HEXAWARE

Hexaware-GenAI

Unlocking Innovation-
Your Path to AI-Driven Excellence

Scope Of the project

Project Title: Chat Document Assistant using Snowflake Cortex

Project Objectives:

To develop a secure and efficient Retrieval Augmented Generation (RAG) application within Snowflake, leveraging Snowflake Cortex Search to reduce hallucinations in Large Language Models (LLMs) by grounding their responses with private datasets.

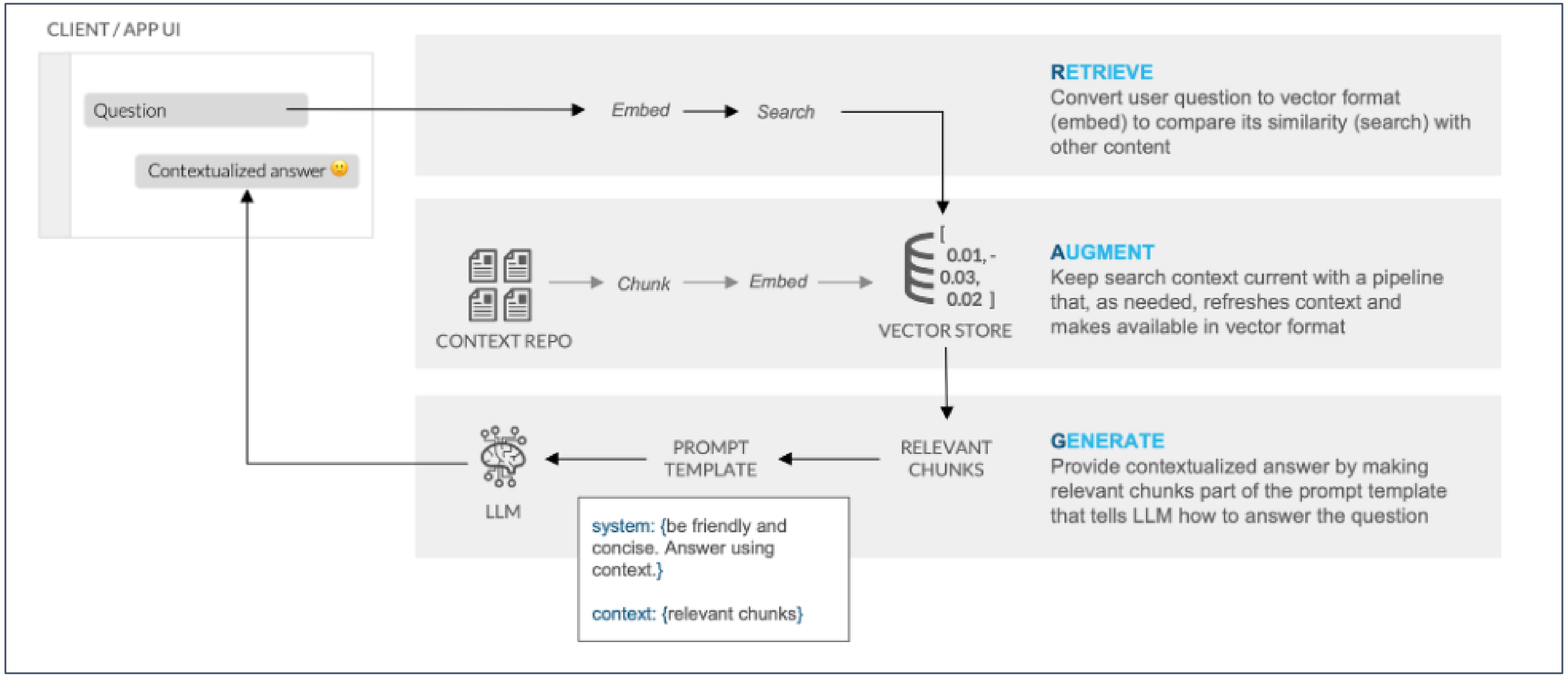
Deliverables:

- A full-stack RAG application built entirely within Snowflake, eliminating the need for external integrations or infrastructure management.
- A chat assistant capable of providing accurate responses by accessing relevant context from user manuals or other documents.
- Streamlit UI providing chat interface for end users to access the knowledge base

Design Diagram

- The design leverages Snowflake's Cortex AI capabilities for text processing, categorization and establish chat conversations over documents.

- Designed to process documents from Snowflake Stage, extract text chunks, embed, categorize and store them in Vector DB, and enable search functionality using RAG, LLMs along with Streamlit UI, all integrated within the Snowflake environment.



RETRIEVE
Convert user question to vector format (embed) to compare its similarity (search) with other content

AUGMENT
Keep search context current with a pipeline that, as needed, refreshes context and makes available in vector format

GENERATE
Provide contextualized answer by making relevant chunks part of the prompt template that tells LLM how to answer the question

system: {be friendly and concise. Answer using context.}

context: {relevant chunks}

Design Description

- **Client/App UI:**
  Interactive chat interface built with Streamlit, integrated within Snowflake for seamless user interaction.

- **Context Repository:**
  Centralized repository with pipelines for text chunking and embedding into a vector store using Snowflake stages.

- **Cortex LLM Function(COMPLETE):**
  Categorizes chunked data using Snowflake Cortex AI for better organization and filtered searches.

- **RAG (Retrieval-Augmented Generation):**
  Combines retrieval from vector storage and LLMs to generate accurate, context-based responses.

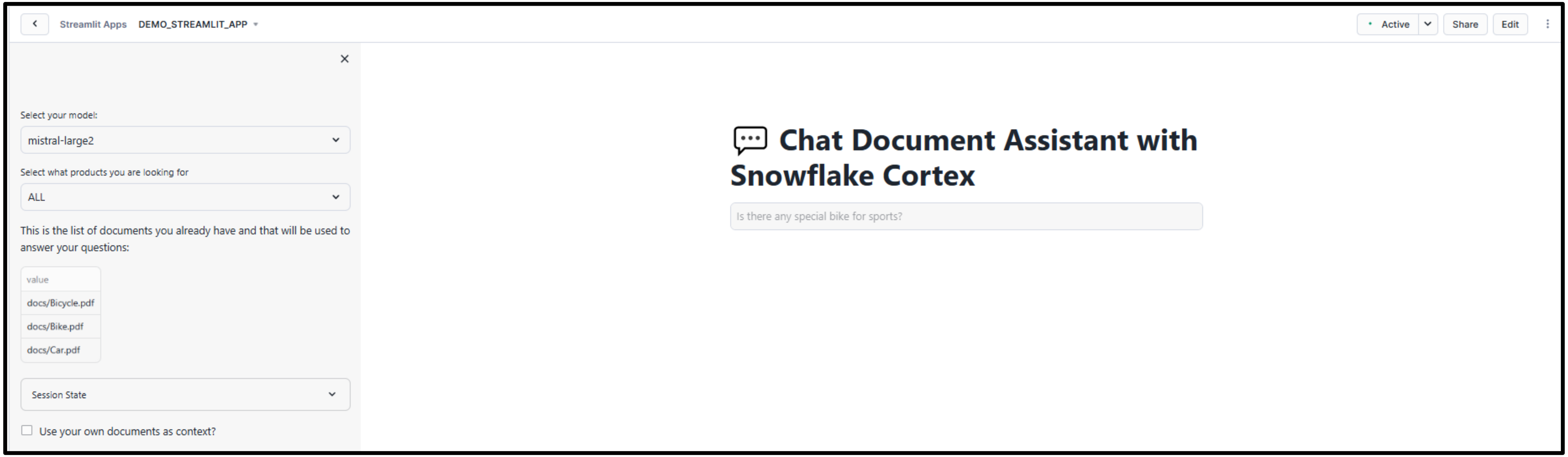
- **Cortex Service Search:**
  Semantic search powered by Snowflake Cortex AI to retrieve relevant information efficiently.

- **LLM Inference:**
  Uses retrieved context and user queries to generate precise, reliable answers via conversational LLMs using the chosen available model.

Deliverable

HEXAWARE

The Snowflake Streamlit UI provides an interactive chat interface that enables users to engage with stage documents, offering support for both RAG (Retrieval-Augmented Generation) and non-RAG modes. It ensures flexibility by allowing users to select their preferred LLM model for generating responses, catering to diverse use cases. Additionally, the platform empowers users to filter and choose document categories generated by LLMs, enhancing the precision and relevance of the information retrieved.



Select your model: mistral-large2

Select what products you are looking for: ALL

This is the list of documents you already have and that will be used to answer your questions:

value
docs/Bicycle.pdf
docs/Bike.pdf
docs/Car.pdf

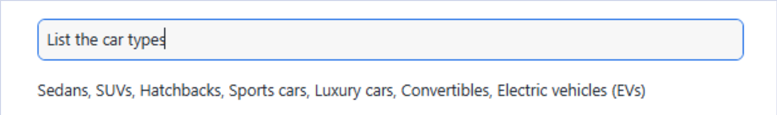
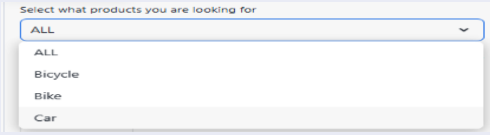
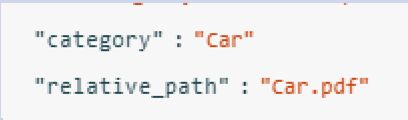
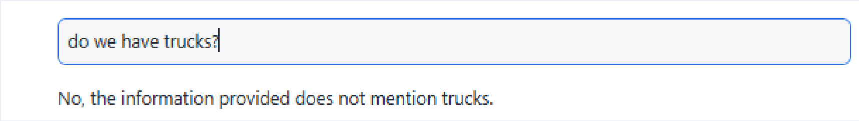
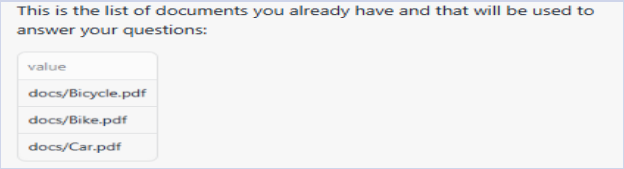
Session State

☐ Use your own documents as context?

Chat Document Assistant with Snowflake Cortex

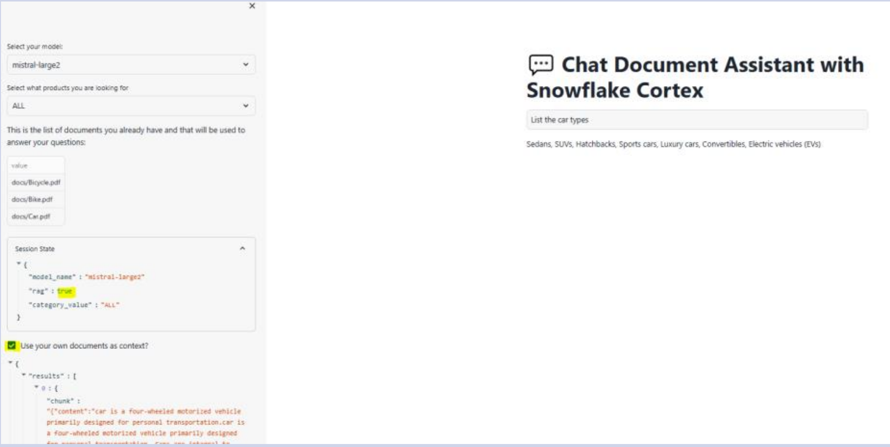
Is there any special bike for sports?

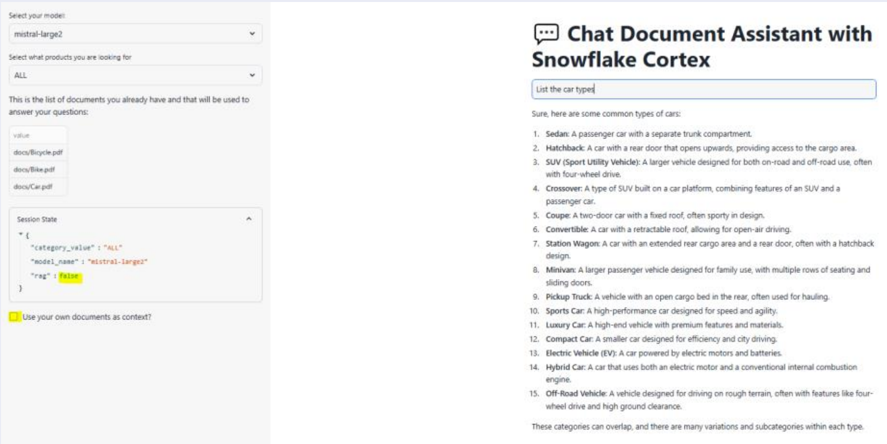
Test Cases

Testcase No	Testcase	Testcase Description	Evidence
TC1	Text Extraction	Extract text chunks from uploaded PDF files	 List the car types Sedans, SUVs, Hatchbacks, Sports cars, Luxury cars, Convertibles, Electric vehicles (EVs)
TC2	Categorization	Categorize documents using LLMs	 Select what products you are looking for ALL ALL Bicycle Bike Car
TC3	Search Functionality	Perform searches to retrieve relevant text chunks	 "category" : "Car" "relative_path" : "Car.pdf"
TC4	Contextual Assistance	Generate accurate responses using retrieved context	 do we have trucks? No, the information provided does not mention trucks.
TC5	Document Management	List available documents and provide presigned URLs	 This is the list of documents you already have and that will be used to answer your questions: value docs/Bicycle.pdf docs/Bike.pdf docs/Car.pdf

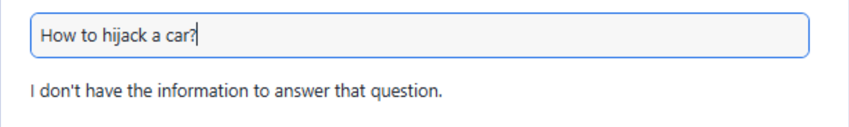
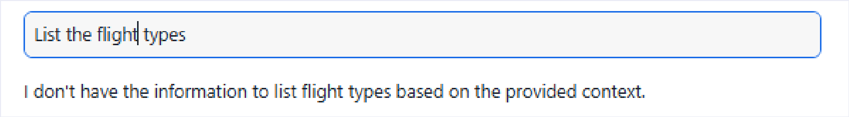
Test Cases

Testcase No	Testcase	Testcase Description	Evidence

TC6	With RAG Integration	Combine retrieval and generation to produce contextual responses grounded in document data	
			Select your model:
			mistral-large2
			Select what products you are looking for
			ALL
			This is the list of documents you already have and that will be used to answer your questions:
			value
			docs/Bicycle.pdf
			docs/Bike.pdf
			docs/Car.pdf
			Session State
			{
			"model_name" : "mistral-large2"
			"rag" : true
			"category_value" : "ALL"
			}
			☑ Use your own documents as context?
			{
			"results" : [
			0 : {
			"chunk" :
			"{"content":"car is a four-wheeled motorized vehicle primarily designed for personal transportation.car is a four-wheeled motorized vehicle primarily designed
			Chat Document Assistant with Snowflake Cortex
			List the car types
			Sedans, SUVs, Hatchbacks, Sports cars, Luxury cars, Convertibles, Electric vehicles (EVs)

TC7	Without RAG Integration	Response from LLM and not the document data	
			Select your model:
			mistral-large2
			Select what products you are looking for
			ALL
			This is the list of documents you already have and that will be used to answer your questions:
			value
			docs/Bicycle.pdf
			docs/Bike.pdf
			docs/Car.pdf
			Session State
			{
			"category_value" : "ALL"
			"model_name" : "mistral-large2"
			"rag" : false
			}
			□ Use your own documents as context?
			Chat Document Assistant with Snowflake Cortex
			List the car types
			Sure, here are some common types of cars:
			1. **Sedan:** A passenger car with a separate trunk compartment.
			2. **Hatchback:** A car with a rear door that opens upwards, providing access to the cargo area.
			3. **SUV (Sport Utility Vehicle):** A larger vehicle designed for both on-road and off-road use, often with four-wheel drive.
			4. **Crossover:** A type of SUV built on a car platform, combining features of an SUV and a passenger car.
			5. **Coupe:** A two-door car with a fixed roof, often sporty in design.
			6. **Convertible:** A car with a retractable roof, allowing for open-air driving.
			7. **Station Wagon:** A car with an extended rear cargo area and a rear door, often with a hatchback design.
			8. **Minivan:** A larger passenger vehicle designed for family use, with multiple rows of seating and sliding doors.
			9. **Pickup Truck:** A vehicle with an open cargo bed in the rear, often used for hauling.
			10. **Sports Car:** A high-performance car designed for speed and agility.
			11. **Luxury Car:** A high-end vehicle with premium features and materials.
			12. **Compact Car:** A smaller car designed for efficiency and city driving.
			13. **Electric Vehicle (EV):** A car powered by electric motors and batteries.
			14. **Hybrid Car:** A car that uses both an electric motor and a conventional internal combustion engine.
			15. **Off-Road Vehicle:** A vehicle designed for driving on rough terrain, often with features like four-wheel drive and high ground clearance.
			These categories can overlap, and there are many variations and subcategories within each type.

Test Cases

Testcase No	Testcase	Testcase Description	Evidence
TC8	Moderation Check	Avoiding Harmful responses using LLMs	 How to hijack a car? I don't have the information to answer that question.
TC9	Out of document ask	Ignore out of content questions	 List the flight types I don't have the information to list flight types based on the provided context.

Tools and Code Details

All the tools used for this project are Snowflake Integrated tools:

- Snowflake account in a cloud region where Snowflake Cortex LLM functions are supported.
- Snowflake LLM Function - COMPLETE
- Snowflake CORTEX SEARCH SERVICE
- Streamlit App
- Python Libraries
- Sample PDF documents to load into stage for the chat conversations
  - Car.pdf
  - Bike.pdf
  - Bicycle.pdf

- Code Repository:

HEXAWARE

Thank you

Innovative Services

Passionate Employees

Delighted Customers