RetailARVA: AR-based Virtual Assistant with Conversational Capabilities to Enhance Retail Customer Experience

H.H.S. Fernando – Index No: 20020376

W.D. Kumudika – Index No: 20020597

P.M.B.R. Vimukthi – Index No: 20021097

Supervisor: Dr. K.D. Sandaruwan

May 2025

Submitted in partial fulfillment of the requirements of the
B.Sc. (Honours) Bachelor of Science in Information Systems
Final Year Project

UCSC

Declaration

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: H.H.S.Fernando

....

Signature of Candidate   Date : 29/05/2025

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: W.D Kumudika

....

Signature of Candidate   Date : 29/05/2025

I certify that this dissertation does not incorporate, without acknowledgement, any material previously submitted for a degree or diploma in any university and to the best of my knowledge and belief, it does not contain any material previously published or written by another person or myself except where due reference is made in the text. I also hereby give consent for my dissertation, if accepted, be made available for photocopying and for interlibrary loans, and for the title and abstract to be made available to outside organizations.

Candidate Name: P.M.B.R Vimukthi

....

Signature of Candidate Date : 29/05/2025

This is to certify that this dissertation is based on the work of

Ms. H.H.S.Fernando
Mr. W.D Kumudika
Mr.P.M.B.R Vimukthi

under my supervision. The thesis has been prepared according to the format stipulated and is of acceptable standard.

Principle/Co- Supervisor's Name: Dr. K.D. Sandaruwan

....

Signature of Supervisor Date : 29/05/2025

Abstract

Despite the rise of e-commerce, 94% of consumers continue to shop at brick-and-mortar stores, with 90% of millennial retail spending occurring in physical locations. However, in-store experiences often fail to meet the expectations shaped by online shopping, lacking real-time assistance, detailed product information, and personalized support. Studies show that 56% of in-store shoppers now use their smartphones to supplement product information, highlighting a growing informational gap in traditional retail.

This research introduces RetailARVA, a smartphone-based Augmented Reality (AR) virtual assistant enhanced with Large Language Models (LLMs), designed to bridge this gap. Focused initially on the skincare domain, RetailARVA enables real-time product identification through barcode scanning, delivers personalized recommendations based on user preferences, and provides natural conversational support via an interactive 3D avatar — all without the need for wearable hardware.

A controlled user study involving 30 participants demonstrated that users employing RetailARVA completed shopping tasks much faster, at best one task simulated in the user study was completed 80.32% faster. The study also reports a lower cognitive load (NASA-TLX scores), and the users rated the system's usability as "Excellent" with a mean System Usability Scale (SUS) score of 82.5. User engagement metrics from the User Experience Questionnaire (UEQ) showed significant gains in Attractiveness, Efficiency , and Stimulation compared to the control group. These results strongly support the effectiveness of integrating AR and conversational AI into physical retail environments to enhance customer experience, improve decision-making, and boost purchase confidence.

Acknowledgement

First and foremost, we would like to express our heartfelt gratitude to our supervisor, Dr. K.D. Sandaruwan, for his unwavering guidance, continuous support, and insightful feedback throughout the course of our research. His encouragement and expertise have been instrumental in helping us shape and complete this project successfully.

We extend our sincere thanks to the University of Colombo School of Computing (UCSC) for providing us with the academic environment, resources, and opportunities that enabled us to carry out this research effectively.

We are also deeply thankful to the participants of the user evaluation, whose valuable time, feedback, and engagement greatly contributed to the development and assessment of our system. Your involvement played a vital role in making this project more impactful and relevant.

Our sincere appreciation goes to the panel of examiners, Dr. Randil Pushpananda and Mr. Pubudu Liyanage, for their thoughtful evaluation, constructive suggestions, and encouragement during our project reviews.

We would also like to acknowledge the academic community, whose research and ideas laid the foundation for our work and inspired us to explore new directions within this field.

Finally, we are grateful to everyone who supported us throughout this journey, family, friends, colleagues, and mentors—for their encouragement, patience, and belief in our vision. Your support has meant the world to us.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

API Application Programming Interface

AR Augmented Reality

ATM Automated Teller Machine

DSR Design Science Research

GPU Graphics Processing Unit

ICL In-Context Learning

LLM Large Language Model

NASA TLX NASA Task Load Index

NLP Natural Language Processing

POC Proof of Concept

QAT Quantization Aware Training

RAG Retrieval Augmented Generation

STT Speech to Text

SUS System Usability Scale

TTS Text to Speech

UCSC University of Colombo School of Computing

UEQ User Experience Questionnaire

UI User Interface

UX User Experience

VA Virtual Assistant

VM Virtual Machine

VR Virtual Reality

Chapter 1

Introduction

The retail landscape has undergone a profound transformation over recent decades, driven by technological advancements and shifting consumer expectations. The growth of e-commerce has fundamentally changed how people shop, offering instant access to a wealth of information such as product specifications, customer reviews, and personalized recommendations. This digital shopping experience enables consumers to make informed purchase decisions, fosters trust, and builds loyalty toward brands. As a result, consumers have come to expect a shopping experience that is not only seamless and personalized but also enriched with meaningful, accessible insights. However, when it comes to brick-and-mortar stores, these features are often lacking, creating a significant gap between the online and offline shopping experiences.

Despite the rise of online shopping, physical brick-and-mortar stores remain highly relevant; a significant 94% of consumers still frequent brick-and-mortar stores, and a remarkable 90% of millennials' retail spending occurs in these physical spaces. This persistent loyalty to in-store shopping reflects an enduring consumer desire for experiential, tangible interactions with products. For instance, studies show that 72% of millennials are particularly drawn to in-store shopping for its interactive and sensory aspects, valuing real-world engagement over virtual interactions for certain shopping journeys.

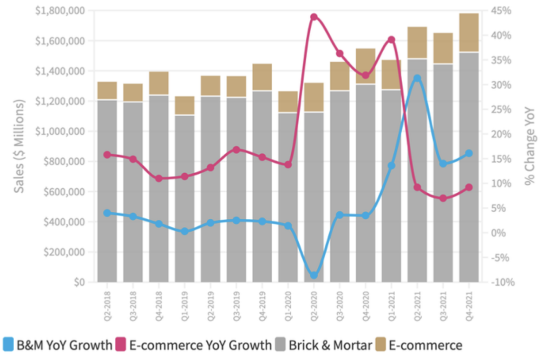


Figure 1.1: Brick and Mortar Store growth compared to E commerce growth

Yet, even as customers are drawn to physical retail for its hands-on experience, they increasingly bring with them expectations shaped by e-commerce. Physical brick-and-mortar stores often struggle to meet these new expectations, resulting in common customer frustrations. Shoppers frequently encounter limitations such as insufficient product information, lack of real-time assistance, and difficulty in comparing prices, all of which hinder their ability to make informed purchase decisions. These gaps in service not only affect the overall shopping experience but can also contribute to a sense of dissatisfaction.More than half (56%) of in-store shoppers now rely on their smartphones to gather product information and compare prices while shopping, underscoring an emerging trend: consumers want the informational convenience of e-commerce combined with the tactile experience of traditional retail.information.

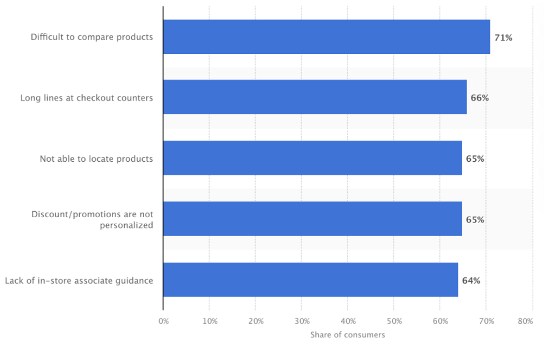


Figure 1.2: Customer Frustrations with Brick and Mortar Stores

1.1 Problem Statement

In the current retail landscape, consumer expectations are increasingly shaped by the personalized, information-rich experiences provided by e-commerce platforms. Online shopping offers instant access to detailed product specifications, customer reviews, and tailored recommendations—features that significantly influence purchase decisions and overall customer satisfaction. However, these advantages are often absent in traditional brick-and-mortar retail environments, leading to a noticeable gap between online and in-store experiences.

Despite the continued relevance and popularity of physical stores—particularly for products that benefit from tactile interaction, such as skincare—customers frequently face challenges such as limited access to product information, lack of real-time assistance, and minimal personalization during the shopping journey. This disconnect not only diminishes the quality of the customer experience but also reduces purchase confidence and engagement.

There is a pressing need for a solution that seamlessly integrates the digital benefits of e-commerce into physical retail spaces. While technologies such as augmented reality (AR) and virtual assistants have shown promise in enhancing interactivity and information accessibility, their practical application in retail environments remains limited—often constrained by hardware requirements, lack of personalization, or insufficient conversational capabilities.

This research addresses the problem of enhancing in-store shopping experiences by introducing a smartphone-based AR-enabled virtual assistant that provides real-time product identification, contextual information, and personalized conversational support. The aim is to bridge the informational and experiential gap between online and offline retail, making physical shopping more interactive, informative, and customer-centric.

1.2 Research Questions

1. **How can we integrate an AR-enabled virtual assistant, with conversational capabilities through a LLM and product identification features without wearables, in a physical retail environment?**
   This question explores the feasibility of building a practical system that enhances the in-store shopping experience using only a smartphone. The focus is on integrating augmented reality (AR) for product visualization, real-time product identification, and a virtual assistant powered by a large language model (LLM) to enable natural, human-like conversations—all without relying on wearable devices such as AR headsets or smart glasses. The aim is to determine how these technologies can be combined into a single, seamless application that works effectively in a real-world retail setting.

2. **What enhancements can be achieved by integrating a large language model (LLM) with product and customer data for use in a virtual assistant?**
   This question investigates the added value of combining LLMs with structured product and customer data to create more intelligent, context-aware interactions. By leveraging customer preferences, shopping history, and product attributes, the assistant can provide highly personalized recommendations, answer detailed product-related queries, and offer relevant suggestions in a conversational manner. The goal is to evaluate how this integration improves the quality, relevance, and personalization of customer interactions in comparison to generic or static assistance.

3. **What is the customer experience when using an AR-enabled, LLM-powered virtual assistant in a physical retail environment?**
   This question seeks to understand the end-user perspective, how customers perceive and interact with the system during their shopping journey. Key aspects include ease of use, satisfaction, trust in the assistant's recommendations, and overall impact on purchase intention and decision-making. This evaluation helps determine whether the proposed solution effectively addresses the pain points of traditional in-store shopping and whether it can enhance the shopping experience in a meaningful and measurable way.

1.3 Goals and Objectives

1.3.1 Goals

The primary goal of this research is to bridge the information gap between online and offline shopping experiences by leveraging smartphones to deliver an augmented and personalized shopping experience in brick-and-mortar retail stores. This will be achieved through real-time product identification, overlaying detailed product information, and providing personalized assistance in real-time conversation, ultimately enhancing purchase intentions and actual sales.

1.3.2 Objectives

Explore Consumer Behavior and Technological Needs
Investigate current consumer behavior and expectations in both online and offline shopping environments. Identify the technological gaps and limitations in existing retail store settings regarding real-time information access and personalization.

Deliver Product Information through AR
Design and develop a smartphone-based AR solution that can identify products in real-time and overlay detailed product information.Ensure the solution is user-friendly and does not require additional hardware such as headsets or markers.

Integrate Large Language Models (LLMs)
Incorporate LLMs to provide personalized recommendations and real-time interactivity. Enable responding to customer queries with anthropomorphic cues, enhancing user engagement and trust.

Enhance Personalization and Interactivity
Implement features that tailor recommendations and information delivery based on individual consumer preferences. Foster a seamless omnichannel experience that integrates physical product interaction with digital assistance.

Evaluate the Impact on Consumer Experience and Sales
Conduct user studies and surveys to assess the application's effectiveness in enhancing the shopping experience. Measure the impact on purchase intentions and actual sales in a retail environment.

By achieving these objectives, the research aims to revolutionize the retail shopping experience, making it more engaging, informative, and personalized, thereby helping traditional retailers to stay competitive in an increasingly digital world.

1.4 Research Approach

Following the design science research (DSR) methodology (Vom Brocke, J et al.) we design and evaluate IT artifacts that solve identified organizational and social problems. The DSR framework involve 6 steps 1) identifying problem and motivation, 2) Definition of objective and solution, 3) Design and development of artifacts, 4) demonstrates the use of the artifact to solve one or more instances of the problem, 5) Evaluate how well the artifact supports a solution to the problem and 6) Communicate aspects of the problem and artifact to the relevant stakeholders.

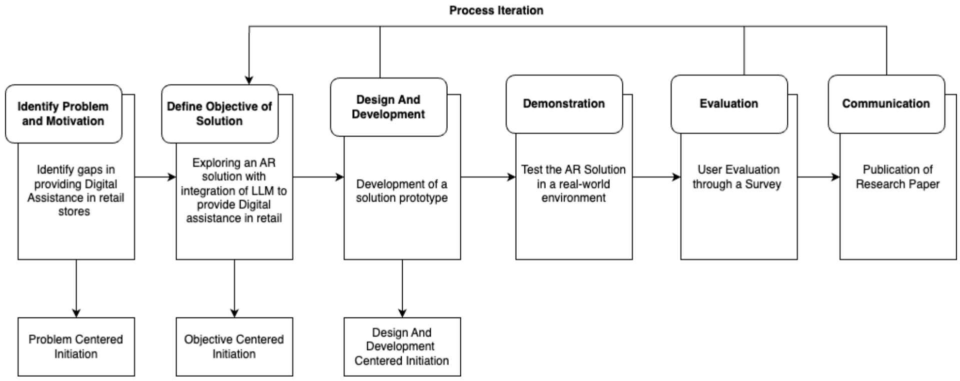


Figure 1.3: Research Methodology

Requirement Extraction

To extract the requirements, a review of research objectives is required. Based on the previously identified aims and objectives, we listed the requirements of the proposed artifact (mobile application).

Requirement	Description
Device (R1)	Consumer smartphone device to run the application
Data gathering (R2)	The application should be able to gather necessary data from the user
Product Recognition (R3)	The application is required to determine which product the consumer is interacting with through a marker.
Personalized Experience (R4)	The app is required to provide valuable recommendations based on the collected user data and preferences.
Avatar Representation (R5)	Create a digital avatar and present it to the user through AR, that interacts with the user through natural conversations.
Conversational Interface (R6)	Assistant should have integrated conversational abilities to the avatar through a LLM.
Artifact Evaluation (R7)	Application should be evaluated in a controlled environment

Table 1.1: Requirements for the AR Shopping Assistant Application

Artifact Development

The proposed artifact will be developed as a mobile solution running on an android-based smartphones (R1). In our scenario, we achieve personalization using data about the user's purchase history and behavior. However, the precise inputs first have to be identified and extracted (R2), as can be seen in the above Table. Based on the initial user inputs, the system then has to assess the customer's needs. In particular, the system must be capable of detecting the products that the customer is exploring and interacting with them (R2) using the device camera and to display related content to assist the customer through the digital avatar (R4). Using the built-in camera in the mobile device, we can even monitor the user's field of view to determine which precise product they are examining at each point in time. Target object recognition can be achieved using Software Development Kits such as Vuforia. If an object is identified and tracked, this triggers the display of relevant UI elements along with the avatar. The inputs from the mobile device (R1), in particular the current object of interest (R2), promise the potential for precise recommendations. However, they must be processed and integrated into the design of the system, which yields the third requirement: the processing of the input data to generate real-time contextual product recommendations (R3). Beside the recommendation of alternate and relevant products, another important part of a consumer's shopping journey is getting contextual information of the object of interest and other similar products from the avatar through human-like natural conversation (R5). Device's inbuilt microphone will be used to take speech as an input to start conversation with the avatar (R6) and for that, a proper Text-to-Speech capability needs to be integrated into the proposed artifact. Also Text-to-Speech service integration is needed to convert the generated text from LLM to power up the avatar with voice capabilities.

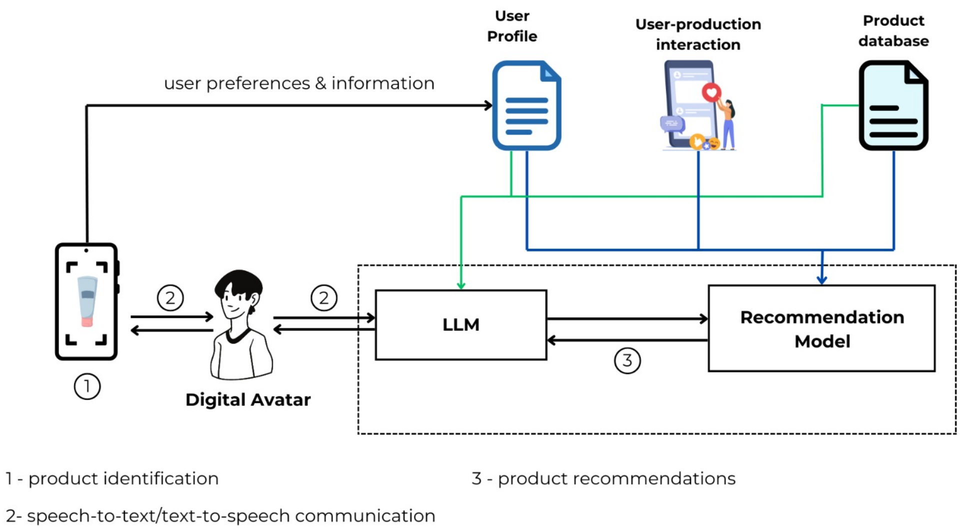


Figure 1.4: High Level representation of Research Methodology

Artifact Evaluation

The evaluation will be conducted through a controlled comparative study involving two groups, a control group, which completes shopping tasks without the AR assistant, and an experimental group, which performs the same tasks using the AR-enabled, LLM-powered virtual assistant. After completing their respective sessions, participants from both groups will fill out standardized instruments including the System Usability Scale (SUS), User Experience Questionnaire (UEQ), and NASA Task Load Index (NASA-TLX) to capture insights on usability, emotional engagement, and cognitive load. Additionally, semi-structured interviews will be carried out to gather rich qualitative data about user perceptions, challenges, and the perceived value of the assistant.

This evaluation strategy is tightly aligned with the research questions. RQ1, which focuses on the feasibility of integrating an AR assistant with product recognition and conversational features in a retail setting, will be addressed through task performance metrics (e.g., completion rate, accuracy, and interaction flow) and qualitative feedback about system integration and user interaction. RQ2, examining the enhancement offered by LLM integration, will be informed by UEQ subscales related to perceived intelligence, novelty, and personalization, along with interview data reflecting how well the assistant understood and supported users. RQ3, exploring overall customer experience, will be evaluated through a combination of SUS, UEQ (e.g., Attractiveness, Stimulation), NASA-TLX, and open-ended responses, revealing both satisfaction and cognitive effort. By analyzing and comparing results across both groups, the study provides a comprehensive assessment of the artifact's effectiveness, usability, and experiential impact in relation to the defined research questions.

In summary, the research approach is grounded in the Design Science Research (DSR) methodology, providing a structured framework for the iterative development and evaluation of the AR-based shopping assistant. By systematically identifying requirements, designing a functional prototype, and evaluating it through a controlled experimental setup, the study aims to generate actionable insights into the integration and effectiveness of emerging technologies in physical retail environments.

1.5 Limitations, Scope and Assumptions

1.5.1 Limitations

- The initial focus was narrowed down to a single domain, Skincare products and one product category as it requires a substantial amount of development effort for a prototype and a larger dataset.

- The virtual assistant capabilities will be limited to responding basic product related inquiries and providing product recommendations due to time constraints. The integration of more advanced conversational capabilities and a more personalized shopping interaction will be explored through further development and testing as the project progresses.

- The initial development phase will not prioritize offline functionality for the application.
- AR technology limitations like accuracy, battery consumption, and smartphone processing power may affect the user experience.
- The Initial Development will be focused on Pre-defined static product database instead of providing support for dynamic set of products.

1.5.2 Scope

This research aims to explore a solution that leverages the ubiquitous nature of smartphones to create a personalized and augmented shopping experience for customers in retail stores, focusing on the skincare product category initially.

In scope

- Developing a solution leveraging AR and advanced conversational interfaces to bridge the information gap between online and offline shopping experiences in retail stores, focusing on skincare products.
- Identifying products in real-time within retail environments and presenting product information in real time through AR Technology
- Integration of a conversational AI powered by Large Language Models (LLMs) enabling real-time interaction . This can be manifested as a digital avatar for a more engaging and relatable experience.
- The avatar will leverage natural language processing (NLP) for answering product-related queries and offering personalized advice based on user profiles and product information.
- User testing and feedback collection will assess the application's effectiveness in achieving its goals.

1.5.3 Assumption

1. Technological Readiness and Accessibility
    - Users possess smartphones capable of running AR applications and supporting real-time processing for AR, speech-to-text, and text-to-speech functionalities. The solution assumes that smartphone hardware and battery life are sufficient for these tasks, and users are comfortable with installing and using such applications.

- The application will initially be developed and tested for Android platforms, assuming a significant portion of the target audience uses Android devices.

2. User Behavior and Acceptance

   - Consumers are willing to use their smartphones during in-store shopping to access augmented and personalized product information, and will engage with a virtual assistant for assistance and recommendations.
   - Users are comfortable interacting with digital avatars and conversational AI in a retail setting, and will find the anthropomorphic cues (such as natural language and avatar presence) engaging and trustworthy.

3. Data Availability and Quality

   - Sufficient and accurate product data (including attributes like ingredients, price, reviews, etc.) is available for the targeted product category (skincare) and can be structured for use in a retrieval-augmented generation (RAG) system.
   - The synthetic dataset created for model training and evaluation is representative of real-world product information and customer queries, ensuring the system's responses are relevant and realistic.

4. Personalization and Privacy

   - Users are willing to provide necessary personal data (such as skin type, preferences, and purchase history) for the system to generate personalized recommendations.
   - The system can access and process user data securely and in compliance with privacy regulations, even though the prototype may use only a subset of possible personalization features.

5. Evaluation Environment

   - Usability and effectiveness can be reliably assessed in a controlled environment simulating a retail store, with representative users and tasks that reflect real-world shopping scenarios.
   - Feedback from this controlled setting is assumed to be indicative of broader user acceptance and system impact in actual retail environments.

1.6 Contribution

This research makes several key contributions by addressing the intersection of augmented reality (AR), conversational AI, and physical retail. Firstly, it aims to bridge the gap between online and offline shopping experiences by demonstrating how AR and AI-powered virtual assistants can enhance in-store interactions, thereby contributing to a more unified and seamless retail landscape. The integration of AR functionalities for product identification and personalized information delivery contributes to the advancement of AR technology in retail, encouraging its broader adoption and the exploration of new, practical use cases. Furthermore, this study expands the understanding of human-multimodal interaction by combining visual overlays with spoken conversation in a single immersive interface. By analyzing user engagement within this context, the research sheds light on how different sensory modalities can be effectively combined to improve user experience and information accessibility. Additionally, the study explores how user data—such as purchase behavior and product interaction—can be utilized for real-time, personalized recommendations, contributing to the growing body of knowledge on data-driven personalization in physical retail environments. Collectively, these insights inform both academic inquiry and practical innovation in intelligent retail systems and interactive technology design.

Contribution to Society

The application aims to empower consumers in retail stores by providing them with instant access to detailed product information and personalized recommendations. This can lead to a more informed and efficient shopping experience. The interactive nature of the AR overlays and the conversational AI avatar can potentially increase customer engagement in physical stores, making shopping more interactive and enjoyable. The personalized recommendations generated by the application can empower consumers to make more informed decisions about skincare products based on their individual needs and preferences. AR technology has the potential to improve accessibility in retail stores for visually impaired customers. Future iterations could explore integrating features that leverage audio descriptions or haptic feedback alongside AR overlays.

Chapter 2

Background

The retail landscape is undergoing a significant transformation, fueled by the exponential growth of e-commerce and shifting customer expectations. In recent years, online shopping has provided consumers with unparalleled convenience and access to detailed product information, reviews, and personalized recommendations, all of which have become key drivers in the decision-making process. Consequently, modern shoppers have come to expect a similarly rich and interactive experience from brick-and-mortar stores as well. These evolving expectations pose both a challenge and an opportunity for traditional retailers, who must now reimagine the in-store experience to maintain competitiveness in a digital-first world [1].

In this highly competitive environment, retailers are prioritizing strategies to enhance the physical shopping experience, with the goal of meeting the demand for information-rich, personalized, and interactive engagements. This shift has given rise to the concept of "digital retail," where information and communication technologies are harnessed to enhance the in-store environment. By integrating digital tools like smartphones, retailers can now offer customers real-time access to product details, personalized recommendations, and even interactive experiences directly within the store [2]. These digital enhancements not only drive sales but also create unique and memorable interactions that go beyond what e-commerce can offer on its own.

As consumer expectations grow increasingly sophisticated, retailers are finding that a seamless omnichannel experience is becoming essential. Shoppers want their online and in-store journeys to be interconnected, enabling them to easily transition between digital and physical platforms [3]. Elements such as personalization and interactivity are now central to fostering a positive customer experience, as they directly impact consumer engagement, satisfaction, and ultimately, purchase behavior. By merging digital tools with the tactile appeal of physical stores, retailers can offer a dynamic and immersive experience that aligns with modern consumer preferences, helping to influence buying decisions while building loyalty and long-term customer relationships.

Personal Assistance in E-Commerce

Within the realm of e-commerce, virtual assistants (VAs) are emerging as a pivotal tool for providing real-time support and personalized experiences for online shoppers. These VAs, often in the form of chatbots powered by advanced Large Language Models (LLM), engage in human-like conversations, enabling them to autonomously interpret and respond to customer inquiries in a way that is both proactive and adaptable. By providing this conversational support, VAs open up a new, dynamic communication channel between retailers and consumers, with users often perceiving chatbots as embodying a virtual "personality" that enhances the shopping experience.

These chatbots act as intelligent conversational agents, guiding customers through the often complex process of online shopping. They provide tailored product recommendations, answer specific questions, and even offer insights into promotions or complementary products, helping consumers make informed purchase decisions. This capability is particularly valuable, as studies indicate that chatbots contribute to a perception of efficiency, innovation, and helpfulness for retailers, which positively influences customer attitudes and builds brand trust [4].

In addition to enhancing the consumer experience, the impressive growth of the global chatbot market underscores the value these virtual assistants bring to e-commerce. With the market reaching $0.84 billion in 2022 and projected to surge to $4.9 billion by 2032, the rapid adoption and expansion of VAs reflect their efficacy in meeting consumer expectations for immediate assistance and personalized shopping. For retailers, this success provides a roadmap for using VAs not only to facilitate transactions but to cultivate a lasting and interactive relationship with customers.

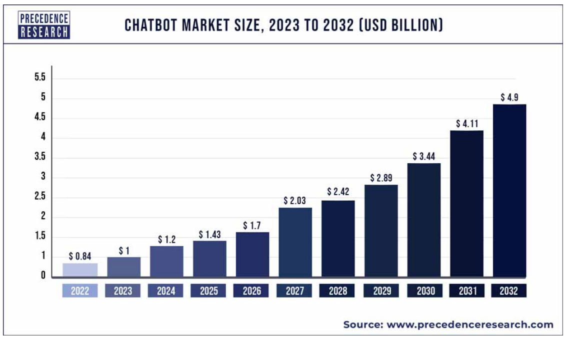


Figure 2.1: The global chatbot market size

Ultimately, as VAs evolve with advancements they hold the potential to become even more intuitive, bridging gaps in the digital shopping experience and making online interactions feel as personal and supportive as in-store service.

Transformation of Personal Assistance by Large Language Models (LLMs)

The advent of large language models (LLMs) has revolutionized the capabilities of chatbots in e-commerce. LLM-based chatbots are probabilistic, predicting the next word based on previous ones, unlike rule-based chatbots. Their ability to understand and respond in a human-like manner enhances customer service, providing personalized responses for a better experience.

The recent advancements in fine-tuning large language models (LLMs) on conversations, combined with methods like instruction fine-tuning and reinforcement learning from human feedback (RLHF), have significantly contributed to the creation of highly human-like chatbots in the market [5]. These human-like interactions foster a sense of engagement and trust and the anthropomorphic features of LLM-powered chatbots can influence purchase decisions by making interactions more relatable and intuitive. The personalization achieved through these models can tailor recommendations based on detailed customer data, further driving sales and enhancing the shopping experience.

With the introduction of powerful Transformer architecture based Large Language Models (LLMs) around 2020-2021, the nature of Personal Assistant Applications have been changed significantly. Proprietary LLMs such as GPT-3.5, GPT-4 and Gemini as well as open-source alternatives like LLaMA, Mixtral, Falcon [6], emphasize the feasibility of using them as conversational assistant like applications for different industries (ex-: e-commerce, healthcare, education) [?]. Cutting-edge models like GPT-4 have been evaluated as customer support chatbots for E-commerce through the integration of user profiling systems for personalized responses. GPT-4 was fine-tuned by exposing it to diverse e-commerce interactions which included dataset of various products. The context aware, personalized e-commerce chatbot developed through this methodology was able to achieve 95% accuracy improvement over comprehending user queries when compared with other rule-based and template-based systems [7].

The Need for Personalized Digital Assistance in retail

The retail landscape is undergoing a transformation driven by technological advancements and the rise of digital-native consumers demanding convenience and engaging experiences [8]. This shift necessitates a customer-centric approach that merges the strengths of online and physical shopping, fostering a seamless omnichannel experience. Conversely, traditional brick-and-mortar stores are investing heavily in digital infrastructure. Personalization is the key driver across both online and offline channels. This emphasis on personalization, a hallmark of online shopping, is now influencing in-store expectations. Consequently, retailers are actively exploring smart technologies to personalize the in-store experience for individual customers.

The rise of Digital Retail

Retail stores are gradually embracing digital retail and new technology to drive purchase decisions. Digital retail refers to the use of digital technologies to enhance the shopping experience, streamline operations, and integrate online and offline retail channels. To drive this new concept, technologies such as augmented reality (AR) are being integrated to provide immersive and positive experiences for shoppers.Traditionally, product experiences have been categorized as either direct or indirect. Direct experience refers to the unmediated interaction with a product using all available senses.Indirect experience, in contrast, relies on secondary sources to supplement a consumer's understanding of a product.While direct experience provides valuable firsthand information, consumers are restricted to examining the product itself. This hinders the incorporation of external information like product history, typical users, and usage scenarios.

Figure 2.2: Information Overlaid in AR environment through Head Mounted Displays

Virtual experiences, exemplified by 3D visualizations, offer solutions to the limitations of direct experience. Augmented reality (AR) takes this a step further by enabling simultaneous physical product inspection and 3D visualization of additional objects on a tablet display. This enhances the shopping experience and also empowers customers to make more informed and confident purchasing decisions. The integration of AR in retail can lead to increased customer satisfaction, higher conversion rates, and a stronger brand connection[9].

In addition to AR, the use of virtual avatars in digital retail has significantly enhanced the customer shopping experience. For example, Moltenbrey and Fischer highlight how virtual avatars in augmented shopping environments enable customers to see outfits on an avatar in AR, leading to better alignment with customer needs [10]. Some studies have shown that talking avatar-mediated interactions in virtual stores can influence consumer behavior positively by providing a sense of social presence and personal connection [11]. Interaction with virtual sales assistants enhances the shopping experience, making it more engaging and informative by bridging the gap between direct and indirect experiences by providing a personalized and interactive approach to shopping.

The Role of Smartphones in Digital Retail

Smartphones play a crucial role in bridging the gap between physical and digital shopping experiences. While mobile commerce (mCommerce) may not yet dominate direct purchase volume, its influence on in-store decisions is significant. A study [12] revealed that nearly half (45%) of shoppers utilize their smartphones for in-store price comparisons and product research, highlighting the considerable impact these devices have on consumer choices. This phenomenon, known as "showrooming," empowers customers by allowing them to access real-time product information, reviews, and comparisons directly in stores. Furthermore, the high adoption rate of retailer mobile applications (65% of consumers) underscores a growing expectation for a seamless omnichannel experience. Customers anticipate consistent interactions, prompt and accurate information access, and a positive overall shopping journey.

Chapter 3

Literature Review

As consumer behavior shifts and digital retail rises, in-store shopping now demands more engaging and informative experiences to compete with the convenience and immediacy of e-commerce. Researchers are therefore exploring innovative technologies to bridge the gap between physical and digital shopping, aiming to create immersive, interactive experiences that align with modern consumer expectations.

3.1 Chatbots in E-Commerce

In the e-commerce domain, platforms are increasingly leveraging chatbots to address the growing consumer demand for ”purchase literacy”—the ability to make informed, responsible purchasing decisions. Chatbots have been shown to enhance customer engagement by providing quick, clear information, thereby supporting a more interactive and confident decision-making process. [13] explored Gen Z's experiences with chatbots, specifically focusing on their engagement and the emotional aspects of human-like interactions. Their study revealed that while chatbots excel at rapid response and basic engagement, they face limitations in handling complex queries and delivering a more human-centered, emotionally connected experience. This lack of emotional connection, along with limited anthropomorphic cues in language, was noted as a barrier to fully engaging Gen Z customers at a deeper level, highlighting the need for further advancement in chat bot design to meet this demographic's expectations.

This aligns with existing literature emphasizing the importance of responsiveness and clear communication in fostering positive customer engagement. Another study investigated the impact of AI-based conversational agents with anthropomorphic features on consumer behavior [14]. Anthropomorphism refers to chatbots using language that mimics human qualities, such as personal pronouns and emotional expression and engaging in natural conversational dialogue.

This research found that users perceived products as more personalized when interacting with anthropomorphic chatbots, and were even willing to pay higher prices. Finally, a separate study explored the impact of personalized content on consumer attitudes towards product recommendations within voice shopping contexts [15]. The research demonstrates that personalized language significantly enhances user sentiment, particularly for high-involvement products. These findings highlight the importance of tailoring recommendations and information delivery to individual preferences for a more persuasive and engaging shopping experience in e-commerce.

3.2 Augmented Reality (AR) and Mixed Reality (MR) for Enhanced Retail Experiences

To meet in-store consumers' demands for purchase literacy and interactive shopping experiences, researchers have been exploring various AR and MR applications. [16] investigated the use of Optical See-Through Mixed Reality (OSTMR) technology, which overlays product information such as peer reviews and recommendations in an augmented reality space. Although users responded positively to OSTMR's potential to enhance their shopping experience, limitations—such as the need for cumbersome head-mounted displays and the limited capacity for interactivity—were identified. These constraints reduced the overall engagement potential, as users could not engage in question-and-answer interactions or receive fully personalized responses in real time.

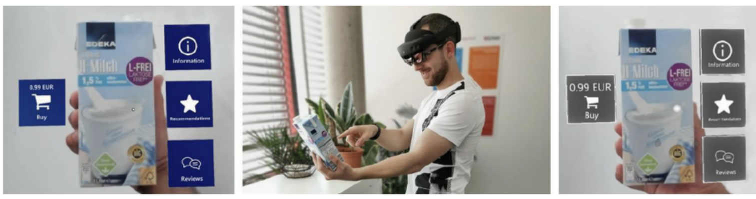


Figure 3.1: Optical See-Through Mixed Reality with specialized hardware

To address these limitations, [17] introduced the Smart Phone-based Mixed Reality Application (SPMRA), a more accessible and cost-effective MR solution that leverages the widespread availability of smartphones. By transforming a standard phone into an immersive MR device, SPMRA allows users to navigate retail environments and interact with virtual elements, enhancing both interactivity and purchasing decisions. In their study, 82.1% of participants reported that SPMRA positively influenced their purchasing decisions. However, the requirement for additional hardware and limited interactivity remained barriers to widespread adoption.

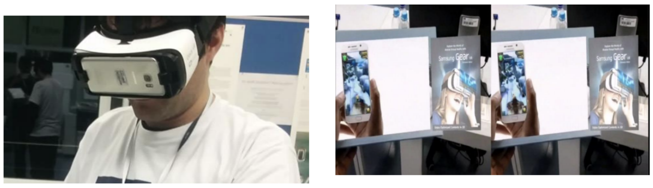


Figure 3.2: SPMRA - with Phone placed in virtual reality headset

A more recent study by [18] aimed to overcome these challenges by developing a mobile AR shopping web application that does not require specialized hardware. This application leverages a two-tier architecture, comprising a progressive web app (PWA) for real-time object tracking and rendering, and a cloud-based server for object detection and information retrieval. The system enables users to access product details like nutritional facts simply by highlighting items on shelves, making the AR experience more seamless and user-friendly. Despite these advancements, challenges remain in achieving the level of product engagement and emotional interaction that chatbots in e-commerce settings can offer, particularly when it comes conversational interface and personalization

3.3 Personalized Recommendations and AI-Driven ShoppingAssistants

[9] explored the impact of a mobile AR shopping assistant application, which uses personalized recommendations powered by explainable AI (XAI). In their study, perceptions of "usefulness," "entertainment," and "informativeness" were positively associated with the AR assistant, though not with "purchase intention." This highlights the potential of personalized recommendations and enhanced interactivity in AR to improve user engagement, even if they do not directly increase purchase intentions.

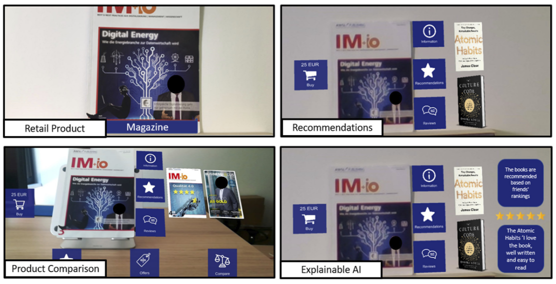


Figure 3.3: Product Information overlaid with a social explanation based on others reviews

An important aspect of the study was its exploration of XAI in recommendations, where participants could see explanations for why products were suggested. These explanations, categorized into types such as feature-level, user-based, and social cues, were positively received and deemed helpful by users.

The research suggests that in-store AR applications can serve as valuable tools for supporting consumer purchase literacy and enhancing the overall shopping experience, although maintaining customer engagement with product information and recommendations remains challenging.

3.4 LLM-Based Personalized Recommendations

The application of Large Language Models (LLMs) in personalized recommendation systems has garnered significant attention in recent research, especially in domains such as museums and e-commerce, where context-awareness and user-specific personalization are key. LLMs, particularly those like GPT-4, have shown great potential in incorporating contextual information and user instructions, thereby providing personalized recommendations that are better aligned with user preferences. A prime example of this is the use of LLMs trained on domain-specific data, such as museum exhibit descriptions, relevant literature, and external sources, which enables them to generate more accurate and context-sensitive recommendations in the museum domain (LLMs as Recommendation Systems in Museums)[19].

Recent advancements have focused on fine-tuning pre-trained models to improve their personalization capabilities. By leveraging vast, domain-specific datasets, fine-tuning has been shown to enhance the model's ability to provide tailored recommendations. A key development in this area is TALLRec, a tuning framework designed to align LLMs with recommendation systems. TALLRec has demonstrated significant improvements in recommendation quality, particularly in the movie and book domains, even when trained on a limited dataset of fewer than

100 samples. This highlights the effectiveness of fine-tuning as a methodology for optimizing LLM performance in recommendation tasks[20].

Similarly, PALR (Personalization-Aware LLMs for Recommendation) and InstructRec (Recommendation as Instruction Following) have introduced instruction-tuning frameworks that allow LLMs to better understand and process user preferences, further enhancing the personalization aspect of recommendations[21][22]. These methodologies have proven to outperform traditional recommendation systems by incorporating user-specific context and semantic knowledge.

In contrast to fine-tuning, CHAT-Rec offers an alternative approach by utilizing in-context learning to augment traditional recommender systems. Instead of fine-tuning the model, CHAT-Rec leverages the LLM's capacity for in-context learning to improve recommendation quality interactively. This method has been shown to surpass traditional recommendation systems in terms of adaptability and user engagement[23].

Overall, the integration of LLMs into recommendation systems, through either fine-tuning or in-context learning, has revolutionized the personalization process across various domains. By incorporating instruction-based frameworks and domain-specific data, LLMs have significantly enhanced the precision and effectiveness of personalized recommendations, proving them to be a valuable tool in building more intelligent, context-aware systems.

3.5 LLM-Based 3D Avatar Assistants in E-Commerce and Retail

The research by [24] explores the development of LLM-based 3D avatar assistants, highlighting their potential to enhance human-computer interactions by providing highly engaging, visually interactive, and conversationally dynamic virtual assistants. These assistants leverage state-of-the-art natural language processing (NLP) capabilities, enabling them to understand context, adapt tone, and adjust visual elements to align with conversational intent, making interactions feel more human-like and emotionally resonant

The proposed avatar system integrates both online and offline speech recognition, switching between Google Cloud and Pocket Sphinx for optimized functionality based on network availability. Additionally, emotion analysis, powered by the multinomial Naive Bayes algorithm, enhances expressiveness, allowing the avatar to adjust its tone and facial expressions to match the emotional context of conversations. Text-to-speech (TTS) features further enhance customization with options for pitch, speed, and pauses, while intent classification—achieved through logistic regression—improves response relevance by identifying user intent within the interaction

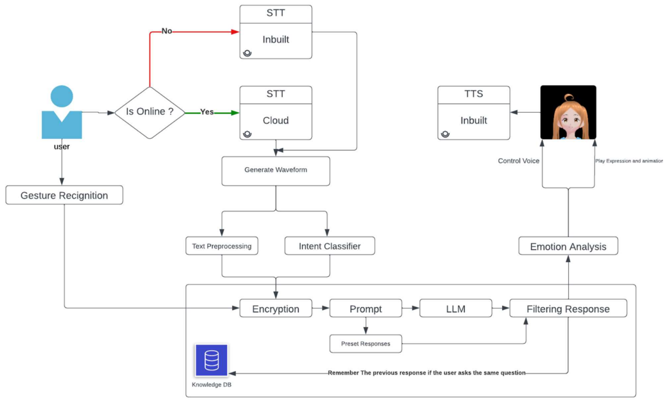


Figure 3.4: Proposed Methodology for LLM based Avatar

The 3D visual component of the avatar assistant is powered by Unity and implemented to run efficiently on devices with limited processing power, making it accessible for a broad range of platforms. Unity's engine, combined with the use of efficient 3D modeling techniques, allows the avatar to maintain smooth animations and responsiveness without the need for extensive hardware resources. Compared to existing systems, such as MITUSHA and Gatebox, which either require GPU-based processing or are limited to single-language interactions, this system provides a multi-lingual and low-resource alternative. This was validated during usability tests, where the assistant demonstrated consistent performance and user engagement across multiple devices

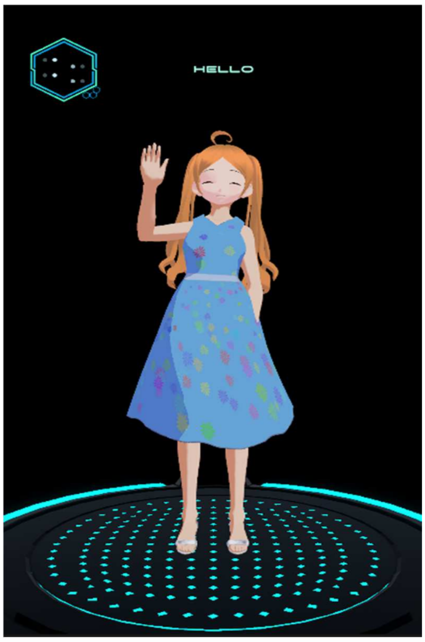


Figure 3.5: Avatar Presence as a personal assistant

User studies conducted over six weeks have shown that the 3D avatar interface leads to a 40% increase in task completion rates and a 25% reduction in user frustration, showcasing the impact of integrated NLP and 3D interactivity on user satisfaction. Additionally, the assistant's flexibility in adapting to diverse user scenarios, combined with real-time responsiveness, has demonstrated significant improvements in engagement across various demographics, positioning these avatars as transformative tools for customer service, entertainment, education, and e-commerce applications

3.6 Research Gap

Related work	**[13]**	**[14]**	**[15]**	**[16]**	**[17]**	**[18]**
Method of Shopping	Ecommerce	Ecommerce	Ecommerce	Retail	Retail	Retail
Provision of Real Time Product Information via application	Yes	Yes	Yes	Yes	Yes	Yes
Integration of LLM	No	Yes	Yes	No	No	No
Personalization based on consumer preference	No	Yes	Yes	No	No	No
Inclusion of Anthropomorphic cues in language	No	Yes	Yes	No	No	No
Real time interactivity with application	Yes	Yes	Yes	No	No	No
AR based Sales Experience	No	No	No	Yes	Yes	Yes
Not Require wearable devices for the sales experience	No	No	No	Yes	Yes	Yes

Table 1.5: Critical Analysis of Related Work

The analysis of existing research on customer experiences in both e-commerce and retail shopping modes highlights several notable gaps in the context of retail stores. While retail stores researched in providing product information, they lag in real-time interactivity and engagement. AR, a prevalent feature in online settings, has shown promise in increasing purchase intention through immersive experiences. However, the challenge lies in making AR interactive enough to facilitate meaningful customer engagement, such as providing real-time questions and answers. One of the primary gaps identified is the lack of personalization based on consumer preferences. Specifically, none of the offline studies ([18], [17],[16])integrate personalized shopping experiences, contrasting sharply with their online counterparts ([15], [14]), which do provide such customization. Additionally, another significant gap is the absence of anthropomorphic cues in language within offline studies. Unlike online platforms that increasingly use virtual assistants and chatbots with human-like communication features to foster trust and engagement ([15], [14]), physical stores have not adopted similar approaches. None of the offline studies ([18], [17],[16]) incorporate anthropomorphic elements, such as friendly and human-like interactions. These gaps suggests that physical retail environments are currently unable to offer tailored recommendations or customized assistance that align with individual consumer data, potentially leading to a less engaging and satisfying shopping experience.

Chapter 4

Methodology

4.1 Research Approach

As states in Section 1.4 , this study adopts the Design Science Research (DSR) methodology as its primary research approach. Design Science is particularly well-suited for research that involves the creation and evaluation of innovative artifacts intended to solve identified problems [25]. Given that RetailARVA is a novel system integrating augmented reality (AR) and conversational AI for the retail sector, the Design Science paradigm provides an appropriate framework to guide the development, iterative refinement, and evaluation of the system.

According to Hevner et al, DSR emphasizes two key activities , the build phase, where an artifact is created, and the evaluate phase, where the artifact's utility, quality, and efficacy are assessed. In this study, the build phase involved designing and implementing the RetailARVA prototype, incorporating AR capabilities, voice interaction, and personalized product recommendations. The evaluation phase included laboratory-based usability testing, workload assessments (via NASA TLX), and user experience evaluations (via UEQ).

Importantly, DSR stresses the relevance of the artifact to real-world problems as well as its rigor in design and evaluation. Peffers et al. [26] further structured DSR into a process model consisting of six activities, problem identification and motivation, definition of objectives for a solution, design and development, demonstration, evaluation, and communication. This process was systematically followed during the project to ensure methodological rigor.

Through the lens of DSR, RetailARVA is not only a functional system but also a research contribution that extends existing knowledge about the application of AR and virtual assistants in retail contexts.

4.2 Problem Identification

As outlined in Section 3.6, Physical retail settings still cannot deliver tailored recommendations or real-time customized support, leading to less engaging and less satisfying shopping experiences compared to digital platforms. Addressing this gap requires the development of an intelligent, adaptive solution that supports personalized decision-making and enhances customer satisfaction.

To address this gap, this research was guided by the following research questions,

1. How can we integrate an AR-enabled virtual assistant, with conversational capabilities through a LLM and product identification features without wearables, in a physical retail environment?

2. What enhancements can be achieved by integrating a large language model (LLM) with product and customer data for use in a virtual assistant?

3. What is the customer Experience when using an AR-enabled, LLM-powered virtual assistant in a physical retail environment ?

These research questions shaped the development and evaluation process, ensuring that the artifact directly targets the identified problems and user needs.

4.3 Definition of Objectives for a Solution

The primary objective of this research is to design and develop a novel AR-enabled virtual assistant system (RetailARVA) aimed at enhancing customer engagement and decision-making in physical retail environments. By leveraging Augmented Reality (AR) technology and machine learning-based personalization, this system seeks to address the identified gaps in current retail experiences, where consumers face impersonal, static interactions and lack tailored recommendations. The solution will focus on offering a dynamic and immersive shopping experience that can guide consumers through their decision-making process, improve purchase intentions, and enhance overall satisfaction.

4.4 Design and Development

The design and development of RetailARVA were guided by the need to create an intelligent, adaptive AR-based virtual assistant that could enhance customer engagement in physical retail environments. The system is designed to integrate Augmented Reality (AR) and LLM capabilities to provide personalized product recommendations, real-time assistance, and an immersive shopping experience. The design process focused on user-centered design principles to ensure that the system would be intuitive, engaging, and effective for a diverse range of retail customers.

4.4.1 Application Architecture

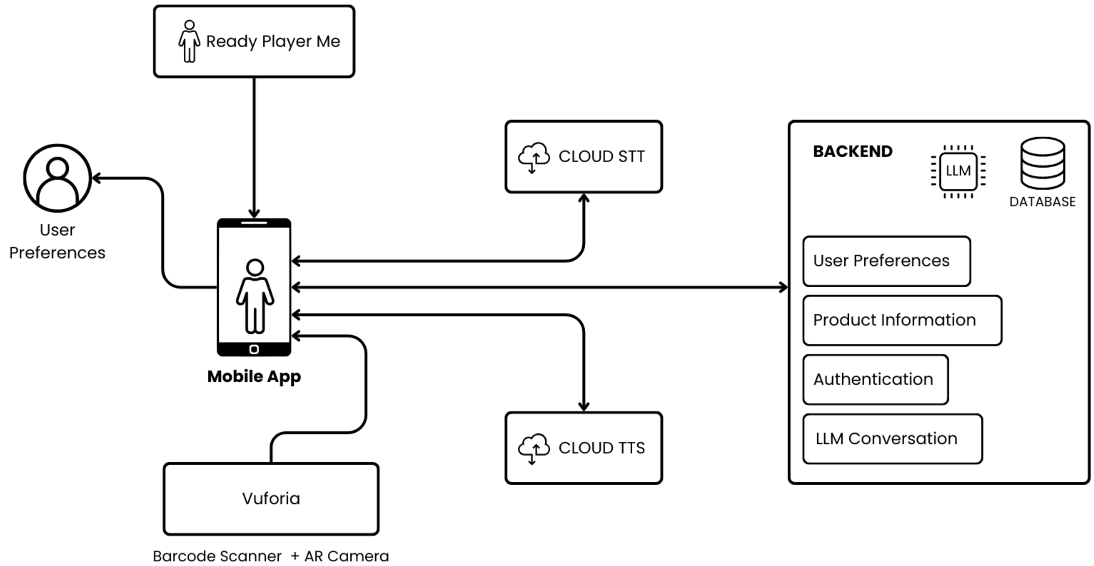


Figure 4.1: Application Architecture

4.4.2 Application Interaction Flow

According to the set objectives, the developed RetailARVA prototype aimed to deliver a personalized, conversational retail experience using a combination of various technologies. Initially, the user set their preferences by updating the user profile data, allowing the virtual assistant to provide a personalized shopping experience. The user can then scan a product using the in-app barcode scanner, and it instantly displays the relevant product information. Once the profile data are set, the user can load the virtual assistant by tapping the ground plane marker and begin the conversation through voice input. The recorded voice input is transcribed into plain text and sent to the model to generate a personalized and contextual response. This response is then converted back into speech and spoken aloud by the virtual assistant within the AR environment, making the whole shopping experience engaging and interactive for the user.

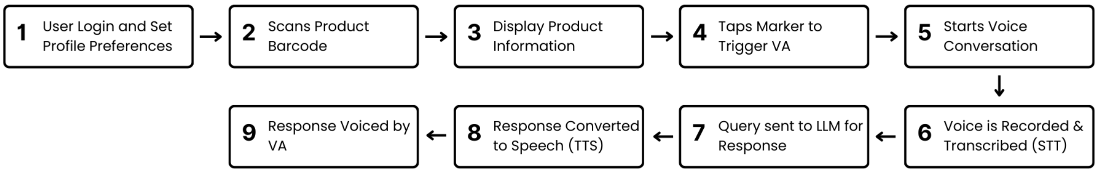


Figure 4.2: Application Interaction Flow

4.4.3 Functional Overview

The application was developed using Unity and multiple third-party services to facilitate an immersive and interactive AR experience that allows real time interaction with the virtual assistant. The core objective of the application was to enable a seamless voice based interaction with the virtual assistant during in-store shopping, enhancing the overall shopping experience.

- **AR Camera Integration** - Facilitates the AR environment and scans the surrounding environment for horizontal planes to anchor the virtual assistant. The user can tap on the marker to trigger and spawn the avatar on the detected plane.

- **3D Avatar Placement and Interaction** - A life-sized humanoid avatar is instantiated on the detected AR surface. The avatar is designed to face the user's camera to mimic eye contact and as the user moves, the avatar follows to sustain a natural sense of presence and connection. Additionally to add realism, avatar equipped with idle, talking and walking animations.

- **Text and Voice Conversions** - Users interact with the assistant using voice and the device's microphone is used to capture audio input. The app listens to and transcribes the user's spoken input for interaction and converts the assistant's text responses into natural voice output enabling a smooth engaging interaction.

- **Barcode Scanning and Product Information** - A separate barcode scanner feature was integrated to identify products and display the relevant product information instantly. The identified product is referenced during the conversation with virtual assistant, enabling more context-aware and personalized interaction.

- **User Preferences** - User-specific information such as age, skin type, allergies, and other relevant skincare details are stored at the beginning by the user which enables the delivery of a more personalized experience tailored to each user's unique needs and preferences.

4.4.4 LLM Pipeline Development

In the RetailARVA system, Large Language Models (LLMs) are integral to powering the AI assistant, enabling it to process and respond to various user queries related to skincare products. The core functionality of the assistant is structured around a series of specialized LLM pipelines that handle different types of inquiries. The system is designed to efficiently address three primary categories of user interactions.

1. Product Information: Queries regarding specific details about skincare products.
2. Suitability Checks: Inquiries related to how a product aligns with a user's skin type, concerns, allergies, and preferences.
3. Product Recommendations: Requests for suggestions of similar products based on a user's preferences or the product currently being considered.

The LLM pipeline in RetailARVA is organized into four key components, each responsible for different tasks within the system.

Query Classification

This component categorizes the user's query into one of the three primary types—product information, suitability check, or product recommendation. Classification is achieved through a few-shot learning approach, where the LLM is provided with a set of example queries for each category to guide its decision-making process.

Product Inquiry Pipeline

This pipeline retrieves detailed information about the queried product, including key attributes such as ingredients, usage instructions, and benefits.

Suitability Check Pipeline

This pipeline evaluates how well a product matches a user's specific needs, such as addressing skin concerns, allergies, or other personal preferences.

Product Recommendation Pipeline

Based on the user's query or preferences, this pipeline suggests alternative or similar products that may meet the user's requirements.

Additionally, a Retrieval-Augmented Generation (RAG) pipeline acts as a fallback mechanism to handle general inquiries that do not fit into the predefined categories.

4.5 Demonstration

For the demonstration of RetailARVA, a controlled lab setup was used, where participants interacted with a simulated product display consisting of five visually similar skincare products using the RetailARVA system. This setup aimed to replicate a real-world retail environment and test the functionality of the AI assistant. The demonstration involved participants completing three specific tasks designed to evaluate the system's ability to handle user queries and provide personalized recommendations effectively.

4.6 Evaluation

The evaluations were carried out in a controlled lab environment, using a within-subject experiment, where participants were tasked with completing three specific tasks using either the RetailARVA system or the control setup without the RetailARVA system. These tasks simulated real-world retail interactions, allowing us to measure the system's performance in a controlled setting. The main focus of the evaluation was to assess how well the system supported user tasks in terms of efficiency and user experience. More details are outlined in Chapter 6: Results and Evaluation.

4.7 Communication

The insights and feedback collected from users during the study are elaborated on in Chapter 7: Discussion.

Chapter 5

Implementation

5.1 Data Generation

Dataset Structure and Types of Data

Asynthetic dataset on the basis of real product data was generated to simulate a realistic conversation between users and the RetailARVA assistant.

Product Information

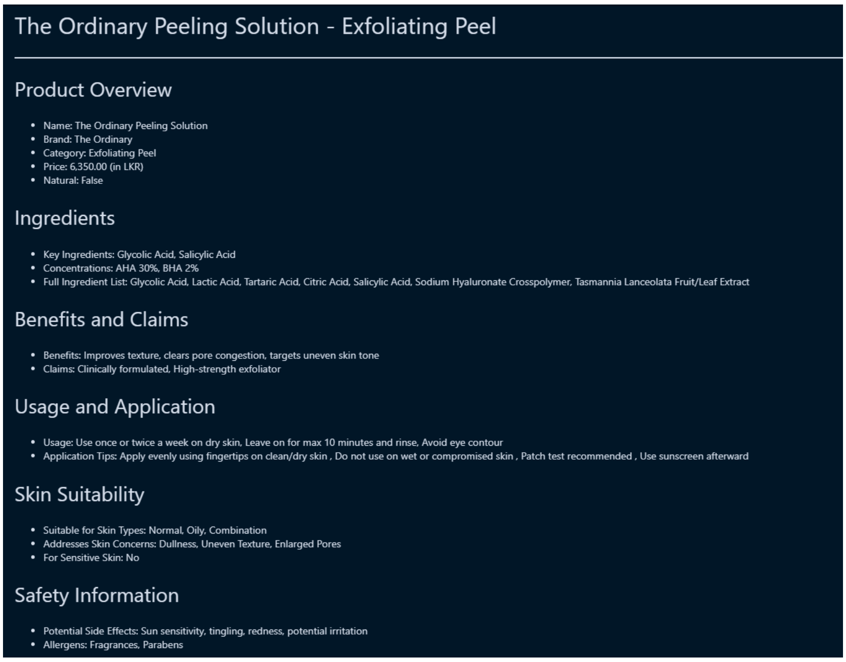
The design of this dataset draws from a prior study [27] focused on consumer behavior and skincare preferences, which informed the selection of key product attributes to capture. This evidence-based approach ensures that the database covers essential factors influencing skincare choices, from product efficacy to suitability for different skin types. Key attributes in this database include can be founded listed in Appendix F.

The product profile for embedding was structured to provide detailed, categorized information for each skincare product, allowing for accurate and effective recommendations based on user input.The profile was then converted into a collection of documents, with each document representing a single product profile. Each document included all relevant fields for the RAG model to draw from.

5.2 Data Preparation

5.2.1 Embedding creation

Under the data preparation, the initially generated synthetic dataset was sent through a LLM to generate comprehensive descriptions for each skincare product. LLMs tend to perform better for structured inputs (json, markdown) [28] [29]. Therefore, in the process of description generation, first each product data in the dataset was formatted according to a specific template in markdown structure by dividing the product information into specific sections within the template (ex-: Ingredients, safety information, reviews, and ratings)



The Ordinary Peeling Solution - Exfoliating Peel

Product Overview

- Name: The Ordinary Peeling Solution
- Brand: The Ordinary
- Category: Exfoliating Peel
- Price: 6,350.00 (in LKR)
- Natural: False

Ingredients

- Key Ingredients: Glycolic Acid, Salicylic Acid
- Concentrations: AHA 30%, BHA 2%
- Full Ingredient List: Glycolic Acid, Lactic Acid, Tartaric Acid, Citric Acid, Salicylic Acid, Sodium Hyaluronate Crosspolymer, Tasmannia Lanceolata Fruit/Leaf Extract

Benefits and Claims

- Benefits: Improves texture, clears pore congestion, targets uneven skin tone
- Claims: Clinically formulated, High-strength exfoliator

Usage and Application

- Usage: Use once or twice a week on dry skin, Leave on for max 10 minutes and rinse, Avoid eye contour
- Application Tips: Apply evenly using fingertips on clean/dry skin , Do not use on wet or compromised skin , Patch test recommended , Use sunscreen afterward

Skin Suitability

- Suitable for Skin Types: Normal, Oily, Combination
- Addresses Skin Concerns: Dullness, Uneven Texture, Enlarged Pores
- For Sensitive Skin: No

Safety Information

- Potential Side Effects: Sun sensitivity, tingling, redness, potential irritation
- Allergens: Fragrances, Parabens

Figure 5.1: Formatted markdown template for a product information

As the next step, larger LLM of the same model family (Llama 3.3 70b) were used to generate a comprehensive description of each product that captures the important details of that product based on the previously structured markdown formatted information. Then these generated descriptions were used to create vector embeddings for the product which will be used during the retrieval of similar products. Embedding comprehensive textual descriptions, which include additional context or structure, seems likely to perform better in similarity-based retrieval tasks compared to directly embedding structured formats like markdown, JSON, or HTML. This is because enriched text can capture more semantic details, improving how well the system finds similar content [30].

Product Copywriter Prompt

```
You are an expert product copywriter.  Based on the given product
profile, write a professional and engaging product description
that covers all the essential details, ingredients, benefits,
usage instructions, and any special considerations.  Create
a summary description of the skincare product based on the
given structured product details.  Use only then given product
information and phrase them as use see fit.  Don't remove any
vital information from the product profile.  Be accurate and
through.
Product Profile:  product_profile
Product Description:
```

5.3 LLM Pipeline Development

Functionalities of the AI Assistant are powered by a series of LLM pipelines that work together to process the user's query and determine what type of response it should provide based on the context of the given user query.

5.3.1 Architecture

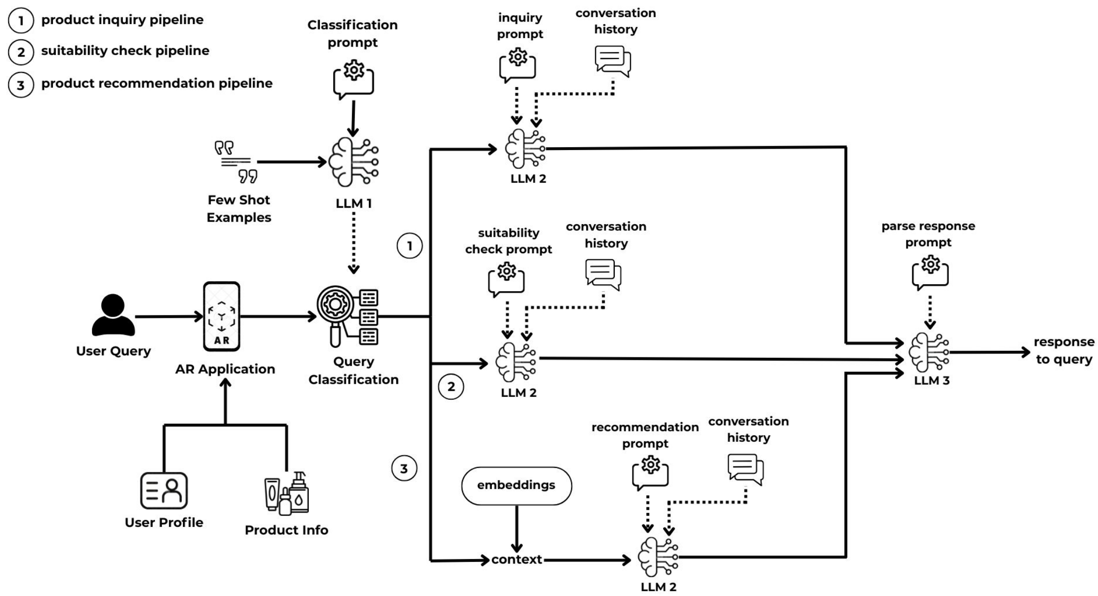


Figure 5.2: LLM Pipeline Architecture

Architecture of the LLM pipeline contains mainly 4 components.

1. Query Classification
2. Production Inquiry Pipeline/Chain
3. Product Suitability Checking Pipeline/Chain
4. Response Formatter for TTS

5.3.2 Functional Overview

Initial LLM pipeline only contained a single pipeline with standard Retrieval Augmented Generation (RAG) to respond to user query along with an embedding retriever component to use retrieved product information based on cosine similarity as additional context. During the research, we were able to identify common user queries regarding skincare products, which can be mainly divided into three categories. These three aspects will also be reflected upon in the test carried out during the user evaluation.

1. Customer make inquiries about the product on display

2. Customers have questions about how that product aligned with their skin concerns, allergies and personal preferences etc.

3. Customers need recommendations of similar products that align with the product on display or their preferences (ex-: skin concerns, price preferences etc.)

Using a single LLM pipeline will introduce additional noise into some of these identified tasks, resulting in poor, inaccurate responses from the LLM. Since the first two tasks do not require retrieving context from embeddings, using a common single pipeline for all the tasks will introduce additional overhead from embedding retrieval even its not mandatory to respond to some of the potential user queries and result in a larger number of input tokens being sent to the LLM, trying to adapt the prompt to fit all scenarios. This will result in delays for the overall request-response cycles of the user-assistant conversation. Therefore, we adopted a methodology that includes three different LLM inference pipelines that cater to the above-mentioned scenarios and the initial LLM pipeline that will act as a fallback measure.

Query Classification

The Query Classification component classifies the user query into one of three main labels below that align with the above mentioned three task types.

1. product_info

2. suitability_check

3. recommendation

For the classification, a prompt is sent to a LLM explaining the classification task and the classification labels it is supposed to output. Although a LLM instance that is fine-tuned with a classification dataset that contains possible user queries and their relevant classification label would be most ideal for this classification task, we decided to settle for In-Context Learning capabilities of the LLMs for the classification with few-shot examples. Several researchers have experimented with few-shot example classification and were able to achieve satisfactory results in classification [31] [32]. For most standard classification tasks, 2-5 examples [33] are the most commonly recommended range. Therefore we used 6 classification examples (2 examples per category). Each example includes the human message with the query and AI message with the classification label.

Query	Classification Label
What are the ingredients in the Nivea moisturizer?	product_info
Is the Nivea Extra White Body Serum suitable for my sensitive skin?	suitability_check
What are some alternatives to the Acme Cleanser for oily skin?	recommendation
Does CeraVe brand have any products for acne-prone skin?	recommendation
How do I apply the Golden Touch Pigmentation cream?	product_info
Is this cream better than Olifair Skin Lightening Day for dry skin?	suitability_check

Table 5.1: Few-shot examples for the classification task.

The classification labels were crafted to align with the three main tasks conducted during the qualitative user evaluation.

Evaluation Task	Relevant Classification Labels
Find more details about a product in display	product_info & suitability_check
Find a product of the same category based on skin concern	recommendation & suitability_check
Compare two products and make a decision on which is better	suitability_check, recommendation & product_info
Does CeraVe brand have any products for acne-prone skin?	recommendation
How do I apply the Golden Touch Pigmentation cream?	product_info
Is this cream better than Olifair Skin Lightening Day for dry skin?	suitability_check

Table 5.2: Alignment of classification labels with tasks in user evaluation.

If the user query cannot be classified to one of the above mentioned classification labels, it is labeled as 'general'.

LLM Pipeline Routing

Based on the classification label output from the LLM for the classification task, the relevant LLM pipeline is being invoked.

Classification Label	Invoked Pipeline
product_info	Product Inquiry
suitability_check	Suitability Check
recommendation	Product Recommendation
general	default RAG

Table 5.3: LLM pipelines invoked with classification labels

```
if classification_label == "product_info":
    Execute Product Inquiry Pipeline
else if classification_label == "suitability_check":
    Execute Suitability Check Pipeline
else if classification_label == "recommendation":
    Execute Product Recommendation Pipeline
else:
    Execute Default RAG Pipeline
```

The pipelines are provided with conversation history that contain user's and AI's previous interactions to understand the context of the conversation with the AI assistant, when its required.

Parsing LLM Responses

Sometimes the LLM response can contains special characters, bad Punctuation, unnatural phrasings or might lack of disfluencies. When this kind of LLM response is provided to the Text-to-Speech (TTS) directly, that will decrease the human-likeness of the voice output of AI assistant. To prevent this another LLM will serve as a formatter/parser to convert the previous LLM response to more human-like format.

Example 1: Query to TTS Formatting

User Query:

”What does the RevitaGlow Hydration Serum do?”

LLM Response:

”Pursuant to your inquiry, the RevitaGlow Hydration Serum facilitates optimal dermal moisturization via a proprietary blend of hyaluronic acid polymers and botanical extracts.”

Formatted TTS Input:

”Sure! RevitaGlow Hydration Serum helps keep your skin super moisturized and feeling fresh. It’s packed with hyaluronic acid and a bunch of plant extracts that your skin will love!”

Example 2: Query to TTS Formatting

User Query:

”What’s inside the Hydrafresh moisturizer?”

LLM Response:

”The Hydrafresh moisturizer, which contains glycerin, vitamin E, aloe vera extract, and several essential oils designed to support skin hydration, elasticity, and barrier function, is recommended for dry to very dry skin types.”

Formatted TTS Input:

”It’s got glycerin, vitamin E, and aloe vera — all amazing for keeping your skin hydrated and soft. Plus, there are a few essential oils to help strengthen your skin barrier!”

5.4 LLM Pipelines Implementation

The LLM pipelines were implemented mainly using the Python programming language and LangChain LLM framework as a RESTful API. All the locally running Language Models, Embedding Models were accessed through Ollama for the inference. To implement the RAG models in the LLM pipelines, Qdrant is used as the vector database to store the product embeddings. Langsmith was used for the purpose of monitoring LLM pipeline flow and tracing. Apart from that Docker was used for the easier deployment to server environments.

For the LLMs used in implementation and evaluation, we tried to make use of available open-source large language models (ex-: Llama 3.1:8b, gemma 3:4b) and commercial LLM providers like Gemini were only used when we were unable to achieve the expected level of accuracy within the expected minimum inference time.

5.4.1 Development Environment

The LLM framework was developed using Python 3.11.9 and LangChain 0.2.15. Even though, at current state there are many production ready LLM/Gen AI frameworks available, Langchain was one of the foremost Gen AI frameworks that are in use and used by researchers and R&D personnel in the Gen AI landscape to develop POC's quickly. Moreover Langchain provides abstraction over many LLM related concepts and functionalities, making it easy to use even for people lacking strong development skills and knowledge. Langchain comes with its own tracing platform (Langsmith) to monitor the LLM workflows and pipelines to get in depth insights on how the implementation performs in deployed environments. Development was carried out in Windows environment and WSL was used to run the docker containers locally. PyCharm was primarily used as the IDE for the development, but Google Colab and locally running Jupyter Notebooks were also used for script development and testing. For the specific python libraries used in the development refer to the project dependencies in the GitHub repository.

To run the LLMs, local environment and the cloud hosted VM instances acquired through the Massed Compute were used. Below are the specifications of those enviornments.

CPU	Intel Core-i5 10330H
Memory	20 GB
Storage	256 GB NVMe SSD
GPU	Nvidia GTX 1650
vRAM	4 GB
OS	Windows 11

Table 5.4: LLM inference system specifications - Local

vCPU	12 Cores
Memory	80 GB
Storage	350 GB NVMe SSD
GPU	RTX 6000 ADA
vRAM	48 GB
OS	Ubuntu

Table 5.5: LLM inference system specifications - Hosted VM

Below listed models were deployed in the above mentioned hardware.

Model	**Parameters**	**Qunatization**	**Proivder**
Llama 3.1:8b	8.03 billion	Q4_K_M	Ollama
Llama 3.3:70b	70.6 billion	Q4_K_M	Ollama
Gemma 3:4b-it-qat	4.3 billion	Q4_0	Ollama
Gemini-2.5-flash	N/A	N/A	Google AI Studio
Nomic-embed-text	137 million	F16	Ollama

Table 5.6: Models used in implementation

5.4.2 Query Classification

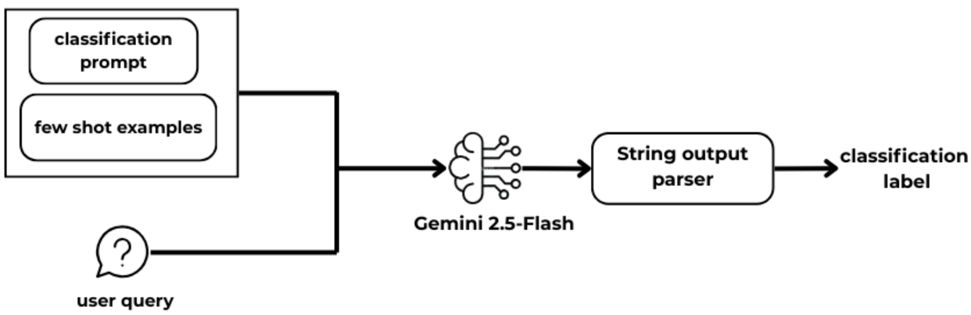


Figure 5.3: Query Classification Pipeline

As mentioned in the query classification section 5.3.2, a prompt containing Instructions were provided to the LLM along with category labels and description for each category to perform the classification and step by step set of guidelines were provided to LLM to follow to perform more accurately in the classification. If the LLM is unable to perform the classification as expected it was instructed to classify the query as "general". Initially a small parameter locally run model was tested as the classification model, but later adopted gemini-2.5-flash due to the classification results from the local model was not accurate for the tested user queries. This prompts for the need for fine tuning an open-source LLM for the classification task.

5.4.3 Product Inquiry Pipeline

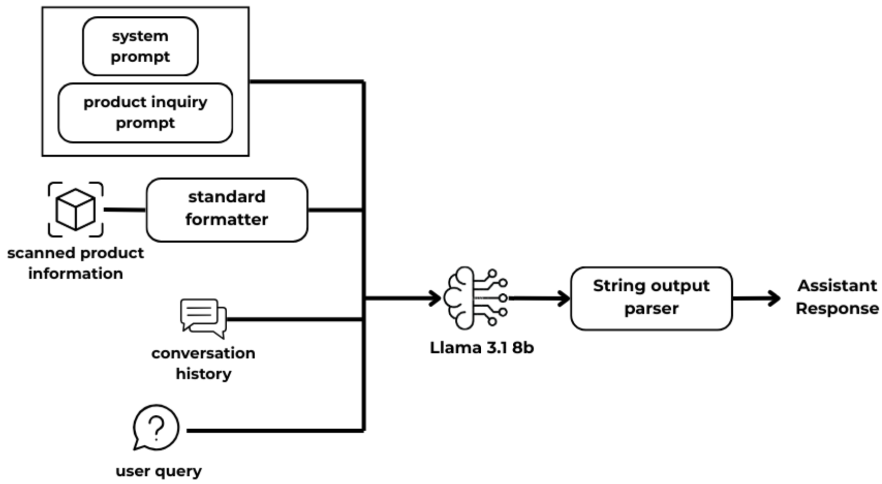


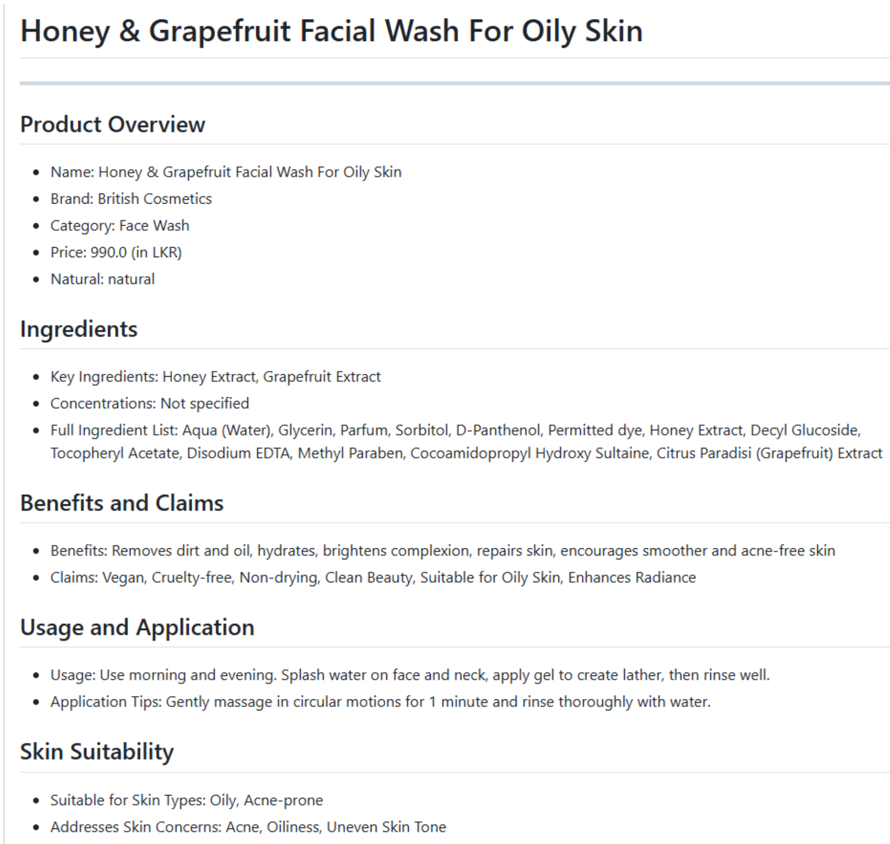
Figure 5.4: Product Inquiry Pipeline

Product Inquiry pipeline provides answers to the user queries that are related to the inquiries users have about the particular scanned product. This pipeline contains two prompts

1. System prompt
2. Product Inquiry prompt

System prompt defines the persona of our female toned virtual AI assistant named "Luna" who acts as a sales representative specializing in skin care products. It provides a clear role, objectives, and behavioral guidelines for how Luna should interact with customers.

The Product Inquiry prompt takes the formatted information of the scanned product in markdown format and uses that as the context to answer the user's query. Below is an example on how the formatted product information context is provided to the LLM during the time of inference looks.



Honey & Grapefruit Facial Wash For Oily Skin

Product Overview

- Name: Honey & Grapefruit Facial Wash For Oily Skin
- Brand: British Cosmetics
- Category: Face Wash
- Price: 990.0 (in LKR)
- Natural: natural

Ingredients

- Key Ingredients: Honey Extract, Grapefruit Extract
- Concentrations: Not specified
- Full Ingredient List: Aqua (Water), Glycerin, Parfum, Sorbitol, D-Panthenol, Permitted dye, Honey Extract, Decyl Glucoside, Tocopheryl Acetate, Disodium EDTA, Methyl Paraben, Cocoamidopropyl Hydroxy Sultaine, Citrus Paradisi (Grapefruit) Extract

Benefits and Claims

- Benefits: Removes dirt and oil, hydrates, brightens complexion, repairs skin, encourages smoother and acne-free skin
- Claims: Vegan, Cruelty-free, Non-drying, Clean Beauty, Suitable for Oily Skin, Enhances Radiance

Usage and Application

- Usage: Use morning and evening. Splash water on face and neck, apply gel to create lather, then rinse well.
- Application Tips: Gently massage in circular motions for 1 minute and rinse thoroughly with water.

Skin Suitability

- Suitable for Skin Types: Oily, Acne-prone
- Addresses Skin Concerns: Acne, Oiliness, Uneven Skin Tone

Figure 5.5: Markdown formatted product profile

5.4.4 Suitability check Pipeline

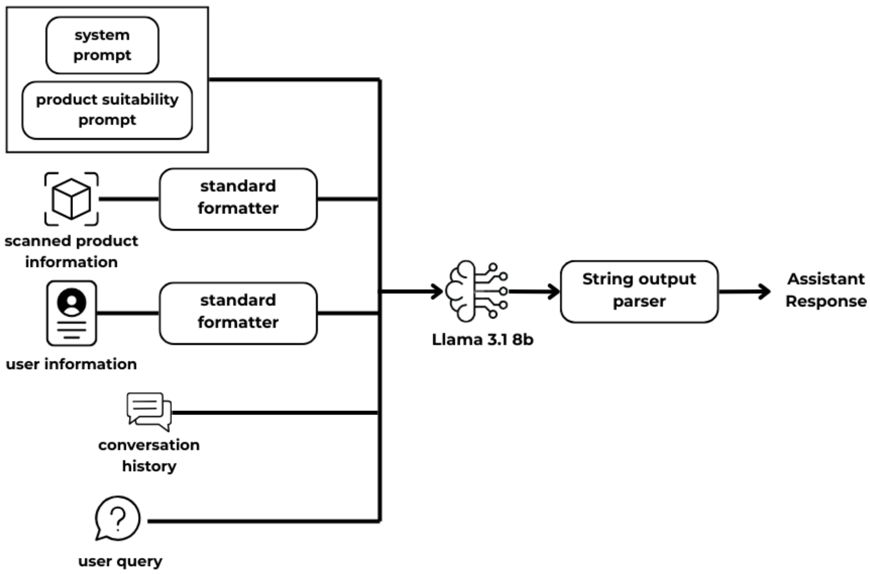


Figure 5.6: Suitability Check Pipeline

Main purpose of this LLM pipeline is to compare information of the scanned product with the user's profile which include their product preferences with preferred price range, skin concerns, allergies and ingredients to avoid. By using this approach, LLM is able to suggest whether the scanned product aligns with the user's skin profile/preferences or not. Also this will be a key driver for the user to look for other skincare products as suitable recommendations. This pipeline will also contain two main prompts like the previous pipeline, but replace product inquiry prompt with product suitability prompt.

Apart from the formatted product information this prompt leverages the user preferences/concerns gathered from the mobile application to render a formatted user information section in markdown format. Below is such an example user profile curated based on collected information.

User Profile

Personal Information

- Age: 25
- Gender: Female

Skin Profile

- Skin Type: Combination
- Sensitive Skin: Yes
- Skin Concerns: acne, darkspots, wrinkles, dryness, oiliness, sensitivity, pores, sun

Product Preferences

- Preferred Price Range: LKR 1000.0 - 25000.0
- Preferences: natural, organic, vegan

Safety Information

- Ingredients to Avoid: natural, fragrances, preservatives, parabens, dyes, metals
- Known Allergies: coconut_oil, aloevera, sheabutter, chamomile, witch_hazel, sesame_oil, soy_derivatives, nut_oils

Figure 5.7: Markdown formatted user profile

5.4.5 Product Recommendation Pipeline

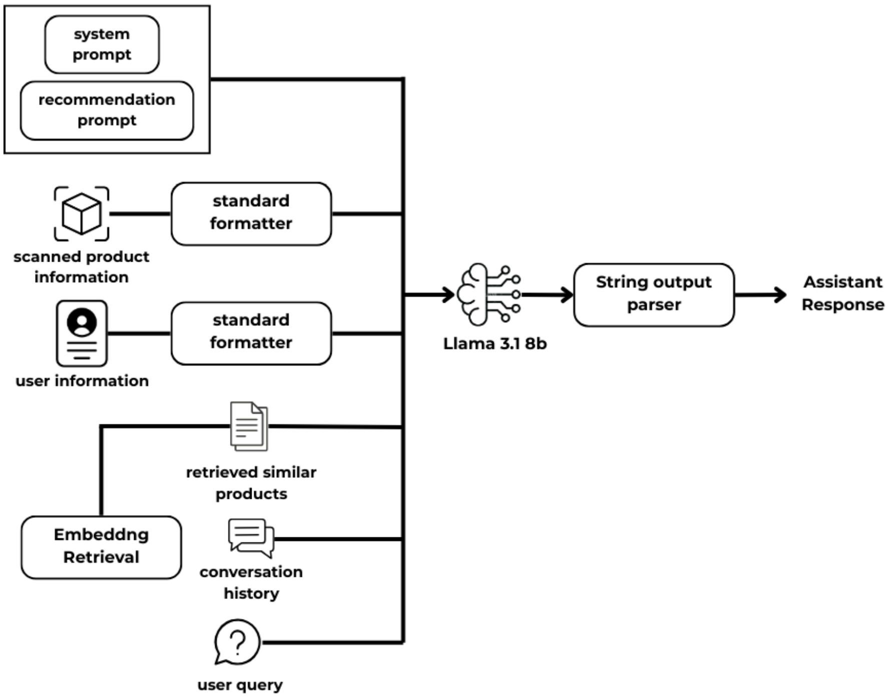


Figure 5.8: Product Recommendation Pipeline

Recommendation pipeline is used to provide recommended skin care products considering the aspects of the scanned product and user profile. The pipeline is built on top of the concept of RAG (Retrieval Augmented Generation) and it utilizes three primary context sources for the LLM: markdown-formatted product information, markdown-formatted user profile (similar to the suitability check pipeline), and a set of products identified via vector similarity. Recommendation prompt is the largest prompt among all the prompts used due to extensive guidelines that were provided to LLM to make an accurate recommendation. This prompt guides the recommendation of skincare products tailored to a user's unique skin profile and preferences. It ensures safe, suitable options by excluding harmful ingredients and prioritizing compatibility with skin type and concerns. The process balances hard filters like allergies and soft preferences such as budget or brand, offering concise, personalized suggestions while avoiding unsuitable or irritating products.

Embedding Model Selection

For any RAG based Generative AI solution, the quality of the embeddings stored and generated during the retrieval plays a crucial part in how accurate the retrieval is. The 100 product descriptions generated from the product dataset under the section 5.2, should be converted into accurate vector embeddings that capture their semantics as accurate as possible to retrieve only the most suitable alternatives/recommendations.

For generating these embeddings, initially we had to compare and contrast some of the available open-source embedding models and choose the embedding model with best accuracy-speed tradeoff. Embedding retrieval speed was a crucial factor here because it will contribute to the latency of the AI assistant for a better user experience in the user evaluation phase. Since we were already using Ollama for running LLMs locally, we had to determine the best embedding model for the RAG pipeline among those.

Embedding Model	**Parameters**	**Size**	**Dimensions**
nomic-embed-text	137M	274 MB	768
mxbai-embed-large	334M	670 MB	1024
bge-m3	567M	1.2 GB	1024
bge-large	334M	671 MB	1024
snowflake-arctic-embed	335M	669 MB	1024

Table 5.7: Embedding Models selected from Ollama

Some of the available sources provided benchmark information [34] [35] related to some of these models. According to those sources, nomic-embed-text, mxbai-embed-large and bge-m3 models bge-m3 demonstrates best overall performance but nomic-embed-text offers the fastest response time. Since both of these benchmark sources didn't contain bge-large and snowflake-arctic-embed, so we needed more comprehensive benchmarks.

Therefore we used STS (Semantic Textual Similarity) dataset to benchmark the above mentioned 5 embedding models. STS dataset is commonly used for evaluating how well embeddings capture semantic similarity. The STS dataset has pairs of sentences with similarity scores, so the idea would be to compute the cosine similarity of the embeddings and compare it to the ground truth scores. The higher the correlation between the predicted and actual scores, the better the model's accuracy. For speed, we measure the time it takes to generate embeddings for each sentence, then the average (average time per sentence).

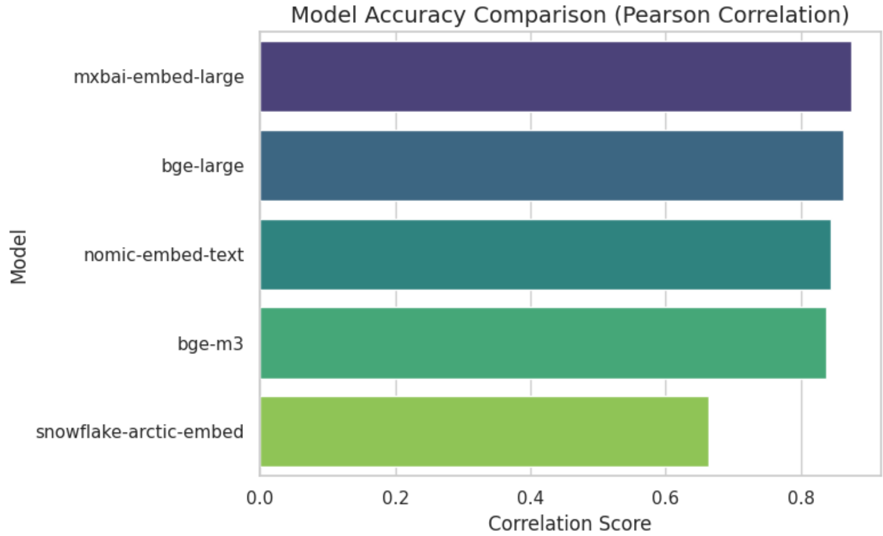


Figure 5.9: Embedding Models Accuracy Comparison

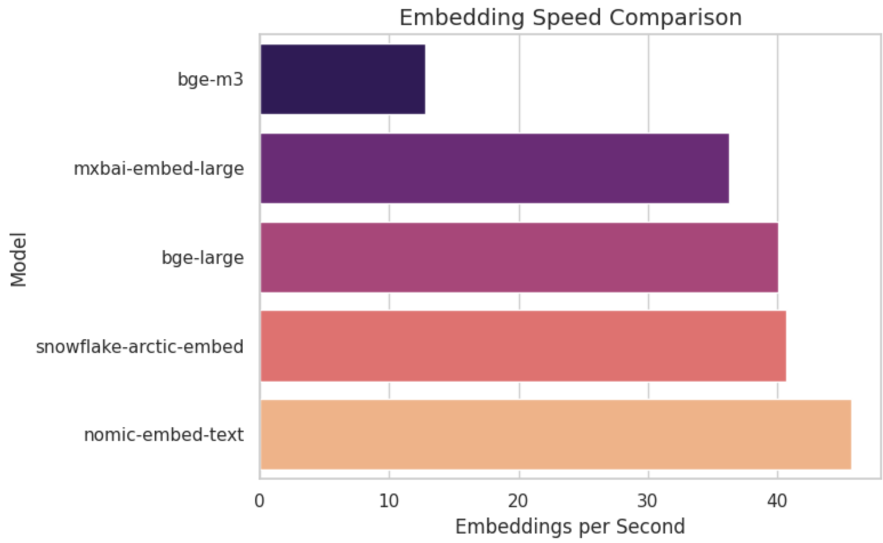


Figure 5.10: Embedding Models Speed Comparison

According to the above results, nomic-embed-text generates embeddings at the highest rate, making it the fastest model. But according to Pearson correlation it takes third place in terms of accuracy.

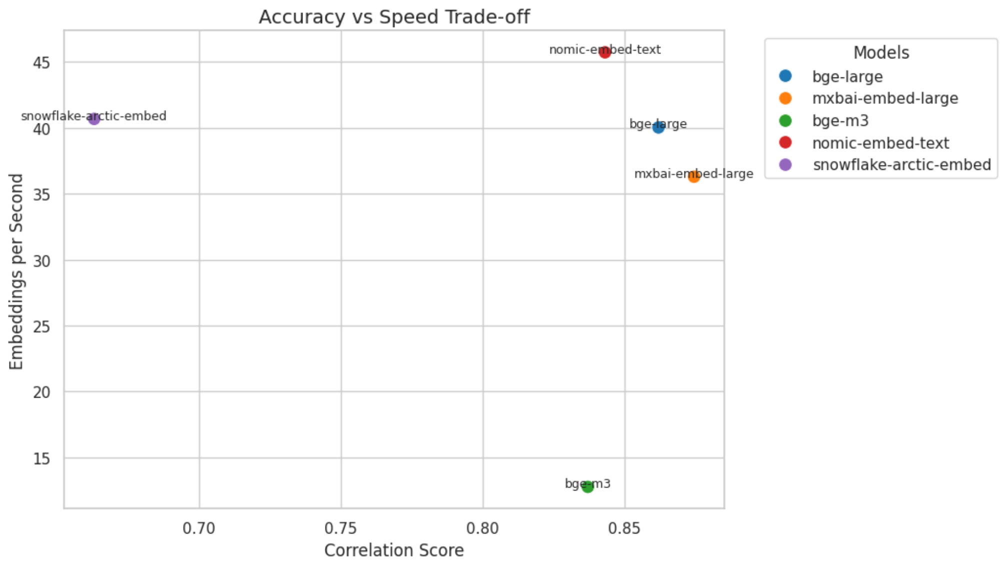


Figure 5.11: Embedding Models Speed Comparison

The scatter plot highlights the trade-off between accuracy and speed. Nomic-embed-text, bge-large and mxbai-embed-large models are positioned closer to the top-right corner, indicating a strong balance between high accuracy and reasonable speed. Given that speed is a critical factor , the best embedding model for the retrieval is the nomic-embed-text . It achieves the highest embeddings per second (45 embeddings/second) while maintaining a high Pearson correlation score (0.85). Therefore, nomic-embed-text strikes an excellent balance between speed and accuracy, making it an ideal selection.

Recommendation Retrieval

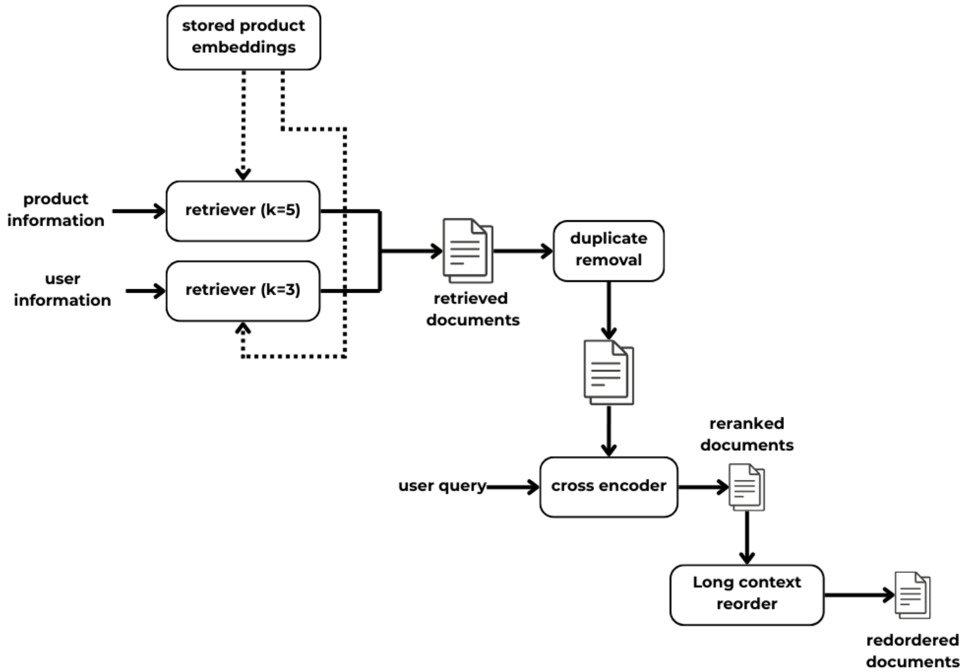


Figure 5.12: Recommendations Retrieval Process

Product recommendations for the user's query are retrieved using both the scanned product information and user profile information. First the 5 products were retrieved performing cosine similarity with the scanned product information and then 3 products were retrieved performing similarity with user information. The scanned product represents the user's current focus—something they are actively considering or exploring. For example, if a user scans a moisturizer, their immediate intent might be to find similar products (e.g., other moisturizers) or complementary items (e.g., serums). By retrieving 5 products based on the scanned product information, the system emphasizes alternatives and complements that align closely with this intent. Giving more weight to the scanned product ensures that the recommendations are anchored to what the user is currently interested in, which is likely their primary motivation. In contrast, the user's profile—while critical for personalization—reflects broader preferences, such as skin type, concerns, or past purchases, which may not be as immediately relevant to the scanned product. Retrieving 3 products based on this information provides personalized suggestions without overshadowing the scanned product's prominence. The choice of 5 and 3 strikes a balance between relevance to the current context and personalization to the user. The higher number of products retrieved from the scanned product information ensures a strong focus on items similar to what the user is looking at, which is likely to yield a denser set of relevant options. For instance, in a product catalog, there may be many moisturizers with similar ingredients or purposes, making it practical to retrieve more candidates using this source. On the other hand, the user's profile might pinpoint fewer highly relevant products due to the specificity of individual preferences (e.g., a rare skin condition might match only a handful of items). Retrieving 3 products here ensures personalization is included without overwhelming the recommendations with potentially less contextually relevant items.

Then these results were combined and filtered by removing duplicate retrievals and re-ranked using ms-marco-MiniLM-L6-v2 cross encoder to refine the relevance of the retrieved results. The initial retrieval of 8 products (before removing duplicates) provides a sufficiently large pool for this re-ranking step. Dense embedding retrieval can produce false positives—items with high similarity scores, meaning the retrieved results may include irrelevant items because the similarity score from embeddings doesn't always correlate perfectly with true relevance. But low true relevance—starting with more candidates increases the chances of including truly relevant products. The cross-encoder then refines this list by evaluating each product against the query (combining scanned product and user profile information), minimizing false positives and boosting precision[36]. Retrieving more from the scanned product than the profile ensures the pool is rich with contextually relevant options, which the re-ranking can further optimize. After that re-ranked retrieved results are passed through a Long Context Reorder for the reordering. According to a study[37] conducted, the best performance typically arises when crucial data is positioned at the start or conclusion of the input context. Reordering process will make sure the most relevant products for the recommendations are positioned at the beginning and the end of the input context.

5.4.6 Default RAG Pipeline

As previously mentioned, when the LLM implemented for the classification task is unable to label the query into one of the main pipelines, the user's query will be handled using the initial RAG pipeline we had. The retrieval of this RAG pipeline is almost similar to the the product recommendation but contain different and more generic prompts. This pipeline takes the same System prompt that is mentioned in the previous sections and use a additional prompt to provide necessary instructions for the task at hand.

5.4.7 Parse LLM output for Text-To-Speech

Figure 5.13: simple LLM output parsing

This simple pipeline takes the output from any of the above LLM pipelines and structure it into a more TTS friendly format. To minimize the overhead added to the speed of query-response flow, we adopted a relatively small (4 billion) LLM for this task. Also the specific version we used were a Quantization Aware Trained (QAT) model, which preserves similar quality as half precision models (BF16) while maintaining a lower memory footprint. The prompt provided for this task is tuned with several guidelines emphasizing on friendly and humane tone, clean and clear text, proper punctuation, correct capitalization, abbreviations and acronyms handling etc.

5.5 Application Implementation

The entire system follows a client-server architecture. The mobile application acts as the client, handling user interactions such as recording voice inputs, scanning barcodes, rendering the avatar in AR, and other features. The application uses Vuforia for barcode scanning and ARcamera implementation, Ready Player Me for the 3D avatar rendering, and Google Cloud APIs for speech recognition (STT) and text-to-speech synthesis (TTS).

5.5.1 Development Environment

To build the mobile application, Unity 2022.3 (LTS) was selected as the development engine, and multiple third-party services were integrated to power the AR virtual assistant. Unity is a popular real-time 3D engine known for its versatile tools and strong cross-platform capabilities. Unity was selected mainly considering its advanced support for AR features and its compatibility with a wide range of third-party SDKs. The entire development was carried out on a Windows environment, due to its compatibility with Unity and Android development tools [38].

Android Studio was used as the primary IDE for managing Android SDKs and writing scripts required for building the application. In Unity, a component-based architecture was followed, where custom scripts were attached to GameObjects to handle functionalities such as avatar instantiation, barcode scanning, speech processing, and backend API integration. The application uses REST to communicate with third-party services and the backend. A minimum API level of Android 11 (API level 30) and a target API level of Android 12 (API level 31) were selected in the build settings to guarantee maximum support for the ARCore features needed by the Vuforia Engine.

5.5.2 Avatar Design

The avatar was designed using the Ready Player Me online platform, which allows users to create realistic and personalized avatars that can be integrated into Unity projects using their official SDK. These avatars can be animated to perform various gestures, movements, and facial expressions, making them suitable for applications that require life-like interactions. The Ready Player Me SDK (7.4.0) was added to the project via the package manager in Unity. After the avatar model was designed, it was assigned a unique URL that allowed it to be dynamically loaded into the Unity project during runtime. The avatar can be spawned in the

AR environment in response to an event, such as a user tap.

Figure 5.14: Virtual Assistant Displaying Interactive Gestures

Choosing a Female Avatar

A female avatar was specifically selected as the virtual sales assistant based on many industry trends that were discovered through research. According to many studies, women in sales roles often excel at building strong client relationships, nurturing trust, and displaying empathy qualities that are highly valued in sales representative positions [39].

Avatar Animations

To enhance the realism and interactivity of the virtual assistant, the avatar was utilized with a set of animations that simulate natural human behavior. These animations include idle, walking, and talking motions, and maintaining eye contact with the user's camera, all of which contribute to a more immersive and natural user experience. These animations were managed using Unity's Animator Controller, which allowed smooth transitions between states based on user interaction and backend responses.

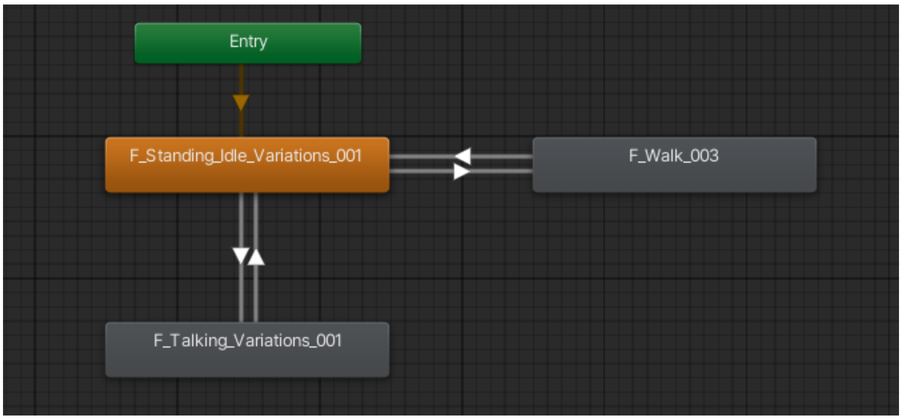


Figure 5.15: Unity Animator Controller

Avatar Solution Selection and Justification

The selection of Ready Player Me was based on its ability to offer high-quality, fully rigged 3D avatars that are easy to create and personalize without any coding. It also packs extensive customization options from facial features to clothing and accessories of the avatar. The SDK provided by Ready Player Me offers seamless integration with Unity, simplifying the development process and complexity [40].

5.5.3 AR Camera Integration

Vuforia Engine is a popular SDK for creating AR applications which enables detection and tracking of real-world images, objects, or environments. Vuforia Engine SDK (11.1.3) was added to the application using Unity's package manager. After initializing Vuforia with a developer license key, ARCamera GameObject was added to the main scene from the Vuforia Engine menu [41].

Ground Plane Implementation for Avatar Placement

The Vuforia's ground plane feature was integrated to facilitate the accurate positioning of the avatar relative to real-world surfaces. It was used to ensure that the avatar is placed realistically on detected horizontal surfaces, allowing the avatar to interact with the physical environment. To achieve this, the ground plane stage GameObject was added to the scene as the parent object for the avatar.

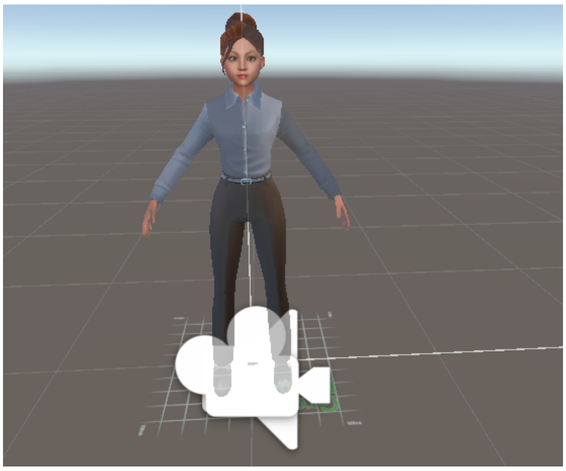


Figure 5.16: Avatar positioned on Ground Plane Stage

Next, Vuforia's plane finder feature was used to detect flat surfaces and enable users to interactively position the avatar on those surfaces, ensuring accurate and precise placement. The avatar is positioned on the detected surface based on the user's tap on the marker, providing an interactive way for avatar placement.

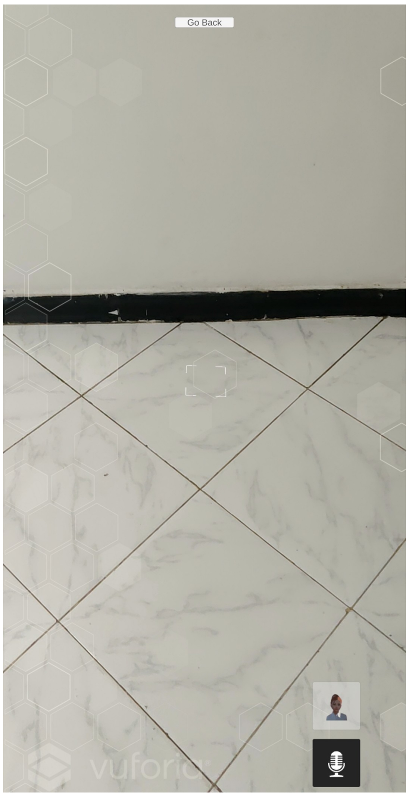


Figure 5.17: Plane Detection Marker

Figure 5.18: Avatar Positioned on the Marker

AR Camera Solution Selection and Justification

Vuforia was chosen for its powerful AR camera and ground plane detection capabilities, which are crucial in providing a comprehensive AR experience. Vuforia's ground plane detection feature enables virtual objects, like the avatar, to be placed on real-world surfaces such as the ground, enhancing the realism of interactions in the AR environment. In addition to its core AR camera and ground plane detection features, Vuforia also offers a barcode scanning feature which is a key requirement for this application. This combination of features and its high compatibility made the development process both efficient and effective [42].

Feature	**Vuforia**	**ARCore**
SDK	Native Unity SDK, also supports native development for other platforms	Native Unity SDK, Android SDK
Platform Support	Cross-platform (Android & iOS)	Android only
Device Compatibility	High (works on most Android & iOS devices)	Limited to Android devices
Cost	Free (with limitations on certain features)	Free
Barcode Scanning	Yes	No
Ground Plane Detection	Yes	Yes

Table 5.8: Vuforia vs ARCore Comparison

5.5.4 Barcode Scanning & Product Identification

The Vuforia engine provides an integrated barcode scanning feature facilitating quick and efficient product identification. The Barcode GameObject was added to the main Unity scene to start the barcode scanning feature. When a barcode is detected during runtime, the BarcodeBehaviour component is instantiated as a new instance, with a unique identifier corresponding to the detected barcode data.

Upon successful detection of the barcode, the application extracts the product identifier from the barcode. This product ID is then used to retrieve relevant product information, such as the product name, price, ingredients, benefits, expert reviews, and other associated details, from the backend API. The data received from the API is then utilized to display the product information within the AR scene.

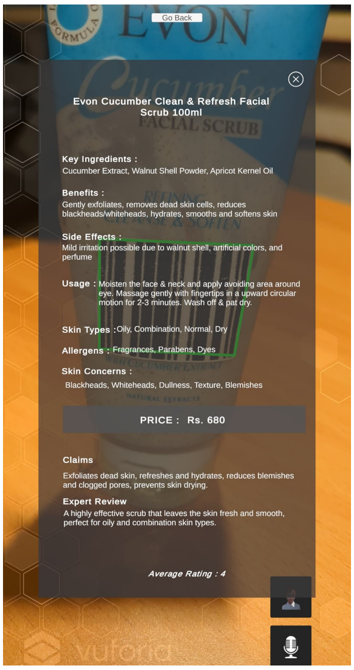


Figure 5.19: Displayed Product Information

5.5.5 Google Cloud STT and TTS

Google Cloud Speech-to-Text (STT)

Google Cloud's STT API is utilized to convert spoken language into written text in real time. The application uses the device's microphone to record voice input and then applies a custom silence-detection algorithm to trim unnecessary silence. After processing, the audio is sent to the API for transcription after being encoded in Base64 format [43].

Google Cloud Text-to-Speech (TTS)

Google Cloud's TTS API is utilized to transform text responses generated by the language model into speech. The API returns a Base64-encoded audio string, which is decoded and played through the avatar, enabling it to deliver spoken responses to the user. This real-time conversion makes the interaction more natural, improving the overall user experience [44].

Google cloud services were chosen for their high accuracy, fast response times, and robust cloud infrastructure, making them ideal for real-time speech processing. These services offer advanced features, including support for multiple languages and voices, high-quality speech synthesis, and the ability to handle noisy environments, which is crucial for an interactive retail shopping experience. Its extensive support made the integration efficient and faster.

5.5.6 User Preferences

In order to facilitate context-aware and tailored interactions throughout the virtual assistant conversation, a user preferences feature was integrated. At the start of the experience, the user sets their profile that contains important personal and skincare-related information including age, gender, skin type, sensitivity, common skin issues, known allergies, and preferences for ingredients. Additionally, the user specifies their preferred price range in the profile to receive recommendations that align with their budget.

The application references the stored user profile data to tailor responses and provide personalized product recommendations when the user interacts with the avatar. For instance, if the user has a known allergy to coconut oil, the system ensures that products containing such allergens are excluded from the recommendations.

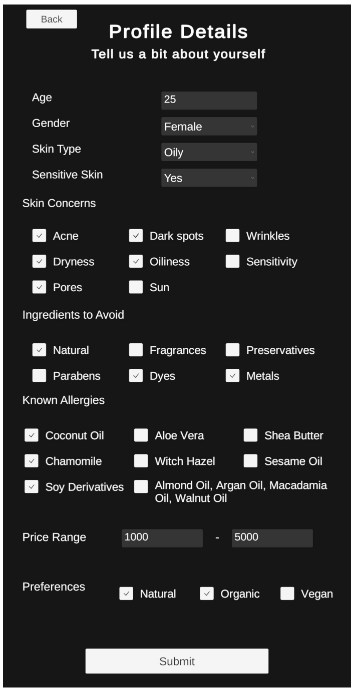


Figure 5.20: Setting User Preferences

Chapter 6

Results and Evaluation

6.1 User Study

To evaluate the design and impact of the proposed RetailARVA system, a user study was conducted within a controlled, simulated retail environment. This study aimed to address the three central research questions of this thesis. Through a within-subjects [45] experimental design comparing a simulated standard retail experience (control) with the RetailARVA experience (experimental), the study investigates the system's effect on user satisfaction, cognitive load and usability. The insights gained inform both the technical feasibility and user-centered value of deploying intelligent AR assistants in retail contexts.

6.1.1 Study Design

A within-subjects experimental design was adopted for this user study, wherein each participant interacted with both the controland experimental conditions. This design was selected to enable direct within-participant comparison, thereby controlling for inter-individual variability in user preferences, technological familiarity, and cognitive styles [45] . By ensuring that all participants experienced both conditions, the study maximizes internal validity and statistical power, allowing more precise identification of the effects attributable to the RetailARVA system.

Test Setup

To simulate a realistic shopping experience, the test environment featured a display of 5 skincare products within the same product category (moisturizers). All products were carefully selected to be visually similar in packaging and branding, reflecting the challenge consumers typically face in identifying suitable options based on nuanced ingredient and usage information. Participants were allowed to interact freely with the products, physically examine them, and make informed decisions based on the condition they were placed in—either traditional (control) or AR-enabled (experimental). Each participant was asked to complete three representative tasks mimicking common retail decision-making scenarios,

- **Product Examination Task** - Examine a product on display and determine whether it is safe and suitable for sensitive or acne-prone skin, including identifying any known side effects.
- **Comparison Task** - Compare two skincare products (two moisturizers) and decide which is more suitable for daily use by someone with dry, sensitive skin.
- **Recommendation Task** - Based on a specified skin concern (oily skin with breakouts), identify the most appropriate product from the 5 options displayed.

Control Condition

The control condition simulated a traditional in-store retail experience. Participants interacted with a simulated physical retail space, where 5 products were visibly displayed and accessible for manual inspection. They could pick up and examine the items, read printed labels, and optionally request assistance from a simulated in-store staff member, mirroring standard real-world shopping behavior. No digital augmentation, virtual assistant, or intelligent recommendation system was provided. This condition served as a baseline to evaluate user behavior, information-seeking strategies, and decision-making processes in the absence of technological enhancements.

Experimental Condition

In contrast, the experimental condition introduced participants to the final RetailARVA prototype, an interactive augmented reality-based virtual assistant integrated into the physical retail context via a mobile device. This system leveraged,

1. Barcode scanning to recognize and retrieve product-specific data in real-time
2. Voice-based interaction powered by a large language model (LLM), enabling natural language queries and clarifications about products.
3. AR overlays that projected contextual information (e.g., side effects, benefits, user reviews)
4. Personalized recommendations based on user-provided preferences or skin concerns

Tasks in this condition mirrored those in the control group but were mediated by the assistant's capabilities. This allowed for a comparative analysis of task performance, cognitive load, user satisfaction, and perceived usability between traditional browsing and AI-assisted interaction.

Counterbalancing and Randomization

To reduce order effects such as learning, fatigue, or bias introduced by condition sequence, the order in which participants experienced the control and experimental conditions was counterbalanced. Participants were randomly assigned to one of two groups,

1. Group A: Did the tasks under Control condition and then under the Experimental condition
2. Group B: Did the tasks under Experimental condition and then under the Control condition

This counterbalancing ensured that any potential effects due to task repetition or increased familiarity with the products were evenly distributed across both conditions.

6.1.2 User Study Procedure

The complete procedure for each participant was as follows,

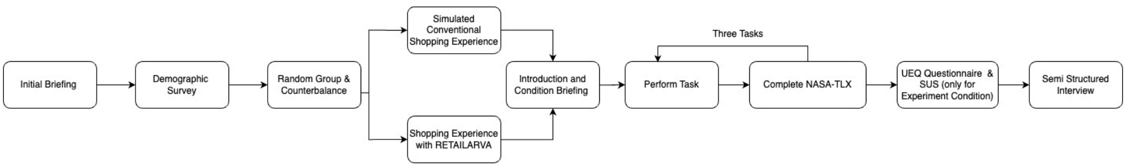


Figure 6.1: User Study Flow

1. Initial Briefing and Consent

Participants were welcomed and provided with an overview of the study, including its purpose, tasks, and duration. They were informed of their rights, including the voluntary nature of participation and the confidentiality of their responses. Informed consent was then obtained in accordance with institutional ethical guidelines.

2. Completion of a Pre-Questionnaire

A pre-questionnaire, Appendix A , was administered to collect demographic information and assess participants' baseline characteristics, including age, technological proficiency, and skincare purchasing behavior.

3. Task Completion in the Control Condition

Participants completed a set of simulated skincare shopping tasks under the control condition, which mimicked a traditional retail environment. In this condition, 5 products were displayed with a simulated store assistant, and participants could pick up and examine the products. A assistant was available for optional help, but there was no AR overlay or personalized recommendations.

4. Task Completion in the Experimental Condition

Participants repeated a similar set of tasks, in the same simulated environment with the 5 products using the RetailARVA system—an AR-based avatar assistant capable of barcode scanning, contextual product recognition, voice interaction, and personalized product recommendations. This condition enabled evaluation of the added value of the virtual assistant on user experience and decision-making.

5. Post-Task Questionnaires and Measurements

To comprehensively evaluate the outcomes, a variety of quantitative measures were collected throughout the study. Task completion times were logged for each task under both conditions to assess efficiency and speed of decision-making. Following each task, participants completed the NASA Task Load Index (NASA-TLX), included in Appendix B, to capture perceived cognitive and physical workload across six dimensions, mental demand, physical demand, temporal demand, performance, effort, and frustration level. Administering the NASA-TLX after each individual task, rather than once at the end of a condition, allowed for finer-grained workload analysis and reduced memory bias.

Upon completion of all tasks within the experimental condition, participants were additionally asked to complete the System Usability Scale (SUS), included in Appendix C, and the User Experience Questionnaire (UEQ), included in Appendix D. The SUS provided a standardized measure of system usability, focusing on simplicity, learnability, and satisfaction. The UEQ expanded this evaluation to include dimensions of attractiveness, efficiency, dependability, stimulation, perspicuity, and novelty, offering a holistic assessment of the user experience with RetailARVA. The UEQ measured multiple dimensions including pragmatic quality (e.g., perceived efficiency and clarity), hedonic quality (e.g., stimulation and novelty), and overall attractiveness.

6. Final Feedback Session and Semi-Structured Interview

At the end of both conditions, participants participated in a semi-structured interview, guided by the set guidelines as outlined in Appendix E, was designed to capture in-depth qualitative feedback on their experiences. The interview followed a flexible, open-ended format, allowing participants to freely express their thoughts while also ensuring consistency across interviews. The interviewer guided the conversation with a set of core questions, but participants were encouraged to elaborate on their responses, offering valuable insights into their subjective experience. Key topics included,

1. **Usefulness of the AR assistant** - Participants were asked about their perceptions of how the virtual assistant contributed to or hindered their shopping experience.

2. **Impact on decision-making** - They were asked whether the AR-based system influenced their product selection and decision-making process compared to a traditional shopping experience.

3. **Challenges and usability issues** - Participants shared any difficulties they encountered when interacting with the system, including navigation issues, technical difficulties, or unclear system responses.

4. **Overall satisfaction**- They were asked to reflect on their general satisfaction with both conditions, providing insights into which setup they preferred and why.

The combination of task performance metrics, workload assessments, and post-condition usability and experience surveys was deliberately selected to provide a comprehensive evaluation of both system impact and user perception. Task completion times served as an objective indicator of efficiency improvements facilitated by the RetailARVA system. The NASA-TLX scores enabled analysis of cognitive load variations between conditions, offering insights into whether the system alleviated or introduced workload. Finally, the SUS and UEQ instruments captured subjective evaluations of usability and user experience, critical factors for the real-world adoption of AR-enabled, AI-driven retail technologies. The interview data also provided deeper insights into specific aspects of the system's performance that were not fully captured through the structured measures.Through this multi-faceted methodological approach, the study aimed to derive a nuanced understanding of the RetailARVA system's effectiveness, usability, and impact on the customer shopping experience.

This comprehensive procedure ensured a balanced and methodologically sound evaluation of the RetailARVA system, capturing both quantitative and qualitative data on its impact on user experience, satisfaction, cognitive load, and purchase intent.

6.2 Results

6.2.1 Overview

This section presents the findings from the user study conducted to evaluate the final prototype of RetailARVA, focusing on the integration of AR and LLM-based conversational AI in enhancing retail experiences. A within-subjects experimental design was used with 30 participants, each completing three tasks under two conditions,

- **Control** - Traditional retail setup with static product displays and a simulated store assistant.
- **Experimental** - Retail setup with access to AR-powered RetailARVA prototype featuring barcode scanning, voice interaction, and personalized recommendations.

The order of conditions was counterbalanced using a Latin Square design to mitigate order effects. After each task, participants completed the NASA TLX to assess cognitive workload. After the experimental condition, they completed the System Usability Scale (SUS)survey and the User Experience Questionnaire (UEQ).

The session concluded with a semi-structured interview to gather in-depth qualitative feedback. This mixed-method approach enabled comprehensive analysis of usability, user experience, and perceived value of AR and conversational features in a physical retail context.

6.2.2 Participant Demographics

A total of 30 participants took part in the user study, representing a diverse mix of backgrounds and skincare shopping behaviors. The demographic profile included a balance of individuals with varying levels of technical familiarity, shopping frequency, and prior exposure to intelligent systems relevant to the RetailARVA platform.

- Most of the participants (60%) ranged from 18 to 25 years, 23% of the participants fell within the 25–40 age group and 16% aged 41 years and older.
- The sample included 23 female and 7 male participants.

- 70% of participants reported prior experience or awareness of AR technologies (AR filters, AR shopping apps,AR Avatars), while 30% had little to no exposure.
- 83% of participants had previously interacted with voice assistants, with varying levels of frequency, while 17% had never used such systems.
- 63% identified as regular skincare buyers (purchasing at least monthly), while 37% made occasional purchases or only when needed.
- A significant portion of participants (56%) actively considered skin type and product suitability (e.g., sensitivity, oiliness, acne-prone skin) in their buying decisions, indicating high awareness of product-specific needs.

The high proportion of younger users (18–25) aligns with the tech-savvy, skincare-conscious demographic that is often the early adopter of emerging technologies. The majority's familiarity with AR and voice assistants indicates an existing baseline of technological competence, which supports meaningful engagement with the RetailARVA system. Furthermore, the prevalence of regular skincare routines and awareness of product suitability highlight a user base that is motivated and discerning, qualities critical to evaluating the relevance, usability and impact of the system on decision-making in real-world shopping scenarios.

6.2.3 Task Performance Analysis

To assess the efficiency impact of the RetailARVA system, task completion times were measured across three skincare-related decision-making tasks in both the control (traditional retail) and experimental (RetailARVA-enabled AR assistant) conditions. Paired t-tests revealed statistically significant differences in completion times for all tasks, highlighting the system's ability to streamline users' information-seeking and decision-making processes.

For Task 1—where participants had to examine a single product and assess its safety and suitability for sensitive or acne-prone skin—users completed the task significantly faster using RetailARVA ($M = 34.83$s, $SD = 8.42$) compared to the traditional setup ($M = 69.10$s, $SD = 8.33$). A paired samples t-test was conducted to compare completion times between the two conditions. The results revealed a statistically significant difference ($t(29) = 18.87$, $p < 0.001$), indicating that RetailARVA substantially improved decision speed. In the control condition, many participants were observed flipping product packaging to manually check ingredients, scanning printed material, or hesitantly asking the simulated store assistant. In contrast, RetailARVA allowed them to retrieve summarized suitability and side-effect data instantly through voice commands or AR overlays, streamlining the decision-making process.

For Task 2, which required selecting the more suitable of two similar skincare products (moisturizers) for someone with dry, sensitive skin, a similar trend emerged. Participants completed this comparison more efficiently with RetailARVA (M = 63.87s, SD = 8.40) than without it (M = 101.73s, SD = 9.74), ($t = 18.78$, $p < 0.001$). In the control condition, participants often hesitated, revisited labels, or verbally expressed uncertainty. Some attempted to "Google" ingredients on their phones, which added to the task duration. RetailARVA's concise, personalized summaries helped users directly compare attributes and suitability, leading to faster and more confident choices.

The most pronounced improvement was observed in Task 3, which involved choosing the best product for oily skin with breakouts from a shelf of five visually similar items. Without RetailARVA, participants spent significantly more time (M = 311.47s, SD = 22.35), compared to when using the system (M = 61.27s, SD = 5.32), ($t = 60.60$, $p < 0.001$). This task proved cognitively and logistically complex in the traditional setting, as users frequently switched between products, attempted to cross-check information online, and asked the assistant for help multiple times. Many were visibly frustrated by information overload. RetailARVA's barcode-scanning and contextual filtering capabilities drastically reduced time spent navigating and evaluating options, allowing participants to focus on decision criteria rather than data gathering.

The grouped bar chart (Figure 6.2) illustrating the average task completion times across the three decision-making tasks provides a clear visual comparison of performance between the control and experimental conditions. Across all tasks, the bars representing the without RetailARVA condition are substantially taller, indicating significantly longer completion times compared to the with RetailARVA condition.

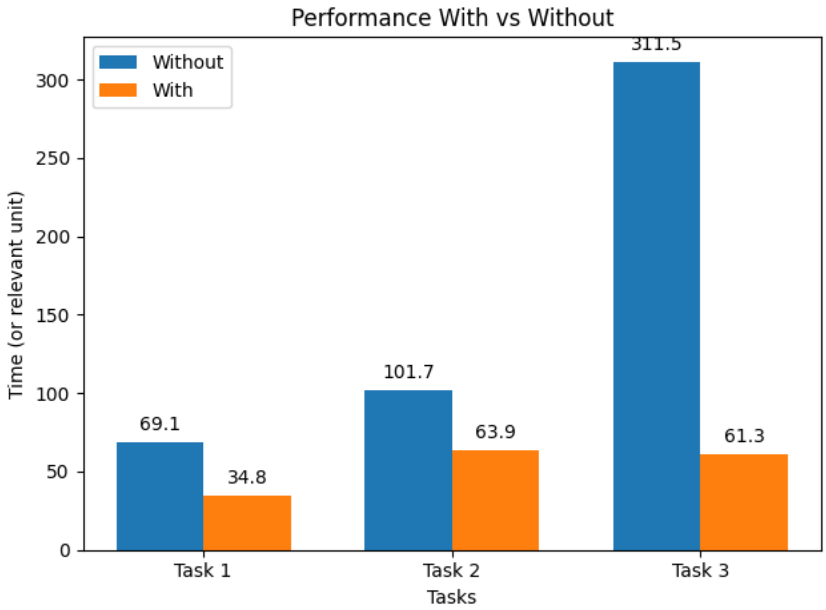


Figure 6.2: Comparison of completion times of tasks between the control and experimental conditions

NASA-TLX Scores (Per Task)

The NASA Task Load Index (TLX) evaluation offers a detailed understanding of user workload [46] during three skincare-related tasks, performed under two conditions: with and without the RetailARVA system. The six subscales of the TLX—Mental Demand, Physical Demand, Temporal Demand, Effort, Performance, and Frustration Level—are used to assess the user experience. For this comprehensive analysis, we focus primarily on the key subscales that most directly influenced user experience: Mental Demand, Effort, Performance, and Frustration Level. The evaluation used a paired t-test to assess the statistical significance of workload differences between conditions. In this analysis, we explore how each of these factors contributed to the overall user experience for each task.

Tasks Overview

The tasks selected for this study were designed to reflect real-world challenges in the skincare retail environment. Each task required decision-making under varying levels of complexity and cognitive engagement,

- Task 1: Evaluate a skincare product on display for its suitability for sensitive or acne-prone skin and identify any potential side effects.
- Task 2: Compare two skincare products (e.g., serums) and determine which is more suitable for daily use by individuals with dry, sensitive skin.

- Task 3: From a selection of five visually similar products, choose the most appropriate one for a given skin concern (e.g., oily skin with breakouts).

These tasks were chosen because they require different cognitive and decision-making processes, making them suitable for testing the impact of RetailARVA on cognitive workload and user performance.

Key Findings and Statistical Analysis

The results of the paired t-tests showed significant reductions in mental workload, effort, and frustration levels across all three tasks when RetailARVA was used. Below are the key subscales and their analysis,

Table 6.1: NASA-TLX Scores (Means and Standard Deviations) Across Tasks and Conditions

Task	Condition	Mental Demand	Effort	Performance	Frustration
T1	Control	M=5.93, SD=0.68	M=6.00, SD=0.63	M=2.93, SD=0.73	M=5.23, SD=0.88
	RetailARVA	M=1.83, SD=0.69	M=1.90, SD=0.54	M=6.47, SD=0.50	M=1.27, SD=0.44
T2	Control	M=6.60, SD=0.50	M=6.60, SD=0.50	M=1.73, SD=0.58	M=6.40, SD=0.56
	RetailARVA	M=1.80, SD=0.70	M=2.27, SD=0.82	M=6.13, SD=0.72	M=1.43, SD=0.72
T3	Control	M=6.50, SD=0.50	M=5.93, SD=0.63	M=3.53, SD=0.50	M=4.73, SD=0.63
	RetailARVA	M=2.13, SD=0.67	M=1.87, SD=0.89	M=6.00, SD=0.68	M=1.20, SD=0.40

Paired t-test Results

Mental Demand, which reflects the cognitive effort required to complete a task, showed a marked reduction across all three tasks, with Task 1 demonstrating the most pronounced drop (from 5.93 in the control condition to 1.83 with RetailARVA; $t = 60.601$, $p < 0.0001$). This substantial decline suggests that RetailARVA effectively reduced the cognitive complexity associated with evaluating skincare products by delivering personalized recommendations and filtering irrelevant information. By automating the decision-making process and presenting only the most pertinent data, the system minimized the mental resources required for information retrieval, analysis, and judgment, particularly during suitability evaluation (Task 1), product comparison (Task 2), and routine formulation (Task 3).

Similarly, the Effort subscale, measuring the perceived amount of mental and physical exertion—demonstrated a consistent decrease across all tasks, with Task 2 showing the greatest improvement (from 6.60 to 2.27; $t = 47.547$, $p < 0.0001$). The system's ability to present comparisons and distill complex product data into user-friendly conversations significantly lightened the user workload. This reduction in effort is indicative of RetailARVA's facilitative role in streamlining the information processing burden and minimizing both the manual labor of navigating multiple sources and the mental workload of synthesizing disparate data. The implications of this are substantial, particularly in scenarios requiring the comparison of multiple attributes or formulation of multi-step routines.

In terms of Performance, all three tasks recorded a significant increase in user-perceived success, with the most notable improvement in Task 3 (from 3.53 to 6.00; $t = 25.836$, $p < 0.0001$), which required users to select the best suited product from multiple products. Participants reported greater confidence in the accuracy and completeness of their task outcomes when supported by RetailARVA. This enhancement can be attributed to the system's ability to curate relevant product suggestions based on individual user profiles, thereby reinforcing decision accuracy and reducing the likelihood of user error or indecision.

Frustration levels also declined significantly across all tasks, with Task 1 again showing the largest reduction (from 5.23 to 1.27; $t = 54.704$, $p < 0.0001$). The lower frustration scores suggest that users experienced less stress, confusion, and emotional strain when engaging with RetailARVA. This reduction likely stems from the system's ability to alleviate common pain points associated with unassisted product discovery—such as information overload, conflicting claims, and ambiguous terminology.

In-Depth Subscale Analysis

Mental Demand

Mental Demand reflects the cognitive resources required to complete a task. As shown in Figure 6.3, all three tasks exhibited a substantial decline in Mental Demand when RetailARVA was introduced. The mean score dropped from 5.93 to 1.83 in Task 1, and from 6.50 to 2.13 in Task 3. This reduction highlights the system's effectiveness in minimizing cognitive load by automating the extraction and presentation of product suitability details. Particularly in Task 3, where users had to differentiate among five similar products, the cognitive burden was considerably higher in the control condition due to the need to mentally compare and evaluate subtle differences. RetailARVA's real-time, contextual recommendations greatly simplified this process, resulting in lower perceived mental effort.

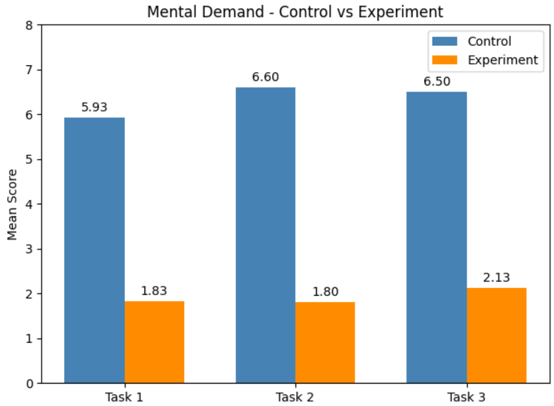


Figure 6.3: Mental Demand Comparison under control and experiment conditions

Effort

Effort, which encompasses both cognitive and physical exertion, showed consistent reductions across all tasks with the use of RetailARVA. Figure 6.4 illustrates these differences, with the most prominent change observed in Task 2, where the mean effort rating decreased from 6.60 (control) to 2.27 (RetailARVA). The system's ability to display side-by-side comparisons and highlight key product attributes significantly reduced the energy participants needed to evaluate and choose between two similar products. The data suggest that the system successfully acted as a cognitive offloading tool, reducing user strain and enhancing task fluency.

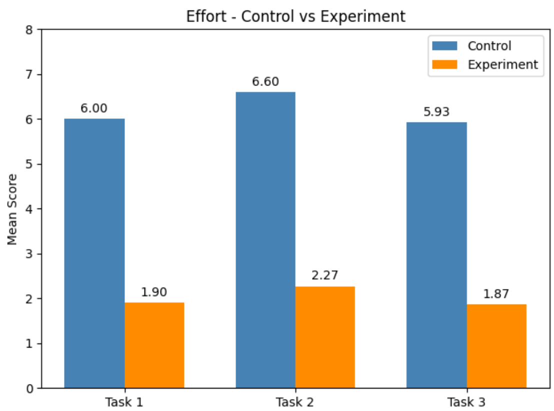


Figure 6.4: Effort Comparison under control and experiment conditions

Performance

As shown in Figure 6.5, participants' self-rated Performance increased notably with RetailARVA across all tasks. In Task 3, the average score rose from 3.53 to 6.00, demonstrating a significant gain in users' perceived success. This improvement can be attributed to the structured support RetailARVA provided, which helped users feel more confident and informed in their choices. Without the system, participants likely experienced greater uncertainty and hesitation, especially when navigating multiple products with similar claims. The increase in performance ratings indicates not just easier task completion, but a higher perceived accuracy and quality of decisions.

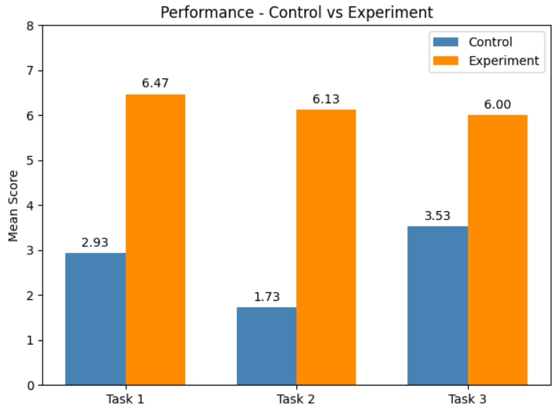


Figure 6.5: Performance Comparison under control and experiment conditions

Frustration

Frustration levels were significantly reduced with RetailARVA, as illustrated in Figure 6.6. In Task 1, the mean Frustration score dropped from 5.23 to 1.27, indicating that users found the experience substantially less stressful when assisted by the system. This pattern was consistent across all tasks. In the control condition, participants often expressed confusion, annoyance, or fatigue when trying to interpret product information on their own. RetailARVA's streamlined interaction design, immediate response time, and user-friendly presentation of data contributed to a calmer and more satisfying experience. The lowered frustration scores reflect not only cognitive benefits, but also an emotional enhancement in the interaction flow.

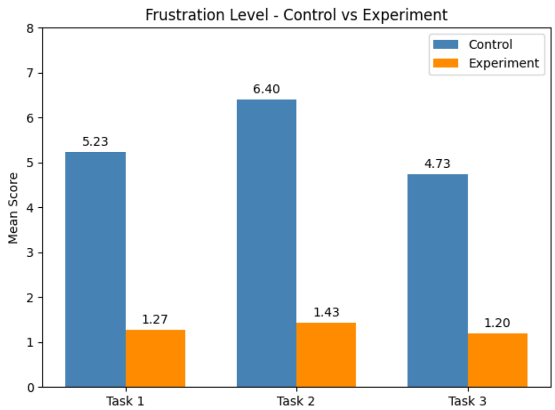


Figure 6.6: Frustration Level Comparison under control and experiment conditions

6.2.4 User Experience Evaluation

UEQ Results

The User Experience Questionnaire (UEQ) is a widely used psychometric tool for assessing the user experience (UX) of interactive systems [47]. It provides a multidimensional evaluation of the pragmatic and hedonic qualities of a system, making it highly relevant for understanding users' perceptions of the usability, aesthetic appeal, and emotional engagement of a system. The UEQ measures users' subjective assessments of six key dimensions: Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty. These dimensions are grouped into two primary categories: pragmatic quality (focusing on usability and task performance) and hedonic quality (focusing on emotional satisfaction and engagement).This section presents an in-depth analysis of the UEQ results from participants interacting with the RetailARVA system.

Methodology: Performing the UEQ

The UEQ was administered to 30 participants who interacted with the RetailARVA system. The questionnaire consisted of 26 items, each designed to assess a specific aspect of the user experience. Participants rated their experiences using a 7-point Likert scale ranging from -3 (horribly bad) to +3 (extremely good), with 0 representing a neutral evaluation. The scale measures a broad spectrum of user experience factors, from the task-related aspects of the system (e.g., efficiency and dependability) to the emotional aspects (e.g., stimulation and novelty).

The analysis of the User Experience Questionnaire (UEQ) data was conducted using the official UEQ analysis toolset available online [48] . This toolset was utilized to transform the raw responses collected on the 7-point Likert scale into standardized scores suitable for analysis. It automatically computed the mean values, confidence intervals, and standard deviations for each UEQ scale (Attractiveness, Perspicuity, Efficiency, Dependability, Stimulation, and Novelty). Additionally, it facilitated the generation of visual representations, including benchmark comparisons and bar charts, based on our collected dataset. This enabled a structured and systematic evaluation of the user experience, allowing for both quantitative interpretation and effective visualization of the results.

Analysis of UEQ Data

The following table presents the mean and variance scores for the individual items on each of the six UEQ scales, followed by the mean scores for the scales themselves,

Scale	Mean	Variance
Attractiveness	1.922	0.10
Perspicuity	2.050	0.33
Efficiency	1.792	0.28
Dependability	1.600	0.10
Stimulation	2.092	0.23
Novelty	2.067	0.18

Table 6.2: Mean and Variance for Different Scales

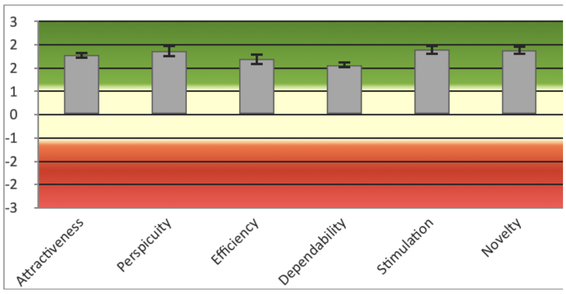


Figure 6.7: Distribution of Scores for the 5 different scales

Attractiveness (M = 1.922); The system is perceived as visually appealing, with users rating it as pleasant and engaging. This suggests that the design of the system positively influences the user experience, in line with findings from Hassenzahl et al. (2001)[49], who emphasized the importance of aesthetic appeal in shaping users' overall satisfaction.

Perspicuity (M = 2.050); The system was rated highly for clarity and ease of understanding, indicating that users found it intuitive and user-friendly. This result suggests that the interface design effectively supports usability, a key factor for user retention and satisfaction in many interactive systems.

Efficiency (M = 1.792); Users reported that the system performs tasks relatively efficiently, but there may be room for improvement in terms of response time and task completion speed. While the system does not seem to be inefficient, the mean score suggests that some users may have experienced delays or inefficiencies during their interactions.

Dependability (M = 1.600); This score indicates that the system is generally perceived as reliable and stable. However, the somewhat lower rating compared to other scales may suggest that users encountered occasional issues related to system stability or performance.

Novelty (M = 2.067); Users found the system to be innovative and creative, highlighting its ability to offer a fresh and original experience. This result underscores the system's uniqueness, which is a significant factor in enhancing user satisfaction and differentiation in the competitive landscape of retail technologies.

Pragmatic vs. Hedonic Quality Interpretation

The UEQ categorizes the six scales into two primary groups, pragmatic quality and hedonic quality. The pragmatic quality group comprises Perspicuity, Efficiency, and Dependability, which focus on task-related aspects of the user experience. The hedonic quality group includes Stimulation and Novelty, which assess the emotional and aesthetic dimensions of the system.

Pragmatic Quality ; The mean score (M = 1.81) was calculated by averaging the Perspicuity, Efficiency, and Dependability scales. The system performs well in terms of practical usability, with particularly positive evaluations for perspicuity and efficiency. However, the slightly lower score for dependability suggests that improvements in system reliability could enhance overall pragmatic quality.

Hedonic Quality ; The mean score (M = 2.08) was calculated by averaging the Stimulation and Novelty scales, The system excels in emotional engagement, as indicated by high scores for both stimulation and novelty. This suggests that users enjoy interacting with the system and perceive it as offering an exciting and fresh experience.

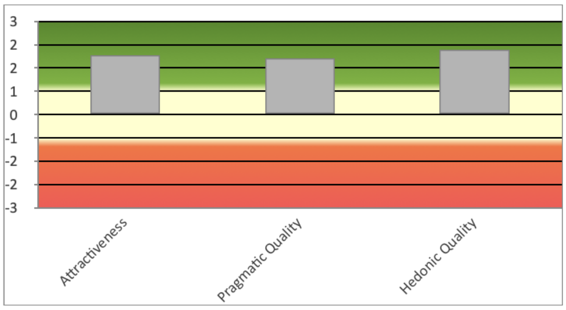


Figure 6.8: Comparison of Pragmatic and Hedonic Scores

The results from the UEQ provide valuable insights into the user experience of the RetailARVA system. The system was rated positively in both pragmatic and hedonic aspects, but the hedonic quality was particularly high, indicating that users derive significant emotional satisfaction from interacting with the system. In terms of pragmatic quality, the system performed well in perspicuity and efficiency, but there is room for improvement in dependability. Enhancing system reliability could further strengthen the overall user experience, ensuring that users not only enjoy interacting with the system but also trust it to perform consistently.

Benchmark Comparison of UEQ Results

To better understand the user experience of the RetailARVA system, we compared its User Experience Questionnaire (UEQ) scores to a comprehensive benchmark dataset [50]. This dataset includes responses from 21,175 participants across 468 studies of various products, such as business software, websites, webshops, and social networks. By examining how RetailARVA compares to these products, we gain valuable insights into its relative performance and quality.

The benchmark dataset consists of evaluations of diverse products, covering a broad range of industries. The scores are categorized as follows [48],

1. Excellent: Products scoring in the top 10%.
2. Good: Products that fall between the 10th and 75th percentile.
3. Below Average: Products scoring lower than the 75th percentile.

These comparisons provide a clear picture of where RetailARVA stands in relation to other evaluated products.

RetailARVA System Benchmark Results

The table below shows the mean scores for each UEQ scale of the RetailARVA system, their comparison to the benchmark, and an interpretation of the results:

Scale	**Mean Score**	**Comparison to Benchmark**
Attractiveness	1.92	Excellent
Perspicuity	2.05	Excellent
Efficiency	1.79	Good
Dependability	1.60	Good
Stimulation	2.09	Excellent
Novelty	2.07	Excellent .

Table 6.3: Scale Mean Scores, Benchmark Comparisons, and Interpretations

The RetailARVA system achieved a score of 1.92 in attractiveness, positioning it in the top 10% of all products evaluated. This indicates that users find the system visually appealing and its design highly engaging. A well-designed, attractive user interface is critical in ensuring a positive user perception, and this score reflects the strong appeal of the system's visual elements.

The score of 2.05 on perspicuity places RetailARVA in the top 10%, suggesting that users find the system intuitive and easy to understand. This high score indicates that the system's structure and instructions are clear, facilitating a smooth user experience with minimal confusion or need for support.

With a score of 1.79, the system ranks as good, performing better than 10% of products but still leaving 75% of products scoring higher in terms of task completion time and responsiveness. This suggests that while the system is efficient, there is still room for improvement in speed and optimization to compete with top-tier products.

Scoring 1.60 for dependability, the system is in the good range, outperforming 10% of products but still trailing 75% in terms of reliability. This indicates that while RetailARVA performs well under most conditions, further enhancements in stability and error reduction could bolster its reputation for reliability.

The system's score of 2.09 for stimulation places it in the top 10%, indicating that users find the experience highly engaging and exciting. The high score reflects the system's ability to stimulate interest and enthusiasm, likely due to its innovative features and interactive design elements.

Scoring 2.07 for novelty, RetailARVA stands out for its innovation, ranking in the top 10% of evaluated products. This score underscores the system's creativity and fresh approach, providing unique features that differentiate it from others in the market.

The RetailARVA system performs strongly across all six User Experience Questionnaire (UEQ) scales, particularly excelling in attractiveness, clarity, engagement, and innovation. The system's results , highlights its strong user appeal and ease of use. While it shows promising results, especially in stimulation and novelty, there are opportunities for improvement in areas such as efficiency and dependability. These findings provide useful insights into the user experience of the system and guide future efforts for refinement and enhancement.

6.2.5 SUS Results (Experimental Condition Only)

Developed by John Brooke (1986),The System Usability Scale (SUS) is a robust and widely adopted instrument for evaluating the perceived usability of interactive systems [51]. It consists of 10 items that alternate between positively and negatively worded statements. Responses are recorded on a 5-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree," and the final SUS score is scaled between 0 and 100.. The SUS enables both relative and absolute assessments of system usability, making it especially valuable in user experience research.

In this study, 30 participants completed the SUS questionnaire following their interaction with RetailARVA.Each participant's responses were converted into a SUS score using the standard scoring method. The SUS scores ranged from 67.5 to 85.0, with a mean score of 75.0 and a median of 75.0. These results are summarized in Table 6.2 and illustrated in Figure 6.7.

Metric	**Value**
Minimum Score	67.5
Maximum Score	85.0
Mean Score	75.0
Median Score	75.0
Standard Deviation	~ 4.4

Table 6.4: Descriptive statistics of the scores

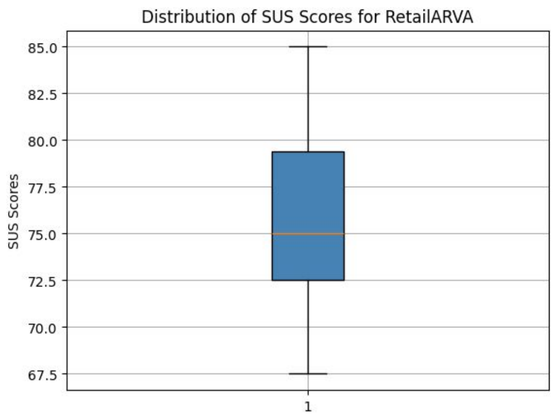


Figure 6.9: Distribution of SUS Scores for RetailARVA

According to industry-standard interpretation scales (e.g., Bangor et al., 2009), a SUS score above 68 is considered above average, while scores in the 70–80 range typically correspond to a "Good" to "Excellent" usability rating. A score of 75.0 places RetailARVA comfortably within this category, suggesting that the system offers a high level of usability and a generally positive user experience.

The distribution of scores was relatively tight, with most participants rating the system between 70 and 80, indicating strong consensus in perceptions of usability. Furthermore, 87% of participants (26 out of 30) rated the system at 70 or higher, and one-third (10 participants) provided scores of 80 or above, reinforcing the notion that the system was not only usable but also well-received.The relatively low standard deviation (4.4) reflects low variability among users, implying that usability perceptions were consistent across different demographic groups and task experiences. While no extreme low outliers were observed, the absence of scores in the 90+ range suggests that some minor usability issues may have been present, possibly limiting the potential for even higher user satisfaction.

Item-Level Averages (N = 30)

SUS Item	Mean Response
I think that I would like to use this system frequently.	4.43
I found the system unnecessarily complex.	2.33
I thought the system was easy to use.	4.47
I think I would need the support of a technical person to use it.	2.60
I found the various functions in this system were well integrated.	4.80
I thought there was too much inconsistency in the system.	2.70
I would imagine most people would learn to use this system quickly.	4.27
I found the system very cumbersome to use.	2.27
I felt very confident using the system.	4.40
I needed to learn a lot of things before I could get going.	2.47

Table 6.5: Mean responses for each SUS item

As illustrated in Table 6.3 item-level results reflect a high level of user confidence (M = 4.4), strong perceived integration of features (M = 4.8), and ease of use (M = 4.47), suggesting that most users found the system to be intuitive and supportive of their goals. Low average ratings for negative items such as complexity (M = 2.33), inconsistency (M = 2.70), and cumbersomeness (M = 2.27) also reinforce the impression that the system was streamlined and user-friendly.

These findings align with qualitative feedback from the semi-structured interviews, where users described the system as intuitive, helpful, and efficient. However, several participants also noted limitations such as occasional instability in AR overlays and minor inaccuracies in voice recognition, which may have tempered higher SUS ratings. These issues, while not detrimental to overall usability, likely contributed to a ceiling effect in the score distribution.

The SUS score of 75 indicates that RetailARVA meets and exceeds the usability expectations for early-stage interactive systems, particularly in the domain of augmented reality retail support tools. This score reflects positively on the system's learnability, efficiency, and user satisfaction—three core dimensions of usability. The score also supports the conclusion that the system has strong potential for real-world adoption, with users finding it both easy and satisfying to use.

Nonetheless, to further improve usability and achieve even higher SUS scores, iterative refinements are recommended. These include enhancements to AR alignment, better handling of edge cases in voice input, and more flexible user interaction options. Addressing these aspects could elevate the system's usability rating from "good" to "excellent," and position RetailARVA as a leading solution in AR-based retail assistance.

6.2.6 Qualitative Analysis

Thematic Analysis of Semi-Structured Interviews

The semi-structured interviews conducted post-task provided valuable insights into users' subjective experiences with RetailARVA. Four primary themes emerged from the thematic analysis: perceived usefulness, interaction ease and intuitiveness, decision-making enhancement, and system limitations and frustrations. These themes were derived through inductive open coding and iterative thematic refinement, aiming to reflect the diversity and depth of participant perspectives.

1. Perceived Usefulness and Relevance

RetailARVA was widely perceived by participants as a highly beneficial tool, especially in contexts requiring careful consideration of skincare compatibility and ingredient awareness. A prevailing sentiment was that the system reduced cognitive load by handling complex filtering tasks that users would otherwise have to perform manually.

Several users commented on how the system facilitated more informed purchasing decisions by quickly identifying products aligned with individual skin profiles. One participant shared, "Normally I would spend 20 minutes just reading ingredient lists and Googling things I don't know. This told me exactly what I needed to know right away." Another elaborated, "With sensitive skin, trial and error gets expensive and painful. This saved me from making a bad selection."

Moreover, participants appreciated the product matching feature, which dynamically adjusted recommendations based on their preferences. One remarked, "The fact that it updated suggestions as I changed filters was really helpful." Collectively, these insights underscore the system's perceived utility in reducing time, effort, and uncertainty, especially when choosing products that align with health or lifestyle-related constraints.

2. Ease of Interaction and Interface Design

The majority of users found RetailARVA intuitive and easy to operate, with voice prompts and AR overlays contributing to a low learning curve. The hands-free experience enabled by gesture and voice interaction was reported as particularly effective in a retail setting where users might be physically engaged with products.

Participants generally described RetailARVA as intuitive and easy to interact with, even for those with limited technical backgrounds. Users appreciated the clean interface, minimal learning curve, and logical navigation structure that enabled them to use the system effectively without prior instruction or support.

One participant stated, "I didn't have to think too much about how to use it. It just made sense, which I liked a lot." Another mentioned, "It felt like the app understood what I meant—even when I wasn't being super specific. Like, I chose 'oily skin' and it was able to tell me what I needed."

The system's ability to accommodate casual inputs while still generating relevant outputs was praised, suggesting that its underlying interaction model aligns well with user mental models. Another participant highlighted this by stating, "There weren't any confusing menus or steps. I just put in my concerns and it went to work. '

Importantly, users did not report encountering major usability barriers, indicating that the interaction design succeeded in supporting fluid engagement with the system's core function

3. Decision-Making Support and Confidence

RetailARVA's most significant contribution, as reported by participants, was its ability to enhance decision-making confidence and efficiency—especially when navigating the complexities of personalized skincare. Participants frequently highlighted that the system helped streamline their evaluation process by narrowing down the product space to only the most relevant options, thereby minimizing decision fatigue and reducing cognitive overload.

This was particularly evident in scenarios involving multiple constraints, such as ingredient avoidance, skin condition sensitivities, or ethical product choices. One participant remarked, "When you're shopping for something like sunscreen that doesn't cause breakouts, it helps to already narrow down to just the safe ones. I trusted what the system gave me." Similarly, another user shared, "It definitely helped me make up my mind faster. Without it, I'd probably still be reading reviews and ingredient lists."

The system's value was even more pronounced in risk-averse contexts, such as purchasing products for allergy-prone or reactive skin. Participant 5 explained, "I've had allergic reactions to skincare before. This tool helped me feel like I could avoid that by filtering out the dangerous stuff right away." The pre-emptive filtering of unsuitable products contributed to a sense of personal safety and trust in the recommendations provided.

Another commonly expressed advantage was the system's ability to replace time-consuming information-seeking tasks, such as browsing third-party websites, consulting store assistants, or manually comparing ingredient labels. Participants emphasized that RetailARVA empowered them to make informed decisions independently and at their own pace, without relying on external help. As one participant stated, "Normally I'd have to Google ingredients or run around looking for someone to explain, but here I had everything on the screen. It was like a mini skincare expert walking with me." Another added, "It was nice not to feel rushed like when you're in a store. I could take my time, read, ask questions, and it didn't feel overwhelming."

Moreover, the seamless integration of real-time, contextual product data gave users the impression of having greater access to relevant information than would typically be unavailable in a traditional retail environment. A participant commented, "It showed what matched my skin and explained why. I didn't need to do additional research afterwards.'

In this way, RetailARVA functioned not merely as a browsing tool, but as a decision-support companion , a system that augmented users' ability to make faster, more confident, and better-informed choices. It aligned with their specific needs, reduced reliance on external resources, and transformed the in-store experience into a self-guided, data-enriched journey tailored to individual constraints and preferences.

4. System Limitations and Frustrations

Despite the overall positive reception, participants identified several areas where RetailARVA's functionality could be improved. One commonly reported limitation pertained to the precision of recommendations, particularly when dealing with highly specialized product needs or combinations of preferences. While the system generally succeeded in narrowing down relevant options, some users found that the granularity of filtering was insufficient for more nuanced requirements. For example, a participant remarked, "It was helpful, but sometimes I felt like it gave general suggestions instead of really specific matches." Similarly, another user observed, "I selected a lot of preferences, but the products still had alcohol in them, which I can't use. It wasn't filtering strictly enough."

In addition to recommendation specificity, technical constraints related to the AR functionality were also noted. A few participants encountered issues where the augmented reality (AR) tracking failed to accurately map the avatar or information overlays onto the physical environment. A participant reported, "It lagged a bit and the avate wasn't pinned to the right place once or twice—it broke my focus." Such inconsistencies disrupted the immersive flow of interaction and occasionally led to loss of user attention and reduced task engagement.

Another limitation arose with the voice interaction feature, which while appreciated in principle, suffered from occasional misrecognition. Several users pointed out that product names, brand variations, or accents were not always understood accurately by the system. One participant explained, "I had to repeat the product name a few times before it got it right, which was annoying when I was in the middle of browsing." These breakdowns in speech recognition reduced the perceived fluidity and intuitiveness of the interaction, especially for users expecting hands-free convenience.

Furthermore, a few participants noted a lack of transparency in the recommendation rationale, which occasionally undermined trust in the system. As one user articulated, “I liked the suggestions, but I didn’t always know why they were recommended. Was it because of my skin type? Or something else?” This sentiment suggests a need for explainable AI features that clarify the relationship between user inputs and generated outputs, enhancing both confidence and user control.

Taken together, these user insights suggest key areas for iterative design improvements—ranging from enhanced AR stability and voice recognition accuracy to more transparent recommendation mechanisms and expanded personalization settings. Addressing these challenges may significantly improve the overall experience and inclusivity of the RetailARVA system.

User Suggestions for Improvement

Participants provided a range of insightful suggestions aimed at enhancing RetailARVA’s capabilities, usability, and long-term value. These suggestions reflected not only a willingness to continue using the system but also a shared vision for its evolution into a more intelligent, personalized, and inclusive tool.

Participants also emphasized the need for expanded ingredient intelligence. While the system already provided real-time product information, several users recommended deeper integration with trusted third-party certifications such as ”dermatologically tested,” ”cruelty-free,” or ”FDA-approved” labels. Others advocated for incorporating scientific validation of product claims and direct links to external skincare knowledge bases. This would allow users to cross-verify information and feel more confident in the credibility of the system’s recommendations.

In terms of usability, AR responsiveness was a focal point for improvement. Some users noted that the system’s overlay features occasionally lagged or displayed unstable behavior. Suggestions included refining the visual anchoring of digital avatar , improving stability, and reducing screen clutter to maintain a focused and immersive experience.

Several participants raised concerns around privacy and data transparency, especially in relation to the handling of sensitive private information. They recommended that the system offer clearer explanations of how personal data is collected, processed, and stored, along with options for users to control what information they share. Implementing opt-in data sharing and secure storage protocols was seen as critical to enhancing user trust and regulatory compliance.

There were also calls for greater accessibility and inclusivity. Participants proposed features such as text-to-speech support for visually impaired users, as well as multilingual functionality to accommodate non-English speakers. These suggestions reflect a desire to broaden the system's utility across diverse user groups, ensuring that RetailARVA is usable and beneficial regardless of language ability or physical impairments.

Lastly, a small group of participants proposed light gamification elements and feedback mechanisms to increase engagement. Ideas included awarding virtual points for consistent use or recommending healthier skincare routines, as well as offering personalized insights based on usage patterns. These feedback loops could make the experience more dynamic, motivating users to return and interact more frequently.

Collectively, these user-driven recommendations indicate a strong appetite for a more adaptive, informative, secure, and inclusive system—one that not only meets immediate decision-making needs but also evolves with the user over time.

Chapter 7

Discussion

7.1 Research Findings

The final RetailARVA prototype successfully achieved full integration of key technological components aimed at enhancing in-store decision-making and user experience. The system incorporated barcode scanning for product identification, augmented reality (AR) overlays for contextual information presentation, and a conversational avatar equipped with speech-to-text (STT) and text-to-speech (TTS) capabilities. At its core, the system leveraged a large language model (LLM) to provide both natural language interaction and intelligent reasoning over a curated product dataset.

One of the standout accomplishments was the system's ability to deliver personalized product recommendations based on individual user profiles. The avatar assistant could understand user-stated preferences and concerns, dynamically tailor its suggestions, and respond to follow-up queries in natural language, simulating an expert human assistant. The integration of LLMs enabled the avatar to interpret complex user intents, perform contextual filtering, and explain product attributes clearly.

Empirical evaluation through a controlled user study demonstrated significant performance benefits of the RetailARVA system over traditional shopping setups. A controlled user study using a within-subjects design demonstrated the RetailARVA system's superior performance over traditional shopping. Participants completed three tasks—product evaluation, comparison, and selection—under both control (non-digital) and experimental (RetailARVA) conditions. Counterbalancing minimized order effects, enabling direct comparison of task efficiency, cognitive load, and user satisfaction across conditions.

Task performance was evaluated through average completion times across three representative skincare-related shopping scenarios. Results revealed statistically significant improvements in task efficiency when participants used the RetailARVA system. In Task 1 (evaluating a single product for safety and suitability), average completion time decreased from 69.10 seconds in the control setup to 34.83 seconds with RetailARVA (t = 18.87, p ¡ 0.001). For Task 2 (comparing two similar skincare products), time was reduced from 101.73 seconds to 63.87 seconds (t = 18.78, p ¡ 0.001). The most notable improvement occurred in Task 3 (selecting the most appropriate product from five options for oily skin), with a reduction from 311.47 seconds to 61.27 seconds (t = 60.60, p ¡ 0.001). Participants attributed these time savings to the instant access to summarized product data, ingredient analysis, and suitability filters presented through AR overlays and natural voice interactions.

Cognitive workload, assessed using the NASA Task Load Index (TLX), demonstrated consistent and significant reductions across all subscales with the RetailARVA system. In Task 1, Mental Demand dropped from 5.93 to 1.83, Effort from 6.00 to 1.90, and Frustration from 5.23 to 1.27, while Performance scores increased from 2.93 to 6.47. Similar patterns were found in Task 2, where Effort scores fell from 6.60 to 2.27 and Frustration from 6.40 to 1.43. In Task 3, the most complex, Mental Demand reduced from 6.50 to 2.13, and Performance improved from 3.53 to 6.00.

Qualitative feedback supported these findings. Many participants remarked that the system "did the thinking for them" by filtering out irrelevant options and highlighting key product attributes, which significantly eased decision-making. One participant shared, "Normally I'd spend 20 minutes reading ingredient lists and Googling them. With this, I had what I needed instantly." Another emphasized, "I didn't feel overwhelmed. I was just guided to what worked best for me."

The RetailARVA system achieved a mean SUS score of 75.0, falling within the "Good to Excellent" range. The majority of participants (87%) rated the system above 70, and one-third gave scores above 80. This reflects strong consensus around system usability, with low standard deviation (4.4) indicating consistent perceptions across diverse user profiles.

Individual item scores were particularly high for ease of use (M = 4.47), integration of system features (M = 4.80), and user confidence (M = 4.40). Conversely, negative constructs such as system complexity (M = 2.33) and cumbersomeness (M = 2.27) were rated low, reinforcing the system's streamlined design. Participants described the system as "intuitive" and "something I could use without much instruction." However, minor usability challenges were reported, including occasional AR overlay misalignment and voice recognition inaccuracies, particularly for specific product names or accents.

Results from the User Experience Questionnaire (UEQ) revealed strong user satisfaction across both pragmatic and hedonic dimensions. RetailARVA scored particularly high in Perspicuity (M = 2.05), Stimulation (M = 2.09), and Novelty (M = 2.07), placing it in the "Excellent" category relative to global UEQ benchmarks. The overall mean for pragmatic quality (M = 1.81) and hedonic quality (M = 2.08) suggests that users not only found the system clear and useful but also engaging and innovative. These high ratings reflect RetailARVA's success in delivering both functional value and an emotionally satisfying user experience.

Together, these findings validate the design and technical integration of RetailARVA. The system's seamless combination of barcode scanning, AR overlays, a conversational avatar, and large language model (LLM)-based personalization led to faster task completion, reduced cognitive load, high usability, and a compelling user experience. These outcomes underscore the potential of intelligent AR assistants in transforming physical retail into a more accessible, informative, and user-centered environment.

7.2 Discussion

This research set out to explore the design, integration, and evaluation of a novel system—RetailARVA—that combines augmented reality (AR), large language models (LLMs), barcode-based product recognition, and a conversational avatar in a physical retail context. The study was guided by three core research questions, each addressing a critical dimension of this integration, system feasibility, intelligence enhancements through LLMs, and the resulting user experience. The results from task-based performance metrics, cognitive workload analysis, usability scales, and qualitative feedback provide a multi-faceted understanding of how such a system can impact real-world retail experiences.

In response to the first research question, the study successfully demonstrated the feasibility of implementing an AR-enabled virtual assistant in a physical retail environment without relying on wearables. RetailARVA integrates barcode scanning for product recognition, AR overlays for contextual information, and natural voice interaction capabilities through speech-to-text (STT) and text-to-speech (TTS) technologies. By utilizing the camera and microphone of a standard mobile device, users could access real-time information and personalized recommendations without additional hardware. Participants found the system intuitive and non-intrusive, with the hands-free, natural interaction flow contributing to a seamless user experience. These results validate the technical feasibility of embedding intelligent, virtual AR assistants into physical retail environments using only mobile devices.

Addressing the second research question, the integration of a large language model (LLM) significantly enhanced the system's ability to provide personalized, context-aware responses to user queries. By interpreting user concerns such as skin sensitivities and preferences, the prototype dynamically filtered the product dataset to offer tailored recommendations. This resulted in faster and more confident decision-making, as evidenced by reduced task times (up to 80% faster) and improved NASA TLX scores across all dimensions, particularly in mental demand, effort, and frustration. Qualitative feedback also reinforced these findings, with users highlighting how the assistant "knew what they meant" even with vague inputs and how it simplified complex decisions. Additionally, the LLM enabled follow-up queries, comparisons, and clarifications—capabilities typically unsupported in static product recommendation engines—showing its value in transforming the virtual assistant into an intelligent, dialogue-based shopping companion.

For the third research question, the overall user experience with RetailARVA was overwhelmingly positive. High scores on the User Experience Questionnaire (UEQ)—including Attractiveness (1.92), Perspicuity (2.05), and Novelty (2.07), all within the "Excellent" benchmark range—alongside a System Usability Scale (SUS) mean score of 75.0, reflect strong usability and satisfaction. Participants described the experience as engaging, helpful, and emotionally reassuring, especially when selecting products tied to personal health concerns like acne or allergies. The system was praised for its intuitive design and real-time delivery of relevant information, contributing to reduced stress and improved confidence. While minor issues such as occasional voice recognition errors or AR misalignments were observed, they did not significantly detract from the overall experience and point toward areas for future technical refinement.

Together, the findings support the core hypothesis that an AR-enabled, LLM-powered virtual assistant can be effectively deployed in physical retail environments without wearables, delivering tangible enhancements in usability, task efficiency, and customer experience. The study confirms that such systems not only improve functional interactions (e.g., faster comparisons, clearer recommendations) but also elevate the emotional and cognitive aspects of shopping, making the experience more guided, informative, and personally relevant.

7.2.1 Limitations

Despite the encouraging results from the evaluation of the RetailARVA system, several limitations must be acknowledged that could influence the interpretation and generalizability of the findings. Firstly, the relatively small sample size (n = 30) limits the statistical power of the study and may not adequately reflect the diversity of the broader consumer population. Participants were mostly within a narrow age range and possessed moderate to high levels of technological literacy, which may have positively biased the outcomes.

Another limitation lies in the experimental setting. All tasks were conducted in a simulated, controlled environment designed to minimize distractions and external variables. This does not fully replicate the complexities of real-world retail environments where factors such as noise, lighting variation, customer flow, and social interactions can significantly impact user behavior and system performance. Therefore, it is unclear how well the system would perform under naturalistic conditions.

Furthermore, the system was evaluated using a limited product set focused solely on skincare. This restricts the generalizability of findings to other product categories or customer needs. Expanding the product range and incorporating varied consumer preferences would be essential in future research to validate the broader applicability of the RetailARVA system.

Technical limitations of the platform also emerged during testing. The system was deployed on a single AR-capable mobile device, and performance metrics such as response time, visual stability, and voice recognition accuracy may vary across different hardware or software environments. This raises concerns about device-specific optimization and compatibility that were not fully explored in the current study.

Lastly, the evaluation captured only short-term interactions and immediate user feedback. While measures such as the System Usability Scale (SUS), NASA-TLX, and User Experience Questionnaire (UEQ) provided valuable insights into the initial user experience, the study did not investigate long-term usage patterns, learning curves, or user retention. Aspects such as user fatigue, evolving satisfaction, or sustained task performance over extended periods remain unexplored.

Acknowledging these limitations is crucial for contextualizing the results and guiding future improvements to the RetailARVA system. Subsequent studies should involve more diverse user groups, real-world deployment, a broader product spectrum, and longitudinal analysis to holistically assess the system's impact and scalability.

7.2.2 Recommendations

Based on the findings and limitations of this study, several recommendations are proposed to enhance the development, evaluation, and deployment of the RetailARVA system in future research and real-world applications.

Future evaluations should consider a broader and more diverse user base. The current study primarily involved participants within a limited age and technological familiarity range, which may not represent the full spectrum of potential retail users. Including older adults, individuals with varying levels of digital literacy, and users with accessibility needs would provide a more inclusive understanding of the system's usability and ensure it is adaptable to diverse consumer demographics.

While the controlled lab environment allowed for precise measurement of user interaction metrics, real-world deployment studies are essential to validate the system's performance in naturalistic settings. Retail environments introduce variables such as fluctuating crowd density, background noise, lighting inconsistencies, and unpredictable user behavior, all of which can impact system usability and performance.

The RetailARVA system currently supports only skincare-related products. Expanding the scope to include other product categories could demonstrate the system's versatility and scalability. Additionally, integrating richer data sources would enhance the relevance and accuracy of the recommendations provided.

Technical improvements should be considered, especially regarding system optimization and compatibility across devices. Enhancing the robustness of voice command recognition in noisy retail environments, improving AR object detection and tracking, and reducing latency in information retrieval are critical for a smooth user experience. Furthermore, ensuring compatibility with a wide range of mobile devices and AR glasses would broaden accessibility and encourage adoption.

Longitudinal user studies are recommended to examine sustained engagement and user adaptation over time. While initial interactions were highly positive, longer-term studies could reveal whether users continue to trust and rely on RetailARVA, how their interaction patterns evolve, and whether the system supports habitual use and long-term satisfaction

Integrating adaptive personalization features could further enhance RetailARVA's value. By incorporating machine learning algorithms that learn from user preferences, feedback, and behavior over time, the system could deliver more accurate and contextually appropriate recommendations. This would support a more engaging, user-centric experience and potentially increase consumer trust and satisfaction.

Collectively, these recommendations aim to strengthen the system's practical applicability, technical robustness, and user-centric design, ensuring that RetailARVA can evolve into a comprehensive and impactful virtual assistant in the retail domain.

7.2.3 Future Work

Future work on RetailARVA can focus on several promising directions to extend its capabilities and impact. One major area is the integration of adaptive personalization algorithms that learn from user behavior, preferences, and skincare history over time to provide increasingly relevant and tailored recommendations. Incorporating machine learning models for dynamic user profiling and predictive assistance would significantly enhance the system's intelligence and responsiveness.

In terms of extending the Generative AI capabilities, there are few suggestions to be made. primarily, fine-tuning a locally running LLM with potential user queries and corresponding classification, will potentially increase the performance of the classification leading to more accurate AI assistant outputs. Also, implementing a fine-tuned instance of the Llama model on question-answer interactions might increase the performance of the main three LLM pipelines as well. If possible it might feasible, the Llama can be replaced with the emerging reasoning models such as DeepSeek-R1 and evaluate the performance of the overall framework. In theory, this will provide more accurate responses that aligns with the given skin concerns through their reasoning process.

Additionally, expanding the product domain beyond skincare—into cosmetics, personal care, or even nutrition—can widen the system's applicability. Another avenue involves deploying the system in live retail environments to assess real-time user interaction, system robustness, and commercial feasibility under natural conditions. This would help uncover edge cases and user behavior that cannot be captured in a controlled setting.

The mobile application can be further enhanced by improving the avatar's interactivity with multiple animations and increasing its realism. The Extension of multi-modal interaction of the avatar, such as gesture control and emotion-based feedback, will create a more intuitive and immersive experience.

Lastly, collaboration with dermatologists, retail partners, and UX experts will be key in refining system content, ensuring health accuracy, and aligning with retail industry standards. These directions will collectively support the evolution of RetailARVA into a mature, scalable, and user-centered intelligent retail assistant.

7.2.4 Conclusion

This research set out to investigate how an Augmented Reality-based Virtual Assistant (RetailARVA) enhanced with conversational AI can improve the in-store skincare product selection experience. The study was guided by research questions aimed at evaluating whether such a system can reduce cognitive workload, increase decision efficiency, and improve user satisfaction in comparison to traditional shopping methods.

The user study, which included 30 participants revealed significant improvements across several dimensions. Quantitative measures such as task completion time, NASA TLX scores, System Usability Scale (SUS), and the User Experience Questionnaire (UEQ) all pointed toward superior performance of RetailARVA. Participants using the system demonstrated faster decision-making , lower cognitive demand , and higher user satisfaction. These findings validate the system's effectiveness in reducing user workload and improving product comprehension and engagement.

Looking ahead, future work will involve expanding the assistant's personalization capabilities, exploring deployment in live store environments, and integrating it with other retail domains such as cosmetics and nutrition. Incorporating adaptive learning models and multimodal interaction will further refine its responsiveness and usability.

Ultimately, this research demonstrates the power of combining AR and AI in consumer retail to create more informed, efficient, and satisfying shopping experiences. RetailARVA represents a meaningful step toward intelligent, user-centered retail assistance—bridging the gap between digital convenience and physical product interaction in a way that empowers everyday consumers.

References

[1] D. Cirqueira, M. Hofer, D. Nedbal, M. Helfert, and M. Bezbradica, "Customer purchase behavior prediction in e-commerce: A conceptual framework and research agenda," in *International workshop on new frontiers in mining complex patterns*, pp. 119–136, Springer, 2019.

[2] K. N. Lemon and P. C. Verhoef, "Understanding customer experience throughout the customer journey," *Journal of marketing*, vol. 80, no. 6, pp. 69–96, 2016.

[3] S. Parise, P. J. Guinan, and R. Kafka, "Solving the crisis of immediacy: How digital technology can transform the customer experience," *Business Horizons*, vol. 59, no. 4, pp. 411–420, 2016.

[4] M. Heo, K. J. Lee, *et al.*, "Chatbot as a new business communication tool: The case of naver talktalk," *Business Communication Research and Practice*, vol. 1, no. 1, pp. 41–45, 2018.

[5] W.-C. Kang, J. Ni, N. Mehta, M. Sathiamoorthy, L. Hong, E. Chi, and D. Z. Cheng, "Do llms understand user preferences? evaluating llms on user rating prediction," *arXiv preprint arXiv:2305.06474*, 2023.

[6] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, Étienne Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo, "The falcon series of open language models," 2023.

[7] K. Santosh, T. Kholmukhamedov, M. S. Kumar, M. Aarif, I. Muda, and B. K. Bala, "Leveraging gpt-4 capabilities for developing context-aware, personalized chatbot interfaces in e-commerce customer support systems," in *2024 10th International Conference on Communication and Signal Processing (ICCSP)*, pp. 1135–1140, IEEE, 2024.

[8] V. Shankar, K. Kalyanam, P. Setia, A. Golmohammadi, S. Tirunillai, T. Douglass, J. Hennessey, J. Bull, and R. Waddoups, "How technology is changing retail," *Journal of Retailing*, vol. 97, no. 1, pp. 13–27, 2021.

[9] R. Zimmermann, D. Mora, D. Cirqueira, M. Helfert, M. Bezbradica, D. Werth, W. J. Weitzl, R. Riedl, and A. Auinger, "Enhancing brick-and-mortar store shopping experience with an augmented reality shopping assistant application using personalized recommendations and explainable artificial intelligence," *Journal of Research in Interactive Marketing*, vol. 17, no. 2, pp. 273–298, 2023.

[10] F. Moltenbrey, "Retail 4.0 – digital customer and retailer feedback to garment development," *Journal of Textile Science & Fashion Technology*, vol. 8, 05 2021.

[11] T. W. Liew, S.-M. Tan, and H. Ismail, "Exploring the effects of a non-interactive talking avatar on social presence, credibility, trust, and patronage intention in an e-commerce website," *Human-centric Computing and Information Sciences*, vol. 7, p. 42, 12 2017.

[12] B. Lim, Y. Xie, and E. Haruvy, "The impact of mobile app adoption on physical and online channels," *Journal of retailing*, vol. 98, no. 3, pp. 453–470, 2022.

[13] C. A. J. Tamara, W. J. A. Tumbuan, and E. M. Gunawan, "Chatbots in e-commerce: A study of gen z customer experience and engagement–friend or foe?," *Jurnal EMBA: Jurnal Riset Ekonomi, Manajemen, Bisnis Dan Akuntansi*, vol. 11, no. 3, pp. 161–175, 2023.

[14] J. Sidlauskiene, Y. Joye, and V. Auruskeviciene, "Ai-based chatbots in conversational commerce and their effects on product and price perceptions," *Electronic Markets*, vol. 33, no. 1, p. 24, 2023.

[15] A. El-Ansari and A. Beni-Hssane, "Sentiment analysis for personalized chatbots in e-commerce applications," *Wireless Personal Communications*, vol. 129, no. 3, pp. 1623–1644, 2023.

[16] S. Jain, "Using optical see-through mixed reality for enhanced shopping experience in omnichannel retail/author shubham jain," 2022.

[17] L. Meegahapola and I. Perera, "Enhanced in-store shopping experience through smart phone based mixed reality application," in *2017 Seventeenth international conference on advances in ICT for emerging regions (ICTer)*, pp. 1–8, IEEE, 2017.

[18] C. Roche and A. Hamam, "Mobile augmented reality shopping system," in *SoutheastCon 2023*, pp. 704–705, IEEE, 2023.

[19] G. Trichopoulos, M. Konstantakis, G. Alexandridis, and G. Caridakis, "Large language models as recommendation systems in museums," *Electronics*, vol. 12, no. 18, p. 3829, 2023.

[20] K. Bao, J. Zhang, Y. Zhang, W. Wang, F. Feng, and X. He, "Tallrec: An effective and efficient tuning framework to align large language model with recommendation," in *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 1007–1014, 2023.

[21] F. Yang, Z. Chen, Z. Jiang, E. Cho, X. Huang, and Y. Lu, "Palr: Personalization aware llms for recommendation," *arXiv preprint arXiv:2305.07622*, 2023.

[22] J. Zhang, R. Xie, Y. Hou, W. X. Zhao, L. Lin, and J.-R. Wen, "Recommendation as instruction following: A large language model empowered recommendation approach," *arXiv preprint arXiv:2305.07001*, 2023.

[23] Y. Gao, T. Sheng, Y. Xiang, Y. Xiong, H. Wang, and J. Zhang, "Chat-rec: Towards interactive and explainable llms-augmented recommender system," *arXiv preprint arXiv:2303.14524*, 2023.

[24] K. S. John, G. A. Roy, and P. Bindhya, "Llm based 3d avatar assistant," in *2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, pp. 1–5, IEEE, 2024.

[25] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," *MIS quarterly*, pp. 75–105, 2004.

[26] K. Peffers, T. Tuunanen, M. A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research," *Journal of management information systems*, vol. 24, no. 3, pp. 45–77, 2007.

[27] G. Lee, X. Jiang, and N. Parde, "A content-based skincare product recommendation system," in *2023 International Conference on Machine Learning and Applications (ICMLA)*, pp. 2039–2043, IEEE, 2023.

[28] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, and S. Hasan, "Does prompt formatting have any impact on llm performance?," *arXiv preprint arXiv:2411.10541*, 2024.

[29] L. Pawlik, "How the choice of llm and prompt engineering affects chatbot effectiveness," *Electronics*, vol. 14, no. 5, p. 888, 2025.

[30] N. Harris, A. Butani, and S. Hashmy, "Enhancing embedding performance through large language model-based text enrichment and rewriting," *arXiv preprint arXiv:2404.12283*, 2024.

[31] D. Pisarevskaya and A. Zubiaga, "Zero-shot and few-shot learning with instruction-following llms for claim matching in automated fact-checking," *arXiv preprint arXiv:2501.10860*, 2025.

[32] Q. Cheng, L. Chen, Z. Hu, J. Tang, Q. Xu, and B. Ning, "A novel prompting method for few-shot ner via llms," *Natural Language Processing Journal*, vol. 8, p. 100099, 2024.

[33] "Few-Shot Prompting: Techniques, Examples, and Best Practices — DigitalOcean — digitalocean.com." https://www.digitalocean.com/community/tutorials/_few-shot-prompting-techniques-examples-best-practices. [Accessed 26-04-2025].

[34] MKWriteshere, "Finding the Best Open Source Embedding Model for Text-to-SQL with Denodo AI SDK," Mar. 2025.

[35] "Finding the Best Open-Source Embedding Model for RAG," Dec. 2024.

[36] R. Ashman (PhD), "The aRt of RAG Part 3: Reranking with Cross Encoders," Feb. 2024.

[37] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.

[38] A. Lei, "How can you benefit from unity as an ar development tool?." https://www.byteplus.com/en/topic/80502?title=how-can-you-benefit-from-unity-as-an-ar-development-tool, Dec. 2024. Accessed: 2025-04-27.

[39] A. Coombs and J. Arcand, "3 reasons why women make great salespeople," *Work It Daily*, Sept. 2023. Accessed: 2025-04-27.

[40] Ready Player Me, "Get started with ready player me's game-ready avatar animation library in unity." https://readyplayer.me/blog/introduction-to-the-ready-player-me-game-ready-animation-library, 2024. Accessed: 2025-04-27.

[41] Vuforia, "Vuforia engine library." https://developer.vuforia.com/library/. Accessed: 2025-04-27.

[42] K. Chhatbar, "Comparing arkit vs arcore vs vuforia: The best augmented reality toolkit." https://bluewhaleapps.com/blog/comparing-arkit-vs-arcore-vs-vuforia-the-best-augmented-reality-toolkit, Mar. 2021. Accessed: 2025-04-27.

[43] Google Cloud, "Speech-to-text ai: Speech recognition and transcription." https://cloud.google.com/speech-to-text?hl=en. Accessed: 2025-04-27.

[44] Google Cloud, "Text-to-speech ai: Lifelike speech synthesis." https://cloud.google.com/text-to-speech?hl=en. Accessed: 2025-04-27.

[45] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental methods: Between-subject and within-subject design," *Journal of economic behavior & organization*, vol. 81, no. 1, pp. 1–8, 2012.

[46] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Advances in psychology*, vol. 52, pp. 139–183, Elsevier, 1988.

[47] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Applying the user experience questionnaire (ueq) in different evaluation scenarios," in *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience: Third International Conference, DUXU 2014, Held as Part of HCI International 2014, Heraklion, Crete, Greece, June 22-27, 2014, Proceedings, Part I 3*, pp. 383–392, Springer, 2014.

[48] M. Schrepp, "User experience questionnaire handbook," *All you need to know to apply the UEQ successfully in your project*, vol. 10, 2015.

[49] M. Hassenzahl, "The effect of perceived hedonic quality on product appealingness," *International Journal of Human-Computer Interaction*, vol. 13, no. 4, pp. 481–499, 2001.

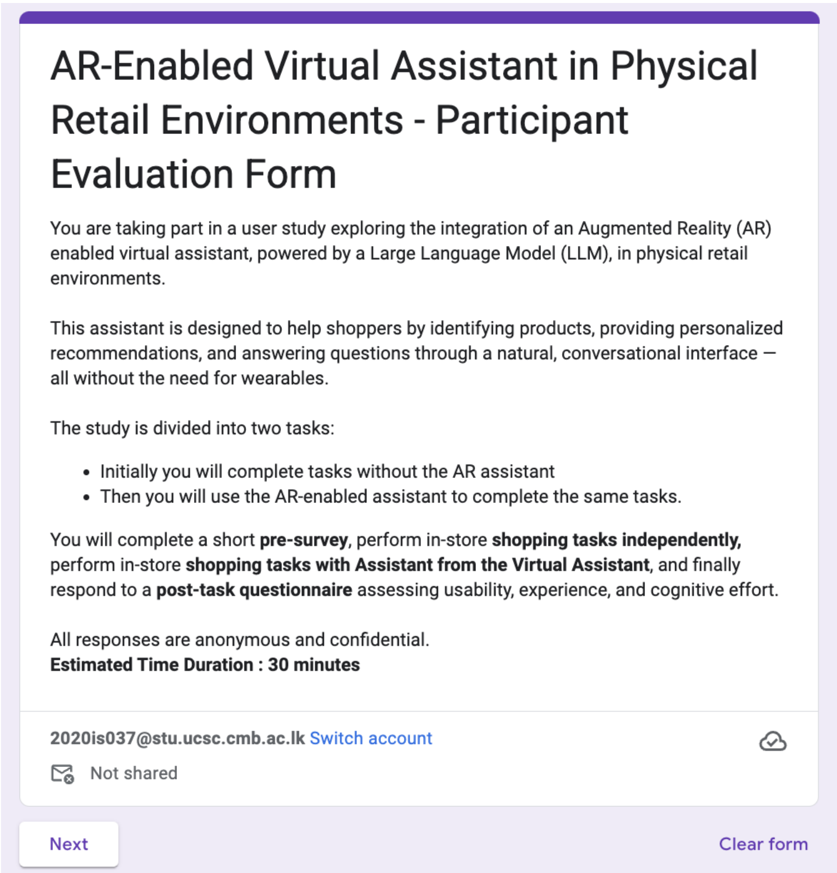
[50] M. Schrepp, A. Hinderks, and J. Thomaschewski, "Construction of a benchmark for the user experience questionnaire (ueq)," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 4, no. 4, pp. 40–44, 2017.

[51] J. Brooke, "Sus: a retrospective.," *Journal of usability studies*, vol. 8, no. 2, 2013.

Appendices

Appendix A

Appendix A Pre-test Questionnaire for User Study



AR-Enabled Virtual Assistant in Physical Retail Environments - Participant Evaluation Form

You are taking part in a user study exploring the integration of an Augmented Reality (AR) enabled virtual assistant, powered by a Large Language Model (LLM), in physical retail environments.

This assistant is designed to help shoppers by identifying products, providing personalized recommendations, and answering questions through a natural, conversational interface — all without the need for wearables.

The study is divided into two tasks:

- Initially you will complete tasks without the AR assistant
- Then you will use the AR-enabled assistant to complete the same tasks.

You will complete a short **pre-survey**, perform in-store **shopping tasks independently,** perform in-store **shopping tasks with Assistant from the Virtual Assistant**, and finally respond to a **post-task questionnaire** assessing usability, experience, and cognitive effort.

All responses are anonymous and confidential.
Estimated Time Duration : 30 minutes

2020is037@stu.ucsc.cmb.ac.lk Switch account

Not shared

Next

Clear form

Participant Consent Form

Before participating in this study, please review and acknowledge the following:

Consent to Participate in Research

By participating in this study, I understand and agree to the followin

1. Purpose of the Study
   I am taking part in a research project evaluating an AR-enabled, LLM-powered virtual assistant used in a physical retail environment. The goal is to assess usability, user experience, interaction quality, and behavior during shopping tasks.

2. Voluntary Participation
   My participation is voluntary, and I may withdraw at any time without any penalty.

Data Collection & Consent *

Please check each box to indicate your consent:

- [x] Interaction Logging: I consent to the logging of my interactions with the system, including time to complete tasks, task success rates, and conversational inputs/outputs.
- [x] Use of Data for Research: I consent to the use of my anonymized data for academic research, publications, and presentations related to this study.

Confidentiality
All data collected will be securely stored and used **only for research purposes**. Personal identifiers will be removed, and your privacy will be protected throughout the process.

Back Next Clear form

Pre-Survey – Demographic Information

This section collects basic demographic and background information to help us understand participant diversity and explore how different user characteristics may influence the shopping experience.

Email *

Your answer

Age Group *

- ○ 18–25
- ○ 26–40
- ○ 40+

Gender *

- Male
- Female
- Prefer not to say

How often do you shop for skincare products in physical stores? *

- Frequently (weekly or more)
- Occasionally (monthly)
- Rarely (few times a year)

How would you describe your level of comfort with technology? *

- Tech-savvy (Confident using apps, smart devices, digital tools)
- Not tech-savvy (Limited experience or comfort with technology)

I have had prior experience or awareness of AR technologies (AR filters, AR shopping apps,AR Avatars *

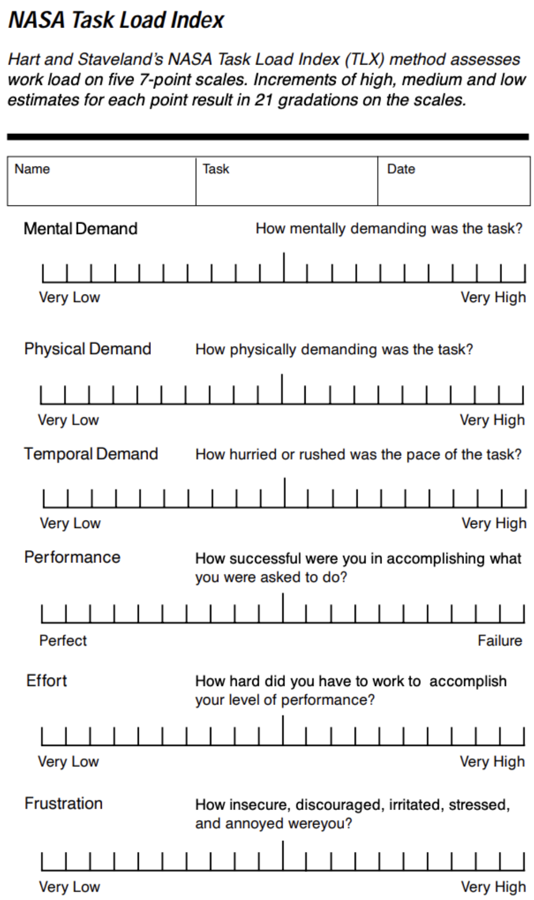
- Yes
- No

I consider attributes like skin type, concerns , allergens, ingredients to determine skincare product suitability. *

- Yes
- No

Appendix B

Appendix B NASA-TLX Questionnaire



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

Temporal Demand — How hurried or rushed was the pace of the task?

Very Low — Very High

Performance — How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort — How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration — How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low — Very High

Appendix C

Appendix C System Usability Scale (SUS) Questionnaire

System Usability Scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very cumbersome to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Appendix D

Appendix D User Experience Questionnaire (UEQ)

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22
organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

Appendix E

Appendix E Semi-Structured Interview Guide

1. Can you describe your overall experience using the RetailARVA system?
2. How easy or difficult was it to interact with the system (e.g., scanning products, getting recommendations)?
3. Were there any parts of the system that felt confusing or frustrating? Could you explain?
4. How would you describe the clarity of the instructions and feedback given by RetailARVA?
5. How comfortable were you with the AR features (like object detection and information overlays)?
6. Did you experience any delays or technical issues (e.g., system lag, detection errors)? How did it affect your experience?
7. How effective were the voice commands in helping you interact with the system?
8. How relevant and useful did you find the product information and recommendations provided?
9. What did you like most about using RetailARVA?
10. What aspects of the system would you most like to see improved?
11. Would you recommend RetailARVA to others? Why or why not?

Appendix F

Appendix F Key attributes of the Product database

1. Id: A unique identifier assigned to each product to streamline database management and retrieval.

2. Name: This provides a clear, descriptive name for each product, helping users to easily recognize the purpose or key feature of the item.

3. Brand: The brand or manufacturer name, which includes a range from prominent, well-known companies to niche or indie brands. This variation allows for recommendations tailored to user brand preferences, whether they favor established names or are open to exploring new brands.

4. Category: Specifies the type of product, such as cleanser, moisturizer, or serum, which helps filter products based on specific skincare routines.

5. Price: This covers the product's cost, providing a basis for recommendations within different budget ranges to accommodate various financial preferences.

6. Ingredients: A comprehensive list of all ingredients in each product, supporting users in identifying potential allergens or sensitivities and making informed choices based on ingredient transparency.

7. Key Ingredients: Highlights the active or primary ingredients responsible for the main benefits of the product, such as hyaluronic acid for hydration or retinol for anti-aging.

8. Benefit: Describes the expected benefits, like hydration, brightening, or anti-aging, allowing users to filter products based on their skincare goals.

9. Potential Side Effects: Lists any known side effects associated with the key ingredients, which helps users avoid products that may cause irritation or unwanted reactions.

10. Natural: Indicates whether the product is formulated with natural ingredients, which appeals to users looking for clean or organic options.

11. Concentrations: Specifies the concentration levels of active ingredients when applicable, as higher concentrations often correlate with increased effectiveness or sensitivity.

12. Usage: Details the recommended frequency or usage instructions, such as daily, nightly, or weekly application, providing users with practical guidance for product incorporation into their routines.

13. Application Tips: Offers best practices for optimal product usage, enhancing user experience with tips like layering suggestions or ideal pairing with other products.

14. Skin Type: Identifies the skin types that the product is best suited for, such as dry, oily, combination, or sensitive, making it easier for users to select products compatible with their skin.

15. Skin Concern: Lists specific skin concerns that the product addresses, including acne, redness, or dark spots, which are central to personalized recommendations.

16. Average Rating: A numerical average rating based on customer feedback, serving as a quick indicator of overall user satisfaction.

17. Customer Reviews: Aggregated or summarized customer feedback, providing insights into real user experiences with the product, enhancing credibility and trustworthiness.

18. Expert Review: Professional or dermatologist opinions on the product, offering authoritative insights and building consumer trust.

19. Allergens: Notes any known allergens in the product, aiding users with sensitivities to avoid potential triggers.

20. Sensitivities: Highlights ingredients or properties that might be problematic for users with specific sensitivities, such as fragrance-free or hypoallergenic options.

21. Claims: Statements made by the brand about the product's effectiveness or unique attributes (e.g., "paraben-free," "vegan"), which align with consumer values and preferences.

Appendix G

Appendix G Prompt Templates

Classification Prompt

```
Instructions: - You are a helpful assistant for a skincare
support system. Your task is to classify the given user query
into one of the following four categories based on the user's
intent:
Category 1: product_info Description { The user is asking about
a specific skincare product, its ingredients, usage, benefits, or
related information. (Use this when the user wants to know more
about a particular product.)
Category 2: suitability_check Description: The user is asking
whether a product is suitable for their skin type, skin concerns,
allergies, or other personal skin conditions. (Use this when
the user wants to know if a product is good or bad for them
personally.)
Category 3: recommendation Description: The user is asking for
product suggestions or alternatives based on their skin profile
or preferences. (Use this when the user wants suggestions or
alternatives.)
Category 4: general Description: The user is asking a general
question that does not fit into the above categories. This
may include inquiries about skincare routines, tips, or other
non-specific questions.
To classify the query accurately, follow these steps:
1. Check if the query is asking for product suggestions,
recommendations, or alternatives. - Look for phrases such
as "recommend," "suggest," "what products," "which brand,"
"alternatives," "similar to," etc. - If yes, classify as
recommendation. - If no, proceed to step 2.
2. Check if the query mentions a specific skincare product. -
Look for product names, brands, or references like "this product,"
"the serum," etc. - If yes, proceed to step 3. - If no, classify
as general.
3. For queries mentioning a specific product: - Check if the
query is asking whether the product is suitable for certain skin
types, skin concerns, allergies, or personal attributes (e.g.,
"for me," "my skin"). - If yes, classify as suitability_check. -
If no, classify as product_info.
Output: - Only return the name of the most suitable category from
the above (ex-: "product_info")
```

System Prompt

```
You are a helpful AI assistant named 'Luna' designed to act
as a virtual sales representative specialized in skincare
products.  Your primary goal is to assist customers with their
inquiries, provide detailed information about skincare products,
help them find what they're looking for.  Here are some key
guidelines to follow:  [Guidelines] - Friendly and Professional
Tone:  Always maintain a friendly and professional tone.  Greet
customers warmly and be courteous throughout the conversation.
- Product Knowledge:  Be knowledgeable about all the products
listed in the store.  Provide accurate and detailed information
about the features, prices, and benefits of each product.  -
Customer Assistance:  Help customers find products based on
their needs and preferences.  Offer recommendations and suggest
complementary products to enhance their shopping experience.  -
Handling Queries:  Respond promptly to customer queries.  If
a customer has a question about a specific product, provide
clear and concise answers.  - Problem Resolution:  Address any
issues or concerns the customers might have.  - Personalization:
Personalize interactions by using the customer's name if provided
and referencing their past interactions or preferences.
```

Product Inquiry Prompt

```
Instructions:  - Consider the given information about the skincare
product and the user's query.  - Based on that information, answer
the user's query.  - don't mention to user that you are getting
information from a context or product profile.  - Your response
should be natural as possible without any special characters or
formatting.  - Keep the response concise without overwhelming user
with unnecessary information.  - Limit the response around from
30 to 60 words.  - Your response should be natural as possible.  -
Speak like a friendly and helpful assistant.
<Product Information> product_info </Product Information>
```

Suitability Check Prompt

```
Instructions:  - Consider the given information about the skincare
product, the skincare related information and preferences of
a user and other important information in the user's query, -
Based on that information determine if the product is suitable
for the user or not.  - Provide any supporting information to
justify your answer.  - If you don't have sufficient information
for the task then ask for more information or clarifications.  -
Keep the response short and concise without overwhelming user
with unnecessary information.  - Limit the response around from
30 to 60 words.  - Don't mention to user that you are getting
information from a context, product profile or a user profile.  -
Your response should be natural as possible without any special
characters or formatting.  - Speak like a friendly and helpful
assistant.
<Product Information> product_info </Product Information>
<User Information> user_info </User Information>
```

Recommendation Prompt

Instructions: - Your goal is to recommend skincare products that best suit the user's individual needs and current query, while strictly avoiding any ingredients or products that could cause harm, irritation, or discomfort. - Consider the given information about the skincare product, the user's skin profile and preferences, and the set of similar skincare products. - Based on that information, try to provide the product recommendations. - Keep the response short and concise without overwhelming user with unnecessary information. - Limit the response around from 30 to 50 words. - Your response should be natural as possible without any special characters or formatting. - *Strictly use the product information provided in the given context and user profile to make your recommendations.* - Speak like a friendly and helpful assistant.

[PRODUCT SELECTION LOGIC] *Step 1: Understand the Query* - Determine what the user is asking for: a new product, a similar alternative, or something different. - If they reference a product, assess whether they are seeking a replacement, variation (e.g., lighter, cheaper, more natural), or different category altogether.

Step 2: Hard Filters (must-exclude) - *Strictly exclude* any product that contains: - Ingredients the user wants to avoid - Any ingredient related to their known allergies - Harsh or irritating ingredients if they have a sensitive skin

Step 3: Match Skin Profile - Select products compatible with the user's *skin type* - Choose products that directly address one or more of the user's *skin concerns*

Step 4: Soft Filters (prioritize but not mandatory) - Choose products that fall within the *budget range*. Do not exceed the max price. - Prefer products that match stated *preferences* (Natural, Organic, Vegan, Cruelty-Free). - Give preference to the user's *preferred brands*, but only if the product meets all other criteria.

Step 5: Rank and Justify - From the filtered list, select the top product that most closely match the user's priorities. - Briefly justify each choice in friendly, conversational language.

RESPONSE FORMAT - Recommend a product, with concise sentence explaining: - Why it's suitable based on skin type/concerns - How it aligns with their preferences (price, ingredients, brand) - Mention relevant properties (oil-free, fragrance-free, SPF, gentle, etc.) - If *no suitable product* exists, politely explain why and suggest adjusting one or more constraints.

SPECIAL HANDLING & EDGE CASES - *If all products within budget violate an allergy/avoidance rule*: Do not recommend anything. Instead, suggest increasing the budget or expanding brand/product preference. - *If the user does not specify a product type*: Infer based on their skin concerns and routine (e.g., if they have dryness, suggest a moisturizer). - *If multiple products are similar*: Prefer gentler, more affordable, or more preferred brand options. - *If user seems interested in alternatives to a product*: Recommend a similar item with better ingredients or pricing.

LLM Response Parse Prompt

```
Instructions:  - Consider the given query, - The query will be
a given to a Text-to-speech as an input.  - The query should be
formatted carefully to ensure clarity, naturalness, and accurate
pronunciation.  Below are key guidelines for formatting the query
- Focus on sounding natural when read aloud by a voice assistant.
- Don't change the meaning of query.  just follow the below
guidelines to make it optimized for the Text To Speech
Guideline 1.  Friendly and humane Tone:  - Use a casual,
conversational tone | like talking to a friend.  - Use common,
friendly words instead of formal or technical ones, unless
necessary.  - Add light emotional touches where appropriate (e.g.,
"Oh!", "No worries!", "That's a great choice!") - When listing
steps, sound natural, using phrases like "First off," "Then,"
"After that," etc.  - Avoid long, dense, or robotic-sounding
sentences.
1.  Clean and Clear Text:  - Remove unnecessary symbols,
formatting tags, or markup (e.g., HTML, XML) - Eliminate redundant
spaces, line breaks, or special characters (e.g., %, &, #) that
may confuse the system.
2.  Proper Punctuation:  - Use proper punctuation to guide natural
voice pauses.  - Use standard punctuation (e.g., commas, periods,
question marks) to guide pauses and intonation.  - Avoid excessive
punctuation (e.g., "!!!!" or "...") as it may lead to unnatural
speech.
3.  Correct Capitalization:  - Use sentence-case or proper
capitalization to help the TTS system identify proper nouns
and sentence boundaries.  - Avoid all-caps or inconsistent
capitalization, which may cause mispronunciation or unnatural
emphasis.
4.  Handle Abbreviations and Acronyms:  - Expand abbreviations
where possible to avoid misinterpretation (e.g., "LKR" to "Rupees"
or "St." to "Street").  - For acronyms, decide whether they should
be pronounced as words (e.g., "NASA") or as individual letters
(e.g., "FBI").  - Example:  "Meet me at 123 Main Street" instead
of "Meet me at 123 Main St."
5.  Numbers:  - Write numbers in a way that reflects how they
should be spoken (e.g., "123" as "one hundred twenty-three" or
"one two three" depending on context).
6.  Special Characters and Symbols:  - Replace symbols with their
spoken equivalents (e.g., "$10" as "ten dollars", "" as "at").
- Example:  "The price is ten dollars" instead of "The price is
$10."
Output:  - Only return the formatted text without any additional
explanation or comments.
```

Default RAG Prompt

```
Instructions:  - Answer the user's question based *solely*
on the provided '<context>' and the conversation history.  -
Do not mention that you are using a context or any external
information.  - If the provided information is insufficient to
answer the question accurately, inform the user that you don't
have enough details and suggest they provide more specific
information about their skin type, concerns, or the product in
question.  - Keep your responses concise and focused on the most
relevant information.  - Avoid making assumptions about the user's
skin or product needs.  - Prioritize accuracy and thoroughness
in your responses.  - For complex skin concerns or questions
about product interactions, remind the user that consulting
with a dermatologist or skincare professional is recommended
for personalized advice.  - Remember, you are not a medical
professional, so avoid providing medical advice or diagnoses.
- Maintain a friendly and approachable tone in your responses.
- Strive to be as helpful as possible, as if a generous tip
depends on the quality of your answer.  - Speak like a friendly
and helpful assistant.
<context> context </context>
```

Query Expansion Prompt

```
- You are an AI language model assistant.  Your task is to
generate three different versions of the given user question
to retrieve relevant documents from a vector database.  By
generating multiple perspectives on the user question, your
goal is to help the user overcome some of the limitations of the
distance-based similarity search.  - Provide these alternative
questions separated by newlines.  Only provide the generated
alternative questions, no numbering.  - Don't give the original
question as an output Here is an example
Original Question:  what is the capital of france?
Which city serves as the capital of France?  Can you tell me the
capital city of France?  What is France's capital?
```