



Bosch Fit.Fest GenAI Hackathon

Team GenStar

Content

Problem Statement	3
Proposed Solution	4
Architecture	13
Performance Analysis	14
References	15

Problem Statement

- Building a **document Q&A** solution to automatically answer questions about documents in natural language using a chat-like interface.
- Automatic **parsing** and storage of content from text, tables, and images in documents.
- **Retrieval** of text, tables, and images in response to user input prompts, providing contextually relevant content from the document.
- **Probing questions** asked towards users to gather detailed information, resolve ambiguities, and stimulate further discussion, ultimately fostering a comprehensive exploration of the topic.



Proposed Solution


- This project tackles the challenge of inefficient document exploration by creating a user-friendly, chat-based Q&A system. Here's a breakdown of the key components:
 - **Data Preprocessing:** Text and table data are extracted from documents.
 - **Table Summarization:** LLM summarizes the extracted tables.
 - **Semantic Chunking:** Textual content is segmented into meaningful chunks using a technique called **semantic chunking**.
 - **Data Storage:** Extracted data (text chunks and summarized tables) is stored in a vector database like Qdrant for efficient retrieval.


- **Query Processing:**
 - A **Retrieval-Augmented Generation (RAG) fusion** method is employed.
 - User provides a natural language query.
 - The LLM generates multiple similar queries based on the original user query.
 - Top K relevant chunks are retrieved from Qdrant based on the generated similar queries.
- **Answer Generation:[Text, Table]**
 - The retrieved text chunks (context) and chat history (if applicable) are combined with the user's original query.
 - Cohere's open-source LLM leverages this information to generate a comprehensive answer.
 - For table reconstruction we use prompt engineering with json based output to generate table.
- **Clarification and Refinement:**
 - The LLM can ask clarifying questions if it feels uncertain about the answer.
 - This back-and-forth interaction helps refine the user's intent and improve answer accuracy
- **Semantic Caching:**
 - By storing previous questions and answer pairs, every time a new question comes in, we look it up on the cache if there's a similar question, and if yes we simply use that response instead of doing another LLM call.
 - This can save lots of cost and latency.

- **Image Retrieval: [Added in Phase-2]**

- We have utilized Google Gemini pro Multimodal model to intelligently extract the **description of the image** and store them along with existing data in VectorDB.
- All the images are **encoded** and stored in MongoDB can generate image on the fly rather than storing it on the disk.
- We have added several attributes such as car_name, image_id etc as **meta data** to each chunk for effective retrieval.
- When user gives a query it retrieves similar chunks as **contexts**. If any of the retrieved contexts belongs to that of image description.
- We take all those image descriptions along with query to make another **LLM call** to intelligently choose the **best image** along with similarity based filtering.
- Once we get the best image and we along with its description we make another LLM call to do final answer **text generation**.

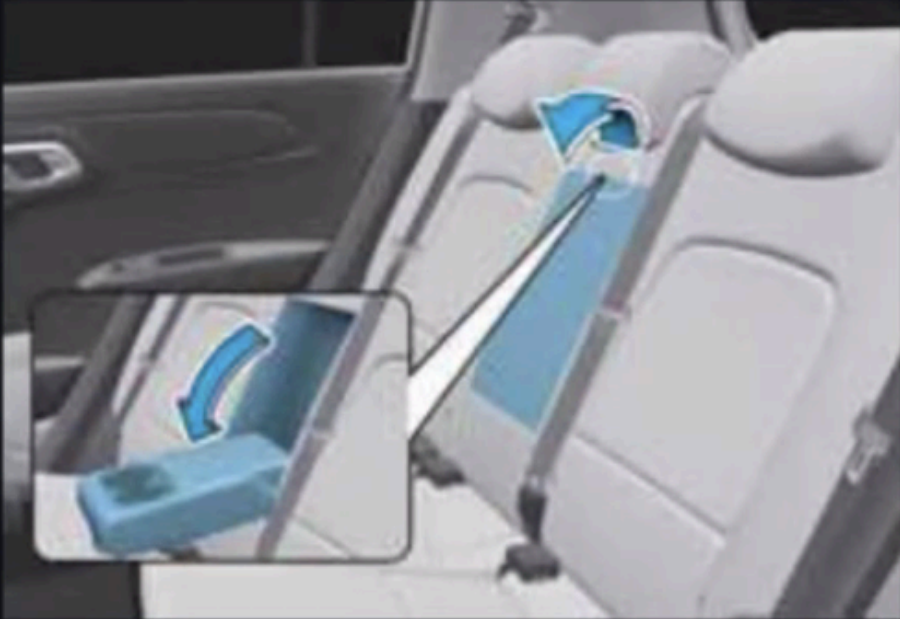
Image Retrieval Results

 arm rest in hyundai exter

 Hello there! Arm rests are a comfortable addition to any car. Here's some simple info on the arm rest in the Hyundai Exter:


- The Hyundai Exter has a rear seat armrest.
- It is located centrally in the back seat.
- It can be pulled down from the seatback.
- Cup holders are also included in the arm rest.
- The armrest can be folded down to make carrying luggage easier.

Hope that helps! Are you interested in learning more about the Exter's features?

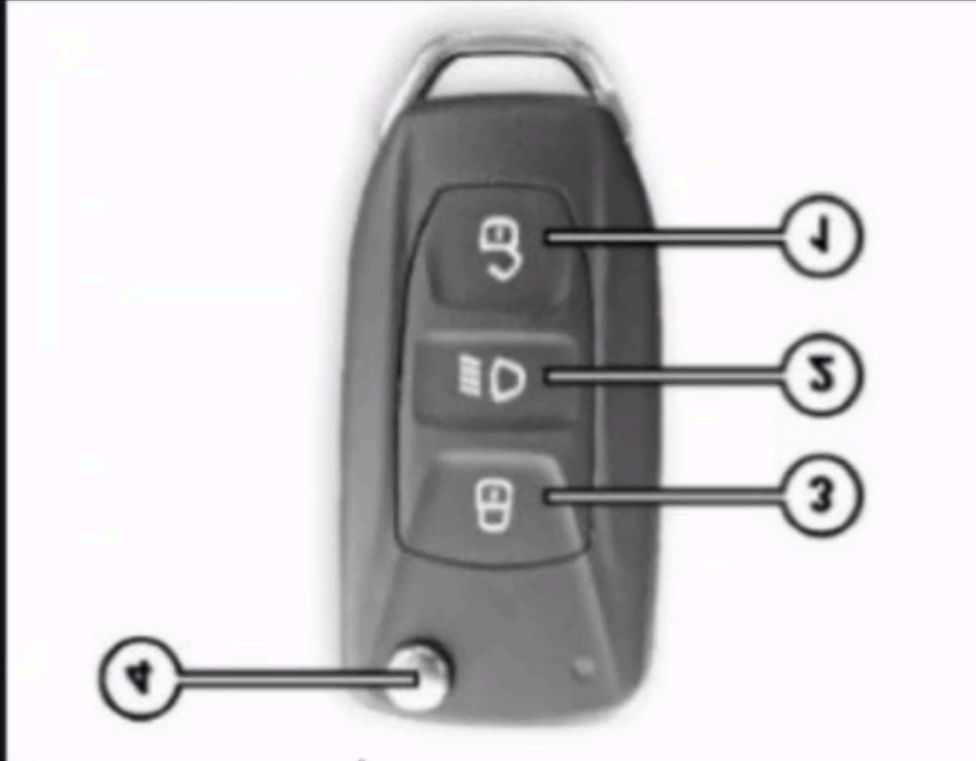


Related Image

What is up? >


 How to use flip key with remote?


- The Flip Key With Remote can unlock all car doors, activate the approach light, and open the tailgate.
- Pressing the unlock button will unlock all doors.
- Pressing the lock button once will lock all the doors of the vehicle.
- To fold out the key blade, press the designated button and fold it back in by pressing the same button again.
- In case of mechanical failure, the key can be used to unlock the doors physically.



Related Image


What is up? >

 Instructions about wearing seat belt for pregnant woman

 Hello there! Pregnant women should follow these guidelines for maximum safety while wearing a seat belt:

- Wear the lap belt across the pelvic bone, fitting snugly across the hips.
- Ensure the seat belt doesn't cross the abdomen.
- Position the shoulder belt across the chest and away from the neck.
- Make sure the seat belt is firm but comfortable.
- Remember, wearing a seat belt is essential for pregnant women, as it provides the best protection for both mother and child in the event of an accident.

I hope that's helpful! Have a great day, and stay safe on the roads.



Related Image

What is up? >

- **Image Based Search: [Added in Phase-3]**
 - We have utilized OpenAI CLIP to extract image embeddings and stored them in VectorDB.
 - User given image we find the CLIP embedding and based on that we find most similar image from the VectorDB. Then We have stored image_id and their corresponding description in MongoDB.
 - So, Once we get the most similar image_id we extract the corresponding image description. we display the description to the user.
- **Image + Text Based Search: [Added in Phase-3]**
 - Same pipeline as above.
 - Once we have description of the most similar image we pass the description along with the user query to LLM with RRF RAG Pipeline.
 - LLM returns the answer based on the context, user_query and image_description.

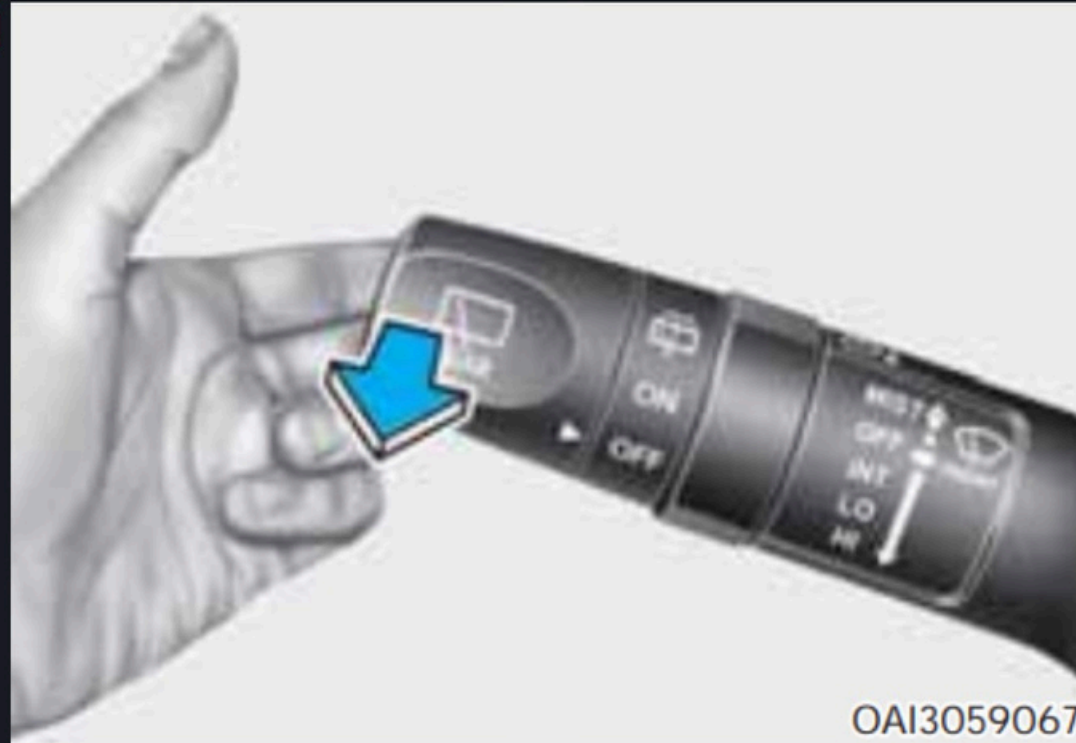
Improvements/ Features added in Phase-2

- Image context retrieval feature
- **Parallelization** of following features:
 - Chat history retrieval
 - Finding suitable images using image summary
 - Checking whether to ask probing questions or not.
- By adding these parallelization we **reduced the latency** of the response from upper bound 30 seconds to 15 seconds.

Improvements/ Features added in Phase-3

- Image based search
 - Image + Text based search
 - Carousel to display the pages in pdf matching the user query.
-
- By adding these parallelization we **reduced the latency** of the response from upper bound 30 seconds to 15 seconds.

Image Based Search Results



Uploaded Image



How to use this wiper?



- To use the wipers, gently pull the windshield washer lever towards you. This will spray fluid onto the windshield and activate the wipers for 1-3 cycles.
- The lever is located on the left side of the steering wheel.
- The wiper will continue to operate until the lever is released.
- The windshield wiper switch has four positions: OFF, INT, LO, and HI.
- INT is for intermittent wiping and is useful in light rain or mist.
- LO and HI adjust the wiper speed.

Windshield wipers

Operates as follows when the ignition switch is turned ON.

MIST : For a single wiping cycle, move the lever upward (or downward) and release it. The wipers will operate continuously if the lever is held in this position.

OFF : The wiper is not in operation

INT : The wiper operates intermittently at the same wiping intervals. Use this mode in light rain or mist. To vary the speed setting, turn the speed control knob. (if equipped)

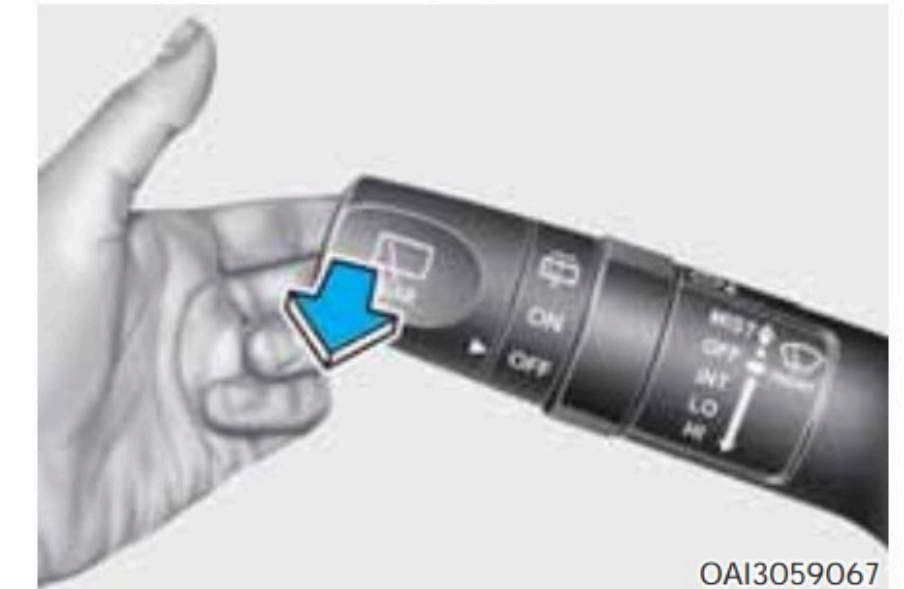
LO : The wiper runs at a lower speed.

HI : The wiper runs at a higher speed.

Information


If there is heavy accumulation of snow or ice on the windshield, defrost the

Windshield washers



In the OFF position, pull the lever gently toward you to spray washer fluid on the windshield and to run the wipers 1-3 cycles. The spray and wiper operation will continue until you release the lever. If the washer does not work, you may need to add washer fluid to the washer fluid reservoir.

Image Based Search Results



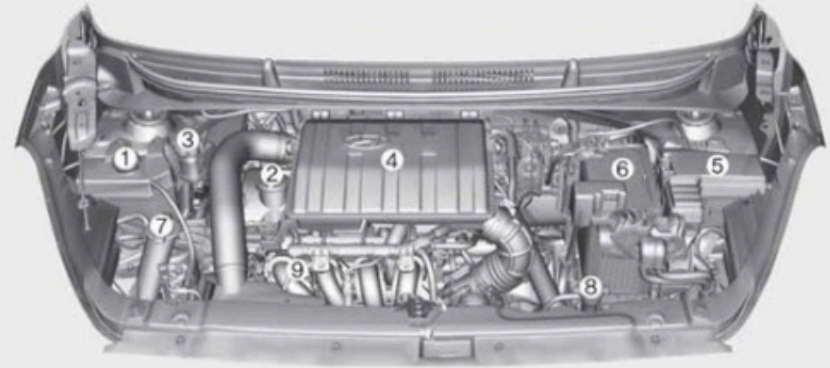
Instrument cluster

Uploaded Image

What does this indicate?

- The indicator light is a seatbelt warning light for the driver and front passenger.
- It illuminates in a sequence and is accompanied by a warning chime to indicate if the driver or front passenger's seatbelt is unfastened.

This is a response regarding the Hyundai Exter's manual. Is there anything else you'd like to know about this particular car model?



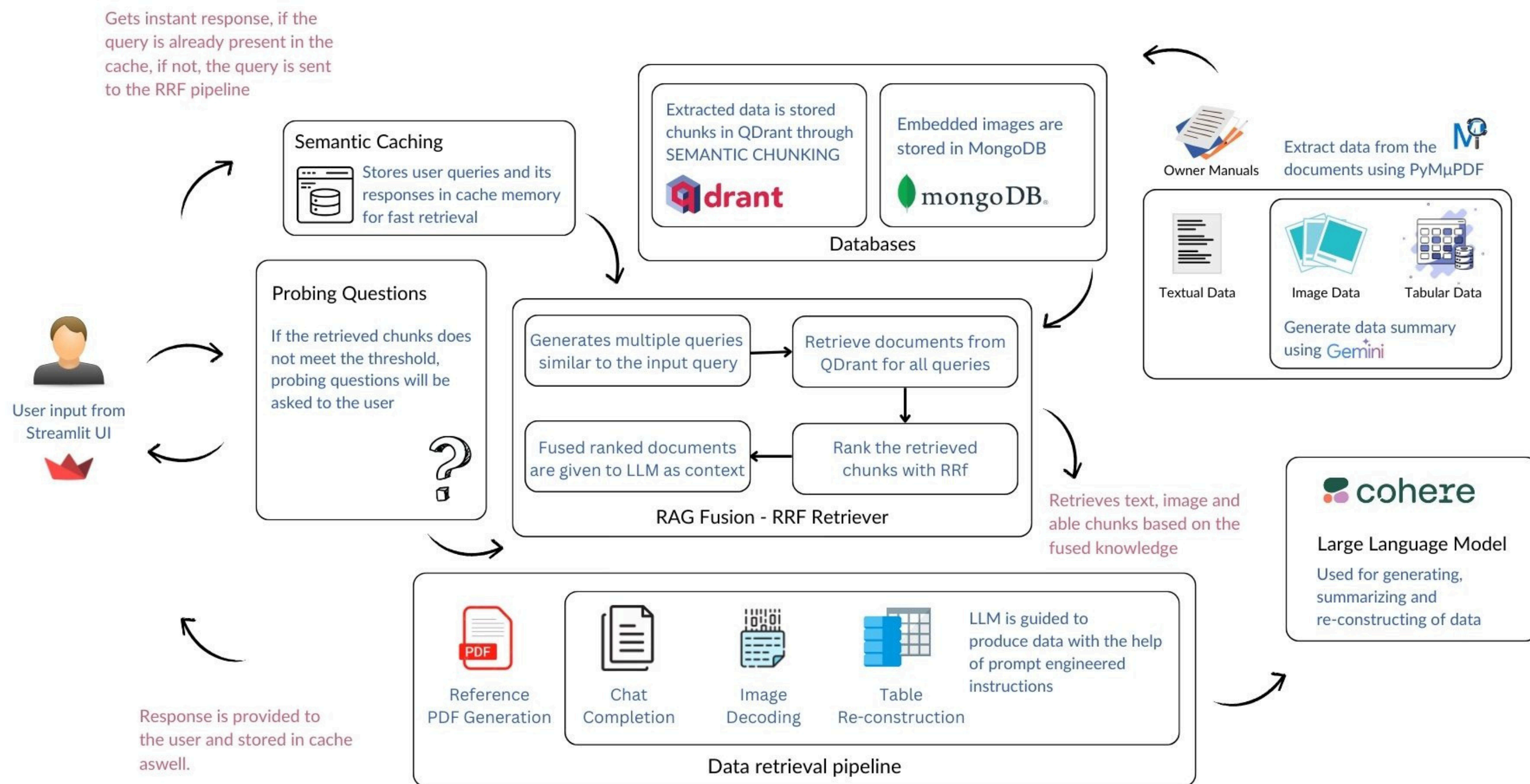
Petrol Engine (Kappa 1.2 MPI)

The actual engine room in the vehicle may differ from the illustration.

Uploaded Image

This Image is extracted from manual of Hyundai Verna. The image is a diagram of an engine compartment of a car, showing the location of various components. The components are labeled with numbers that correspond to a list of parts and their page numbers in a manual. These components include the engine coolant reservoir, radiator cap, brake/clutch fluid reservoir, air cleaner, engine oil dipstick, engine oil filler cap, windshield washer fluid reservoir, fuse box, and battery. The diagram is in grayscale and the car is viewed from the front, with the hood open. The illustration is intended to be used for reference purposes and may not accurately reflect the actual engine compartment of a specific vehicle. The text indicates that the actual engine compartment in the vehicle may differ from the illustration. The text also indicates that some components, such as the brake/clutch fluid reservoir, are optional.

Architecture



Performance Analysis

- The **Latency** of the chatbot ranges between 10 - 15 seconds depending on the speed of internet connection.
- We evaluated the **accuracy** of probing questions mechanism and found it to be effective when we set threshold of cosine similarity between query and context as 0.3 .

Github Repo Link: [click here](#)

– Deployed Chatbot application: [click here](#)

References

1. Article on Advanced RAG Techniques by Pinecone.
2. Research Paper on RAG Fusion by Zackary Rackauckas.
3. Notebook by Greg Kamradt on Semantic Chunking.
4. Cohere API Documentation.
5. LangChain Documentation.
6. Qdrant Documentation.
7. Reduce LLM Cost and Latency with Semantic Cache.