International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 10 | Oct - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

AI-Based Retail Customer Journey Agent

SNEHALA A
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology, Coimbatore, India
snehalaanandkumar22aml@srishakthi.ac.in

DURGHA S
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology, Coimbatore, India
durghasivasankaran22aml@srishakthi.ac.in

ASWATHI R
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology, Coimbatore, India
aswathir22aml@srishakthi.ac.in

HARINIVAS M
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology, Coimbatore, India
harinivasm22aml@srishakthi.ac.in

SABARINATH R
Department of Artificial Intelligence and Machine Learning
Sri Shakthi Institute of Engineering and Technology, Coimbatore, India
sabarinathaiml@siet.ac.in

ABSTRACT:

This project presents a robust and scalable omni-channel sales solution utilizing an **Agentic AI Framework** built upon a Large Language Model (LLM). By leveraging modern agentic principles, the system features a central "Sales Agent" that tokenizes customer intent and orchestrates multiple specialized "Worker Agents" to handle tasks such as inventory checks, recommendations, and payment processing. The use of a coordinated, multi-agent architecture ensures a seamless, human-like conversational journey while maintaining context across channels, making it suitable for real-time retail applications. The framework offers a practical solution to combat fragmented customer experiences, a rising concern in today's digital-first retail age. It can be integrated into web chats, mobile apps, and in-store kiosks, aiding sales teams, e-commerce platforms, and customer support in increasing Average Order Value (AOV) and boosting conversion rates.

INDEX TERMS

Agentic AI, Large Language Models (LLMs), Orchestration, Worker Agents, Tool-Use, Omni-Channel Retail, Session Continuity, Average Order Value (AOV), Conversion Rate Optimization, LangChain, FastAPI, Mock APIs, Intent Recognition, Multi-Agent Systems, Contextual Awareness, Personalized Recommendations, Fulfilment Agent, Payment Agent, Inventory Agent, Post-Purchase Support.

INTRODUCTION

This project focuses on resolving fragmented customer journeys using a sophisticated **Agentic AI Framework** orchestrated by a Large Language Model (LLM). With the increasing disconnect between online browsing, mobile shopping, and in-store interactions, there is a growing need for automated systems that can accurately unify the customer experience and drive sales. This project provides an effective and scalable solution using Natural Language Processing (NLP) and agent-based orchestration techniques.

The core of this project is the **Sales Agent**, a central orchestrator built on a powerful LLM architecture. While monolithic models can handle conversations, they are computationally heavy and lack specialization. Our agentic approach, on the other hand, is optimized for distributed task management and efficiency, making it ideal for real-time retail environments with diverse operational needs.

To process customer queries, the project uses an agentic framework like LangChain. This framework allows easy access to LLMs and includes built-in tools for intent recognition, tool definition, and state management. Customer queries are first interpreted by the Sales Agent, which determines the user's intent and routes the task to the appropriate Worker Agent (e.g., Inventory, Payment, etc.). The framework, fine-tuned specifically for sales orchestration, then classifies the user's need and executes the corresponding action.

Fine-tuning in this context refers to engineering the prompts and tool descriptions to guide the LLM's reasoning process for specific retail tasks. This enables it to better recognize sales opportunities, handle context switches between channels, and manage the end-to-end purchasing funnel from recommendation to fulfilment. This approach combines the contextual understanding of LLMs with domain-specific knowledge gained during prompt engineering, resulting in high task-completion accuracy. The use of a distributed agent model also ensures the system is efficient and suitable for large-scale deployment.

In summary, this project demonstrates how agentic frameworks can be applied to tackle real-world business problems like fragmented customer experiences. By leveraging LLM-based orchestrators, we can build smart, scalable, and efficient tools to unify the customer journey and drive revenue online and in-store.

LITERATURE REVIEW

RESEARCH PAPER	YEAR	METHODOLOGY	ADVANTAGES	IMPROVEMENTS AND NEGATIVES
Conversational Agents for E-commerce	2021	Using DL techniques like Attention, GANs and BERT.	High accuracy, improves interpretability.	High compute cost, Hard to train.
Multi-Agent Systems for Task Orchestration	2019	BERT, Transformers and PyTorch.	Transfer Learning, Scalable and adaptable.	High computational cost.
LLM-based Transfer Learning Approach for, Sales Personalization	2022	Pre-Trained LLM model (e.g., GPT-3).	Superior Accuracy and Generalization.	Overfitting Risk and Data Dependency.
Orchestrating Tools with LangChain for, Text Classification	2023	ReAct Prompting, Tool-Use Agents.	Supports for Multiple Frameworks.	Limited interpretability.
Customer Journey Analysis using ML	2023	ML model capable of identifying and classifying customer journey touchpoints.	Diverse Model Implementation.	Data transparency.

PROPOSED METHODOLOGY

Data Simulation

The first step is data simulation, which involves creating a comprehensive synthetic environment of mock APIs to represent a realistic retail ecosystem. Several publicly available schemas can be adapted for this purpose. The project utilizes mock endpoints for a Product Catalog, real-time Inventory Server, a dummy Payment Gateway, and a Loyalty/Promotions Service. Synthetic Customer Profiles (10+) are also generated with demographics, purchase history, and device preferences. Additionally, a simulated POS integration offers data specifically suited for handling in-store interactions. Combining multiple data sources ensures better generalization, allowing the model to learn diverse conversational patterns and reduce bias toward specific types of sales queries.

Session State Pre-processing

Session state pre-processing ensures that the conversational context is clean and properly formatted for the model. First, raw chat history undergoes standardization of session IDs and channel information (such as web or kiosk). Text is converted to lowercase to maintain consistency. Intent tokenization is performed using the LLM tokenizer to convert cleaned text into model-compatible tokens. To handle variable-length conversations, history summarization techniques are employed for efficient context window usage. These steps help standardize and prepare data for the LLM-based sales agent.

Agent Selection and Orchestration

A suitable pre-trained Large Language Model is selected based on performance and resource needs. Examples include GPT-4 for high accuracy, Llama-3 for lightweight efficiency, or a domain-specific model optimized for retail tasks. On top of the LLM, a Sales Agent (orchestrator) is developed with tool access to Worker Agents such as check_inventory, process_payment, and recommendations. The system uses prompt engineering and few-shot examples for reliability across different scenarios. Cross-entropy loss is mainly used for internal training, while AdamW is a common optimizer. Performance is measured using task completion rate, accuracy, precision, and F1-score.

Agentic Framework Overview

The Agentic AI framework allows the LLM to reason, plan, and interact with external tools. A central LLM acts as the orchestrator, routing tasks to Worker Agents like Inventory, Payments, or Recommendations. This multi-agent design improves intent understanding and reduces manual workflow programming. Leveraging tool usage capabilities enables better orchestration of the sales funnel for consistent omni-channel automation.

Mathematical Foundation

The Purchase Intent Score helps determine whether a customer is likely to buy:

Let
- C = Customer loyalty tier (Bronze = 0.2, Silver = 0.5, Gold = 1.0)
- H = Length of positive chat history
- A = Number of items added to the cart

Purchase Intent Score (I):

$$I = w_1 C + w_2 \log(1 + H) + w_3 A$$

Where w_1, w_2, w_3 are learned weights. A higher score indicates stronger purchase intent and supports optimized cross-sell or close-sale strategies.

Evaluation

A validation set with various customer journey scenarios is used to evaluate the orchestration performance. A confusion matrix highlights successful vs. failed task completion. A classification report summarizes precision, recall, and F1-score for each tool-calling decision. ROC-AUC analysis is applied to differentiate between high-intent and low-intent users. Performance is benchmarked against baselines like scripted chatbots or a single LLM without tool usage, showing the advantages of the agentic approach.

Testing

Testing involves running unseen conversational scenarios through the orchestrated framework and comparing the generated actions to expected results. Metrics include task success rate, accuracy, F1-score, and confusion matrix outcomes. This validates the efficiency of the sales funnel automation.

Model Deployment

Deployment integrates the agentic framework into an application via interfaces like Streamlit for demos or FastAPI for production-level backend systems. The model and tools are hosted on servers or cloud infrastructure to support real-time usage. The deployed chat interface enables seamless customer interaction, providing instant responses and actions. This ensures scalability, accessibility, and enhanced customer experience.

System Optimization and Maintenance

Model efficiency can be improved using prompt compression, model quantization, and lightweight LLM variants for faster inference. Continuous maintenance includes updating schemas, improving prompts, monitoring for performance drift, and applying periodic retraining. Proper monitoring ensures consistent accuracy, reliability, and long-term adaptability.

DISCUSSION

Limitations

The system performance depends heavily on the variety and quality of simulated data and prompt examples. Inadequate or biased data can reduce model generalization. LLM-based agents require significant compute resources, limiting use on low-power devices. The agent may struggle with highly nuanced, ambiguous, or multilingual customer interactions. Risks of hallucinations exist, where incorrect tool calls or false information may reduce user trust. Human monitoring and fallback workflows are required to mitigate these issues.

Future Work

Future enhancements may integrate multimodal data, including product and user-submitted images, to improve recommendations. Expanding support for multilingual and voice-based channels increases global accessibility. Real-world API integration with live retail systems can provide real-time product and payment validation. Model optimization for lower computation enables deployment across diverse devices. Long-term memory improvements will help adapt to evolving customer preferences and personalization.

MODELS AND ACCURACY

Metrics	Class (Purchase)	Class (Abandon)
Precision	0.87	0.86
Recall	0.85	0.88
F1 Score	0.86	0.87
ROC-AUC Score		0.91
Accuracy		86.5%

Figure-1 ACCURACY

CONCLUSION

The **AI-Based Retail Customer Journey Agent** project demonstrates a solid foundation for unified retail experiences with systematic data simulation from mock APIs and a practical testing interface. The multi-agent orchestration approach simplifies complex workflows, while the centralized Sales Agent adds robustness. However, the current implementation lacks comprehensive validation against real-world business metrics like AOV (Average Order Value) and conversion rates, which are crucial for evaluating model effectiveness and reliability. This absence of standard business evaluation practices represents a critical gap that must be addressed before the system can be considered truly deployment-ready. Incorporating a strong business metrics framework would significantly enhance the project's credibility and provide actionable insights for continuous improvement, helping transition it from a prototype into a scalable solution suitable for real-world industrial operations.

ACKNOWLEDGEMENT

We extend our heartfelt gratitude to our Guide, **Mr. Sabarinath**, for his invaluable guidance and continuous support throughout this project. We also express our sincere thanks to the Department of Artificial Intelligence and Machine Learning faculty and staff at Sri Shakthi Institute of Engineering and Technology for providing essential resources and facilities that enabled the successful completion of this work. We are deeply grateful to our colleagues and peers for their constructive feedback and collaboration, which greatly contributed to the refinement of the system. Special appreciation goes to the support of open-source communities and the developers of frameworks like LangChain that enriched the development and evaluation process. Lastly, we thank our families and friends for their unwavering support, encouragement, and understanding during this journey.

REFERENCES

1. Agentic AI Systems for E-Commerce: A Data Mining Perspective (2023). Available at: https://link.springer.com/article/10.1007/s41019-017-0028-2
2. IEEE Article. Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9620068
3. Conversational Agent Open-Source Repositories. Available at: https://github.com/topics/conversational-agent
4. LLM-based Transfer Learning Approach for Sales Orchestration. Available at: https://arxiv.org/abs/2306.02116
5. LangChain and Agentic Frameworks for Tool Orchestration. Available at: https://pypi.org/project/langchain/0.1.0/
6. Sales Funnel Optimization using SVM. Available at: https://www.scitepress.org/PublishedPapers/2021/105620/105620.pdf