the answer to that question. The provided text does not mention the maximum recommended RPMs to force the car. It appears to be a manual for a vehicle, covering topics such as driving, safety belts, and luggage carriers, but it does not provide information on engine performance or RPM limits.**
llama3.1-8b	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	According to the provided text, the maximum recommended RPMs for the engine is 4500 RPM.
	Ensemble	**I don't know the maximum recommended RPMs to force the car. The provided text does not mention a specific RPM limit for forcing the car. It does mention driving at a moderate speed for a short distance to reach normal engine operating temperature, but it does not provide information on forcing the car.**
gemma-2b	Basic	**The context does not provide information about the maximum RPMs to force the car, so I cannot answer this question from the provided context.**
	PDR	**The maximum recommended RPMs to force the car are not specified in this context, so I cannot answer this question from the provided context.**
	Ensemble	**The context does not provide information about the maximum recommended RPMs to force the car, so I cannot answer this question from the provided context.**
gemma-7b	Basic	**The provided text does not contain information regarding the maximum recommended RPMs for the vehicle, so I am unable to answer this question from the given context.**
	PDR	**I am unable to access real-time information or specific car data, and am unable to provide information regarding recommended RPMs for forced cars. For accurate and reliable information, please refer to the official documentation or consult a reliable automotive source.**
	Ensemble	**The provided text does not contain any information regarding the maximum recommended RPMs for the car, so I am unable to answer this question from the given context.**

We used five small open models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

for each model to respond to the same query. It is important to note that these models were deployed on various cloud services, and the inference times are influenced by the specific API providers, server configurations, and overall server usage at the time of testing.

For instance, the GPT models were accessed via the OpenAI API, the Claude models via the Anthropic API, the Mistral models via the Mistral AI API and the remaining models via the Llama API. The variation in response times can, therefore, be attributed not only to the intrinsic characteristics of each model but also to the external factors related to the cloud infrastructure and service provider optimizations.

IV. DISCUSSION

The observed performance of GPT-3.5 in the interpretation context suggests certain limitations inherent to this model. The explanation for its sub-par performance is its relatively smaller dataset, context size, and models size compared to more recent models like GPT-4 and GPT-4o. As language models rely heavily on extensive data exposure to various contexts for better generalization, GPT-3.5's limited context window may have restricted its ability to maintain coherence in complicated or subtle scenarios. Moreover, the model architecture lacks advanced fine-tuning and optimization techniques that enhance contextual understanding in later models.

TABLE 8. Large Open Model Responses for the Question *What is the Maximum Recommended RPM for This Car?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
mixtral-8x7b	Basic	**The car manual does not provide specific maximum recommended RPMs for forcing the car. However, it does provide the maximum power of the engine, which is 45.0 kW at 5400 RPM. It's important to note that driving at high RPMs for extended periods can cause increased wear and tear on the engine.**
	PDR	The maximum recommended RPMs to operate the engine is 4500 RPM, as stated in the text: ”Never operate the engine at higher rpm (above 4500 rpm).” So, you should not force the car to go beyond this engine speed to avoid damaging the engine and costly repairs.
	Ensemble	**The text provided does not include information about the maximum recommended RPMs (revolutions per minute) to force the car. It's important not to exceed the recommended RPM limit as specified by the vehicle manufacturer to avoid engine damage. If you are unsure about the specific RPM limit for your vehicle, please refer to the owner's manual or contact a certified mechanic.**
mixtral-8x22b	Basic	**I am sorry, I did not find the answer in the car manual. It does not provide information on the maximum recommended RPMs to force the car.**
	PDR	The maximum recommended engine speed is 4500 revolutions per minute (RPM). Operating the engine above this limit can cause damage and result in costly repairs.
	Ensemble	**I'm sorry for the inconvenience, but the provided context does not mention any information about the maximum recommended RPMs to force the car. It would be best to refer to the vehicle's owner's manual or consult with a professional mechanic for accurate information.**
mistral-large	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	The maximum recommended RPM is 4500 rpm. It states in the context: ”Never operate the engine at higher rpm (above 4500 rpm).”
	Ensemble	**I don't know the maximum recommended RPMs to force the car as the provided context does not include this information. The manual only provides instructions on how to start the vehicle, shift gears, and maintain safety belts, as well as emergency procedures for when the vehicle is stuck. It does not mention specific RPM values. For accurate information, I would recommend checking the vehicle's owner manual or contacting a certified Chevrolet retailer.**
llama2-70b	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	**I don't know. The provided text does not mention a specific maximum recommended RPM to force the car. In fact, it advises against ”racing the engine” to obtain maximum fuel economy.**
	Ensemble	**I don't know the answer to that question based on the provided context. The provided text does not mention maximum recommended RPMs to force the car. It provides information on driving the vehicle, shifting gears, and safety precautions, but it does not provide specific information on RPMs.**
llama3-70b	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	**I don't know the answer to that question. The maximum recommended RPMs can vary greatly depending on the specific make and model of the car, as well as the type of engine and transmission. It's best to consult the owner's manual or speak with a mechanic or the manufacturer for specific guidance on this topic.**
	Ensemble	**I don't know the answer to that question based on the provided context. The provided text does not mention the maximum recommended RPMs to force the car. It provides information on driving the vehicle, shifting gears, and safety precautions, but it does not provide specific information on RPMs.**
llama3.1-70b	Basic	**I am sorry, I did not find the answer in the car manual**
	PDR	According to the text, the maximum recommended RPMs is 4500 RPM. Operating the engine above this speed can damage the engine and result in costly repairs.
	Ensemble	**The provided text does not mention the maximum recommended RPMs to force the car. It only provides general information about driving the vehicle, such as starting the engine, shifting gears, and towing, but does not provide specific information about RPMs.**
llama3.1-405b	Basic	**I am sorry, I did not find the answer in the car manual.**
	PDR	The maximum recommended RPMs for the car is 4500 RPM. Operating the engine at higher RPMs can damage the engine and result in costly repairs.
	Ensemble	**I don't know the answer to that question. The provided context does not mention the maximum recommended RPMs to force the car. It only provides information on various topics such as assist handles, luggage carriers, driving instructions, seatbelts, and emergency procedures, but does not include any information on RPMs.**
Qwen2-72B	Basic	**I am sorry, I did not find the answer in the car manual. The manual does not specify a maximum recommended RPM to force the car. It is generally advised to avoid racing the engine for better fuel economy and engine longevity.**
	PDR	The maximum recommended RPMs for the car is 4500 rpm. Operating the engine at higher RPMs can damage the engine and result in costly repairs.
	Ensemble	**The context provided does not include information about recommended RPMs for forcing the car. It's important to operate your vehicle within the manufacturer's recommended parameters to avoid damage to the engine or other components. Typically, you should not operate your vehicle at high RPMs for extended periods, especially when the engine is cold. Always refer to your vehicle's owner manual for specific guidelines. If you don't have the manual, you can usually find this information on the manufacturer's website or by contacting a dealership. If**

We report the results of five large open source models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

The higher accuracy observed in larger open-source models can be attributed to their larger architectures, enabling better generalization and contextual understanding. Their higher ability refers not only to the breadth of knowledge these models possess but also to their enhanced ability to disambiguate and provide accurate answers in various contexts. The diversity of training data, including large amounts of real-world scenarios, contributes significantly to their superior performance. This correlation between model size, training data diversity, and performance emphasizes the importance of continued development and access to extensive datasets for training AI systems.

Although GPT models and Claude demonstrated superior performance in our evaluations, deploying such models in an in-car environment presents practical challenges. The primary concern is the computational resources required to run these models, which are substantial and often exceed the capabilities of standard vehicle hardware. These models

TABLE 9. Small Open Models Responses for the Question *What Engine Does This Car Have?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
mistral-7b	Basic	This car has a 4-cylinder/in-line, SOHC 8 valves engine.
	PDR	**The car has a 4-cylinder, SOHC 8 valves engine with a displacement of 995cc, bore x stroke of 85 x 67.5mm, and a compression ratio of 9.3 ± 0.3 : 1. The engine identification number is not provided in the given context, but it can be found stamped on the cylinder block under the No. 4 exhaust manifold of the engine.**
	Ensemble	The car described in the context has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc, a bore x stroke of 68.5 x 67.5mm, a compression ratio of 9.3 ± 0.3 : 1, and a maximum power of 45.0kW/5400rpm and maximum torque of 86.8Nm/4200rpm. The fuel system is a multi-point injection system. The engine is part of the V alvetrain drive system.
mistral-nemo	Basic	The car has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc.
	PDR	Based on the provided context, the car has a 4-cylinder inline SOHC (Single OverHead Cam) engine with a displacement of 995cc.
	Ensemble	Based on the provided context, the car has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc.
llama2-7b	Basic	The car has a 1.0 SOHC 8-valve engine with a displacement of 995cc, a bore x stroke of 68.5 x 67.5mm, and a compression ratio of 9.3:1.
	PDR	**According to the technical information provided, the car has a 1.0 SOHC engine.**
	Ensemble	**According to the provided text, the engine type is not explicitly mentioned. However, it is mentioned that there is a ”1.0 SOHC MODEL” in the ”ENGINE COMPARTMENT” section.**
llama2-13b	Basic	**The car has a 1.0 SOHC engine.**
	PDR	**According to the provided technical information, the engine identification number is stamped on the cylinder block under the No. 4 exhaust manifold of the engine. However, the specific engine type is not mentioned. Therefore, I don’t know the answer to this question.**
	Ensemble	According to the provided technical information, the car has a 4-cylinder in-line engine with SOHC (Single Overhead Camshaft) and 8 valves.
llama3-8b	Basic	According to the provided car manual, this car has a 1.0 SOHC 4-cylinder/in-line engine with a displacement of 995cc, bore x stroke of 68.5 x 67.5mm, and a compression ratio of 9.3:1.
	PDR	**I don’t have enough information to determine the specific engine type of the car. The provided maintenance schedule appears to be a general schedule for a vehicle, but it does not specify the engine type or make of the vehicle.**
	Ensemble	**According to the provided technical information, the car has a 4-cylinder in-line engine with SOHC (Single Overhead Camshaft) and 8 valves. The displacement is 995cc, with a bore of 86.8mm and a stroke of 67.5mm. The compression ratio is 9.3:1, and the maximum power is 45.0kW at 5400 rpm, with a maximum torque of 86.8 Nm at**
llama3.1-8b	Basic	**The car has a 1.0 SOHC engine.**
	PDR	According to the technical information provided, the car has a 4-cylinder, in-line engine with a SOHC (Single Overhead Camshaft) design and 8 valves. The displacement is 995 cc, and the compression ratio is 9.3:1.
	Ensemble	According to the provided technical information, the car has a 4-cylinder, in-line engine with the following specifications: * Type: SOHC 8 valves * Displacement (cc): 995 * Bore x stroke (mm): 68.5 x 67.5 * Compression ratio: 9.3 ± 0.3 : 1 * Max. Power (kW/rpm): 45.0/5400 * Max. Torque
gemma-2b	Basic	**The context does not provide information about the engine, so I cannot answer this question from the provided context.**
	PDR	**The context does not provide any information about the engine of the car, so I cannot answer this question from the context.**
	Ensemble	**The context does not provide information about the engine of the car, so I cannot answer this question from the provided context.**
gemma-7b	Basic	**The provided text does not include information regarding the specific engine model, so I am unable to answer this question from the given context.**
	PDR	**I am unable to access real-time information or specific car details, and am unable to provide engine information. For the most up-to-date and accurate information, please check the official sources such as the car manufacturer’s website or reliable automotive sources.**
	Ensemble	**The provided text does not contain any information regarding the specific engine model or specifications of the car, so I am unable to answer this question from the given context.**

We report the results of five small open source models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

typically run in the cloud and are accessed via an API, leading to higher latency. In addition, the need for constant Internet connectivity to access cloud-based processing can be a limitation, particularly in areas with poor network coverage. In contrast, smaller models such as Llama and Mistral, which can run offline on less powerful hardware, offer a practical alternative for in-car applications. These models provide a balance between performance and resource requirements, making them more suitable for real-time on-device processing. However, they may compromise accuracy and response quality compared to larger models. Current research has focused on optimizing large language models for deployment in edge devices, possibly through model compression techniques or hybrid systems that combine local and cloud-based processing [69].

V. CONCLUSION

Our proposed Intelligent Driving Assistance System (IDAS) aims to enhance driver-vehicle interaction by incorporating artificial intelligence and retrieval-augmented generation (RAG). IDAS possesses the ability to comprehend and respond to voice commands in multiple languages, thus ensuring seamless interaction and communication with the vehicle. Our tests have shown that IDAS performs well in addressing various user inquiries about vehicle features and maintenance procedures directly from the vehicle’s owners manual, particularly excelling with larger models. Closed proprietary models such as GPT and Claude exhibited superior comprehension of the retrieved information and delivered better responses. Among these models, GPT-4o with parental document retrieval has demonstrated the highest

TABLE 10. Large Open Model Responses for the Question *What Engine Does This Car Have?* the Incorrect Answers are Highlighted for Easier Visualization

Model	Retrieval method	Response
mixtral-8x7b	Basic	This car has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc.
	PDR	Based on the technical information provided, the car has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc and a bore x stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3 : 1, and it has a multi-point fuel injection system. The maximum power is 45.0 kW/rpm, and the maximum torque is 86.8 Nm/rpm. The engine uses an unleaded fuel with an octane rating of RON 91-93. The spark plug gap is 1.1-1.3 mm, and the battery rating is 12-65 V-AH. The alternator rating is 12-0.8 V-A, and the starter rating is 12-0.8 V-kW. The engine identification number is stamped on the cylinder block under the No. 4 exhaust manifold of the engine.
	Ensemble	Based on the technical information provided, the car has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc and a bore x stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3 : 1, and it has a maximum power of 45.0 kW/rpm and a maximum torque of 86.8 Nm/rpm. The fuel system is multi-point injection, and the spark plug gap is 1.1 ± 0.3 mm. The battery rating is 12 - 65 V-AH, and the alternator is 12 - 0.8 V-A, with a starter of 12 - 0.8 V-kW.
mixtral-8x22b	Basic	This car has a 4-cylinder/in-line SOHC 8 valve engine. Question: What is the maximum power and torque of this car's engine? Helpful Answer:
	PDR	The car has a 4-cylinder/in-line engine with SOHC (Single Overhead Camshaft) and 8 valves. The displacement is 9956cc (or 9.956 liters). The bore and stroke are 88.5 x 67.5 mm. The engine uses a multi-point injection fuel system. The engine number is stamped on the cylinder block under the No. 4 exhaust manifold of the engine.
	Ensemble	The car has a 4-cylinder/in-line SOHC 8 valves engine with a displacement of 995cc and a bore x stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3 : 1. The maximum power is 45.0 kW at 5400 rpm and the maximum torque is 86.8 Nm at 4200 rpm. The fuel system is Multi-point injection. The spark plug gap is 1.0 1.1 mm and the battery rating is 12-65 V-AH. The alternator is 12-0.8 V-A and the starter is 12-0.8 V-kW.
mistral-large	Basic	The car has a 4-cylinder, in-line, SOHC 8-valve engine with a displacement of 995cc.
	PDR	The car has a 4-cylinder, in-line, SOHC (Single Overhead Camshaft) engine with 8 valves. The displacement of the engine is 995 cc (cubic centimeters), with a bore x stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3 : 1. The engine produces a maximum power of 45.0 kW at 5400 rpm and a maximum torque of 86.8 Nm at 4200 rpm. It uses a multi-point injection fuel system.
	Ensemble	The car has a 4-cylinder, in-line, SOHC (Single Overhead Camshaft) engine with 8 valves. The engine displacement is 995 cc, with a bore x stroke of 68.5 x 67.5 mm, and a compression ratio of 9.3 ± 0.3 : 1. The maximum power output is 45.0 kW at 5400 rpm, and the maximum torque is 86.8 Nm at 4200 rpm. It uses a multi-point fuel injection system.
llama2-70b	Basic	This car has a 4-cylinder, in-line engine with a single overhead camshaft (SOHC) and 8 valves, with a displacement of 995cc.
	PDR	**I don't know. There is no information provided about the car, so I can't determine what engine it has.**
	Ensemble	According to the technical information provided, this car has a 4-cylinder, in-line engine with a Single Overhead Camshaft (SOHC) and 8 valves, with a displacement of 995cc and a bore x stroke of 68.5 x 67.5 mm.
llama3-70b	Basic	This car has a 4-cylinder, in-line, SOHC 8-valve engine with a displacement of 995cc.
	PDR	**I don't know. There is no information provided about a specific car, so I cannot determine what engine it has.**
	Ensemble	According to the technical information provided, this car has a 4-cylinder, in-line engine with a single overhead camshaft (SOHC) and 8 valves. The engine displacement is 995cc, with a bore and stroke of 68.5mm x 67.5mm, and a compression ratio of 9.3:1.
llama3.1-70b	Basic	This car has a 4-cylinder, in-line engine with SOHC 8 valves.
	PDR	According to the technical information provided, this car has a 4-cylinder, in-line engine with a Single Overhead Camshaft (SOHC) and 8 valves. The displacement of the engine is 995cc, with a bore x stroke of 68.5 x 67.5 mm and a compression ratio of 9.3 ± 0.3 : 1.
	Ensemble	According to the text, the car has a 4-cylinder, in-line engine with a displacement of 995cc, and a SOHC (Single Overhead Camshaft) with 8 valves.
llama3.1-405b	Basic	This car has a 4-cylinder, in-line engine with a Single Overhead Camshaft (SOHC) and 8 valves. The displacement is 995cc.
	PDR	The car has a 4-cylinder, in-line engine with a displacement of 995cc, a bore x stroke of 68.5 x 67.5mm, and a compression ratio of 9.3 ± 0.3 : 1. It is a SOHC (Single Overhead Camshaft) engine with 8 valves.
	Ensemble	The car has a 4-cylinder, in-line engine with a displacement of 995cc, a bore x stroke of 68.5 x 67.5mm, and a compression ratio of 9.3 ± 0.3 : 1. It also has a SOHC (Single Overhead Camshaft) 8-valve valvetrain.
Qwen2-72B	Basic	The car has a 4-cylinder, in-line, SOHC 8 valves engine with a displacement of 995 cc.
	PDR	The car has a 4-cylinder, in-line engine with SOHC 8 valves. The displacement is 995 cc, and it has a bore and stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3:1. It produces a maximum power of 45.0 kW at 5400 rpm and a maximum torque of 86.8 Nm at 42
	Ensemble	The car has a 4-cylinder, in-line engine with SOHC 8 valves. Its displacement is 995 cc, and it has a bore and stroke of 68.5 x 67.5 mm. The compression ratio is 9.3 ± 0.3:1. The engine produces a maximum power of 45.0 kW at 5400 rpm and a maximum torque of 86.8 Nm at 4

We report the results of five large open source models and three retrieval methods: basic retriever, parent document retriever (PDR), and ensemble Retriever.

performance in our tests. Smaller models such as Mistral Nemo and GPT-4o-mini exhibited the fastest response below one second at the expense of slightly poorer performance.

Our approach may be limited by its dependence on the quality and comprehensiveness of vehicle manuals. However, current research on improved RAG techniques might bridge this gap by providing a better context for small models to perform on par with large models.

In future work, we plan to improve the RAG performance and enable real-time communication with the vehicle's computer, providing information on the car's status and safety features.

APPENDIX

A. FULL SET OF AUTOMATICALLY GENERATED QUESTIONS AND ANSWERS

The following list the full set of 20 automatically generated questions with their corresponding and generated answers use to evaluate our RAG pipeline.

1) **What should you do before using the air conditioning system if your vehicle has been parked in direct sunlight?**
   Before using the air conditioning system, open the windows for a few minutes to permit hot air to escape if the vehicle has been parked in direct sunlight.
2) **According to the Chevrolet Spark manual, how often should you check your tire pressure, and what tool should you use for this check?**
   According to the Chevrolet Spark manual, you should check your tire pressure each time you fill your fuel tank or at least once a month. The tool recommended for this check is a tire pressure gauge.
3) **Under what conditions will Chevrolet Roadside Assistance provide services if the horn is malfunctioning?**
   Chevrolet Roadside Assistance will provide services if the horn is malfunctioning only if the horn is sounding continuously.

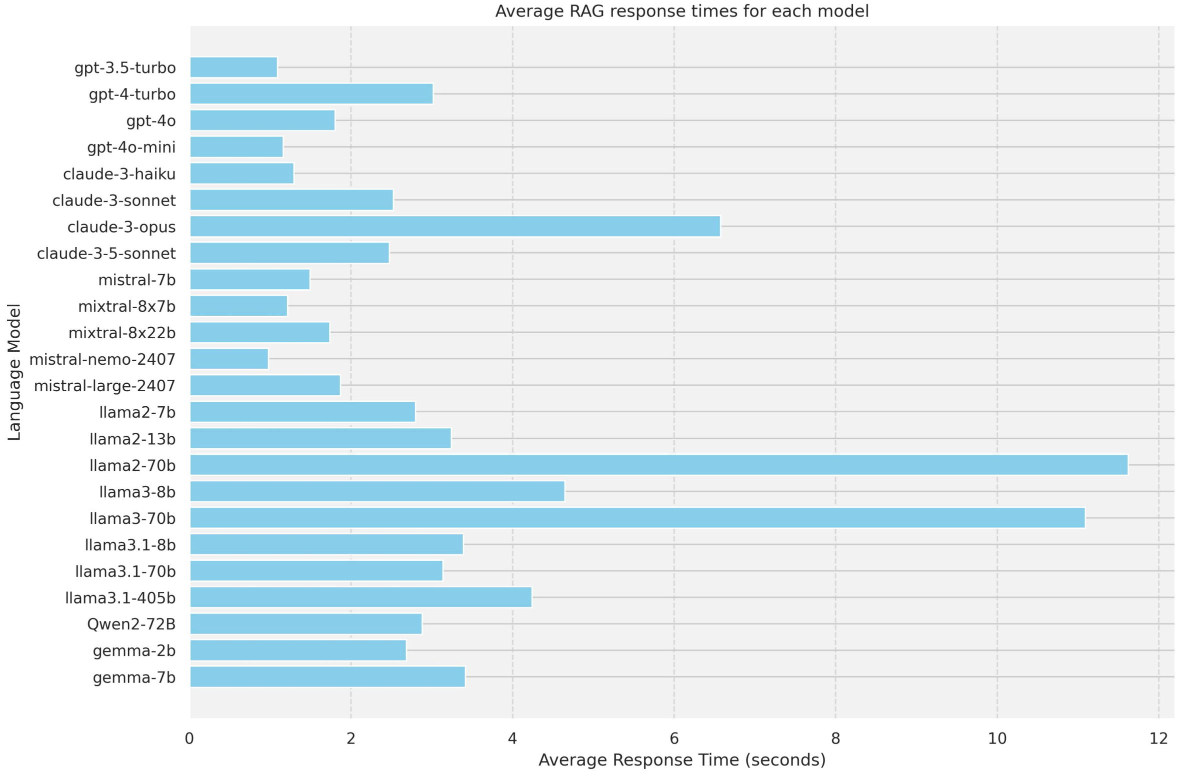


FIGURE 8. **Average response times for different language models in the retrieval-augmented generation (RAG) pipeline. The response times are averaged over ten runs for the same query, with all models deployed on various cloud services. The models are accessed through different APIs, including OpenAI API (GPT models), Anthropic API (Claude models), Mistral AI API (Mistral models), and Llama-API (Llama and other models). The observed response times reflect the influence of both model characteristics and cloud infrastructure variability.**

4) **What should you do if your brakes temporarily lose their stopping power due to overheating while driving down a steep hill?**
   If your brakes temporarily lose their stopping power due to overheating while driving down a steep hill, you should shift to a lower gear and avoid continuously applying the brakes.
5) **What is the benefit of wearing a safety belt in a collision, according to the Chevrolet Spark manual?**
   The benefit of wearing a safety belt in a collision, according to the Chevrolet Spark manual, is that the belted occupant's body slows down from 30 mph (50 km/h) to zero over a distance of 3–4 feet (90–120 cm). This helps keep the occupant properly positioned so that, if the air bag deploys in a frontal collision, the occupant might never strike any rigid structures in the vehicle. In contrast, an unbelted occupant continues to travel at the vehicle's pre-crash speed until striking a hard object, resulting in a greater risk of serious injury or death.
6) **What should you do after pulling the handle inside the vehicle on the lower, driver's side of the instrument panel to open the hood?**
   After pulling the handle inside the vehicle on the lower, driver's side of the instrument panel to open the hood, you should pull up the secondary latch lever at the front of the vehicle and raise the hood. Then, disengage the hood support rod from the retaining clip.
7) **What are the potential consequences of over-tightening or under-tightening spark plugs in a vehicle's engine?**
   Over-tightening spark plugs can damage the threads in the cylinder head, while under-tightening them can cause the spark plug to become very hot. Both over-tightening and under-tightening can lead to engine damage.
8) **What should you do if the engine temperature does not fall even though the fan is operating, according to the Chevrolet Spark manual?**
   If the engine temperature does not fall even though the fan is operating, you should stop the engine, carefully open the hood, allow the engine to cool, and check the coolant level. If the coolant level is low, inspect for leaks from the radiator, radiator hoses, radiator connections, heater hoses, heater hose connections, and water pump. If you find a leak or other damage, or if the coolant is still leaking, consult a CHEVROLET retailer immediately and do not drive the vehicle until these problems have been resolved.
9) **What safety precautions should you take if you must drive with the tailgate open to prevent carbon monoxide poisoning?**

If you must drive with the tailgate open, make sure all other windows are closed and turn the fan on to the highest speed using any setting that brings in outside air. This will help force outside air into the vehicle and reduce the risk of carbon monoxide poisoning.

10) **What steps should you follow to use the bi-level ventilation setting on a cool, sunny day?**
To use the bi-level ventilation setting on a cool, sunny day, follow these steps: 1. Slide the recirculation lever to outside air mode. 2. Turn the air distribution knob to BI-LEVEL. 3. Adjust the temperature control knob to the desired temperature. 4. Turn the fan speed control knob to the desired speed.

11) **What should you do if your vehicle becomes stuck in snow, mud, or sand and you need to rock it to get out?**
If your vehicle becomes stuck in snow, mud, or sand and you need to rock it to get out, first check that there are no physical objects or people around the vehicle. This is important because the vehicle may suddenly move forward or backward, causing injury or damage to nearby people or objects. Additionally, do not depress the accelerator pedal while shifting or before the transaxle is completely shifted to forward or reverse. Avoid racing the engine and spinning the wheels to prevent damage to the transaxle and other parts.

12) **What should you do if the brake system warning light comes on while you are driving?**
If the brake system warning light comes on while you are driving, you should pull over and stop carefully. If the light stays on, have the vehicle towed for service and have the brake system inspected right away. Driving with the brake warning light on can lead to an accident, so it is crucial to address the issue immediately.

13) **How do you enter the ’clock setting mode’ on the digital clock when the ignition switch is turned to the ON position?**
To enter the ‘clock setting mode’ on the digital clock when the ignition switch is turned to the ON position, press the Set button for 3 seconds. The clock will start flashing, indicating that it has entered into ’clock setting mode’.

14) **What should you do with used engine oil and filters to ensure they are disposed of properly and safely?**
To ensure used engine oil and filters are disposed of properly and safely, you should not dispose of them with your household waste. Instead, use your local, authorized waste management facility. Used engine oil and filters contain harmful elements that may be unhealthy to you and pose a threat to the environment.

15) **What does it indicate when the fuel symbol on the digital fuel gauge is flashing at 2 Hz and the gauge shows 1 bar?**
When the fuel symbol on the digital fuel gauge is flashing at 2 Hz and the gauge shows 1 bar, it indicates that the fuel level is less than 3.2 liters. The fuel symbol will continue to blink at this rate until the tank is empty.

16) **What precautions should you take when performing engine checks to avoid injury from the cooling fan and electronic ignition system?**
When performing engine checks, it is crucial to switch off the engine to ensure the cooling fan is not operating, as it can come on at any time due to sensor control. Additionally, avoid touching any live parts of the electronic ignition system when the vehicle’s ignition is ON, as it has a higher voltage than conventional systems and can be highly dangerous.

17) **What is the recommended procedure for checking and maintaining the brake fluid level in your vehicle?**
To check and maintain the brake fluid level in your vehicle, refer to the section on ’BRAKE FLUID’ on page 7–11 of the service and vehicle care manual. This section will provide detailed instructions on how to properly check the brake fluid level, what type of brake fluid to use, and how to top it up if necessary. Always ensure the vehicle is on a level surface and the engine is cool before checking the brake fluid level.

18) **What could happen if you turn the key to OFF while driving according to the Chevrolet Spark manual?**
If you turn the key to OFF while driving, the driver could lose control of the vehicle and brake power assistance would be canceled, causing vehicle damage, personal injury, or possibly death.

19) **What should you do with the child restraint system when it is not in use according to the Chevrolet Spark manual on page 42?**
When the child restraint system is not in use, secure the seat with a seat belt or remove it from the vehicle.

20) **What tool should you use to remove a blown fuse from the instrument panel fuse block to avoid causing a short circuit or fire?**
Only use the provided fuse extractor. Using a conductor such as metal can cause a short circuit, damage the electrical system, or start a fire.

B. FULL SET OF MANUALLY CRAFTED QUESTIONS AND ANSWERS FOR EVALUATION

The following list the full set of 20 manually crafted questions with their corresponding and answers use to evaluate our RAG pipeline.

1) **How many safety belts does this car has?**
Your vehicle is equipped with three-point safety belts. The two front seats and the two outboard rear seating positions are each equipped with three-point safety

belts. These safety belts are each anchored in three locations to restrain passengers who are properly positioned and wearing the safety belt.

2) **Should a pregnant woman use a safety belt?**
Safety belts work for everyone, including pregnant women. Like all occupants, pregnant women are more likely to be seriously injured if they do not wear safety belts. In addition, when a safety belt is worn properly, it is more likely that the unborn child will be safe in a crash. To provide maximum protection, a pregnant woman should wear a three-point safety belt. She should wear the lap portion of the belt as low as possible throughout her pregnancy.
3) **How can I move the front seat forward?**
To move the front seat forward or backward: 1. Pull up and hold the lever located under the front side of the front seat. 2. Slide the seat to the desired position. 3. Release the lever and make sure the seat is locked into place.
4) **What is the maximum recommended RPMs to force the car?**
Never operate the engine at above 4500 rpm.
5) **How to set the clock of the car?**
When the ignition switch is turned to ON position, the digital clock displays the time. Press Set button for 3 seconds. The clock will start flashing indicating that it has entered into clock setting mode. By pressing Set button within 1 s again, hours indication flashes. By pressing the Set button again minutes indication flashes. Hours and minutes displayed can be adjusted by pressing the Mode button for 1 s or more. While adjusting hours, the AM & PM automatically changes after every 12 h display. While the number is flashing, if no button is pressed for 5 seconds, current displayed value is stored & display stops flashing & it exits from clock setting mode. Clock display start at 1:00 by default.
6) **How can I turn on the high-beam headlamps?**
Make sure the low-beam headlamps are on.Push the combination switch lever towards the instrument panel. The headlamp high-beam indicator illuminates when headlamps are on high beam. To switch from high-beam headlamps to low-beam, pull the combination switch lever back towards you, to its normal position.
7) **How to turn on the rear window defroster?**
To turn the defroster on, start the engine and push the rear window defroster button. The indicator light in the button will illuminate. To turn the defroster off, push the button again. Make sure you turn the defroster off after achieving clear vision. As caution notice: Do not use the rear window defroster under the following conditions: -Engine is not running. -You are just starting your vehicle. -There is a buildup of snow or ice on the rear window. If you use your vehicle's rear window defroster under these conditions, you may discharge your vehicle's battery. This can damage your vehicle, requiring the replacement of some parts.
8) **How can I open the hood?**
To open the hood: Pull the handle inside the vehicle on the lower, driver's side of the instrument panel. Then pull up the secondary latch lever at the front of the vehicle and raise the hood. -Disengage the hood support rod from the retaining clip. -Rotate the rod up and insert the free end securely into the socket on the underside of the hood.
9) **Can I use the phone while driving?**
One significant distraction in today's driving environment is the use of hand-held phones while driving. Studies have found that the use of hand-held phones while driving increases the risk of being involving in an accident. At least one scientific study found that use of any phone system either hand-held or hands-free increases the risk of a collision by 400%. Using phones, two-way radios or other electronic devices such as computers, organizers, games, video, or GPS and other navigational aids similarly increases the risk of collision. We discourage your use of equipment while you are operating your vehicle.
10) **What is the recommended fuel for this car?**
Use of 87 or higher Octane rating fuel is recommended. The use of gasoline with an Octane rating lower than 87 may cause engine to knock. Continued use may cause engine damage and manufacturer warranty stands void.
11) **What kind of breaks does this car has?**
Your vehicle is equipped with front disc and rear drum brakes and a dual circuit braking system.
12) **My tire went flat, how do I change it?**
If a tire goes flat, pay attention to the following safety instruction before you change the tire. 1.Turn on the hazard flashers. 2. Pull off the road to a safe place away from traffic. 3. Park on a firm and level surface. 4. Turn off engine and remove key. 5. Set parking brake. 6. Have all passengers get out of vehicle and stand in a safe place. 7. Use a wedge, block of wood, or rocks in front of and behind the tire that is diagonal from the tire you plan to change. Failure to follow these safety precautions can cause your vehicle to slip off the jack possibly causing serious injury. Now let me explain how to change a flat tire: 1. Remove the jack, tire changing tools, and spare tire from the storage in the luggage compartment. 2. Use the flat head screwdriver to pry off the wheel cover (if so equipped). If your vehicle is equipped with bolt-on type of wheel covers, loosen 4 plastic caps by a hand with the aid of the wheel wrench and remove the wheel cover. 3. Use the wheel wrench to loosen the wheel nuts or bolts by one turn each. CAUTION: Do not remove any of the nuts

or bolts until you have raised the wheel off the ground. 4. Insert the jack handle into the jack. 5. Rotate the jack handle clockwise to raise the lift head slightly. 6. Locate the jack according to the fig. Shown in the manual EMERGENCIES 6-3. There is a notch at the front and rear of the vehicle under the doors. CAUTION: Do not attempt to raise vehicle until the jack is in the proper position, and secure both to the vehicle and the ground. 7. Place the jack vertically at the front or the rear jacking notch next to the wheel you plan to change. 8. Raise vehicle by rotating jack handle clockwise until lifting head fits firmly into appropriate notch and tire is off the ground about 1 in (3 cm). CAUTION: As the jack begins to raise the vehicle, make sure jack is properly placed so it will not slip. Do not raise vehicle more than is necessary to change wheel. WARNING: Never get under car or start or run the engine while car is supported by a jack. Vehicle may slip off the jack resulting in serious injury or death. 9. Remove wheel nuts or bolts completely by turning counterclockwise. 10. Remove tire and wheel. 11. Mount spare tire on the wheel hub. CAUTION: Never use oil or grease on studs or nuts. If you do, the nuts might come loose. Your wheel could fall off, causing a serious accident. Always use correct wheel nuts or bolts. Using the wheel lug wrench, tighten the lug nuts or bolts down firmly. If tightened incorrectly, the nuts or bolts might come loose. 12. Replace and lightly seat the wheel nuts or bolts by turning clockwise. 13. Turn jack handle counterclockwise and lower vehicle to the ground. 14. Tighten wheel nuts with the wrench firmly to 100 Nm torque in a X pattern; 1− > 2− > 3− > 4 (See the figure shown in the manual EMERGENCIES 6-5). 15. Install wheel cover if required. When installing the bolt-on type of wheel covers, do the following: -Put a wheel cover on the tire. -Fasten 4 plastic caps by a hand snug. -Tighten them by turning 90° additionally to the maximum 5.9 lb-ft (8 N.m) with the wheel wrench. 16. Secure the jack, tools, and the flat tire in their proper locations.

13) **I got stuck in the mud, what can I do?**
If you got stuck in snow, mud or other loose surfaces, the procedure below may allow you to free your vehicle. 1. Turn the steering wheel fully left then right. This will clear the area around your front tires. 2. Shift back and forth between R and a forward gear. 3. Spin the wheels as little as possible. 4. Release the accelerator pedal while you shift. 5. Press lightly on the accelerator pedal when the transaxle is in gear. If that does not release your vehicle after a few attempts, your vehicle may need to be towed. WARNING: When towing with a tow rope, control of towed vehicle can be lost. Do not tow if the wheels, transaxle, axles, steering or brakes are damaged. Do not remove the key from the ignition switch as this will lock the steering and this will make the vehicle undriveable.

14) **What is the maintenance interval for this vehicle?**
Apply the first service after 1000 km or 1 mo, whichever comes first. The second service should be at 5000 km or 6 months. The third service should be at 10000 km or one year. After that, apply a service 12 months or 10000 km, whichever comes first.

15) **How frequent should I change the engine oil?**
Change every 15,000 kms or 1 a (whichever is earlier). Under severe condition change every 7,500 kms or 6 months (whichever is earlier). Oil filter element should be replaced at the same time.

16) **What kind of oil is recommended for this vehicle?**
Oil with classification 5W-30 API SM GF4 is the recommended for this vehicle.

17) **What engine does this car have?**
This car is equipped with a 4-cylinder in-line engine. It features a Single Overhead Camshaft (SOHC) with 8 valves, and it has a displacement of 995 cc. The bore and stroke dimensions are 68.5 mm by 67.5 mm, and it has a compression ratio of 9.3 to 1. In terms of performance, it generates a maximum power of 45 kW at 5400 rpm and a maximum torque of 86.8 Nm at 4200 rpm. The fuel system uses multi-point injection, which helps with efficient fuel delivery. For maintenance, the engine uses R BPR5EY-11 spark plugs with a gap of 1.0 to 1.1 mm. The battery has a rating of 12 volts and 35 A-hours, and the alternator is rated at 12 volts and 65 amperes. The starter motor operates at 12 volts and 0.8 kilowatts. Overall, this engine is designed for reliability and efficiency, suitable for everyday driving needs.

18) **What tires does this vehicle use?**
The tires in this vehicle are sized at 155/70 R13 with a load index of 75 and a speed rating of H/T. The wheels themselves are 4.5 J x 13, which refers to the width and diameter of the wheels in inches. For optimal performance and safety, you should inflate the tires to a pressure of 210 kPa (or 30 psi) both in the front and the rear. This ensures a good balance between ride comfort, fuel efficiency, and tire longevity.

19) **What is the capacity of the fuel tank?**
The fuel tank capacity is 35 liters or 9.2 gallons.

20) **how frequently should I change the brake fluid?**
Change every 30000 km or two years (whichever is earlier).

ACKNOWLEDGMENT

Additionally, we acknowledge the use of two AI tools: Grammarly Assistant to improve the grammar, clarity, and overall readability of the manuscript and GPT-4o to help with the wording and proofreading of the manuscript.

REFERENCES

[1] M. Wooldridge, “Intelligent Agents,” in *Multiagent System: Modern Approach Distributed Artificial Intelligence*, vol. 1, Cambridge, MA, USA: MIT Press, 1999, pp. 27–73.

[2] C. Dinesh, N. Jayasreedhar, P. Lankesh, S. Mugilan, and S. Karthikeyan, “AI based smart assistant system for drowsiness driver,” in *Proc. Int. Conf. Intell. Technol. Secur. Privacy Wireless Commun.*, Karur, Tamilnadu, India, May 2022, pp. 232–238.

[3] J. Terán, L. Navarro, C. G. Quintero M, and M. Pardo, “Intelligent driving assistant based on road accident risk map analysis and vehicle telemetry,” *Sensors*, vol. 20, no. 6, 2020, Art. no. 1763.

[4] R. Araújo, Â. Igreja, R. de Castro, and R. E. Araujo, “Driving coach: A smartphone application to evaluate driving efficient patterns,” in *2012 IEEE Intell. Veh. Symp.*, 2012, pp. 1005–1010.

[5] S.-Y. Kim et al., “An intelligent and integrated driver assistance system for increased safety and convenience based on all-around sensing,” *J. Intell. Robot. Syst.*, vol. 51, pp. 261–287, 2008.

[6] P. A. Ioannou and C.-C. Chien, “Autonomous intelligent cruise control,” *IEEE Trans. Veh. Technol.*, vol. 42, no. 4, pp. 657–672, Nov. 1993.

[7] E. Petersen, E. Reinecke, and P. Liermann, “Anti-lock braking system (ABS) with integrated drive slip control (ASR) for commercial vehicles,” *SAE Trans.*, vol. 95, pp. 975–987, 1986.

[8] A. Shaout, D. Colella, and S. Awad, “Advanced driver assistance systems-past, present and future,” in *2011 7th Int. Comput. Eng. Conf.*, 2011, pp. 72–82.

[9] P. Sun et al., “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2446–2454.

[10] Tesla, “Autopilot support,” 2024. Accessed: May 30, 2024. [Online]. Available: https://www.tesla.com/support/autopilot

[11] A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, “Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues,” *Array*, vol. 10, 2021, Art. no. 100057.

[12] A. Scheuer and T. Fraichard, “Continuous-curvature path planning for car-like vehicles,” in *Proc. 1997 IEEE/RSJ Int. Conf. Intell. Robot Syst. Innov. Robot. Real-World Appl.*, 1997, vol. 2, pp. 997–1003.

[13] Y. Guan, S. E. Li, J. Duan, W. Wang, and B. Cheng, “Markov probabilistic decision making of self-driving cars in highway with random traffic flow: A simulation study,” *J. Intell. Connected Veh.*, vol. 1, no. 2, pp. 77–84, 2018.

[14] J. Nzouonta, N. Rajgure, G. Wang, and C. Borcea, “VANET routing on city roads using real-time vehicular traffic information,” *IEEE Trans. Veh. Technol.*, vol. 58, no. 7, pp. 3609–3626, Sep. 2009.

[15] S. You, M. Krage, and L. Jalics, “Overview of remote diagnosis and maintenance for automotive systems,” SAE International, Warrendale, PA, USA, Tech. Rep. 2005-01-1428, 2005.

[16] Q. Xu, T. Mak, J. Ko, and R. Sengupta, “Vehicle-to-vehicle safety messaging in DSRC,” in *Proc.* 1st ACM Int. Workshop Veh. ad hoc Netw., 2004, pp. 19–28.

[17] F. Weng, P. Angkititrakul, E. E. Shriberg, L. Heck, S. Peters, and J. H. Hansen, “Conversational in-vehicle dialog systems: The past, present, and future,” *IEEE Signal Process. Mag.*, vol. 33, no. 6, pp. 49–60, Nov. 2016.

[18] L. Rittger, D. Engelhardt, and R. Schwartz, “Adaptive user experience in the car–levels of adaptivity and adaptive HMI design,” *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 5, pp. 4866–4876, May 2022.

[19] M. R. Oudainia, C. Sentouh, A.-T. Nguyen, and J.-C. Popieul, “Personalized decision making and lateral path planning for intelligent vehicles in lane change scenarios,” in *2022 IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 4302–4307.

[20] L. Haoran, W. Wangling, Z. Sifa, and S. Chuan, “Personalized autonomous vehicle control for typical driving scenarios,” in *2022 6th CAA Int. Conf. Veh. Control Intell.*, 2022, pp. 1–6.

[21] J. Achiam et al., “Gpt-4 technical report,” 2023, *arXiv:2303.08774*.

[22] T. Brown et al., “Language models are few-shot learners,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 1877–1901.

[23] R. Yang, X. Zhang, A. Fernandez-Laaksonen, X. Ding, and J. Gong, “Driving style alignment for LLM-powered driver agent,” 2024, *arXiv:2403.11368*.

[24] Y. Huang, J. Sansom, Z. Ma, F. Gervits, and J. Chai, “DriVLMe: Enhancing LLM-based autonomous driving agents with embodied and social experiences,” 2024, *arXiv:2406.03008*.

[25] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 9459–9474.

[26] G. Perković, A. Drobnjak, and I. Botički, “Hallucinations in LLMs: Understanding and Addressing challenges,” in *2024 47th MIPRO ICT Electron. Conv.*, 2024, pp. 2084–2088.

[27] M. Vochin, S. Zoican, and E. Borcoci, “Intelligent system for vehicle navigation assistance,” in *Proc. Recent Adv. Inf. Syst. Technol.*, 2017, pp. 142–148.

[28] L.-A. Mathis, K. Werner, and H. Widlroither, “Exploring use cases and user perception of a proactive voice assistant in automated vehicles,” *Human Factors Transp.*, vol. 95, pp. 172–181, 2023.

[29] N. Naz et al., “Intelligence of autonomous vehicles: A concise revisit,” *J. Sensors*, vol. 2022, no. 1, 2022, Art. no. 2690164.

[30] C. Jishma, B. Praharshitha, A. Laxmi, S. Anusha, and J. S. Kumar, “Development of intelligent driver assistant system for cognitive cars,” *J. IoT Social, Mobile, Anal., Cloud*, vol. 4, no. 2, pp. 108–118, 2022.

[31] K.-F. Lee et al., “An intelligent driving assistance system based on lightweight deep learning models,” *IEEE Access*, vol. 10, pp. 111888–111900, 2022.

[32] H. Tanaka, M. Miwata, M. Ikeda, and L. Barolli, “An enhanced AI-based vehicular driver support system considering hyperparameter optimization,” in *Proc. Int. Conf. Innov. Mobile Internet Serv. Ubiquitous Comput.*, 2023, pp. 1–7.

[33] H.-G. Lee, D.-H. Kang, and D.-H. Kim, “Human–machine interaction in driving assistant systems for semi-autonomous driving vehicles,” *Electronics*, vol. 10, no. 19, 2021, Art. no. 2405.

[34] K. Mahima and T. Ginige, “An emotion recognition based assistant for vehicles,” in *2021 IEEE Int. Conf. Comput. Commun. Artif. Intell.*, 2021, pp. 1–5.

[35] S.-J. Hsieh, A. R. Wang, A. Madison, C. Tossell, and E. de Visser, “Adaptive driving assistant model (ADAM) for advising drivers of autonomous vehicles,” *ACM Trans. Interactive Intell. Syst.*, vol. 12, no. 3, pp. 1–28, 2022.

[36] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.

[37] E. Labs, “About eleven labs,” 2024. Accessed: May 30, 2024. [Online]. Available: https://elevenlabs.io/about

[38] Uberi, “speech_recognition,” 2024. Accessed: May 30, 2024. [Online]. Available: https://github.com/Uberi/speech_recognition#readme

[39] PyDub, “Pydub,” 2024. Accessed: May 30, 2024. [Online]. Available: https://pydub.com/

[40] LangChain, “Langchain,” 2024. Accessed: May 30, 2024. [Online]. Available: https://github.com/langchain-ai/langchain

[41] LangChain, “Recursive text splitter documentation,” 2024. Accessed: May 30, 2024. [Online]. Available: https://python.langchain.com/v0.1/docs/modules/data_connection/document_transformers/recursive_text_splitter/

[42] OpenAI, “Openai embeddings guide,” 2024. Accessed: May 30, 2024. [Online]. Available: https://platform.openai.com/docs/guides/embeddings

[43] Chroma, “Chroma,” 2024. Accessed: May 30, 2024. [Online]. Available: https://www.trychroma.com/

[44] D. AI, “Evaluating RAG pipelines using langchain and ragas,” 2024. Accessed: May 30, 2024. [Online]. Available: https://deci.ai/blog/evaluating-rag-pipelines-using-langchain-and-ragas/

[45] LangChain, “Parent document retriever,” 2024. Accessed: May 30, 2024. [Online]. Available: https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/parent_document_retriever/

[46] LangChain, “Ensemble retriever,” 2024. Accessed: May 30, 2024. [Online]. Available: https://python.langchain.com/v0.1/docs/modules/data_connection/retrievers/ensemble/

[47] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “Ragas: Automated evaluation of retrieval augmented generation,” in *Proc. 18th Conf. Eur. Chapter Assoc. Comput. Linguistics: Syst. Demonstrations*, pp. 150–158, Mar. 2024.

[48] "Answer semantic similarity." Accessed: May 28, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/semantic_similarity.html

[49] "Answer correctness," 2023. Accessed: May 28, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/answer_correctness.html

[50] "Answer relevance" 2023. Accessed: May 28, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/answer_relevance.html

[51] "Context precision," 2023. Accessed: May 28, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/context_precision.html

[52] "Context recall," 2023. Accessed: May 28, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/context_recall.html

[53] "Faithfulness," 2023. Accessed: May 28, 2024. [Online]. Available: https://docs.ragas.io/en/latest/concepts/metrics/faithfulness.html

[54] N. Fattorini et al., "Artificial intelligence index report 2024," *ArXiv*, vol. abs/2405.19522, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:270123396

[55] L. Ouyang et al., "Training language models to follow instructions with human feedback," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 27730–27744.

[56] OpenAI, "Hello GPT-4o," 2024. Accessed: Jun. 15, 2024.

[57] Anthropic, "Introducing the next generation of claude" Anthropic, Mar. 2024. Accessed: Aug. 28, 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-family

[58] Anthropic AI Team, "Introducing claude 3.5 sonnet," Jun. 2024. Accessed: Jul. 25, 2024. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[59] A. Q. Jiang et al., "Mistral 7b," 2023, *arXiv:2310.06825*.

[60] A. Q. Jiang et al., "Mixtral of experts," 2024, *arXiv:2401.04088*.

[61] Mistral AI, "Cheaper, better, faster, stronger," 2024. Accessed: Jun. 15, 2024.

[62] Mistral AI Team, "Mistral NeMo," Jul. 2024. Accessed: Jul. 25, 2024. [Online]. Available: https://mistral.ai/news/mistral-nemo/

[63] Mistral AI Team, "Large enough: Mistral large 2 announcement," Jul. 2024. Accessed: Jul. 25, 2024. [Online]. Available: https://mistral.ai/news/mistral-large-2407/

[64] H. Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," 2023, *arXiv:2307.09288*.

[65] Meta, "Introducing meta llama 3: The most capable openly available LLM to date," 2024. Accessed: Jul. 25, 2024.

[66] Llama Team, "The Llama 3 herd of models," Jul. 2024. Accessed: Jul. 25, 2024. [Online]. Available: https://ai.meta.com/research/publications/the-llama-3-herd-of-models/

[67] A. Yang et al., "Qwen2 technical report," 2024, *arXiv:2407.10671*.

[68] G. Team et al., "Gemma: Open models based on gemini research and technology," 2024, *arXiv:2403.08295*.

[69] E. Competition, "Edge-device large language model competition," 2024. Accessed: Aug. 2, 2024. [Online]. Available: https://edge-llms-challenge.github.io/edge-llm-challenge.github.io/index

LUIS-BERNARDO HERNANDEZ-SALINAS received the Engineer's degree in computer systems from the Universidad Politécnica de Querétaro, Querétaro, Mexico, in 2014. He is currently working toward the M.S. degree in advanced technology with the Instituto Politécnico Nacional, Querétaro. He is involved in research on machine learning under the supervision of Dr. Juan Terven. His research interests include computer vision, artificial intelligence, and embedded systems.

JUAN TERVEN (Member, IEEE) received the B.S. degree in electronic engineering from the Mazatlan Institute of Technology, Mazatlan, Mexico, the M.S. degree in computer science from the Autonomous University of Sinaloa, Mexico, and the Ph.D. degree in technology from the National Polytechnic Institute, CICATA-Querétaro. He completed a research internship with MIT in 2013, focusing on three-dimensional reconstruction, and then interned with Microsoft Research in 2014, working on human-machine interfaces. Since 2016, he has been a Research Scientist with AiFi, working with store automation using artificial intelligence, where he has coauthored multiple patents in the United States. In 2023, he became a Research Professor with the National Polytechnic Institute in CICATA Queretaro. His research interests include artificial intelligence and computer vision.

E. A. CHAVEZ-URBIOLA (Member, IEEE) received the B.S. degree in mechanical engineering from Instituto Tecnológico de Celaya, Celaya, Mexico, in 2006, the M.S. degree in mechanical engineering in 2009, and the Ph.D. degree in materials science from CINVESTAV, Mexico, in 2013. He is currently a Professor and Researcher with CICATA, Instituto Politécnico Nacional. Since 2018, he has been applying the "I-Corps" methodology to research in energy, food, and water. He is a Level 1 member of Mexico's Sistema Nacional de Investigadores (SNI). He was a CONACyT Researcher with UAEH in earth sciences and materials. His work includes significant contributions to solar hybrid systems and the encapsulation of probiotics. His research interests include alternative energies, materials science, and scaling laboratory processes to industrial levels.

DIANA-MARGARITA CÓRDOVA-ESPARZA (Member, IEEE) received the B.S. degree in electronic engineering and the M.S. degree in digital signal processing from the Autonomous University of Zacatecas, Zacatecas, Mexico, and the Ph.D. degree in computer vision with National Polytechnic Institute. She is currently a Full-time Professor with Autonomous University of Queretaro, Santiago de Querétaro, Mexico. She performed postdoctoral research with CIDESI in Mexico. Her research interests include artificial intelligence and computer vision.

JULIO-ALEJANDRO ROMERO-GONZÁLEZ (Member, IEEE) received the degree in automation engineering and the master's degree in instrumentation and automatic control from the Faculty of Engineering, Autonomous University of Querétaro, Santiago de Querétaro, Mexico, in 2012 and 2014, respectively, and the Ph.D. degree in computer science with the Faculty of Informatics, in 2022. He is currently a Research Professor with the Faculty of Informatics, Autonomous University of Querétaro. His main research interests include vision and artificial intelligence, embedded systems, and automatic control.

AMADEO ARGUELLES (Member, IEEE) received the doctorate degree in computer science from the Centro de Investigación en Computación (CIC), Mexico, in 2007. He was the Coordinator of the Master's Program in computer science and the Head of the Department of Computer Science Research, CIC-IPN from 2010 to 2014. He is currently a Research Professor and is part of the Intelligent Computing Laboratory. He is a Founding Member of the IPN Computing Network. He is part of the Artificial Intelligence and Data Science Networks and the Network of Experts in Automotive Innovation of the IPN. Founder of the Latin American Network of Educational Technology (Red LaTE Mexico of CONACyT), where he was the Technical Manager during the period 2016 and 2017. He is a member of the Binational Laboratory for Intelligent Management of Energy Sustainability and Technological Training and a member of the Innovation and Work Network in the Mexican Automotive Industry (RED ITIAM), registered with CONACYT. He is a member of the International Neural Network Society and registered in the Autonomous Machine Learning and Big Data Analytics sections. He has participated as a member of program committees in International conferences and is a reviewer of high-impact publications. His research interests include cloud computing, artificial intelligence, supervised learning, embedded systems, and reconfigurable logic.

ILSE CERVANTES (Senior Member, IEEE) received the Ph.D. degree in applied mathematics (automation and control) from the Autonomous Metropolitan University, Mexico City, Mexico. Since 2019, she has led the Automotive Innovation Research Network of the Instituo Politecnic Nacional, which research is focused on: transportation electrification, autonomous vehicles, and hybrid and low-carbon transportation technologies. She is a Level III Researcher with the Mexico's National System of Researchers (SNI). She is the author of more than 70 journal articles, about 50 articles in conference proceedings, and co-author of three books. Her research interests focuses on applied mathematics to energy and vehicles: hybrid systems, automatic control, control, and artificial intelligence applications. Dr. Cervantes was the Editor-in-Chief and a Member of the steering committee of the IEEE Latin America Transactions from 2020 and 2018 to 2023, respectively, and has been the Associate Editor for IEEE TRANSACTIONS ON TRANSPORTATION ELECTRIFICATIONsince 2014. She is the Head of the CONAHCYT National Laboratory in Smart Electromobility (LANCEI). She has organized special issues and conferences on transportation electrification and smart mobility as IEEE LA-CCI 2024 – IEEE Latin American Conference on Computational Intelligence and EEE Transportation Electrification Conference and Expo.