Introducing

# RETAILIA

**Team Members:**

Manzur Elahi Shaik

Harish Inavolu

Sowmya Katla

Suprathika Vangari

# Agenda

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

# 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.

# 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.

# 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.

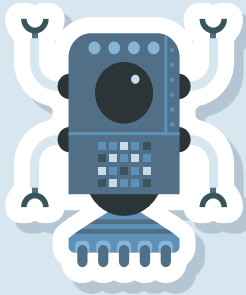
- **Enhancing International Graduate Student Support:** Saha and Saha (2024) developed an AI-driven chatbot using GPT-3.5 turbo and RAG to support international graduate students. This research demonstrates RAG's potential for providing tailored support and highlights the importance of domain-specific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

- **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.

# LLM - Overview

- An LLM (Large Language Model) is an advanced AI system that excels in processing, comprehending, and generating human-like text.
- It is typically a deep neural network trained on massive text datasets, allowing it to learn intricate language patterns.
- LLMs' strength lies in their ability to process and synthesize large amounts of data. This characteristic is particularly valuable in cybersecurity and, by extension, retail, where data can be vast and siloed.
- Overall, LLMs are the foundation for Retailia's innovative approach to retail customer service.



# Understanding Retailia’s Core Components

**Llama 3** is a core language processing engine.

Llama 3 was chosen for its ability to process complex language and generate human-like text, making it suitable for creating engaging and informative customer interactions.

**Agentic RAG** extends traditional RAG with autonomous agents for more intelligent decision-making.

Agentic RAG combines information retrieval from external databases with LLM-powered language generation

# How Retailia Handles Queries

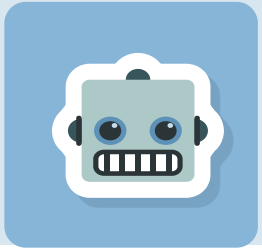
**SQL Agent** employs more complex queries requiring real-time data.

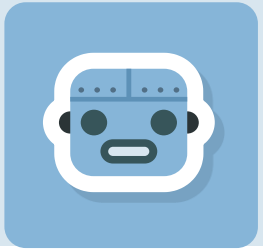
The SQL Agent interacts with the **MySQL Database** housing eCommerce database to retrieve relevant information.

**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.

# 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.

# From User Input to Response Generation

User Input: Customer submits the query to chatbot.

Query Processing & Routing: Query is processed and routed to either SQL agent or FAQ agent.

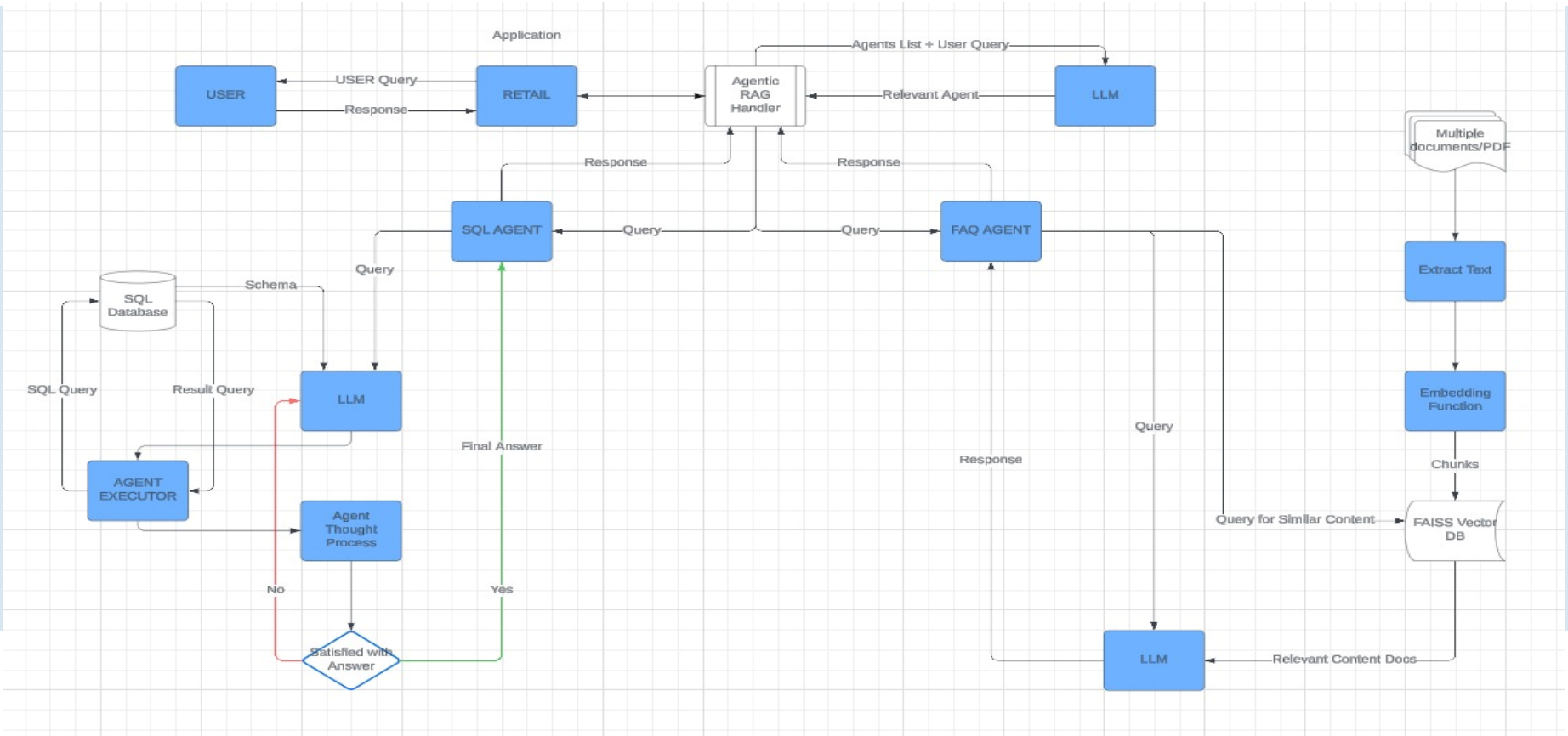
SQL Agent configured to handle database schema for real-time user-specific data access.

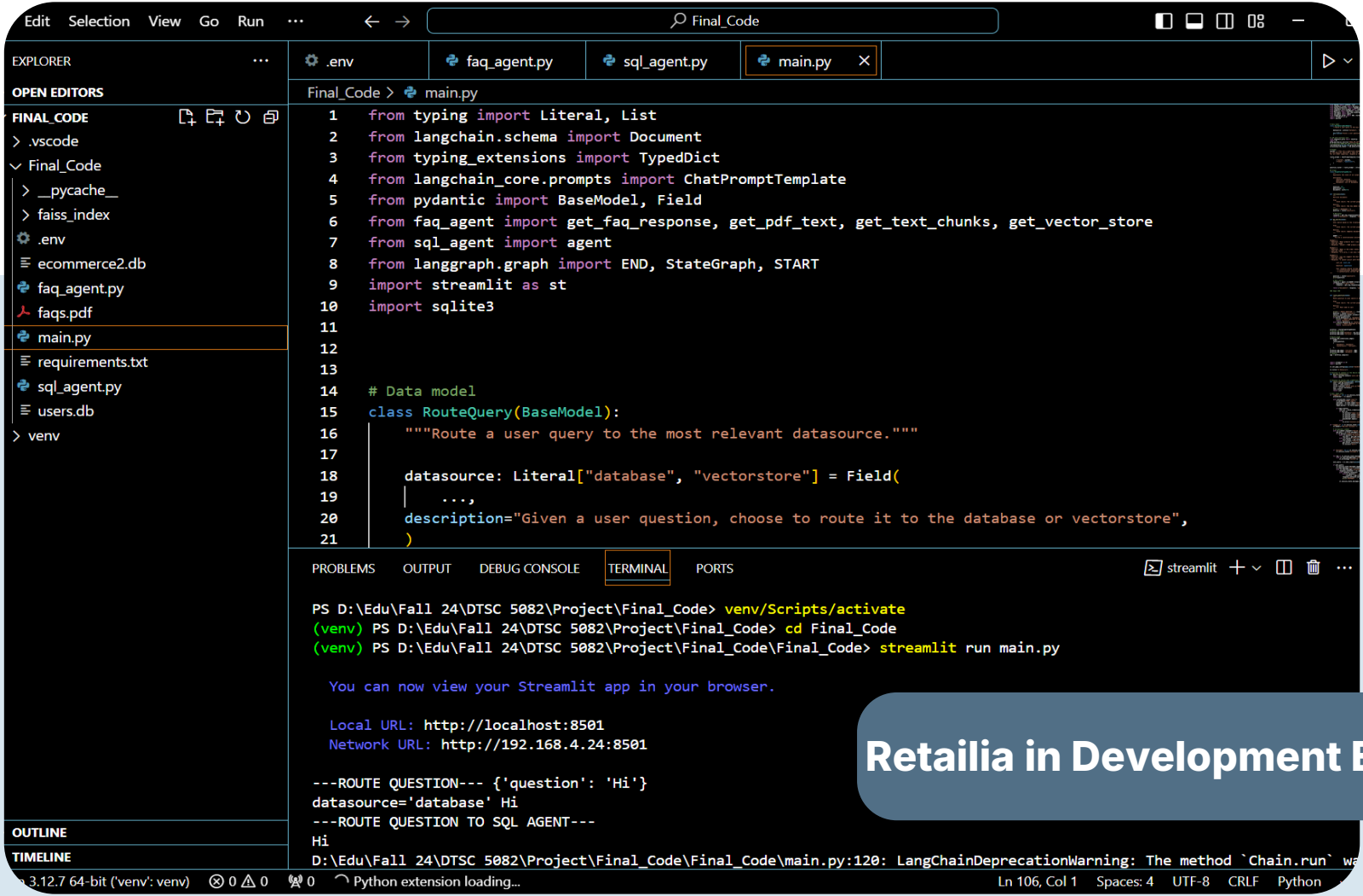
FAQ Agent relies on LangChain and RAG to fetch responses from a trained knowledge base.

Data Retrieval: The designated agent retrieves necessary data, and the raw response is formulated.

Response Generation: Leveraging the LLaMA 3 LLM, Retailia generates a natural language response, integrating retrieved information and tailoring it to the specific user query.

# Components and Architecture

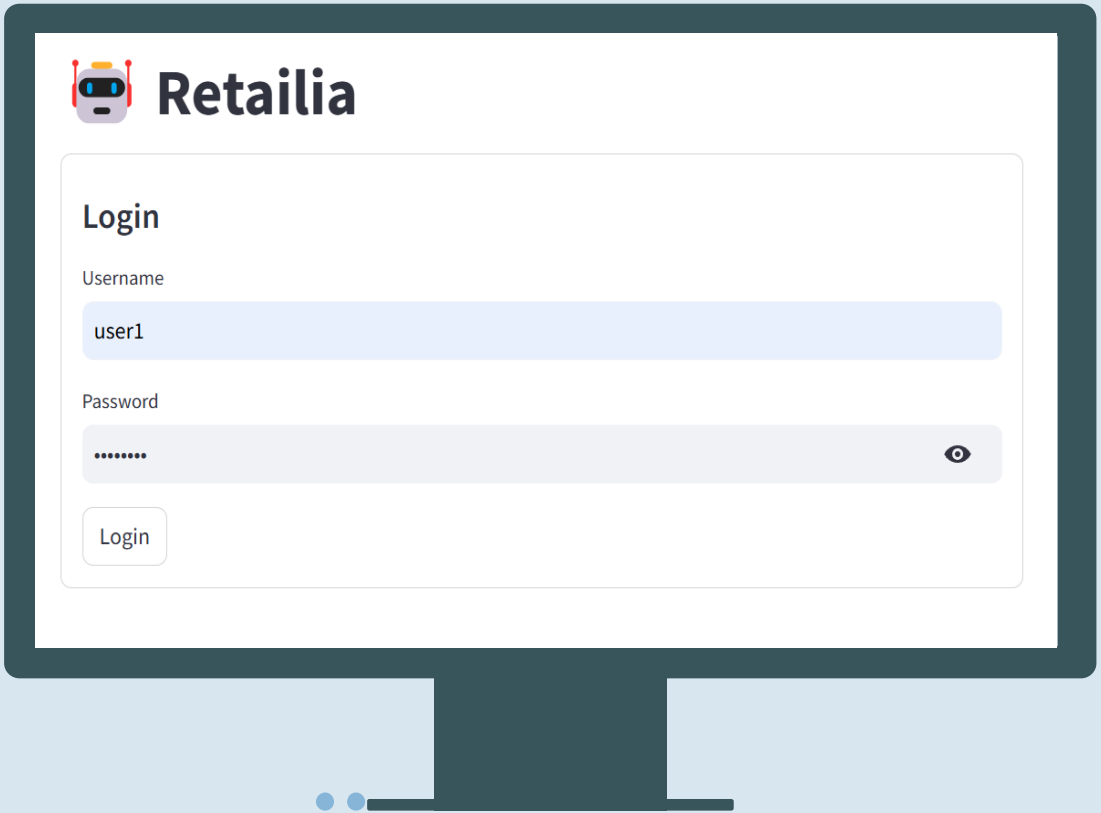





# Retailia in Development Environment

## Sneak peek

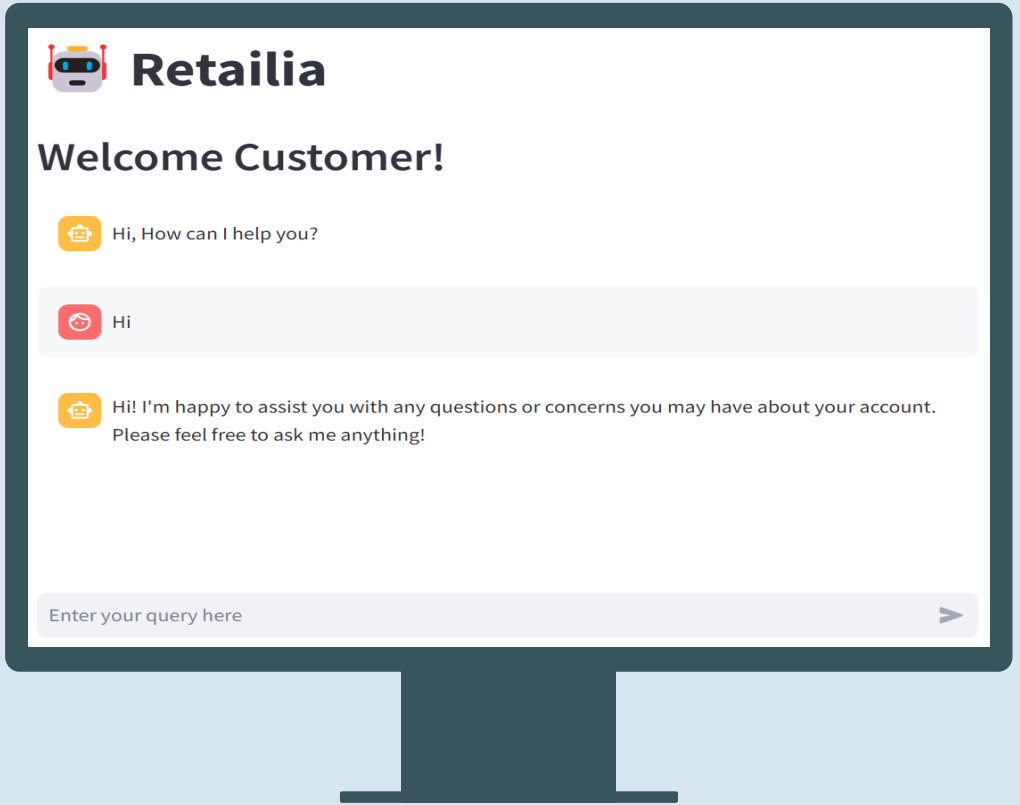
# Interfacing with Retailia: Streamlit’s Role




User Login

## Sneak peek

# Interfacing with Retailia: Streamlit's Role




**Chatbot Interface**

# Challenges

**While building 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.

# Contributions

Retailia project

Integrate | Automate | Invite / 1

Main Table

New task | Search | Person | Filter | Sort | Hide | Group by

Completed

| Task | Status | Team Members | Due date |
|---|---|---|---|
| Developed `faq_agent.py`,optimizing data retrieval with FAISS index for improved FAQ handling | Done | Shaik Manzur Elahi | Sep 1 - Nov 13 |
| Integrated Llama3 with Agentic RAG and designed intuitive Streamlit interface,elevating user experience | Done | Harish Inavolu | Sep 1 - Nov 13 |
| Integrated Llama3 with Agentic RAG and designed intuitive Streamlit interface,elevating user experience | Done | Suprathika Vangari | Sep 1 - Nov 13 |
| Implemented `sql_agent.py`,ensuring robust SQL database connectivity and streamlined data processing | Done | Sowmya Katla | Sep 1 - Nov 13 |

# 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!