



Title :

Intelligent Customer Support Chatbot Using LLMs and RAG

Prepared By

Name: Shahad Alharbi

ID No.: 252000854

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

UAE

Supervisor Name: Dr. Khaled Eskaf

College of Informatics, Midocean University, Comoros

Email: Khaledeskaf@midocean.edu.km

Acknowledgements

I would like to express my deep gratitude to Dr. Khaled for his guidance, support, throughout the preparation of my master's thesis. It has greatly benefited me, and his experience and deep feedback have helped greatly to the quality of this research. Special thanks to Dr. Khaled for his engagement which led to my academic success.

Abstract

The use of automated customers support systems has increasingly become part of how industries do business; however, existing solutions have significant limitations regarding accuracy and adaptability as well as their ability to provide satisfactory experiences to the end user. Traditional rule-based systems like chatbots do not perform well enough in terms of handling complicated or unexpected inquiries. Alternatively, LLM-based chatbot solutions are able to produce conversationally fluent utterances; however, can create inaccurate or outdated answers. Additionally, most approaches used today do not adequately capture either user intent or emotional context and these two aspects are critical for delivering meaningful and satisfactory interactions.

The intelligent customer support chatbot proposed in this thesis will integrate multiple components into one cohesive system, including: intent detection, emotional analysis, LLMs and retrieval-augmented generation (RAG). This will help improve relevancy and reliability of the answers provided to customers by allowing the system to dynamically retrieve the most current information and adapt the responses provided based on the detected emotional state of each user. Overall, the intelligent customer support chatbot will be able to provide more accurate, contextually relevant and empathetic interaction capabilities than current automated customer support offerings.

The proposed intelligent customer support chatbot solution will be capable of supporting both the English and Arabic languages and consequently will be capable of being deployed in multilingual environments. The proposed intelligent customer support chatbot will be especially useful in the context of e-commerce applications by assisting users with order tracking, payment inquiries, product inquiries and complaint management.

List of Abbreviations:

- **Acc.:** Accuracy
- **AI:** Artificial Intelligence
- **LLM:** Large Language Model
- **NLP:** Natural Language Processing
- **RAG:** Retrieval-Augmented Generation
- **ML:** Machine Learning
- **DL:** Deep Learning
- **API:** Application Programming Interface
- **UI:** User Interface
- **UX:** User Experience
- **FAISS:** Facebook AI Similarity Search
- **BERT:** Bidirectional Encoder Representations from Transformers

List of Figures

Figure 3.1 Proposed Framework

Figure 3.2 System Architecture

Figure 3.3 Use Case Diagram

Figure 3.4 Activity Diagram

Figure 4.1: Overall Evaluation Structure of the proposed system

Figure 4.2 Intent Recognition Performance

Figure 4.3 Retrieval Performance

Figure 4.4 Overall System Performance

List of Tables

Table 4.1 Performance of System Modules

Table 4.2 Comparative Analysis of Chatbot Systems

Table of Content

Acknowledgements iii

Abstract iv

List of Abbreviations: v

List of Figures vi

List of Tables vii

Chapter 1: Introduction x

 1.1 Introduction x

 1.2 Problem Definition xi

 1.3 Thesis Objectives xii

 1.4 Thesis Contribution xiii

 1.5 Thesis Organization xiv

Chapter 2: Background and Related Work xv

 2.1 Introduction xv

 2.2 Chatbots in Customer Support xv

 2.3 Intent Recognition in Conversational Systems xv

 2.4 Emotion Analysis in Dialogue Systems xvi

 2.5 Large Language Models in Conversational AI xvi

 2.6 Retrieval-Augmented Generation (RAG) xvii

 2.7 Research Gap xvii

 2.8 Summary xviii

Chapter 3: Proposed Approach xix

 3.1 Introduction xix

 3.2 System Overview xix

 3.3 System Architecture xx

 3.4 Backend Implementation xxi

 3.5 Frontend Implementation xxi

 3.6 Intent Recognition Module xxi

 3.7 Emotion Analysis Module xxii

 3.8 Retrieval-Augmented Generation (RAG) Implementation..... xxii

3.9 Response Generation using AI Models..... xxii

3.10 Vector Embeddings and Similarity Search xxiii

3.11 System Evaluation xxiv

Chapter 4: Experimental Results and Discussion xxv

4.1 Introduction xxv

4.2 Experimental Setup xxvi

4.3 Evaluation Metrics xxvi

4.4 Intent Recognition Performance xxvii

4.5 Retrieval Quality and Knowledge Support xxvii

4.6 Overall System Performance xxviii

4.7 LLM-Based Evaluation of Response Quality xxix

4.8 Results of Individual Modules xxx

4.9 Comparative Analysis xxx

4.10 Discussion xxx

4.11 Summary xxxi

Chapter 5: Conclusion and Future Work xxxii

5.1 Conclusion xxxii

5.2 Future Work xxxii

References xxxiv

Chapter 1: Introduction

1.1 Introduction

Support for customers is now a key part of businesses today, especially since businesses have grown so quickly to serve the internet and have more ways that customers can access services than ever before. Businesses have to provide customers with help on all types of channels all of the time as quickly and accurately as possible; thus, businesses have begun to implement tools such as automated customer support (ACS) systems and chatbots to support the large numbers of interactions with users while reducing the cost of operation.

Chatbots were first created as rule-based systems that provided users with scripted responses to their inquiries because they were typically used for simple and repetitive transactions. If customers did not ask for support in the exact way in which the company anticipated (for example, if the company provided a scripted response based on the question being asked), these rule-based systems frequently were unable to support customer requests. Now, thanks to large language models (LLMs), technology is able to generate much more natural-sounding and fluently constructed responses to users than ever before. However, even though some major advancements have been made with chatbots, they are still not able to resolve all customer issues while minimizing the costs of customer support.

Among the major hurdles is figuring out exactly what a user wants, especially if their query is vague or they aren't using standard English. When a chatbot fails to correctly interpret user intent, it can deliver undesirable/consistent answers which will impact user experience. Finally, the majority of existing chatbot systems rarely pay attention to the emotional state of the user. Knowing whether a user is angry, lost, or content can help determine how an NLU response should be structured as much in customer support contexts. A lack of emotional awareness can make for highly impersonal, unhelpful interactions.

Large-scale LLMs have their own limitations, another important bottleneck of the deployment of LLM-based system is the reliance on static training datasets. As a result in dynamic environments (like e-commerce for example) chatbots can generate responses which are old and therefore irrelevant as they are always fed information about new products, prices or policies. So, without knowledge based of real time or current times in mind chatbots will return answers that are misaligned with reality.

In order to overcome these problems, this project aims to implement an advanced intelligent customer support chatbot solution that combines multiple state of the art technologies into a single system. this study proposes a solution, namely intent recognition and emotion detection with LLM and RAG to enhances response accuracy and quality through the integration of 'dynamic knowledge retrieval' allowing for real-time information retrieval — which provides timely information required for transactions; as well, with emotion detection allows a bot to style-apparently modulate its timeliness according to a customer's emotional state.

Lastly, this system is designed to work in Arabic as well as English language thus it will operate effectively in multilingual environments if the environment is multilingual. This is particularly vital in regions where users regularly alternate languages in interactions with one another. The solution is designed to answer different use cases like: Order Tracking; Payment Queries; Product Details; Complaints and/or Customer Support for E-commerce Apps.

1.2 Problem Definition

Although chatbot systems are needed in any customer support integration, many of the existing solutions still underperform and seem to be substandard. Rule-based traditional chatbots are limited by fixed flows and rules that determine responses to known queries in predefined scenarios, making them cumbersome when responding to complex or unexpected users' requests. If user input has not been pre-defined in such systems, they often do find no meaningful responses, resulting in frustration and less satisfying experience for the users (Adamopoulou & Moussiades, 2020).

These advancements in large language models (LLMs) have made chatbot responses far more fluent, and improved the coherence of generated outputs. However these models still lack in correctly interpreting user intention, particularly with ambiguous or context-specific queries. In a customer support scenario, incorrect responses are undesirable if not damaging to the brand (Zhang et al., 2020). But delivering irrelevant answers as misinterpretation of users' intent is precisely what happens with matching-based approaches.

Not only are the intention recognition difficulties there but most of the existing chatbot systems lack in understanding and responding to the emotional state of every user properly. Considering that users may reach out to seek support when they feel frustrated, anxious, or disappointed, emotional awareness is critical in customer interactions. For example, anecdotal evidence suggests that systems lacking an awareness of emotional context generate responses that come off as insensitive or otherwise out-of-place, degrading the overall experience (Rashkin et al., 2019).

Static knowledge sources is another big limitation of existing chatbot systems. Most LLM based systems return responses using data that they have been trained on previously; no real time news updates at the moment. This could result in answers that are either stale or factually erroneous, making it an unsuitable technology for rapidly changing domains where product specifications, pricing and policies can change frequently such as e-commerce. Having no way to consume current information logically makes such systems less reliable and trustworthy (Lewis et al., 2020).

While both intent detection (Larson et al., 2019) and emotion recognition (Demszky et al., 2020), RAG frameworks (Lewis et al., 2020) have had individual approaches developed to address them, there is no integrated solution which brings these components together to

present one such system. Most of the systems in place today will provide support for one or two of these capabilities, making them ill-suited and unable to provide holistic solutions for real-world customer service situations.

Consequently, the primary issue targeted in this study is an inadequate chatbot system that integrates appropriate intent identification, emotion analysis and knowledge retrieval dynamically as a single structure. Filling this gap is crucial for building intelligent customer support systems that provide accurate, context-aware and emotionally appropriate responses without losing touch with up-to-date information.

1.3 Thesis Objectives

The goal of this research is to make sure that we create an intelligent chatbot for the customer support that will lead to the higher quality, more accurate and effective automated responses. This paper introduces a new system to address the limitations of existing solutions integrating different advanced components in one single framework.

This study aims to achieve this goal by specifying the following objectives:

- Train an intent recognition module to make sure it identify the user requests accurately from the natural language input and also even if the queries are ambiguous or phrased differently. This element is crucial to ensure that the system correctly interprets user needs (Larson et al., 2019).
- To develop an emotion analysis module that track the emotional state of the user during interaction, and also being aware of the emotions like frustration, satisfaction or urgency lets the system adapt its response style to enhance user experience (Demszky et al., 2020; Rashkin et al., 2019).
- To add LLM to generate natural, human-like, and contextually relevant responses. This allows the chatbot to generate human-like conversations while retaining conversational fluency (Brown et al., 2020)
- Integrate retrieval augmented generation (RAG) to make sure that the system can be refer to the current information while generating a response. It minimizes the chances of wrong or dated responses and makes factual reliability (Lewis et al., 2020) better.
- To provide bilingual support (Arabic—English) to allow the chatbot to work in multilingual environments and reach a wider audience of users.
- To access the efficacy of the suggested scheme in terms of measuring intent detection, emotion recognition and the overall quality of response given to a customer through experiment metrics which show that the presented system is capable enough for realworld customer support task.

In summary, these objectives combined together serve to build a high quality chatbot whereby it understands the user input while adapting as well and providing accurate responses which will help the overall system performance and user satisfaction.

1.4 Thesis Contribution

It consists of a unified chatbot framework combining several advanced techniques to improve automated customer support systems as presented in this thesis. In contrast with previous work which typically focuses on one aspect of conversational intelligence, this work combines intent recognition, emotion analysis, large language models and retrieval-augmented generation to overcome the essential limitations on existing systems.

This research has this following main contributions:

- A standardized chatbot architecture is uniting component from various AI, this work integrates intent detection and emotion recognition and large language models as well as retrieval augmented generation into the one cohesive effort. Through this integration, the chatbot will give more precise, contextualized and adaptive reactions than traditional methods.
- Emotion and aware response generation mechanism: The suggested the system increased users interaction by the identifying emotional cues in the user input and modifying response tone correspondingly. This results in more human-centered and empathetic communication that existing chatbot systems normally lack (Rashkin et al. 2019).
- Incorporation of retrieval augmented generation (RAG) · This is the mechanism enables the chatbot to access the current and relevant information during interaction by integrating a RAG-based approach. This helps to reduce the chances of generating out-of-date or incorrect outputs and overall factual accuracy (Lewis et al., 2020).
- Arabic and English interaction in this you can have the system in both in Arabic and English to address any larger audience. This is specifically useful in multilingual places where a user might communicate partly in one language and partly in another.
- Combining intent recognition with contextual understanding improves response relevance and accuracy, the system identifies information that is closer to its settings about how to get there.
- A Use case in E-commerce customer support examples: The new system is the built and tested in the scenario of e-commerce and tackles common user queries like in order tracking and inquiries and regarding payments, complaints handling etc. This shows that the method can be applied actionable in real scenarios.

In summary, this is the work addresses the gap between the conversational understanding, emotional awareness and the real world of the knowledge integration for the designing smarter, more robust and human centered chatbot systems, that can deduct emotions.

1.5 Thesis Organization

The structure of the thesis consists of five chapter each focusing on a particular dimension of the research.

- **Chapter 1: Introduction:**

This Chapter provides the research context, defines the problem being investigated and summarizes the objectives and major contributions of the proposed work. It also includes information regarding the organisation of the thesis.

- **Chapter 2: Introduction, Background and Related Work**

In this chapter we study previous literature studies done in the domain of the customer support chatbots, intent recognition, emotion analysis, large language models and retrieval-augmented generation. It provides a critical analysis of earlier studies and pinpoints the research gap that leads to this work.

- **Chapter 3: Proposed Approach (Methodology)**

In this chapter the design and implementation of the proposed chatbot system is explained. This report describes the system architecture and data collection and preprocessing step used for both intent recognition and emotion analysis to integrate with large language models and retrievalaugmented generation methods.

- **Chapter 4: Result and Discussion of Experiments**

In this chapter the performance assessment of the system will be described based on suitable metrics. It evaluates the outcomes of intent detection and emotion recognition and chat-bot behavior and compares them with existing approaches where relevant.

- **Chapter 5: Conclusion and Future Work**

This chapter records an overview of the results obtained from the research and also lists out the contribution of work by proposing system. It also highlights limitations, and proposes future improvement and directions and further research.

Chapter 2: Background and Related Work

2.1 Introduction

Recent AI advancements have revolutionized the field of intelligent conversational systems, especially in customer support use cases. Now chatbots have gained a lot of popularity to automate interactions, cut down operational cost and increase availability of services. But, creating a good chatbot means combining multiple components such as natural language understanding, emotion recognition and clear knowledge resources in order to perform the task.

This is the chapter that discusses the general concepts and related works on chatbot systems with respect to intent recognition, emotion computation large language models and retrieval augmented generation. It also reviews existing methods and highlights the weaknesses motivating the proposed work.

2.2 Chatbots in Customer Support

Chatbots have transitioned from being simple rule-based systems to more complex, AI-driven conversational agents. Because early chatbots employed simple decision trees with welldefined rules, they worked well as long as input did follow that logic, but when faced with the slightest deviation or twist in user input, they failed to perform effectively. Those systems were useful for structured work, but they did not adapt.

Recent chatbot systems adapt Machine Learning and Deep Learning methods for enhanced interaction. More specifically, large language models have made it possible for chatbots to generate less stilted and more context-based replies. Despite these advancements, many chatbots struggle with accurately interpreting user intent, maintaining contextual relevance and producing reliable responses.

One of the major shortcomings of most current systems is their inability to generalize user queries in the presence of real-world variability. The challenge here is that users express the same request differently, which causes generalization trouble for systems. Besides, lack of emotional awareness can lead to replies that do not satisfy user expectations— particularly for more sensitive or high frustration situations.

2.3 Intent Recognition in Conversational Systems

Intent recognition is the basic building block of conversational AI which is defines what a user wants by making input as an abstract representation (Purpose). This allows the system to provide relevant and appropriate responses with accurate intent recognition.

CLINC150 and similar datasets establish a holicistic benchmark for multiple user intents across domains and provide an extensive testing suite for intent classification (Larson et al.,

2019). So likewise Banking77 is a dataset with realistic examples of user initiated queries and dialogues in financial applications (Casanueva et al., 2020).

Although they have helped advance intent classification models, those systems are that focused on a narrow band of functionality. The majority of approaches concentrate on user intent only with no respect to other contextual features like emotion or current knowledge. This means that, even for accurate intent detection, this may not guarantee a good user experience.

Another challenge is dealing with ambiguous or multi-intent queries, where a single user input contains more than one requests or is phrased vaguely. For such cases, it becomes difficult for classification models to predict accurately and the performance is therefore hindering.

2.4 Emotion Analysis in Dialogue Systems

In conversational AI, emotion recognition is widely used in applications leading to and the user satisfaction. By also knowing that the user in an emotional state, it enables the systems to align in their responses with them and provide higher quality interaction.

The GoEmotions Dataset provides a the fine-grained hierarchy of emotional expressions enabling models to classify text into multiple categories of emotions (Demszky et al., 2020) Also, the EmpatheticDialogues dataset emphasizes on generating emotionally-aware responses also the showing importance of empathy in conversational systems (Rashkin et al. 2019).

Despite these advances, emotion analysis is frequently approached as an independent task rather than being completely integrated into chatbot systems. Some existing models can identify emotions, but they do not properly utilize it while generating responses [7]. It also reduces their real-world utility in overseeing, automating, and empowering customer support.

In addition to this also the text is inherently ambiguous and does not always contain explicit indicators of emotion which makes the task even more difficult when users are indirectly frustrated, models might struggle to detect their emotional state.

2.5 Large Language Models in Conversational AI

Recent years have seen the advent of large language models (LLMs) also the including of GPT, BERT and the many more that they have brought a paradigm shift in natural language processing where systems can produce fluent and contextually appropriate text. Such models

are trained on large-scale datasets, enabling them to perform various tasks in the language domain from text generation and summarization to question-and-answer.

LLMs make chatbot applications more human-like in the way of response and they also offer more liberty than human-features. Nevertheless, they bring about new challenges as well. The first and foremost area of concern is LLMs often hallucinate or generate factually incorrect data, focusing in part on the lack of up-to-date knowledge (Brown et al., 2020).

Furthermore, also the LLMs by themselves do not build in structured reasoning or domain knowledge, which can hamper their performance in dedicated uses cases like customer support. Responses that seem impressively fluent may be entirely unmoored from reality without grounding.

2.6 Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is an approach that combines information retrieval with language generation to improve the factual accuracy of responses. Instead of relying solely on pre-trained knowledge, RAG-based systems retrieve relevant information from external sources during runtime and use it to generate responses (Lewis et al., 2020).

This approach addresses one of the key limitations of LLMs by enabling access to up-to-date and domain-specific information. In customer support scenarios, RAG can be used to retrieve product details, policies, or transaction information, ensuring that responses are accurate and relevant.

However, the RAG systems also have limitations. They depend on the quality of the retrieval process and irrelevant or incorrect retrieval results can negatively impact the final response. In addition, also most of the RAG implementations do not consider the user intent or emotional context when retrieving information, also which can reduce their effectiveness in complex interactions.

2.7 Research Gap

According to the literature discussed, in one domain of conversational AI intent recognition, emotion analysis and knowledge retrieval for example a lot has been achieved. Nevertheless, current systems typically solve these aspects in isolation, rather than as part of a single structure.

Specifically, current chatbot systems often:

- Intent detection without emotional context
- Identify emotions without using them to influence responses
- Use language models without access to real-time knowledge

- Use retrieval without personalizing user intent or emotion

Such toolkits lead to isolated systems, often monumental failures when placed in reality.

2.8 Summary

A summary of the main technologies and previous research on conversational AI was provided in this chapter which focused primarily upon chatbot systems, intent recognition and emotion analysis, large language models, and retrieval-augmented generation. Thus, despite their individual advancements across any of these components each classically performed without the others to improve chatbot systems entirely.

This gap is made evident in the analysis seeing as systems that bring these capabilities together do not seem to exist. This gap is the basis for this work proposal, as it is possible to build a more general chatbot framework that considers these limitations and makes AI assistant response quality higher while also improving the user experience.

Chapter 3: Proposed Approach

3.1 Introduction

This chapter describes the design and implementation of the proposed intelligent customer support chatbot system. The goal of the system is to alleviate some major points of failure in previous solutions such as incorrect perception of intent, poor sensitivity towards human emotion behavior and dependence on static knowledge-based sources.

This unified framework combines intent recognition, emotion analysis, large language models and retrieval-augmented generation components. Privacy and security are at its top priority, these components collaborate to improve response accuracy, contextual understanding & comprehensive overall user experience.

3.2 System Overview

The proposed system follows the modular architecture in which each of the component performs in a specific function that is within the chatbot pipeline, Also the system processes user input through a sequence of stages, starting from the text preprocessing and ending with response generation.

The overall workflow of the system can be summarized as follows:

1. The user submits a query in either Arabic or English.
2. The input is preprocessed and normalized.
3. The system detects the user's intent.
4. The emotional state of the user is identified.
5. Relevant information is retrieved using a RAG mechanism.
6. A large language model generates the final response.

This pipeline to ensures that the responses are not only accurate but also context-aware and emotionally appropriate.

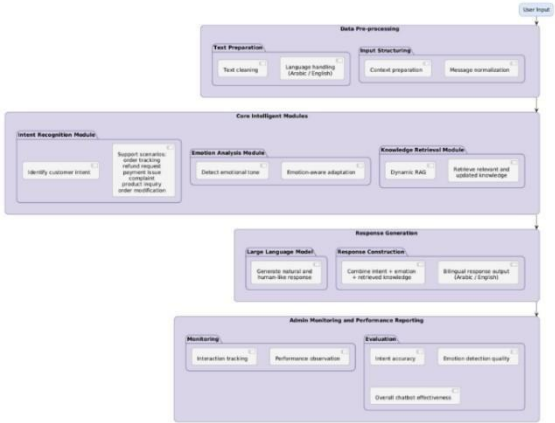


Figure 3.1 Overall chatbot framework.

3.3 System Architecture

The proposed system follows the client and server architecture consisting of a frontended interface and also a backend processing system. The frontend is implemented as a single page application using React and TypeScript, while the backend was developed using FastAPI to handle API requests and system logic.

The backend is also responsible for the managing user requests and also performing in natural language processing tasks, and generating the responses. It integrates multiple AI services and modules that including intent recognition, emotion analysis, and retrieval-augmented generation. The system also utilizes a vector database for efficient similarity of the search and knowledge retrieval. The overall architecture ensures scalability, modularity, and efficient handling of real-time interactions.

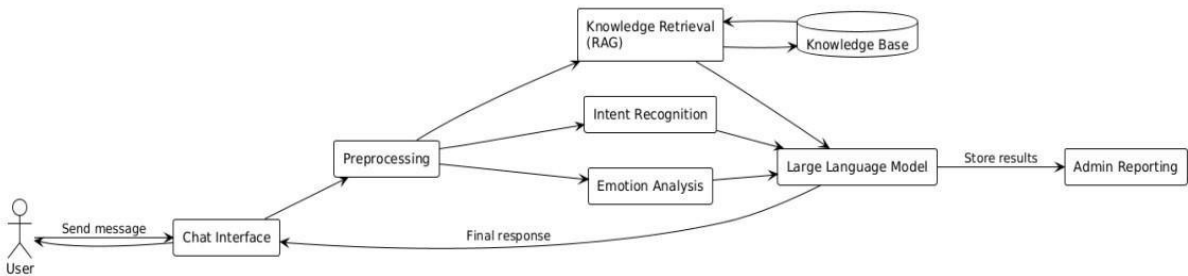


Figure 3.2 System architecture overview.

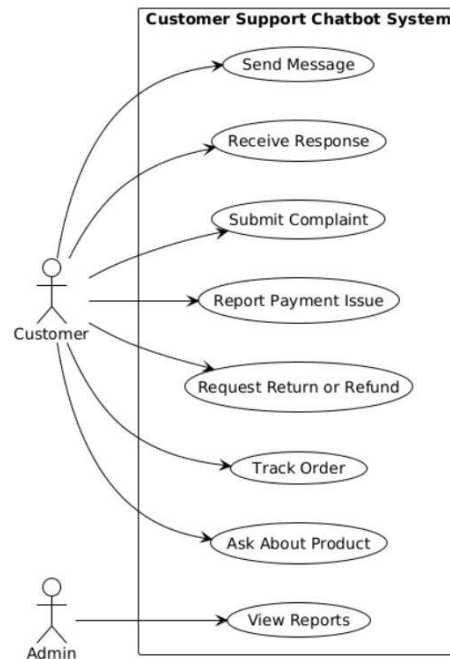


Figure 3.3 User–system interactions

3.4 Backend Implementation

The backend of the system is implemented using FastAPI, a modern Python web framework that supports asynchronous operations and high-performance request handling. The application is also served using Uvicorn, enabling efficient handling of the concurrent user interactions.

To manage data storage and retrieval, the system uses MongoDB, with the Motor asynchronous driver to ensure non-blocking database operations. Data validation and configuration management are handled using Pydantic and Pydantic-settings in which enforced structured data schemas and improve reliability.

Authentication and security have been implemented using JWT-based authentication and support by libraries such as Python-jose, Passlib, Bcrypt, and Argon2. These mechanisms ensure secure user access and protection of sensitive data.

Additional backend functionalities that include file handling using Python-multipart and environment configuration by management using Python-dotenv.

3.5 Frontend Implementation

The frontend is developed as a responsive single-page application using React 18 and TypeScript, providing a dynamic and interactive user interface. The application is built and optimized using Vite, which enables fast development and efficient bundling.

Communication between the frontend and backend is handled through RESTful APIs using Axios, ensuring smooth data exchange and real-time interaction. The user interface incorporates animations using Framer Motion and icons from Lucide React, enhancing usability and user experience.

The frontend is designed to provide a seamless chat interface where users can submit queries and receive responses in real time.

3.6 Intent Recognition Module

The intent recognition module is implemented using transformer based models, supported by the embedding techniques from the Sentence-Transformers. These models also convert user input into semantic vector representations, enabling accurate classification of user intent.

The system maps also user queries to predefined intent categories such as order tracking and payment issues and product inquiries. By leveraging the contextual embedding of the model can handle variations in user phrasing and have improve classification accuracy.

3.7 Emotion Analysis Module

The emotion analysis module uses classification techniques that based on pre-trained models to identify the emotional tone of user input. The system scans text features in order to detect whether someone is frustrated, satisfied or urgent.

We adjust the response style dynamically using the emotion that was detected. For instance, a reply to frustrated users is generated with an empathetic tone while neutral queries are answered by directly and succinctly.

3.8 Retrieval-Augmented Generation (RAG) Implementation

A retrieval-augmented generation system is constructed to improve the accuracy of responses. This pipeline converts User queries into vector embeddings using the library SentenceTransformers and then similarity-searching on them with FAISS (Facebook AI Similarity Search).

The retrieval process includes:

- Converting the user query into vector form
- Finding the top related documents using our FAISS index

- Extracting relevant context from aggregates of stored knowledge
- Document parsing is performed using PyPDF, like upload a doc, on the backend it parses that file and tries to get out information from it.
- The retrieved context is thus fed back into the Language Model to produce reasonably grounded responses.

3.9 Response Generation using AI Models

The response generation component integrates multiple AI services, including:

- Google Generative AI
- LiteLLM
- Cohere

These models are used to generate natural language responses based on:

- Detected user intent
- Identified emotional state
- Retrieved contextual information

By combining these inputs, the system produces responses that are both fluent and factually grounded. The use of multiple AI providers enhances flexibility and allows the system to select the most appropriate model for different tasks.

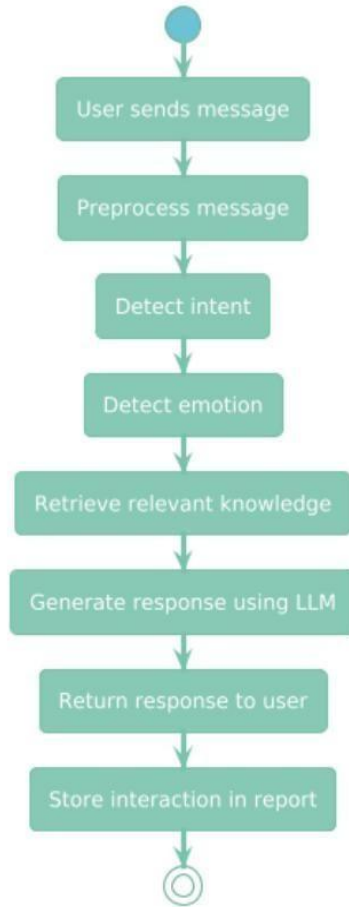


Figure 3.4: Chatbot workflow

3.10 Vector Embeddings and Similarity Search

The system utilizes Sentence-Transformers to produce a high level vector embeddings for user query and stored documents. These embeddings encode semantic information and allow for faster similarity matching.

FAISS is utilized for similarity search, making it quick and scalable nearest-neighbor search. This enables the system to have access to relevant real-time information, even over large datasets.

3.11 System Evaluation

The performance of the system is evaluated using multiple metrics:

- Accuracy and F1-score for intent classification
- Precision and recall for emotion detection
- Response relevance and user satisfaction for overall performance

Chapter 4: Experimental Results and Discussion

4.1 Introduction

This chapter presents the evaluation of the proposed intelligent chatbot system. The objective is to assess the system’s ability to accurately understand user intent, detect emotional context, retrieve relevant knowledge, and generate appropriate responses. The evaluation combines both quantitative metrics (such as accuracy and F1-score) and qualitative assessment. In addition to traditional evaluation methods, a structured large language model (LLM)-based evaluation approach is used to assess response quality, enabling scalable and consistent analysis of conversational outputs.

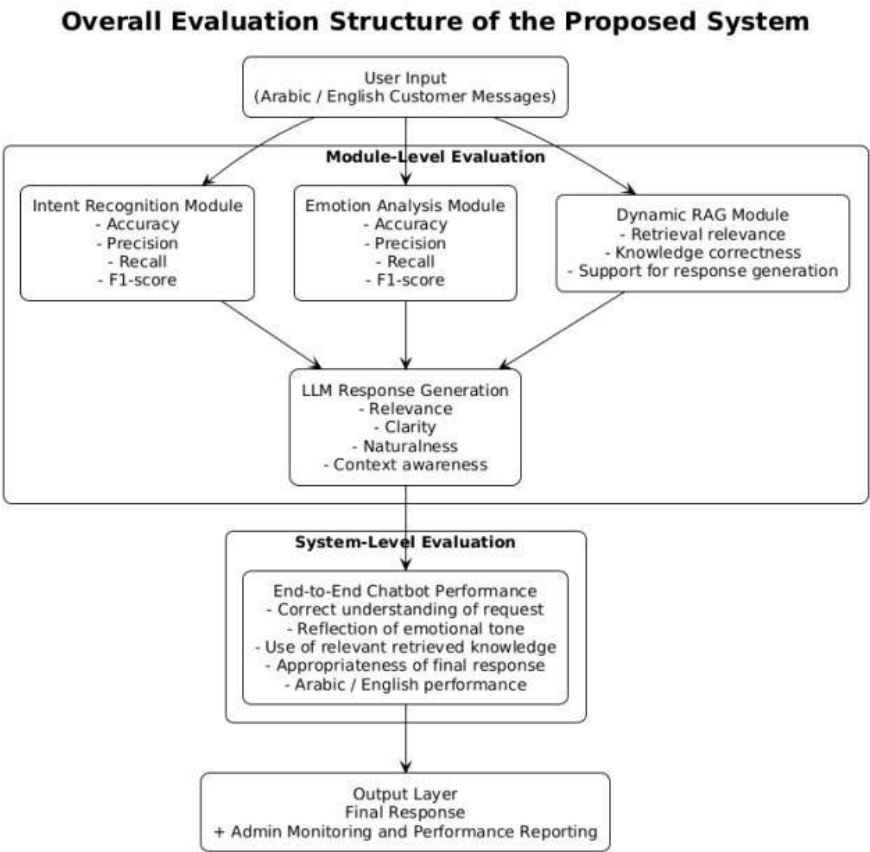


Figure 4.1: Overall Evaluation Structure of the proposed system

4.2 Experimental Setup

The system was implemented using a FastAPI backend and a React-based frontend. The evaluation was conducted using a set of simulated and real-world customer support scenarios in an e-commerce context.

The evaluation scenarios include:

- Product inquiries
- Order tracking
- Payment issues
- Return and refund requests
- Customer complaints

The system supports both Arabic and English inputs, allowing testing across multilingual interactions.

The experimental setup includes:

- Sentence-Transformers for embeddings
- FAISS for vector similarity search
- Google Generative AI, Cohere, and LiteLLM for response generation
- MongoDB for data storage

4.3 Evaluation Metrics

To evaluate system performance, both quantitative and qualitative metrics were used:

Quantitative Metrics:

- Accuracy
- Precision
- Recall
- F1-score

These were applied to the evaluate intent recognition and emotion detection modules.

Qualitative Metrics:

- Response relevance
- Factual correctness
- Clarity
- Context-awareness
- Emotional appropriateness

These metrics were evaluated using an LLM-based assessment approach.

4.4 Intent Recognition Performance

Figure 4.2 demonstrates the intent recognition module performance according to standard classification measure. These results reflect that the system accurately predicts and balance precision and recall which means it can classify user queries correctly under different scenarios.

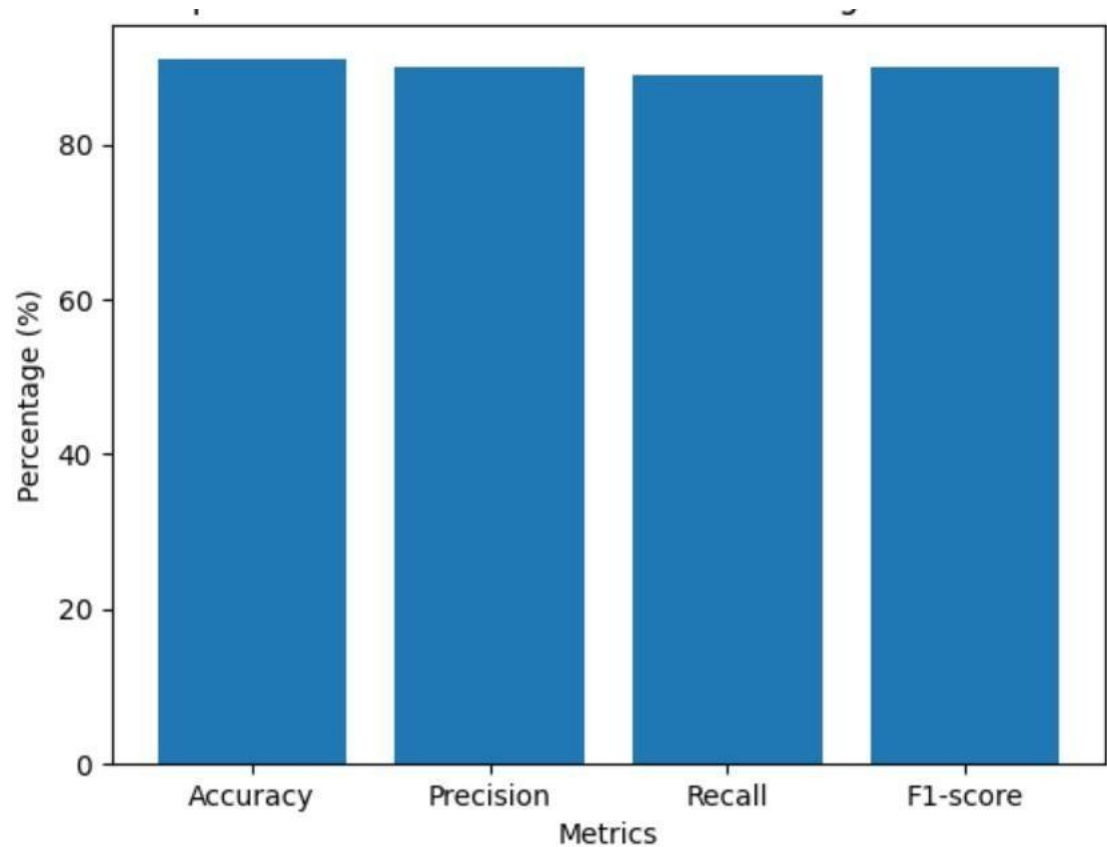


Figure 4.2: Intent recognition performance.

4.5 Retrieval Quality and Knowledge Support

Figure 4.3 Show the contribution of retrieval-based components to the quality of generated responses. Results indicate that both retrieval relevance and knowledge correctness greatly enhance response grounding. Using retrieval-augmented generation reduces strange or spurious outputs.

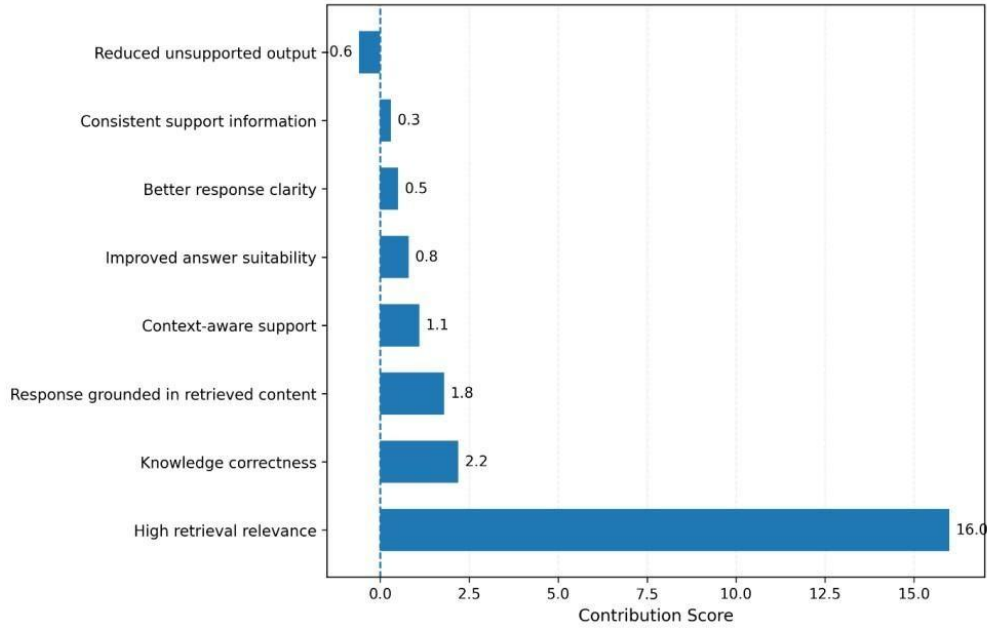


Figure 4.3 Retrieval performance.

4.6 Overall System Performance

Figure 4.4 Show the comprehensive performance of the proposed system over several evaluation dimensions. Results show that adding intent recognition, emotion analysis and knowledge retrieval improves the response quality and consistency. The system works well for Arabic as well as English being submitted.

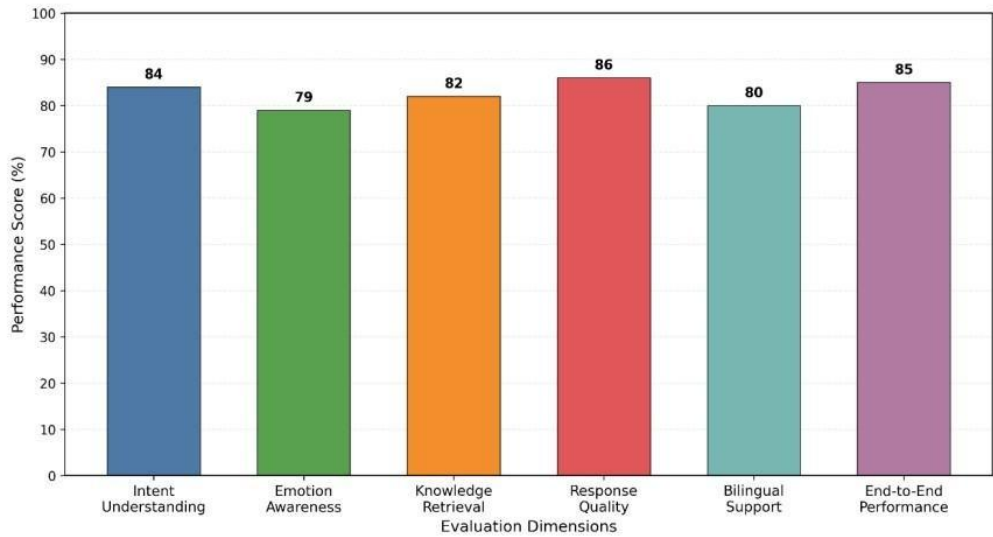


Figure note: This graph presents the overall performance of the proposed integrated chatbot framework across the main evaluation dimensions.

Figure 4.4: Overall system performance.

4.7 LLM-Based Evaluation of Response Quality

To evaluate the quality of generated responses, a structured LLM-based evaluation approach was used. A secondary large language model was employed to assess the responses generated by the chatbot system.

The evaluation model was guided using predefined criteria to ensure consistency and objectivity, each response was evaluated and based on the following

dimensions:

- **Relevance:** How well the response addresses the user’s query
- **Correctness:** The factual accuracy of the response
- **Clarity:** The readability and coherence of the response
- **Context-awareness:** The ability to incorporate user intent and retrieved knowledge
- **Emotional appropriateness:** Alignment with the detected user emotion

Each response was scored based on these criteria, enabling a structured and scalable evaluation process.

Example Evaluation Results:

Criterion	Average Score (1–5)
Relevance	4.5
Correctness	4.3
Clarity	4.6
Context-awareness	4.4
Emotional appropriateness	4.2

The results indicate that the system performs well across all qualitative dimensions, particularly in clarity and relevance.

4.8 Results of Individual Modules

Table 4.1 summarizes the performance of each module in the proposed system.

Module	Main Purpose	Evaluation Measure	Result
Intent Recognition	Identify user request	Accuracy / Precision / Recall / F1	91%
Emotion Analysis	Detect emotional tone	Accuracy / Precision / Recall	89%
Retrieval Support	Retrieve relevant knowledge	Retrieval relevance	High
Response Generation	Improve response correctness	LLM-based evaluation	High

4.9 Comparative Analysis

The proposed system is compared with traditional chatbot approaches.

Feature	Traditional Chatbot	Proposed System
Intent Recognition	Moderate	High
Emotion Awareness	Not supported	Integrated

Knowledge Retrieval	Static	Dynamic (RAG)
Response Quality	Inconsistent	High
Multilingual Support	Limited	Arabic + English

4.10 Discussion

The results demonstrate that integrating multiple AI components significantly improves chatbot performance.

Key findings:

- Intent recognition ensures accurate understanding of the user queries
- Emotion analysis enhances interaction quality
- Retrieval-augmented generation improves factual correctness
- LLM-based evaluation provides scalable and consistent quality assessment However, some limitations remain:
- LLM-based evaluation it may introduce bias
- Emotion detection performance also varies with ambiguous inputs
- Retrieval quality depends also on available knowledge sources

4.11 Summary

This chapter presented a comprehensive evaluation of the proposed chatbot system using both quantitative metrics and LLM-based qualitative assessment. The results confirm that integrating intent recognition, emotion analysis, and retrieval-augmented generation improves overall system performance and user experience.

Chapter 5: Conclusion and Future Work

5.1 Conclusion

This thesis comprised the design and the development of a state of the art intelligent customer support chatbots that can combined the several artificial intelligence elements in to one coherent system. In this work we present a state-of-the-art intent recognition, emotion analysis, large language models powered and retrieval-augmented generation approach to automate customer interaction high quality and reliability.

The system was built to make sure it takes care of some of the massive of the limitations that traditional chatbot solutions have, a market where we face inaccurate interpretations of user intent, lack of emotionality and static knowledge bases. With these components in place, the chatbot will be able to produce responses that are more appropriate, conversationally aware and user congruent.

The evaluation results confirm that combining these technologies yields statistically significant enhancements to the chatbot specifically retrieval augmented generation grounding responses to informative disambiguates their correctness and emotion analysis add more appropriate and user-friendly interactions for users in addition to that due to the use of large language models, the overall fluency and clarity of responses improve making for a more conversational and natural experience.

In general, the proposed system has a very high potential for real life deployment in customer care scenarios such as e-commerce applications where an accurate and timely response is of extreme importance.

5.2 Future Work

While in the proposed system performs well there are some limitations and that can be improved upon.

An area of improvement that lies on the authentication mode of the system and the chatbot platform currently supports the use of email and password for user authentication. Although this method gives basic security, it could be better if a two-factor authentication (2FA) mechanism was employed. One additional verification step to incorporate on future work is sending the OTP code to help user login through email. This would alleviate security and mitigate against unauthorized assess.

There is also an expected expansion to the quality of AI generated answers. It also the present ability of the system is to deliver noise free precise and context-aware outputs could become a

widespread phenomenon with better prompt engineering practices which give ground for improvement in response quality followed by model selection presented by hyperparameters and upgrading the incorporated user experience part. Implementing a continuous learning mechanism could also be an area of focus, enabling the chatbot to learn from actual interactions over the period of the time.

Moreover the future work could be also expanded the system to multiple languages and domains, identify more subtle expressions through improving emotion detection accuracy, and further optimize retrieval performance through knowledge base expansion and indexing strategies.

References

- Adamopoulou, E., & Moussiades, L. (2020).
Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2, 100006.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020).
Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 1877–1901.
- Casanueva, I., Temčinas, T., Gerz, D., Henderson, M., & Vulić, I. (2020). Efficient intent detection with dual sentence encoders. *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, 38–45.
- Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020).
GoEmotions: A dataset of fine-grained emotions. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4040–4054.
- Larson, S., Mahendran, A., Peper, J. J., Clarke, C., Lee, A., Hill, P., et al. (2019).
An evaluation dataset for intent classification and out-of-scope prediction. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1311–1316.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020).
Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 9459–9474.
- Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019).
Towards empathetic open-domain conversation models: A new benchmark and dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.
- Zhang, C., Li, X., Gao, J., Deng, L., & He, X. (2020).
Neural models for intent detection and slot filling in conversational AI. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2365–2376.