Introducing

# RETAILIA

By:
Suprathika Vangari

# Agenda

- Introduction
- Motivation
- Objectives
- Literature Review
- LLM Overview
- Core Components
- Workflow & Results
- Challenges & Future Directions
- Conclusion

# Motivation

- Rapid Digital Transformation
- Consumer Expectations
- AI Integration in Customer Service
- Inefficiencies in Human-Driven Support
- Limitations of Rule-Based Chatbots
- Lack of Personalization
- Improved Efficiency and Automation
- Enhanced Understanding and Contextual Responses

# Objectives

Chatbot that effectively responds to both product based and user related queries.

Implement dynamic agent selection to automatically route user queries.

Integrate a knowledge base, constructed from manuals and PDF documents and SQL databases

Utilize Generative LLM and Agentic RAG technologies.

# Reimagining Customer Interactions with AI

As digital interactions become the norm, the expectations for instant, relevant, and engaging customer service have skyrocketed. Traditional customer service models are struggling to keep pace, creating a pressing need for innovative solutions. This is where Retailia comes in.

It leverages the latest in Large Language Models and Agentic RAG systems to deliver responses that are not just timely but contextually enriched and conversational. Retailia can handle both routine queries and complex requests, providing dynamic, up-to-date information while maintaining a natural conversational flow.

# Research and Insights: Literature Review

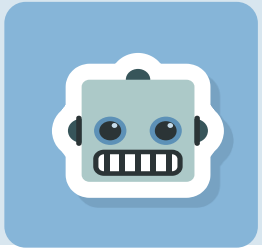
Our Literature review synthesized AI applications in e-commerce, focusing on evolutions in Retrieval-Augmented Generation (RAG), Natural Language Processing (NLP) in chatbots, and Large Language Models (LLMs) like LLaMA in Retail Industry.

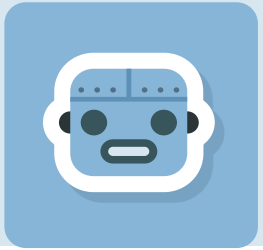
- **Enhancing International Graduate Student Support:** Saha and Saha (2024) developed an AI-driven chatbot using GPT-3.5 turbo and RAG to support international graduate students. This research demonstrates RAG's potential for providing tailored support and highlights the importance of domain-specific data sources for accurate responses.

- **Improving LLMs for Domain-Specific Tasks:** Santos et al. (2024) explored using RAG and fine-tuning techniques to enhance LLM performance in specialized tasks. The research emphasized the importance of data retrieval and demonstrated how combining RAG with fine-tuning leads to better performance in text generation and summarization.

- **Effective Handling of Sensitive Queries:** A study by S. Vakayil et al. (2024) explored using RAG with LLaMA-2 to create an empathetic chatbot for sexual assault victims, demonstrating the capacity of LLMs to provide sensitive and contextually appropriate responses.

# Data at work





## eCommerce Database

Contains product information and user-specific data such as cart, orders, ratings, and feedback.

## FAQ Knowledge Base

Contains FAQs documents which are list of frequently asked questions and corresponding answers.

# How Retailia Handles Queries

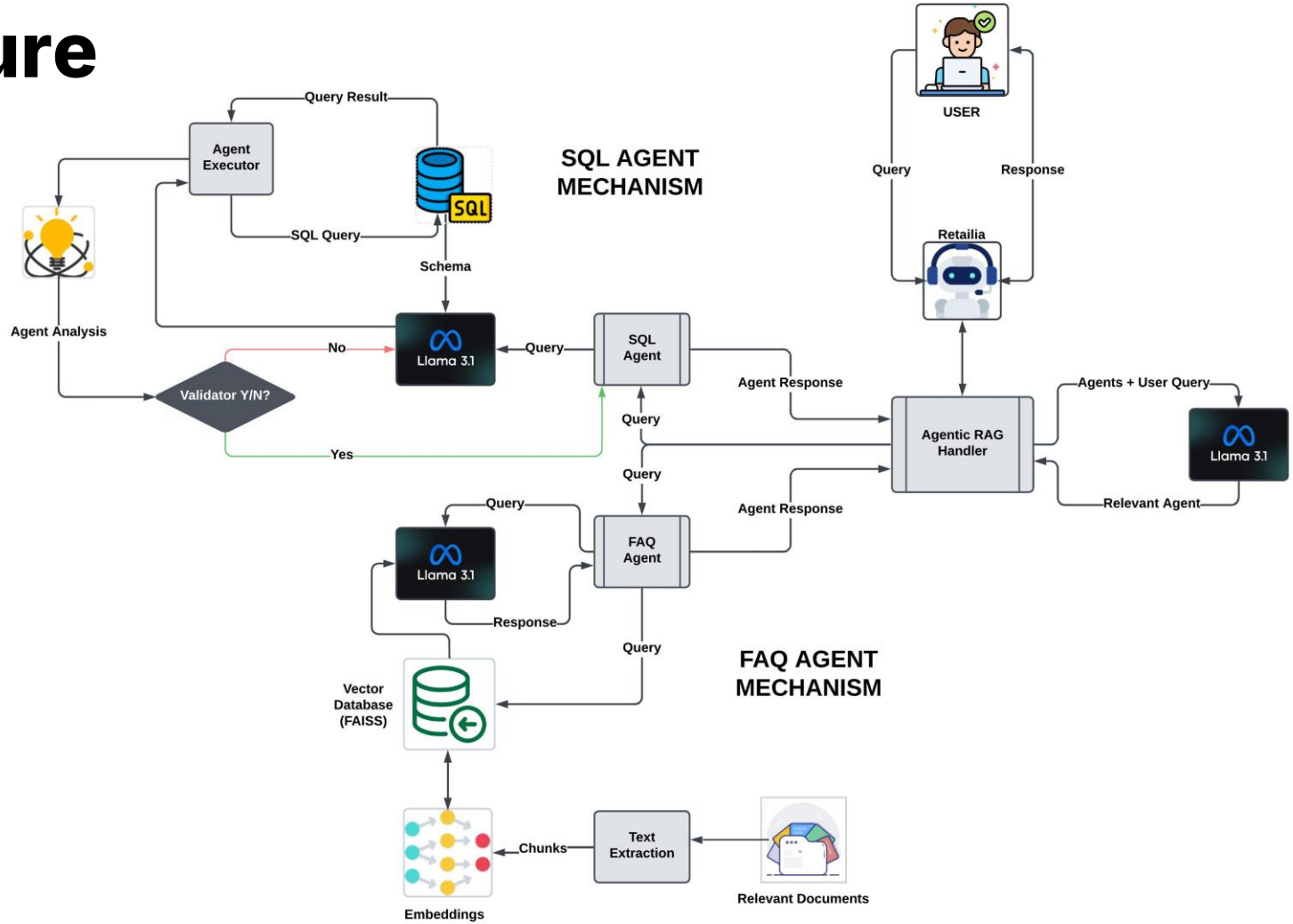
**SQL Agent** employs more complex queries requiring real-time data.

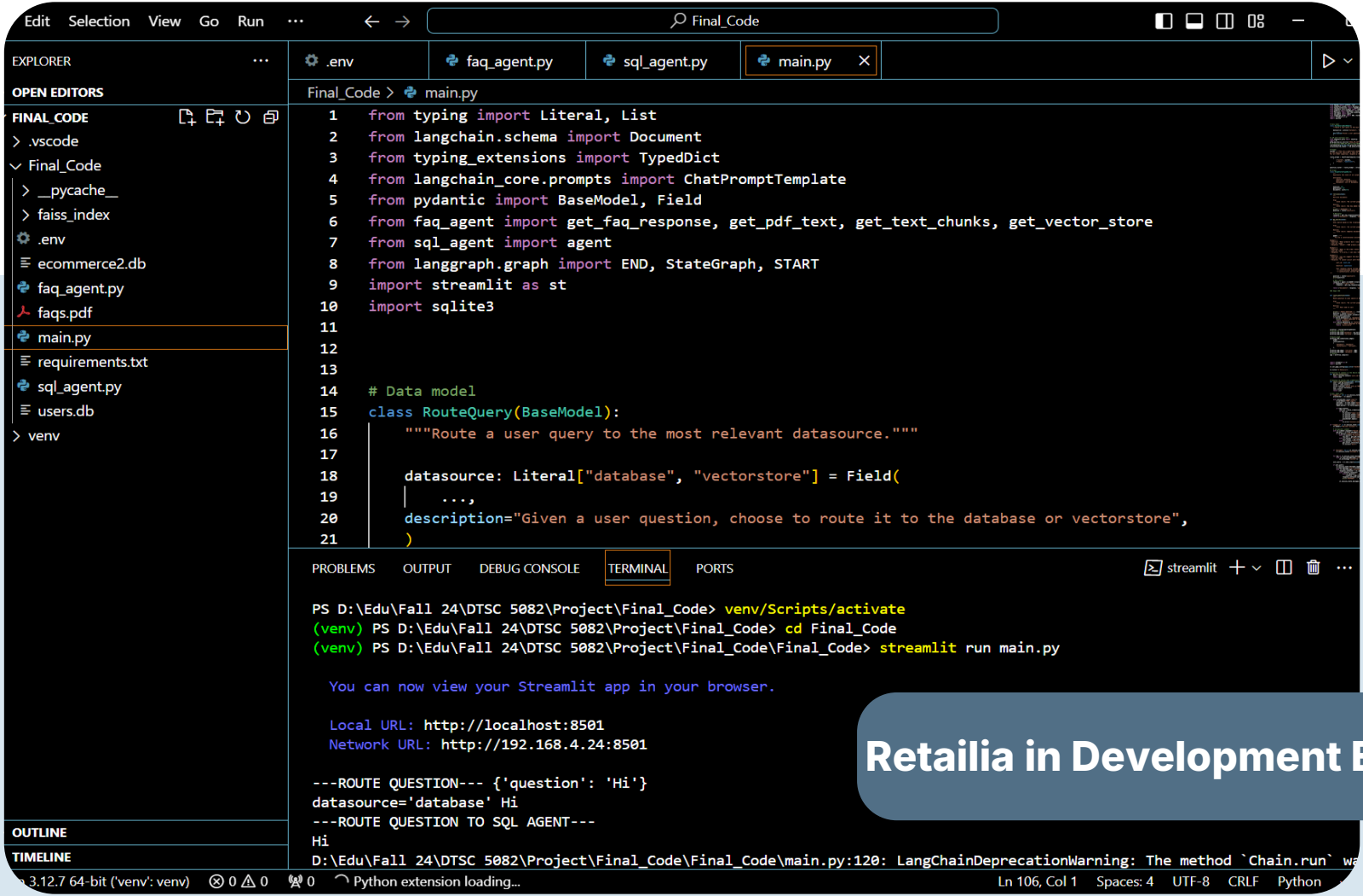
The SQL Agent interacts with the **MySQL Database** housing eCommerce database to retrieve relevant information.

**FAQ Agent** takes center stage for FAQ related queries.

It utilizes a **FAISS index**, a powerful tool for efficient similarity search within a vector database.

# Architecture

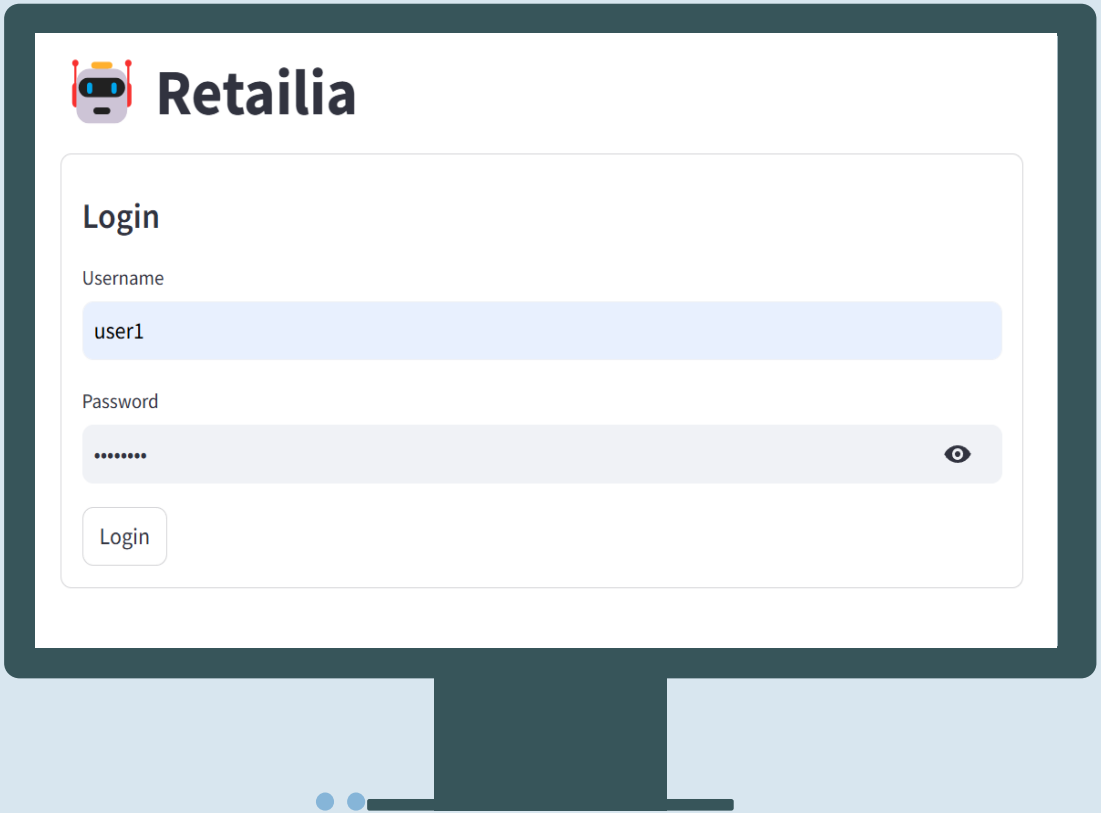





# Retailia in Development Environment

Sneak peek

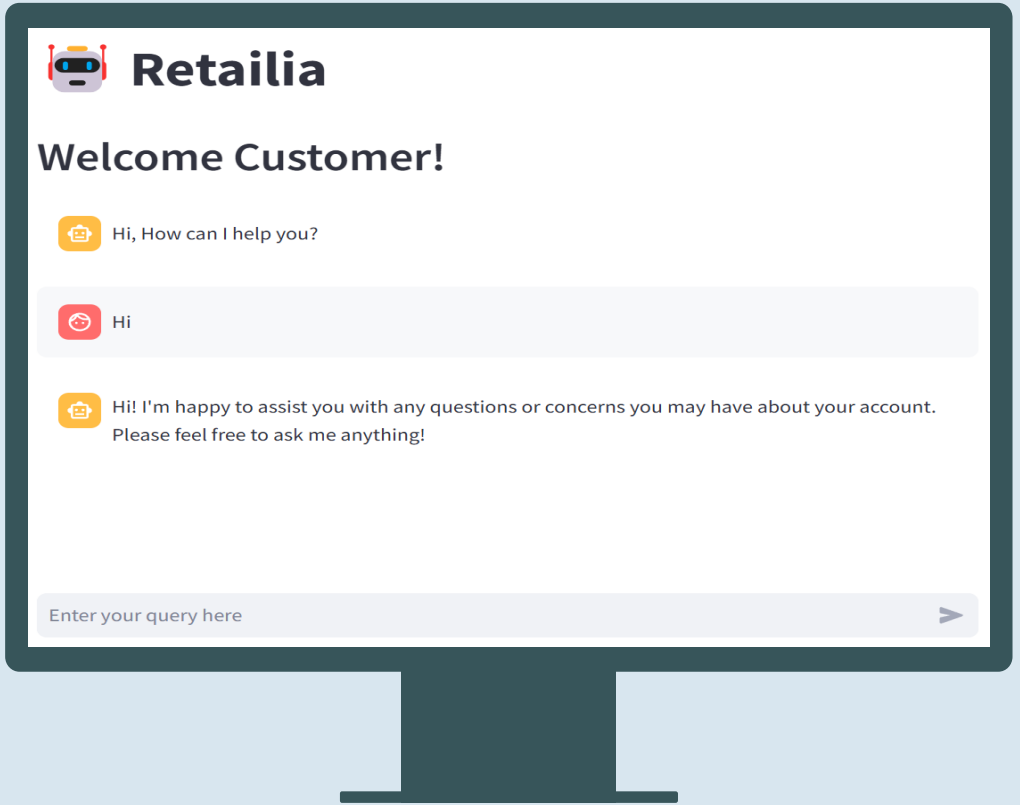
# Interfacing with Retailia: Streamlit's Role




User Authentication

## Sneak peek

# Interfacing with Retailia: Streamlit's Role




Chatbot Interface

# Routing to FAQ Agent

Would I get any discount on my first order?

Yes, you would receive a 10% discount on your first purchase when signing up for their newsletter.

```
---ROUTE QUESTION--- {'question': 'Would I get any discount on my first order?'}
datasource='vectorstore' Would I get any discount on my first order?
---ROUTE QUESTION TO FAQ AGENT---
---RETRIEVE---
RES :  Yes, you would receive a 10% discount on your first purchase when signing up for their newsletter.
RESPONSE :  {'question': 'Would I get any discount on my first order?', 'documents': 'Yes, you would receive a 10% discount on y
our first purchase when signing up for their newsletter.'}
```

# Routing to SQL Agent

```
---ROUTE QUESTION--- {'question': "What's the status of my latest order"}
datasource='database' What's the status of my latest order
---ROUTE QUESTION TO SQL AGENT---
What's the status of my latest order
D:\Edu\Fall 24\DTSC 5082\Project\Final_Code\Final_Code\main.py:120: LangChainDeprecationWarning: The method `Chain.run` was depr
ecated in langchain 0.1.0 and will be removed in 1.0. Use :meth:`~invoke` instead.
  response = agent.run(QUERY.format(user_id=st.session_state['user_id'], question=question))


> Entering new SQL Agent Executor chain...
Thought: I should look at the tables in the database to see what I can query.  Then I should query the schema of the most releva
nt tables.
```

```
*/I have the schema of the orders and carts tables. Now I should think about what the user is asking and how I can answer their
question.

Thought: The user is asking about the status of their latest order. I don't see a column for order status in the orders table, b
ut I can still provide the latest order information for the user.

Action: sql_db_query_checker
Action Input: SELECT * FROM orders WHERE user_id = 1 ORDER BY order_date DESC LIMIT 1SELECT * FROM orders WHERE user_id = 1 ORDE
R BY order_date DESC LIMIT 1Action: sql_db_query
Action Input: SELECT * FROM orders WHERE user_id = 1 ORDER BY order_date DESC LIMIT 1[(43, 1, 995.82, '2024-10-15 20:14:09')]I n
ow know the final answer
Final Answer: Your latest order is from 2024-10-15 20:14:09 with a total amount of 995.82.

> Finished chain.
RESPONSE :  {'question': "What's the status of my latest order", 'documents': 'Your latest order is from 2024-10-15 20:14:09 wit
h a total amount of 995.82.'}
```

# Challenges

**While developing Retailia;**
**Addressing Data Bias:** AI systems can inherit biases present in their training data.
**Efficient Query Routing:** As the volume of queries increases, Retailia's ability to efficiently route queries to the appropriate agent (FAQ or SQL) becomes paramount.

**In Real world setting;**
**Privacy and Security:** Retailia's access to sensitive customer data necessitates robust security measures and strict adherence to privacy regulations to prevent data breaches or misuse.
**Job Displacement:** The automation of customer service tasks by Retailia could raise concerns about potential job displacement for human agents.

# Growth Potential

- **Personalized Recommendations:** Retailia can be enhanced to provide personalized product recommendations in future by leveraging customer data and purchase history.
- **Multi-Modal Interactions:** Future enhancements could incorporate multi-modal interactions, allowing customers to interact using voice commands and images.
- **Advanced Sentiment Analysis:** Advanced sentiment analysis can be implemented to ensure that chatbot responds based on the emotional tone of user queries.

# References


- Saha, B., & Saha, U. (2024). Enhancing international graduate student experience through ai-driven support systems: A llm and rag-based approach. *2024 International Conference on Data Science and Its Applications (ICoDSA)*, 300–304. https://doi.org/10.1109/ICoDSA62899.2024.10651944
- Santos, H., & Khalil, A. (2024). Unleashing the potential of llm in ml: Techniques for fine-tuning, adaptation, and practical deployment with chatgpt. *Baltic Multidisciplinary Journal*, *2*(2), 179–184. https://balticjournals.com/index.php/baltic/article/view/49
- Vakayil, S., Juliet, D. S., J, Anitha., & Vakayil, S. (2024). Rag-based llm chatbot using llama-2. *2024 7th International Conference on Devices, Circuits and Systems (ICDCS)*, 1–5. https://doi.org/10.1109/ICDCS59278.2024.10561020
- *Agents*. (n.d.). Retrieved November 21, 2024, from https://www.kaggle.com/whitepaper-agents
- *Embeddings & vector stores*. (n.d.). Retrieved November 21, 2024, from https://www.kaggle.com/whitepaper-embeddings-and-vector-stores
- *Foundational large language models & text generation*. (n.d.). Retrieved November 21, 2024, from https://www.kaggle.com/whitepaper-foundational-llm-and-text-generation

# THANK YOU!