Fine-Tuning OpenAI GPT Chatbot in Western Saudi Dialect: A Case Study of Taibah University

Maimounah Alhujaili, Ruqayya Abdulrahman
College of Computer Science and Engineering, Taibah University, Al-Madinah, Saudi Arabia

***Abstract*—The current era is characterized by technological advancement and innovation, which affect various sectors. Numerous remarkable and alluring computer programs and applications have surfaced, including ones that aim to replicate human behavior. A chatbot is an example of an Artificial Intelligence (AI) computer program that uses natural language to mimic human conversations in voice or content. Even though a lack of Arabic chatbots, most of these chatbots use Modern Standard Arabic instead of Arabic dialects. This research presents the development and evaluation of a chatbot designed to respond to academic inquiries from university students using the Western Saudi dialect. A traditional Support Vector Machine (SVM) baseline model was first implemented to establish a reference point for performance. Subsequently, a fine-tuned version of Generative Pre-trained Transformer (GPT) 3.5-Turbo-0125 was developed using a culturally specific system prompt to enhance the model's understanding of regional language and academic contexts. Evaluation was conducted through a multi-dimensional framework combining human assessments, BERTScore semantic similarity measurements, and GPT-4-based automatic judging. With human assessors determining that 85% of GPT-3.5's replies to 132 messages of test data were appropriate, the transformer-based model clearly outperformed the SVM baseline, which had an accuracy of 42.86% on 20 messages of test data. These findings highlight the importance of cultural and contextual fine-tuning in building effective conversational agents for dialectal Arabic communities. The research contributes to the growing field of localized AI by demonstrating how advanced language models can be adapted to serve specialized linguistic and academic needs.**

Keywords*—*Artificial intelligence (AI); large language model (LLM); generative pre-trained transformer (GPT); Modern Standard Arabic (MSA); Western Saudi dialect

I. Introduction

In contemporary society, Artificial Intelligence (AI) influences daily life by contributing to the development and assessment of advanced applications. AI helps people to automate various tasks in multiple fields. One of the most exciting AI applications that have become widespread is a chatbot. A chatbot is a popular term for conversational agent. It is software that mimics human conversations in voice or content mode by reacting based on contributions made. Also, it is the AI to engage in conversation and interaction with users in natural language that simulates human behavior. The utilization of chatbot is increasing in different fields with different languages. The main goal of utilizing chatbots are to simulate interpersonal interactions [1].

Based on researches [2], the history of chatbot started in October of 1950. Alan Turing proposed the first idea of a chatbot. Alan Turing questioned if a computer program could converse with a group of humans without their recognizing that their conversation partner was artificial. Many people believe that this question, known as the "Turing test", is the one that inspired the creation of chatbots. Chatbots have improved over the years, and the first utilization of AI with a chatbot was in 1988 for the Jabberwacky chatbot. A growing increase in the use of chatbots and the enhancement of AI techniques used with chatbots was observed.

Chatbots use The International Organization of Legal Metrology (OIML) to represent knowledge. Chatbots can replace humans in repetitive tasks like customer service and e-help desk answering and providing prompt responses. Additionally, chatbots are utilized by operating systems as virtual assistants, such as Siri on Macs and Cortana on Windows. These days, chatbots are becoming more and more popular in the business world because of their enormous potential to automate customer care and reduce employee time [3].

A chatbot could be implemented in any human language. Unlike the abundance of studies on English chatbots, there is a lack of studies on Arabic chatbots because of the language's challenges [4]. Arabic is obviously different from languages like English in several aspects. For Instance, the Arabic alphabet consists of 28 primary characters, thirteen of them contain dots like (خ (Kha) - ج (Jeem) - ت (Ta) - ف (Fa)), while fifteen do not, such as (ح (Hha) - م (Meem) - ع (Ayn) - د (Dal)), and it written and read from right to left.

The difficulties in Arabic languages negatively effect on applying Natural Language Processing (NLP) on Arabic languages [5]. Even though the Arabic research on applying Arabic language on NLP technique has increased in recent years, the utilization of Arabic dialect is very rare. Most of Arabic natives communicate with dialect Arabic instead of Modern Standard Arabic (MSA). Also, some of them faced problems when they expressed themselves by MSA. For those reasons, the Arabic dialect chatbot will be more helpful than MSA chatbot.

A. Problem Identification

The development of chatbots that can understand and respond to regional dialects has been an emerging field in NLP and machine learning. However, despite significant advances in multilingual NLP models and transformer-based architectures (such as GPT models), there remains a notable gap in the application of these technologies to specific regional dialects, particularly within the Saudi Arabian context.

Existing research in chatbot development has largely focused on general-purpose language models, or models tailored to widely spoken dialects and languages, such as MSA. However, there is a lack of focus on specific regional dialects like the Western Saudi dialect, which is rich in local expressions, cultural nuances, and vocabulary unique to regions such as Madinah, Makkah, Jeddah, and Taif. Most existing chatbots fail to address the linguistic and cultural diversity found in these regions, resulting in suboptimal performance when handling local dialects. While some studies have explored Arabic dialectal chatbot systems, there is limited work that specifically targets the Western Saudi dialect, especially in the context of university students.

This research is necessary, as it addresses a significant gap in the existing literature by developing a chatbot specifically tailored to understand and respond to the Western Saudi dialect, which has been underrepresented in the field of NLP. While there has been substantial work on general-purpose chatbots and models designed for widely spoken dialects or MSA, few studies have focused on region-specific dialects like the Western Saudi dialect, especially in the context of university students. By developing a chatbot that can respond to questions unique to Taibah University, including course registration, academic regulations, and campus information, this research seeks to give students a useful solution. This project will not only fill the gap in dialect-specific chatbot development but also enhance institutional efficiency and student assistance by relieving staff workloads and providing students with quicker access to information.

B. Research Questions

This research is structured around three primary research questions that aim to explore the role of Arabic dialects—specifically the Western Saudi dialect—in the development and evaluation of intelligent chatbot systems: 1) How does the Western Saudi dialect influence the performance of a Support Vector Machine (SVM)? This question investigates the effect of incorporating dialectal variations on the accuracy and efficiency of traditional machine learning models. 2) What is the difference in performance between a chatbot utilizing an SVM model and a chatbot based on a GPT model? The comparison is supposed to assess the merits and shortcomings of both methods in comprehending and generating dialectical Arabic responses. 3) How can a GPT-based model be optimized to enhance academic support for Saudi university students? The final question focuses on adapting and fine-tuning large language models to provide more effective and contextually appropriate assistance in educational settings.

C. Research Objective

The research aims to implement a Western Saudi dialect chatbot that understands the users and responses effectively with same Western Saudi dialect. The chatbot will communicate with users in the educational field. This research will fill the gap in Arabic dialects research area and especially the Saudi dialect.

The remainder of this research is organized as follows: Section II provides the background, including the definition of chatbots, chatbot methodologies, the Transformer framework, and a history of chatbot evolution. Section III outlines the literature review, with a particular focus on the most relevant prior work in the field. While Section IV discusses the research methodology used in this research. Section V describes the implementation and experimentation settings, including the development of the emails retrieval-based model, the baseline model, and the GPT-based model. Section VI reports the experimental results, and Section VII discusses the key findings and their implications. Finally, Section VIII concludes the study and outlines potential directions for future research.

II. Background

Based on the dictionary, the definition of chatbot is ‘‘A computer program designed to simulate conversation with human users, especially over the Internet’’ [6]. A chatbot may uses Sentiment Analysis (SA) and NLP to communicate with humans or other chatbots in different languages by text or voice. It has multiple terms for interactive agents, for instance, artificial, smart bots, conversation entities, and digital assistants. In recent years, AI chatbots have become exceptionally popular in replacing human responses [7].

A. Chatbot Approaches

Chatbot approaches are classified into two main categories: rule-based chatbot and Machine Learning (ML) chatbot [8]. Rule-based chatbots apply pattern-matching algorithms to compare user input to a rule pattern and choose a prepared response from a list of options. The choice of rules and the answer format may also be influenced by the context. The automatic, repetitive, and lack of originality and spontaneity in the responses is a drawback of the pattern-matching approach. However, because a more thorough syntactic or semantic analysis of the input text is not carried out, there is a quick reaction time. The first chatbot ELIZA is based on this approach. However, these Chatbots are not able to respond to patterns that do not fit the prewritten script[9].

On the other hand, ML approach divided into retrieval-based chatbot and generative chatbot [10]. Retrieval-based chatbot works by comparing user input to a predetermined list of answers stored in a knowledge base. After understanding user intents and queries using NLP techniques, they obtain the best pre-written response from their database and display it to the user. In contrast, large text datasets from books, journals, the internet, and other sources are used to train generative chatbots. They process and produce text using Deep Learning (DL) models and NLP. The chatbot preprocesses and tokenizes information, dividing it into smaller pieces known as tokens, before a user submits a message. After creating an initial representation of the user's message using these tokens, it predicts the subsequent words or tokens based on its training data and acquired linguistic patterns to produce a response.

B. Transformer in Chatbot

In 2017 [11], the authors presented the transformer, a deep learning architecture that uses self-attention processes to process sequential input. Transformers enable simultaneous processing of input sequences, in contrast to conventional Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs), which process sequences step by step. Modern Large Language Models (LLMs) like Generative Pre-trained Transformer (GPT) and BERT are built on top of this design. The encoder and decoder, the two primary components of the

Transformer, are made up of several identical layers that allow for effective learning and representation of sequential input [12].

Processing the input sequence and producing meaningful representations are the encoder's responsibilities. It is made up of several identical layers, usually six in the original transformer, that carry out the following tasks. The input embedding Layer converts words into dense vector representations (word embeddings). Then, it captures semantic meaning but does not retain positional information. Multi-head attention refers to the simultaneous operation of several attention mechanisms that record various word relationships. This layer applies softmax after using a triangle mask to set future words' attention scores to negative infinity. The Multi-head self-attention layer enables each word to attend to all other words in the sentence. The mechanism works by computing three vectors for each word. Query (Q) represents what the word is looking for, Key (K) represents the content of all words, and Value (V) contains the actual information to be transferred. These are used in the scaled dot-product attention equation in the Eq. (1) [11]:

$$Attention\,(Q,K,V) = softmax\,\left(\frac{Qk^t}{\sqrt{d_k}}\right) V \qquad (1)$$

Fig. 1 illustrates the internal structure of the multi-head attention mechanism used in transformer models. It begins by projecting the input into Q, K, and V vectors through separate linear transformations. Multiple attention heads are applied in parallel, each computing scaled dot-product attention independently. The outputs from all heads are then concatenated and passed through a final linear layer, allowing the model to capture a richer representation of the input by attending to information from different subspaces.

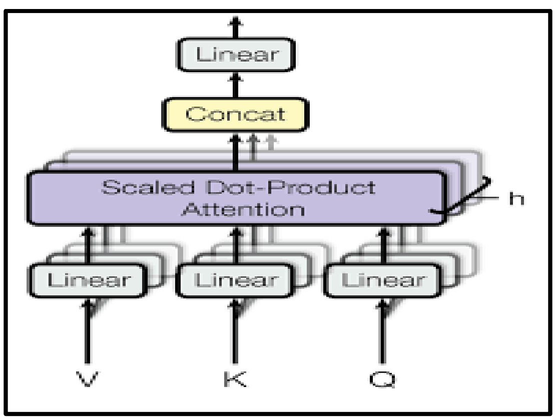

Fig. 1. Multi-head attention consists of multiple attention layers operating in parallel [11].

Layer normalization and residual connection improves training stability and gradient flow. Each word representation from the self-attention layer is passed through a feed-forward network, consisting of two linear transformations with a non-linearity in between (typically ReLU). This layer allows the model to learn complex patterns beyond attention-based relationships. The decoder receives a contextualized word embedding as the output from the final encoder layer [13].

On the other hand, the decoder is responsible for generating the output sequence based on the encoder's output. Each decoder has six identical layers, just like the encoder. It does, however, have an extra attention mechanism which is encoder-decoder multi-head attention layer. In this layer, the decoder can concentrate on the encoder's output and produce pertinent answers thanks to the multi-head attention layer. The decoder provides the Q, while the encoder output provides the K and values V. The layer guarantees the words produced by the decoder are pertinent to the input sequence in terms of context. The final output of the decoder side is mapped to the vocabulary size by passing it by a fully linked linear layer. For every word, probabilities are generated using a softmax function. In the output sequence, the word with the highest probability is chosen as the following word [14].

Fig. 2 illustrates the architecture of the transformer model, which is composed of an encoder and a decoder. Each component consists of stacked layers that incorporate attention mechanisms and feed-forward networks. Positional encoding is applied to both inputs and outputs to retain sequence information, and the final output is generated through a linear layer followed by a softmax function.

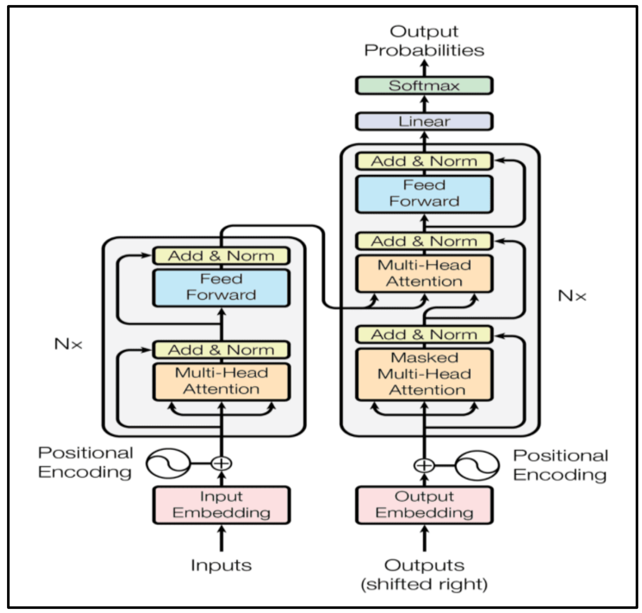

Fig. 2. The transformer model architecture [11].

C. Chatbot Historically

The development of chatbot has been improved significantly over the decades. Fig. 3 summarizes the historical development of chatbots, which will be discussed in detail below. It provides a visual timeline of major milestones, illustrating the evolution of chatbot technologies from early rule-based systems to modern AI-powered models.

- In 20th century: (1950 to 1999)

Alan Turing's published his famous article "Computing Machinery and Intelligence" in 1950. In the article, he discussed a question "Can machines think?", and he proposed a measure of intelligence. This measure called now Turings test. It determines whether the machine acts like human in a real-time written conversation and the human judge cannot absolutely distinguish (based on the conversation only) between the machine and the real human. Based on many experts, the

Turings test is considered to be the generating principle for chatbots [15].

In 1966, the first chatbot was developed by Weizenbaum called ELIZA. ELIZA is simulating conversation with therapist. Early users felt they were speaking with someone who understood their input because of the program's scripted responses that use "pattern matching" and substitution methods [16]. Many early users were persuaded by ELIZA's intelligence and comprehension; however, Weizenbaum may have insisted differently. ELIZA is incapable of maintaining lengthy discussions or picking up on context from the interaction, which is A drawback of ELIZA. It also has limited expertise means that it can only discuss a specific range of topics [17].

PARRY chatbot appeared in 1972 at Stanford University by psychiatrist Kenneth Colby. It simulated a patient with schizophrenia [18]. Because PARRY is designed to have a "personality" and an improved controlling structure, it is thought to be more advanced and intelligent than ELIZA. In an experiment conducted in 1979, five psychiatrists utilized PARRY to determine whether a patient was a real schizophrenia patient or a chatbot. Hence, ten diagnoses were provided by psychiatrists. Two diagnoses were made correctly by the first psychiatrist, and two inaccurate ones by the second one.

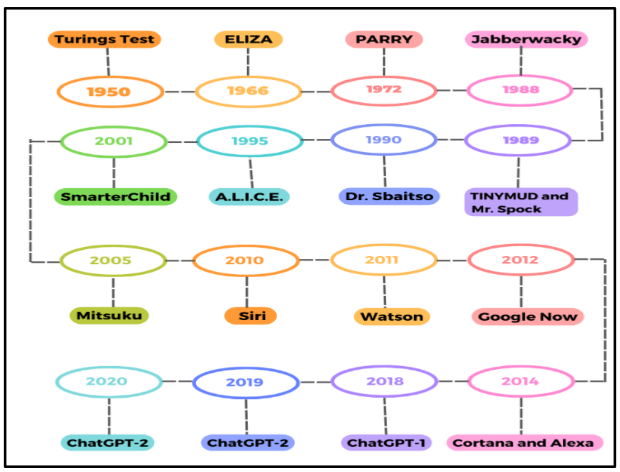

Fig. 3. Chatbot history timeline.

While the other two concluded that the participants were chatbots, the third believed that the subjects were actual patients [19]. However, because there is some degree of confusion in the speech of people who have schizophrenia, the small sample size of five doctors makes it difficult to interpret the results. PARRY is generally regarded as a chatbot with limited language comprehension and emotional expression skills. Additionally, it cannot learn from discussions and responds slowly.

Artificial intelligence was first applied in the field of chatbots in 1988 in the Jabberwacky chatbot. Jabberwacky was created using CleverScript, a spreadsheet-based language that made it easier to create chatbots. It responded by using contextual pattern matching to go back and review past conversations. Nevertheless, Jabberwacky is unable to process large amounts of data quickly or efficiently [20].

One of the first attempts to create a conversational AI character in an online virtual world based on textual was called TINYMUD [21]. James Aspnes created it in 1989, and Carnegie Mellon University's TINYMUD server hosted its initial implementation of it. It was a kind of early online multiplayer game in which players could interact and communicate by sending text commands to one another. Mr. Spock was able to use a straightforward pattern matching algorithm to reply to simple conversational prompts and questions. There was a finite list of possible answers. As they explored the virtual text world together, players could communicate with Mr. Spock by sending him text commands like "say to Spock" or "ask Spock about". Mr. Spock was an early attempt at an artificial intelligence (AI) figure who could converse with people in a virtual setting, albeit being incredibly limited. It introduced the concept of NPC (non-player character) bots, which are widely used in virtual worlds and games today. Later conversational agents in MUDs, MOOs, and finally massively multiplayer online games were made possible through the development of TINYMUD.

Based on [22], the first chatbot to engage people in conversation was Dr. Sbaitso, which was introduced in the early 1990s. It was created specifically for MS-DOS and assumed the role of a psychotherapist, using voice to mimic a counseling session. By enhancing the pattern matching and replacement skills of earlier chatbots, Dr. Sbaitso was able to identify phrases and produce pertinent spoken responses. Despite its limitations, it was a step forward in conversational skills as it combined synthesized speech and text pattern matching to simulate a human-like dialogue. The basis for later chatbots' more complex verbal exchanges was established by Dr. Sbaitso's combination of voice interaction and basic language processing.

A.L.I.C.E. is an approximately more sophisticated bot developed by Richard Wallace in 1995 that is inspired by ELIZA [23]. The bot produced outcomes by pattern matching inputs against (input) (output) pairs kept in knowledge base documents. Artificial Intelligence Markup Language (AIML), an extension of XML that is still in use today, was used to write these documents. The Loebner Prize, which honors the most intelligent chatbot each year and attempts to run the Turing Test, has been won by ALICE three times [15].

- In 21st Century : First decade (2000 to 2009)

SmarterChild chatbot was developed by ActiveBuddy in 2001. It was an inspiration in achieving widespread acceptance. When inquired about something, it could supply information about a wide range of subjects and functioned on instant messaging networks. SmarterChild is one of the first chatbots to be integrated into messaging apps, demonstrated how AI could function as a virtual assistant by retrieving information and responding to user inquiries [24]. It might have been integrated to AOL Instant Messenger and MSN Messenger, reaching over 30 million users. SmarterChild inspired chatbots on messaging networks, opening the door for later AI assistants such as Samsung's S Voice and Apple's Siri. Its easy integration with famous messaging apps showed how chatbots may be used to offer services through dialogue-based messaging settings. Through introducing practical chatbot interactions into regular messaging, SmarterChild demonstrated the potential for millions of users to have access to chat-based assistants in the future [25].

Mitsuku chatbot was developer in 2005 by by Steve Worswick. It is an award-winning chatbot with a playful, adolescent girl demeanor. It has been the recipient of the Loebner Prize five times between 2013 and 2018. Judges for the Loebner Prize interact with a combination of chatbots and humans to identify which are the bots; Mitsuku has consistently deceived the judges in this regard. It is the most widely used AIML-based chatbot. It is hosted on Pandorabots and employs NLP techniques like heuristic patterns and default categories to have natural conversations across many platforms. It uses pre-written patterns and Bot modules to use natural language processing and is made for multiple platforms, including Telegram and Twitter. Although practical, Mitsuku needs new AIML categories to direct user input and manage mismatched prompts. It has been shown to be a useful conversation partner in multiple languages [21].

- In 21st Century: Second decade (2010 to 2019)

Apple developed Siri in 2010. It paved the path for personal assistants. It integrates with audio, video, and image files and allows users to utilize voice commands for questions and conversations using Messenger. With continuous use, Siri learns to recognize the users' language preferences, queries, and requests while providing recommendations and addressing user requests through a variety of online services. Siri has limitations, even if it is a powerful system. It demands an internet connection. Although it is multilingual, it is incapable of supporting many languages, and the only language in which navigation instructions are offered is English. Additionally, it has trouble understanding the person speaking because of their strong accent or background noise [26].

IBM developed a chatbot named Watson in 2011 [17]. Watson had the ability to decipher genuine human language well enough to defeat two previous champions in the game show "Jeopardy", in which players had to predict the questions based on responses they were given. Years later, Watson helped companies develop more advanced virtual assistants. Furthermore, Watson Health was created to assist medical professionals in diagnosing illnesses. Watson's limitation is that it only supports English, though.

When Google Now was first launched in 2012, it was intended to provide information to users based on their location, preferences, and time of day [27]. The next version of Google Now is Google Assistant, which was developed in 2016. It anticipates user needs and provides information to them with a more conversational, amiable interface that combines a deeper level of artificial intelligence. It lacks personality, though, and because it connects to the user's Google Account directly, some of the questions could be considered private [28].

Microsoft has introduced Cortana personal assistant in 2014. It is capable of recognizing voice commands to perform things like emailing, obtaining required information, and making reminders, there certainly have been reports of security holes that allow malware installation. Amazon developed Alexa voice assistant in the same year In order to promote IOT accessibility. One innovation is the Alexa Skills Kit, which allows developers to create custom skills for Alexa, though there are security risks associated with using Alexa. Although Alexa is focused on controlling home devices and Cortana was designed by Microsoft to be a personal assistant, both have raised security concerns over voice command functionality. Notwithstanding continuous security concerns, the introduction of Cortana and Alexa signaled a significant development in the field of AI voice assistants [29].

- In 21st Century: Third decade (2020 to current)

Based on [30], OpenAI introduced ChatGPT in multiple versions that started with GPT-1 in 2018, GPT-2 in 2019, and GPT-3 in 2020. ChatGPT showcases research on training neural networks on massive text data sets to acquire robust language skills, highlighting the most recent developments in employing large language models, such as GPT-3.5 and 4 for conversational AI. It is a prime example of state-of-the-art methods, including scaled-up self-supervised learning, that generate flexible models with robust reasoning and contextual understanding abilities. ChatGPT offers an interactive platform that allows users to experiment with and assess the advancements in research being done on creating dialogue agents with sophisticated LLMs optimized for multi-turn talks.

DeepSeek Chatbot is an advanced AI language model developed by DeepSeek, a Chinese AI research company [31]. It was first introduced to the public in early 2024, gaining attention for its strong performance in coding, reasoning, and multilingual capabilities, competing with models like OpenAI's GPT-4 and Meta's LLaMA. The chatbot is built on a transformer-based architecture, supporting 128K context length, which allows it to handle long documents efficiently. DeepSeek initially released open-weight models (such as DeepSeek LLM 7B/67B) before launching its conversational AI. As of mid-2024, DeepSeek Chatbot became available broadly, both in the free and paid versions, as a major player in the market for AI assistants. Its emergence is part of the global expansion of generative AI tools through an imitation of the rapid evolution of open and potent AI models beyond Western technology companies.

III. Literature Review

A thorough analysis of the main findings of existing research on chatbots is given in this section. It starts with a summary of earlier research that has looked at related subjects, especially in relation to conversational systems. The section goes on to list the main research gaps that this research attempts to fill, emphasizing how little attention has been paid to regional dialects, such the Western Saudi dialect, in both academic and real-world contexts.

A. Overview of Related Work

The authors in this study [27] provide an intelligent chatbot that utilizes advanced technologies for instance Python, Flask web framework, and NLP. This chatbot aims to produce a student's information access and investigate academic administration. The study's results emphsize that, the traditional programming paradigms and rule-based systems have enabled users to seamlessly fetch inclusive academic information by NLP queries. With quick access to extensive students' performance information, the administrators and teachers are able to simplify decision making that can improve methods for enhancing academic performance. Instantaneous insights into students' development allows teachers to recognize areas that need improvement. The chatbot has created a standard for creative and approachable solutions inside educational establishments.

For the purpose of helping university students with academic and administrative issues, the study [28] introduces the development and execution of a College Assistance Chatbot. The chatbot effectively comprehends and processes student queries by utilizing deep learning models, especially an Artificial Neural Network (ANN), and NLP techniques. The architecture of the system is carefully developed to efficiently handle communication, guaranteeing immediate and accurate outcomes. For the neural network training, a large dataset was obtained from university resources and pre-processed. With a series of experimental campaigns, the performance of chatbot was thoroughly assessed. After 1000 epochs, it achieved an accuracy of 80.07%. The findings show that the chatbot can effectively respond to a variety of student requests, offering a scalable way to improve student support services. This novel methodology shows great promise for wider use in educational contexts and provides notable improvements over conventional techniques.

By utilizing the GPT model, the study [29] investigates the design, development, and assessment of an instructional chatbot for a university course. Testing the viability and efficacy of a GPT-based chatbot performing as a personal tutor for students enrolled in the Sociology of Education course is the main objective of the study. In addition to outlining a prototype for developing comparable educational chatbots, the researchers discuss to comprehend the advantages and disadvantages of GPT in educational contexts.

In order to foster critical thinking, the chatbot was developed to respond to inquiries, explain ideas, and involve students in Socratic discussion. The chatbot was created to be a tutor to students in two main implementation steps. First, building a knowledge base to improve subject-matter expertise. Second, configuring behavior, the chatbot was designed with an accurate prompt to prioritize answers from its knowledge base and refrain from depending on outside resources. The results of the study include that the GPT-based chatbot could provide precise answers, encourage interactive learning, and effectively adjust to the demands of the students. It was convenient and accessible because it was available all the time. While the limitations are relying on a very good quality training dataset, occasionally misinterpreting context, and having trouble completely recognizing complex ideas.

While most of the educational chatbot help students in self-learning, the study [30] explores how chatbots can enhance team-based learning independent on operating as a virtual instructor to assist students in working together for successfully complete their projects. Tubo (tutoring bot) was implemented as a component of a website platform for collaborative learning in which students work in groups to accomplish assignments while Tubo serves as a team member. Tubo offers a number of features, including onboarding by welcoming students and introducing the classroom, assignment guidance by outlining tasks and providing directions, using multiple-choice questions to reinforce important ideas, progress monitoring involves keeping tabs on student's development and responding to often requested queries, and maintaining students' motivation through small conversation and encouragement. At the end of the experience of Tubo, students enjoyed Tubo's interactive and enjoyable diversion from traditional classroom instruction. A few pupils said they could envision themselves using Tubo on their own. On the other hand, a few students lost interest in Tubo's instruction because they thought it was overly strict and repetitious, and some of them struggled with task division and frequently worked in switches rather than cooperatively.

Based on the purpose of reducing educational inequities, particularly in the area of English language instruction, where rural children frequently lack access to resources and qualified teachers. The authors in this study [32] introduce BuddyBot chatbot to help achieve Sustainable Development Goal 4 (SDG4), which is to guarantee that everyone has access to high-quality, inclusive education. By giving instant feedback on grammar, vocabulary, and sentence structure, the chatbot attempts to provide a customized learning experience. BuddyBot is constructed by utilizing GPT-2 model and Google's FLAN-T5 model. In order to comprehend and react to learners' inquiries, correct grammar, and offer language feedback, both models were refined using specific datasets. The introduced chatbot performed strongly on fact-based and open-ended questions, as well as brief and lengthy chats. The results show a decrease in training loss from 0.778 to 0.414 across 8 epochs and an improvement in accuracy.

Considering the growing requirement in industries such as Small and Medium-sized businesses (SMEs), this study [33] investigates essential technologies for chatbots to produce responses. The study trains an enterprise chatbot knowledge base using deep learning frameworks. The study presents a chatbot that utilizes a Sequence to Sequence (Seq2Seq) model with Long Short-Term Memory (LSTM) to enhance SMEs industry. LSTM was used as the encoder and decoder, while the Seq2seq model merged with the target corpus to improve propriate the roles of customer service. At the beginning of neural network processing, the chatbot dataset completed preprocessing, word vector training, and dialogue Q&A conversion. Users were able to submit queries and view the chatbot's generated answers through a Flask web interface. The introduced chatbot website can dynamically hot-load and train the model without restarting the chatbot after deployment. This feature enables ongoing enhancement and flexibility in response to changing client demands.

"Nabiha", a chatbot designed by [34] to engage in conversations with Information Technology (IT) students at King Saud University using the Saudi Arabic dialect. Consequently, Nabiha marks the first Arabic chatbot utilizing Saudi dialect. To enhance accessibility, Nabiha is available across multiple platforms including Android, Twitter, and the web. Students can interact with Nabiha by downloading the application, engaging with her on Twitter, or visiting her website. The effectiveness of Nabiha was assessed by IT department students, yielding somewhat satisfactory results considering the challenges posed by the Arabic language, especially the Saudi dialect.

In [35], the authors suggested and created the smart guiding chatbot that utilized Saudi dialect, a text-based application for

Jeddah's tourism industry. Users could always reach the chatbot for the most recent information on locations, events, weather, and recommended places to visit. They provided a technical overview of the materials NLP, ML, and Rasa.ai are necessary to develop a chatbot. The study also covers how the chatbot uses the probabilistic model LSTM to classify, process, and predict data in order to identify the best match. The suggested chatbot can comprehend user demands and reply promptly, according to the results.

The other research, which has proposed a chatbot in Arabic dialect, was presented by [36]. The purpose of the authors is to create an intelligent chatbot system that can lessen this kind of burden in Arabic. Specifically, this approach facilitates the use of a Jordanian dialect of Arabic with students, especially those attending the Al-Zaytoonah Private University of Jordan. As a result, the suggested system is a basic Jordanian chatbot that speaks in the local dialect. Several artificial algorithms are used in the Intelligent Arabic chatbot system's architecture to assess and comprehend a range of visitor queries. This system is a website that quickly responds to inquiries from users. It also has an easy-to-use interface that answers inquiries about the placement cell, test cell, academics, admissions, users' attendance records, and grade point averages, among other things.

In [5], the authors describe the creation and deployment of Bashayer, a task-oriented chatbot integrated inside the WhatsApp app. Its goal is to assist Saudi Arabian postgraduate students with their study methodologies and motivation. A single-subject experimental design utilizing a quasi-experimental setup was used with a group of 60 postgraduate Saudi students. The findings of the descriptive analysis of the data gathered indicated that postgraduate students who used the Bashayer chatbot system achieved encouraging outcomes. Compared to the control group, participants in the experimental group using Bashayer showed higher levels of motivation to learn. In addition, when using the chatbot, participants used more cognitive and metacognitive learning processes than the control group. The findings provide encouragement for the creation of Bashayer-like chatbots that will aid in the effective learning of postgraduate students. These findings fill a study void and add to the body of knowledge regarding the application of chatbots in postgraduate educational settings.

Table I summarizes these studies, focusing on the findings and limitations of each study.

This research is necessary, as it addresses a significant gap in the existing literature by developing a chatbot specifically tailored to understand and respond to the Western Saudi dialect, which has been underrepresented in the field of NLP. While there has been substantial work on general-purpose chatbots and models designed for widely spoken dialects or MSA, few studies have focused on region-specific dialects like the Western Saudi dialect, especially in the context of university students. By developing a chatbot that can respond to questions unique to Taibah University, including course registration, academic regulations, and campus information, this research seeks to give students a useful solution. This project will not only fill the gap in dialect-specific chatbot development but also enhance institutional efficiency and student assistance by relieving staff workloads and providing students with quicker access to information.

IV. Methodology

This section describes the research methodology that has been followed to achieve all the research objectives. Throughout this section, a description of the research methodology has been provided. Each phase ends with results obtained from their activities, and the results of each phase are used as a starting point for the next phase. The research methodology consists of six interrelated stages, namely:

- Problem Identification (stated earlier).
- Algorithms
- Data Collection
- Implementation Plan
- Evaluation Methods

A. *Algorithms*

The chatbot will utilize the LLM specifically, GPT model. This section briefly describes the LLM and GPT. LLM is a Statistical language model that estimates the probability of word sequences, facilitating a range of activities like code creation, question answering, and translation. To generalize from context, they are trained on large datasets and mostly rely on transformer designs. LLM learns a probability distribution for word sequences, see Eq. (2). This allows it to predict the next word given the previous context. Furthermore, it is designed to generate text by sampling from the learned probability distribution, facilitating tasks like summarization, essay writing, or even solving problems [37].

$$P(W1, W2, \dots, WL) \tag{2}$$

Autoregressive techniques, in which the model predicts the next word one step at a time, are frequently used by LLMs, see Eq. (3):

$$P(Wn + 1 | W1, W2, \dots, Wn) \tag{3}$$

The main factors of transformer architecture are transformer models which are rely on the attention mechanism, which enables the model to consider the relative relevance of various input phrases while making predictions, see Eq. (1). In the equation, Q, K, VQ, K, V represent queries, keys, and values derived from input embeddings.

Also, positional encoding utilizes to encode word order, positional embeddings are appended to word embeddings because transformers are intrinsically permutation invariant. Additionally, scalability transformers are more effective than recurrent designs because they enable parallel processing.

Self-supervised learning is used to train LLMs to anticipate next or masked words in text sequences. The goal is to reduce a loss function, such as cross-entropy, see Eq. (4). According to empirical scaling laws that relate loss to model and dataset size performance is enhanced with larger models and datasets.

TABLE I STUDIES SUMMARIZATION TABLE

Research	Approach	Model	Language	Key findings	Limitation
[27]	Retrieval based	ML	English	Enhancing academic improvement strategies	cost considerations, user acceptance, and the need for continuous updates.
[28]	Retrieval based	NLP, ANN	English	Handling a wide range of student queries and providing a scalable solution for enhancing student support services	-
[29]	Generative	GPT	English	Providing precise answers, encouraging interactive learning, and effectively adjusting to the demands of the students	having difficulty comprehending context and relying on the quality of the training data to generate accurate responses
[30]	Rule based	-	English	Enhancing the team learning method	Some student think that the chatbot became somewhat repetitive and boring over time
[32]	Generative	GPT-2, Google's FLAN-T5	English	achieving 80% accuracy in dialogue generation, demonstrating proficiency in question answering with an 83% accuracy	-
[33]	Retrieval based	Seq2Seq, LSTM	English	Effective preprocessing, Enhancing chatbot performance with LSTM	-
[34]	Retrieval based	-	Arabic, Saudi dialect	The initial Nabiha chatbot experiment's findings were fairly satisfactory.	Lack of dataset
[35]	Retrieval based	NLP, LSTM	Arabic, Saudi dialect	Effective interaction, and immediate responses	-
[36]	Retrieval based	NLP	Arabic, Jordanian dialect	Quickly responding to inquiries from users, and an easy-to-use interface	Additional experiments and adjustments are necessary.
[5]	Retrieval based	-	-	Supporting postgraduate students' motivation and learning strategies	Small students sample, short duration of testing

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\ P(w_i \mid w_{1:i-1}) \qquad (4)$$

GPT model is a type of LLM which use transformer architecture to anticipate the next word in a sequence. Pretraining on large corpora to identify general linguistic patterns and fine-tuning on particular tasks are the two phases of training GPT models. With GPT-3 (175 billion parameters) bringing few-shot and zero-shot learning and GPT-4 (estimated 1.4 trillion parameters) showcasing sophisticated reasoning skills, GPT has seen significant evolution since its launch in 2018. Although it is still constrained by hallucinations, logical fallacies, and computational needs, its autoregressive methodology, scalability, and generalization across tasks are important aspects [38].

B. Data Collection

Since the accuracy of the results and conclusions is greatly impacted by the quality of the data collected, the data collection phase is one of the most important phases in any research or analytical study. The data was collected from a variety of sources during this phase to guarantee thoroughness and diversity, successfully matching the goals of the research.

Formal requests have been made to gather student inquiries that was send to academic advisers as part of the research on creating a GPT-based approach to assist Saudi university students. The Deanship of Information Technology at Taibah University has received the request. In order to examine common issues, commonly asked questions, and the general framework of academic advising exchanges, the goal of these requests was to obtain access to student inquiries that was send to academic advisers from eservices portal of Taibah University.

A Taibah University student Telegram group for the faculty of computer science (CCSET) was employed as one of the primary data sources. This platform was a powerful source of qualitative data since it offered insightful information about student conversations, concerns, and shared experiences. The author gathered pertinent information by examining the group discussions and exchanges, which helped clarify students' academic and administrative difficulties.

Moreover, information received in emails pertaining to the Course Equivalency Committee has been gathered. These emails contained information related to the policies concerning course credit transfers, decision-making processes, and the criteria for determining course equivalency requests. This source was particularly useful in understanding how the university manages credit recognition and how students interact with it.

The other major source of information included Academic Advising Committee emails. The emails provided data regarding the advising and counseling extended to students, including advising policy, mentorship programs, and academic services to help students in their educational planning. From perusing these messages, the university's academic advising mechanism and how the system assisted the students have been provided.

Also, the names and email addresses of faculty members at Taibah University was obtained from the Deanship of Information Technology. This source was essential in ensuring faculty members across various departments had access to accurate and up-to-date contact details.

By relying on these previously diverse sources, the balanced dataset that supports a multifaceted analysis of the academic experience at Taibah University has been collected. The reliance on several sources ensures a balanced perspective, drawing from student perceptions as well as official university policies.

C. Implementation Plan

The implementation strategy focuses on developing two chatbot systems that are specifically designed to interact using the Western Saudi dialect, ensuring that the responses align with the linguistic and cultural nuances of the target user group—university students in the western region of Saudi Arabia. Traditional machine learning performance is measured against the first chatbot, which is constructed using an SVM model. The GPT-3.5 model, a transformer-based generative language model renowned for its sophisticated contextual comprehension and fluency, is used by the second chatbot in contrast.

The objective of this dual-model setup is to conduct a systematic comparison between conventional machine learning techniques and state-of-the-art transformer models in the context of dialect-specific natural language understanding. Additionally, a separate retrieval-based model using Term Frequency-Inverse Document Frequency (TF-IDF) and Cosine Similarity is implemented to handle queries related to faculty email addresses. In order to prevent the possibility of fabricated or hallucinated email information, this retrieval component is purposefully separated from the generative system. This ensures factual accuracy and upholds user trust.

D. Evaluation Methods

In this research project, a multi-dimensional assessment method was employed to critically assess the chatbot's performance. Three complementing methods were selected: human evaluation, BERTScore, and automatic GPT-4-based judgment. Human evaluation was used to elicit subjective factors such as naturalness, tone, and cultural appropriateness, yielding useful real-world observations particularly for the Western Saudi dialect. For quantitative measurement of semantic similarity among generated responses and reference answers independently, BERTScore was utilized as a context-sensitive alternative to typical surface-level measures. Further, GPT-4 also served as an automatic judge to enable scalable and uniform assessment across large datasets without trading practical practicability at the cost of linguistic insensitivity loss. By merging human judgment, semantic assessment, and machine testing, the technique provided a robust and rich assessment of the chatbot's linguistic, cultural, and functional capacity.

First, human evaluation was conducted to capture subjective aspects of response quality that are difficult to measure algorithmically, such as appropriateness, cultural alignment, and conversational naturalness. Human annotators assessed the relevance, fluency, and tone of the chatbot's replies based on a predefined set of guidelines tailored to the Western Saudi dialect context. This form of evaluation provided valuable insights into the real-world usability and acceptability of the chatbot among the target audience.

To evaluate the accuracy of the chatbot responses using human judgment, a human evaluation process was conducted involving five independent annotators. Each annotator reviewed the chatbot's responses and provided a binary judgment—either "Yes" if the response was considered appropriate, or "No" if it was not. For each response, the final decision was determined through majority voting, where the most frequent judgment (i.e., at least 3 out of 5 votes) was taken as the overall evaluation of that response. Once the final decisions were established for all responses, accuracy was computed by dividing the number of responses that received a majority "Yes" by the total number of responses evaluated. This approach ensures a reliable and consensus-based assessment of the chatbot's performance.

Second, BERTScore was employed to objectively measure the semantic similarity between the chatbot's generated responses and the ground-truth reference responses. Unlike traditional metrics such as BLEU or ROUGE that rely on exact n-gram matches, BERTScore employs deep contextual embeddings to identify whether two sentences are identical in meaning, even if they are differently phrased. This was particularly important given the linguistic richness and flexibility of the Western Saudi dialect, where multiple phrasings could be semantically correct.

In BERTScore, precision measures how well the chatbot's response captures the meaning of the reference answer. In Eq. (5), precision does this by checking each word in the chatbot's response and finding the most similar word in the reference, using BERT embeddings. A high precision means that most words in the chatbot's response are semantically similar to words in the reference answer.

$$P = \frac{1}{|X^\wedge|} \sum_{x^\wedge_j \in X^\wedge} max_{x_i \in X} \; sim(x_i , x^\wedge_j) \qquad (5)$$

In the BERTScore precision equation, each symbol plays a specific role in capturing the semantic alignment between the chatbot's response and the reference answer. The symbol $X^\wedge$ represents the set of tokens (words) in the generated response, while X denotes the set of tokens in the reference (ground truth) response. For every token, the equation searches for the most similar token x_i in the reference using cosine similarity of BERT embeddings, represented as $sim(x_i , x^\wedge_j)$. The maximum similarity for each generated token is then summed, and the total is divided by the number of generated tokens $|X^\wedge|$, resulting in the average semantic precision across the response.

On the other hand, recall in BERTScore checks how well the chatbot's response covers the important content of the reference answer. In Eq. (6), recall looks at each word in the reference and finds the closest matching word in the chatbot's response. A high recall means the chatbot included most of the important meanings from the reference answer.

$$R = \frac{1}{|X|} \sum_{x_i \in X} max_{x^\wedge_j \in X^\wedge} \; sim(x_i , x^\wedge_j) \qquad (6)$$

The recall equation in BERTScore operates in the reverse direction of precision. Here, X still denotes the set of reference tokens, and $X^{\wedge}$ is the set of generated tokens. For each reference token, the model computes the maximum similarity with any token $x^{\wedge}{}_{j}$ in the generated response. This similarity is again based on cosine distance between BERT embeddings, expressed as $im(x_i\ , x^{\wedge}{}_j)$. The total of these maximum similarities is then divided by the number of reference tokens $|X|$, producing a measure of how completely the generated response covers the semantic content of the reference.

Finally, the BERTScore F1 combines precision and recall to give a single balanced score. As shown in Eq. (7), F1 score tells us how well the chatbot's response matches the reference in both accuracy and coverage of meaning. A high F1-Score means the chatbot response is both relevant and complete from a semantic perspective.

$$F1 = \frac{2PR}{P+R} \tag{7}$$

Third, GPT-4 was used as an automatic judge to provide scalable and consistent evaluation across a large number of test samples. GPT-4 was prompted to assess whether the chatbot's response was contextually appropriate and semantically accurate based on the given input message. The accuracy was calculated in the same way as human evaluation, by dividing the number of responses that were appropriate by the total number of responses. By using a state-of-the-art LLM as an evaluator, the research was facilitated with high linguistic sensitivity and judged consistency and reduced the need for a lot of human effort at the expense of evaluation quality.

By integrating these three mutually complementary evaluation methods, the performance assessment of the chatbot was made more balanced in nature, encompassing both human-centered perspectives and machine-based semantic and contextual accuracy. This provided an overall view of the strengths and shortcomings of the chatbot, thus enhancing the validity of the experiment results.

The selection of human evaluation, BERTScore, and GPT-4-based judging is particularly well-aligned with the specific objectives and challenges of this research. The primary goal is not merely to assess the technical accuracy of the chatbot's responses, but to ensure their cultural, linguistic, and contextual appropriateness within the Western Saudi dialect environment — an aspect that conventional metrics alone cannot fully capture.

Human judgment plays a crucial role in this regard, making fine-grained judgments founded on real-world conversational practices and user expectations. Given the sociolinguistic variation and familiarity built into the Western Saudi dialect, human judges are necessary to validate the chatbot's ability to converse naturally and respectfully with the target group.

Complementing this subjective perspective, BERTScore was chosen for its capacity to assess semantic similarity with high sensitivity to context and meaning. In dialectal settings, where different surface forms can express identical intents, traditional exact-match metrics would severely underestimate model performance. BERTScore circumvents this risk by taking into account if the answers generated convey the intended meaning, thus giving a more linguistically fair and representative score.

Furthermore, GPT-4 as an automatic judge introduces a scalable and efficient dimension to the assessment. Its advanced language understanding enables it to make near human-like assessment judgments on large quantities of data without the prohibitive time and costs of manual annotation. This scalability is especially relevant with the sizes of the test sets here and the need for consistent evaluation metrics on thousands of samples.

Overall, the selected evaluation metrics collectively are well-suited because they accommodate the research's new linguistic complexity, deployment practicability, and research objectives. Overall, they measure the chatbot's performance wholeheartedly, not just its correctness, but its meaningful conversation in a specific cultural and dialectal context.

V. Implementation and Experimentation Settings

This section provides the implementation method of the chatbot tailored to the Western Saudi dialect, specifically for the context of Taibah University. A baseline model using an SVM architecture and a more sophisticated model based on the GPT framework were both used to evaluate performance. The GPT-based model uses transformer-based language understanding, which is better suited to capture the subtleties of dialectal Arabic, whilst the SVM model provides a classificational modeling method. The two models were both trained with a customized dataset representing the western Saudi dialect linguistic properties, and the hyperparameters as well as the preprocessing methods used were deliberately chosen to give an informative and objective comparison. More information regarding the tools, the frameworks, training configurations, and the testing techniques used in this research is given in the next subsections.

A. Experiment 1: Email Retrieval-Based Model

To handle faculty email inquiries with greater reliability and to avoid the potential risk of generating inaccurate or non-existent email addresses, a separate retrieval-based model was developed. This model is built using TF-IDF and Cosine Similarity to retrieve verified faculty emails from a structured dataset. The system ensures that only authorized and approved email addresses are returned by depending on exact matches rather than generating output. The GPT-3.5-Turbo-0125 chatbot will eventually be merged with this retrieval-based email architecture. The hybrid system will automatically route email-related queries to the retrieval model, while general academic or administrative questions will be handled by the generative GPT-3.5-Turbo-0125 model. This architecture enhances both the accuracy and trustworthiness of responses by combining retrieval precision with the conversational fluency of transformer-based generation.

The TF-IDF approach is used to measure each text message in numerical terms by assigning weights to words based on their importance. Especially, TF refers to the number of times a word repeats in a message, and IDF reduces the weight of common words that appear in most messages. This provides a high-dimensional vector for each message that reflects its unique linguistic features. In order to find the best fit for a user's query, the input is converted by the same TF-IDF vectorizer and matched against the stored message vectors by Cosine Similarity, which measures the angle between vectors. The most similar message is chosen, and its corresponding response is returned. This method is computationally efficient, doesn't require model training, and performs well when the data set is very clean and there are stable patterns in the language.

An important function was developed to serves as a decision-making mechanism that intelligently routes user queries to the appropriate response model. Specifically, it checks whether the user's input contains keywords related to email inquiries (e.g., "email", "بريد" ,"إيميل"). If such keywords are detected, the function delegates the task to a retrieval-based model built on TF-IDF and cosine similarity to return an exact match from a predefined list of official faculty email addresses. This approach is deliberately chosen to avoid generating or "hallucinating" non-existent email addresses, a risk commonly associated with generative models like GPT. Conversely, if the query is unrelated to emails, it is forwarded to a fine-tuned GPT model trained to respond in the Western Saudi dialect. This dual-model architecture ensures both accuracy in factual information retrieval and fluency in conversational responses.

B. Experiment 2: Baseline Model

In this experiment, a baseline model was implemented using SVM to evaluate the feasibility of building a dialect-specific chatbot capable of understanding and responding in the Western Saudi dialect. The SVM model was selected due to its effectiveness in handling classification tasks with limited training data and its interpretability as a classical ML method. This experiment served as a foundational comparison for the subsequent evaluation of a more sophisticated GPT-based model. By establishing the performance ceiling of a traditional model, the research aimed to determine the added value of utilizing LLMs in regional Arabic dialect applications.

The dataset used in this experiment consisted of paired question–answer entries written in the Western Saudi dialect, preprocessed to remove emojis, special characters, and unnecessary whitespace. The data was divided into 80% for training and 20% for testing. Text preprocessing included tokenization, lowercasing, and TF-IDF vectorization to convert the textual data into numerical feature representations. An SVM classifier with a linear kernel was then trained on the resulting feature matrix.

C. Experiment 3: GPT Model

The third experiment explores the performance of an advanced large language model, GPT-3.5-Turbo-0125, developed and hosted by OpenAI. Unlike traditional ML models such as SVM, GPT-based models leverage deep transformer architectures pre-trained on massive corpora. This capability enables the model to generate contextually relevant and linguistically coherent responses, even in dialectal variants with limited annotated resources. The purpose of this experiment is to evaluate the effectiveness of GPT-3.5-Turbo-0125 when fine-tuned on a custom dataset of dialogues in the Western Saudi dialect, and to compare its performance against the previously established SVM baseline.

For fine-tuning the same dataset that was utilized in the SVM, has also been utilized in this phase. The dataset of Western Saudi dialect messages and their corresponding responses was converted into a format compatible with OpenAI's fine-tuning pipeline. The input-output pairs were structured as prompt–completion sequences, with each user message serving as a prompt and the ideal chatbot reply as the completion. The training process was conducted via OpenAI's fine-tuning API, utilizing the base model gpt-3.5-turbo-0125. Once the custom model was trained, evaluation was performed using a test set consisting of unseen message-response pairs. Additionally, some test cases were assessed using an automated judge—GPT itself—by requesting it to determine whether the generated response was appropriate, based on a scoring system that reflected qualitative judgment. This setup enabled both quantitative and human-aligned evaluation.

VI. Results and Discussion

This section presents a detailed analysis of the experimental results and interprets their significance in the context of the research. The performance of the proposed chatbot model is examined using relevant evaluation metrics to assess its ability to understand and respond in the Western Saudi dialect. A comparative study is carried out against a baseline SVM model in order to better position the results and emphasize the advantages of transformer-based techniques over traditional methods for ML. The ramifications of the results and their applicability to the goals of the research.

A. Results of Email Retrieval-Based Model

The model achieved an Exact Match Accuracy of 99.96%, indicating that it performs with a very high level of precision in retrieving the correct email address based on the provided Arabic display names.

TABLE II RESULTS OF SVM RESPONSES OF USER'S MESSAGES

Message		Response	
Arabic	***English***	***Arabic***	***English***
لو نزلي حرمان وعندي اعذار ينشال؟	If I receive a deprivation notice but have valid excuses, is it possible to have it lifted?	ايوا	Yes
كم الغياب المسموح لمحاضرات عددها٣٠؟	How many absences are allowed for 30 lectures?	ايوا	Yes
أبغى ايميل دكتور ماهر	I want Dr. Maher's email	لازم تعطيني الإسم الثنائي لعضو هيئة التدريس عشان أبحث بشكل دقيق	Please provide the full name of the faculty member in order to perform an accurate search.

B. Results of SVM Baseline Model

Table II shows the results of the SVM-based chatbot's performance in handling different inquiries. Each row in the table represents a message from the user and the corresponding response generated by the chatbot.

The response of message "لو نزلي حرمان وعندي اعذار ينشال" (If I receive a deprivation notice but have valid excuses, is it possible to have it lifted) indicates that the chatbot correctly understood the user's query about whether an academic penalty could be waived in the presence of valid excuses. The chatbot provides a simple affirmative response "ايوا" (yes), confirming the possibility of such an action.

The model's weakness in answering the question " كم الغياب المسموح لمحاضرات عددها ٣٠" (How many absences are allowed for 30 lectures?) is evident in its vague and imprecise response of "ايوا" (yes). This indicates that the SVM model failed to fully understand the context of the question, as it should have provided a specific answer regarding the allowed absenteeism percentage for a given number of lectures. Instead, it gave an unclear response.

In the third response, the chatbot correctly asked the user for the full name of the faculty member "دكتور ماهر" (Doctor Maher) in order to provide an accurate result.

Table III provides a summary of the SVM model's performance metrics on the test dataset, which comprises 20 messages and the responses that go with them.

TABLE III EVALUATION METRICS RESULTS OF 20 MESSAGES AND ANSWERS OF TEST DATA FOR SVM

Evaluation metric	Result
Accuracy	0. 4286
F1-score	0. 8709
Precision	0. 8806
Recall	0. 8631
latency	0.2 sec

The model was able to accurately predict the answers for almost 43% of the test data, according to its accuracy of 42.86%. Given that the model is trained in predefined classes and assessed on unseen data, this accuracy may seem moderate, but it actually represents the difficulty of the task. The test set contained unseen responses, which the model might not have been trained to handle well, which could account for the comparatively poor accuracy. Given how much the SVM model depends on the data it was trained on, this points to a possible drawback.

In terms of F1-score, the model achieved a value of 87%, suggesting a strong balance between precision and recall. The precision of 0.8806 indicates that when the model predicted a response, it was correct approximately 88% of the time, minimizing false positives. Meanwhile, the recall of 0.8631 shows that the model was able to identify about 86% of the true positive responses, indicating that it was successful in capturing most of the correct answers in the test set. With a latency of 0.2 seconds per prediction, the model can respond in real time, which makes it appropriate for chatbots and other applications that need fast feedback. Even with its moderate accuracy, the overall performance indicates that, when trained on the different formations of data, the SVM model may effectively detect accurate responses.

C. Results of GPT-3.5-Turbo-0125 Advanced Model

The sample results from the GPT-3.5-Turbo-0125 model that is provided in the Table IV show its ability to provide contextually accurate and relevant responses to a variety of student inquiries. For instance, when asked about permissible absenteeism for 30 lectures, the model gave a specific response, stating that 25% absenteeism is allowed, which is a precise and useful answer. Other responses, such as requests for emails or explanations on university policies, are similarly helpful, providing the necessary details or requesting additional information when needed.

The fine-tuned GPT-3.5-Turbo-0125 model demonstrated a significant improvement in handling dialogue generation tasks compared to traditional ML approaches. Upon interacting with user prompts written in the Western Saudi dialect, the model consistently produced coherent, contextually appropriate responses, and linguistically aligned with the informal and natural conversational style characteristic of the region. The system message used on each interaction gave the model a welcoming, culturally appropriate tone and essentially replicated the pattern of speech of students in Madinah, Jeddah, Mecca, and Taif. Throughout the testing phase, the model exhibited an excellent ability to comprehend highly varied input phrasings, including colloquial expressions, truncated sentences, and context-implicated questions, without requiring explicit rephrasing or clarification.

Furthermore, the qualitative observations output of the model highlighted its strong capacity for conforming to delicate language use, such as responding appropriately to rhetorical questions, handling indirect requests, and engaging with culturally embedded expressions. Unlike traditional models that rely on fixed feature extraction techniques, GPT-3.5-Turbo-0125 leveraged its deep contextual understanding to generate responses that not only matched the intended meaning of the prompts but also maintained the conversational fluency expected in natural human dialogue. The consistency, richness, and contextual awareness observed in the outputs confirm the model's potential for real-world deployment in dialect-specific chatbot applications. These results strongly support the adoption of LLMs when aiming to develop conversational agents capable of interacting naturally and effectively within specific cultural and linguistic contexts.

Table V contains the evaluation metrics that provide important information about the performance of the GPT-3.5 model, which has been optimized to comprehend and reply to student questions in the Western Saudi dialect. The Human Evaluation Accuracy of 0.8473 demonstrates that the model's responses were deemed accurate by the five evaluators, consisting of three students from the Computer Science department, a Computer Science graduate, and a technical employee. This great accuracy shows that the model was able to correctly analyze and answer questions with a degree of agreement among experts, even with its specialized training.

TABLE IV RESULTS OF GPT-3.5-TURBO-0125 MODEL RESPONSES OF USER'S MESSAGES

Message		Response	
Arabic	*English*	*Arabic*	*English*
لو نزلي حرمان وعندي اعذار ينشال؟	If I receive a deprivation notice but have valid excuses, is it possible to have it lifted?	ايوا	Yes
كم الغياب المسموح لمحاضرات عددها ٣٠؟	How many absences are allowed for 30 lectures?	الغياب المسموح 25٪ من مجموع المحاضرات	The allowed absence is 25% of the total number of lectures.
أبي ايميل دكتور ماهر	I want Dr. Maher's email	لازم تعطيني الإسم الثنائي لعضو هيئة التدريس عشان أبحث بشكل دقيق	You need to provide the full (first and last) name of the faculty member so I can search accurately.
بسأل في التقييم بالموقع يطلع عند الدكتورة اسمي؟	If I submit a course evaluation on the website, will my name appear to the instructor?	لا ما يطلع اسمك بالتقييم	No, your name does not appear in the evaluation.
انا اول سنة واخذ برمجه طبيعي اني احسها جدا صعبة؟	This is my first year, and I'm taking programming. Is it normal to find it very difficult?	طبيعي جداً ، اغلب الطلاب يحسون بصعوبة في مادة البرمجة لكن المهم تراجعها أول بأول	It's completely normal. Most students find programming difficult, but the key is to review it regularly.
التاجيل قفل ؟	Is the postponement period closed?	على حسب التقويم	It depends on the academic calendar.

TABLE V EVALUATION METRICS RESULTS OF ALL TEST DATA FOR GPT-3.5-TURBO-0125

Evaluation metric	Result
Human Evaluation Accuracy	0.8473
GPT-4 Evaluation Accuracy	0.60
F1-score	0.706
Precision	0.71
Recall	0.705
latency	2 sec

On the other hand, the GPT-4 Evaluation Accuracy of 0.60 reflects a somewhat lower performance from the GPT-4 model in evaluating the chatbot's responses. The difficulty of having a non-human evaluator evaluate responses involving complicated dialects and region-specific settings is highlighted by the fact that, despite its power, GPT-4 was less in line with the perceptions of human evaluators. Combining Precision (0.71) and Recall (0.705), the F1-score of 0.706 indicates a well-balanced model performance. It was able to effectively identify correct responses while minimizing errors. Precision suggests that 71% of the responses identified as correct were truly correct, while Recall shows that the model captured about 70.5% of all the correct responses. Lastly, these metrics demonstrate the model's accuracy and relevance in processing queries in the Western Saudi dialect.

VII. DISCUSSION OF FINDINGS

As shown in Table VI, the difference in responses between GPT-3.5-Turbo-0125 and SVM for the message "كم الغياب المسموح لمحاضرات عددها ٣٠" (How many absences are allowed for 30 lectures?). It is noticeable that GPT-3.5-Turbo-0125 performs better than the SVM model in terms of giving more precise and thorough answers. In particular, GPT-3.5 gave an exact response of 25% absenteeism to the question on acceptable absences for 30 courses, demonstrating a deeper comprehension of academic policies. The SVM model, on the other hand, had trouble answering this kind of question and provided ambiguous answers like "ايوا" (Yes), which lacked the relevant information. Particularly when contrasted with the more straightforward and uninformative results produced by the SVM model, GPT-3.5's capacity to manage intricate queries and deliver precise, context-aware responses is remarkable. This demonstrates GPT-3.5's superior understanding of context and ability to generate specific and meaningful responses.

TABLE VI RESPONSE COMPARISON BETWEEN SVM AND GPT-3.5-TURBO-0125

	Message	Response
SVM	كم الغياب المسموح لمحاضرات عددها ٣٠	ايوا
GPT-3.5-Turbo-0125	كم الغياب المسموح لمحاضرات عددها ٣٠	الغياب المسموح 25٪ من مجموع المحاضرات

The comparative evaluation between the SVM baseline model and the fine-tuned GPT-3.5-Turbo-0125 model clearly demonstrates the superiority of transformer-based approaches in handling dialect-specific dialogue tasks. The SVM baseline, trained on TF-IDF features with an RBF kernel, achieved a modest accuracy of 42.86% on the test set that contains only 20 messages. Its performance was limited by its dependency on surface-level lexical similarities and the insufficiency of generalization to all the random and garrulous expressions that belong to the Western Saudi dialect.

Human assessors determined that 84.73% of the chatbot's responses were both culturally appropriate and contextually accurate. In addition to the human evaluation, BERTScore provided further validation of the model's performance, achieving an F1 score of 71% and confirming its ability to capture the intended semantic meaning. These findings were reinforced by GPT-4, used as an automated assessor, which judged 60% of the responses as acceptable. Collectively, these outcomes illustrate the deep contextual understanding, semantic flexibility, and cultural sensitivity of transformer-based models, particularly when compared to classical ML classifiers. This distinction becomes even more evident when addressing the linguistic richness and complexities associated with regional Arabic dialects.

The outcomes of this research carry important implications for the broader field of computer science, particularly in the domains of NLP, dialogue systems, and dialectal language modeling. The significant performance gap observed between the SVM baseline model, and the fine-tuned GPT-3.5-Turbo-0125 model highlights the critical role that large-scale pretraining and context-sensitive fine-tuning play in building effective conversational agents. These results emphasize that typical ML techniques, while helpful for certain structured classification tasks, are insufficient to deal with the richness, dynamism, and cultural nuance of real-world dialogue, especially in underrepresented language varieties such as the Western Saudi dialect.

More generally, the successful deployment of a chatbot that answers academic queries in culturally responsive and linguistically sensitive terms demonstrates the growing capacity of AI to support broad and niche communities. It demonstrates how adapting LLMs to regional and domain environments can significantly enhance accessibility, inclusivity, and user satisfaction. This outcome is particularly relevant where MSA dominates the technology interfaces to the detriment of local dialect speakers. Thus, this research contributes to the broader movement toward more personalized and culturally sensitive AI technologies.

Moreover, this work moves the frontier of knowledge and practice by illustrating an effective way of combining fine-tuning techniques, dialectal prompt engineering, and multi-dimensional assessment methods to attain high-quality conversational systems. It also provides empirical evidence that cultural and linguistic localization is not only possible but also highly effective when developing AI solutions for real-world applications. By systematically comparing the performance of the chatbot through human evaluation, semantic similarity metrics, and automatic metrics, this study sets a new benchmark for the evaluation of dialogue systems in low-resource language settings. Lastly, the research paves the way for future studies to bridge the gap between state-of-the-art language modeling technologies and the rich linguistic diversity within and across Arabic-speaking societies.

VIII. Conclusion

This research effectively created a culturally sensitive chatbot that could respond to academic questions in the Western Saudi dialect, providing significant contributions to Arabic NLP and dialogue system research. The comparison between the SVM baseline and fine-tuned GPT-3.5-Turbo-0125 model unequivocally exhibited the greater contextual understanding, semantic versatility, and cultural suitableness of transformer-based models. By integrating human judgment, BERTScore computation, and GPT-4-based analysis, the research provided a holistic and balanced assessment of the chatbot's performance. The findings demonstrate that AI technology is more inclusive and relevant than only its technological advancements when it is modified for underrepresented languages. This work sets a new benchmark for developing conversational agents in low-resource dialects by combining culturally appropriate system prompts, fine-tuning strategies, and stringent evaluation processes.

The research has limitations, especially in obtaining adequate and representative data for the Western Saudi dialect, despite the encouraging results. The complexity of Arabic and the variation within the dialect across different cities also posed challenges to building a consistent and natural chatbot experience.

Future work should focus on expanding dialect-specific datasets, addressing intra-dialectal variation, and exploring adaptive models that better generalize across regional differences. To further improve the caliber and versatility of AI-driven dialogue systems, future research might build on this foundation by investigating other dialects, expanding domain-specific applications, and improving assessment frameworks.

Acknowledgment

I would like to express my heartfelt gratitude to my project supervisor for her invaluable guidance, continuous support, and encouragement throughout the course of this research.

I am deeply thankful to my family and friends for their unwavering support, motivation, and patience during my academic journey.

I would also like to extend my sincere appreciation to Taibah University for providing an outstanding educational environment.

References

[1] S. Nithuna and C. A. Laseena, “Review on Implementation Techniques of Chatbot,” Proc. 2020 IEEE Int. Conf. Commun. Signal Process. ICCSP 2020, pp. 157–161, 2020, doi: 10.1109/ICCSP48568.2020.9182168.

[2] M. T. ZEMČÍK, “A Brief History of Chatbots,” DEStech Trans. Comput. Sci. Eng., no. aicae, 2019, doi: 10.12783/dtcse/aicae2019/31439.

[3] S. P. Reddy Karri and B. Santhosh Kumar, “Deep learning techniques for implementation of chatbots,” 2020 Int. Conf. Comput. Commun. Informatics, ICCCI 2020, pp. 20–24, 2020, doi: 10.1109/ICCCI48352.2020.9104143.

[4] Y. Saoudi and M. M. Gammoudi, “Trends and Challenges of Arabic Chatbots: Literature Review,” Jordanian J. Comput. Inf. Technol., vol. 9, no. 3, pp. 261–286, 2023, doi: 10.5455/jjcit.71-1685381801.

[5] A. M. Al-Abdullatif, A. A. Al-Dokhny, and A. M. Drwish, “Implementing the Bashayer chatbot in Saudi higher education: measuring the influence on students’ motivation and learning strategies,” Front. Psychol., vol. 14, 2023, doi: 10.3389/fpsyg.2023.1129070.

[6] “Chatbot |definition of chatbot in english by Lexico,” Lexico Dictionaries. [Online]. Available: https://www.dictionary.com/browse/chatbot.

[7] A. . dkk Trivedi, “Chatbot generation and integration: A review,” Int. J. Adv. Res., vol. 5, no. 2, pp. 1308–1311, 2019, [Online]. Available: www.IJARIIT.com.

[8] S. Singh and H. K. Thakur, “Survey of Various AI Chatbots Based on Technology Used,” ICRITO 2020 - IEEE 8th Int. Conf. Reliab. Infocom Technol. Optim. (Trends Futur. Dir., pp. 1074–1079, 2020, doi: 10.1109/ICRITO48877.2020.9197943.

[9] S. Hussain, O. Ameri Sianaki, and N. Ababneh, A Survey on Conversational Agents/Chatbots Classification and Design Techniques, vol. 927, no. October. Springer International Publishing, 2019. doi: 10.1007/978-3-030-15035-8_93.

[10] J. Skrebeca, P. Kalniete, J. Goldbergs, L. Pitkevica, D. Tihomirova, and A. Romanovs, “Modern Development Trends of Chatbots Using Artificial Intelligence (AI),” ITMS 2021 - 2021 62nd Int. Sci. Conf. Inf. Technol. Manag. Sci. Riga Tech. Univ. Proc., no. October, pp. 1–6, 2021, doi: 10.1109/ITMS52826.2021.9615258.

[11] B. J. Cahn, “CHATBOT : Architecture , Design , & Development,” 2017.

[12] J. Weizenbaum, “ELIZA-A computer program for the study of natural language communication between man and machine,” Commun. ACM, vol. 9, no. 1, pp. 36–45, 1966, doi: 10.1145/365153.365168.

[13] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” Mach. Learn. with Appl., vol. 2, no. November, p. 100006, 2020, doi: 10.1016/j.mlwa.2020.100006.

[14] K. M. Colby, S. Weber, and F. D. Hilf, “Artificial Paranoia,” Artif. Intell., vol. 2, no. 1, pp. 1–25, 1971, doi: 10.1016/0004-3702(71)90002-6.

[15] J. O. N. F. Heiser and R. C. Parkison, “CAN PSYCHIATRISTS DISTINGUISH A COMPUTER SIMULATION OF PAR , ANOIA FROM THE REAL THING ? THE LIMITATIONS OF TURING-LIKE TESTS AS MEASURES OF THE ADEQUACY OF SIMULATIONS * The computer program is a model of a mythical 28 year old , single , Caucasian , nat,” vol. 15, pp. 149–162, 1979.

[16] N. Jeevanandam, “Exploring Jabberwacky - a chatbot emulating a human conversation.,” INDIA AI. [Online]. Available: https://indiaai.gov.in/article/exploring-jabberwacky-a-chatbot-emulating-a-human-conversation

[17] M. S. Ali, A. Salam, A. Khan, A. Ali, and A. Ullah, “History of generative Artificial Intelligence (AI) chatbots : past , present , and future development,” arxiv, 2024.

[18] B. Harsh and A. Ms.Geeta, “Review on Different Types of Chatbots,” Int. Res. J. Mod. Eng. Technol. Sci., no. 05, pp. 1347–1349, 2021.

[19] R. Tamrakar and N. Wani, “Design and Development of CHATBOT : A Review,” no. April, 2021.

[20] T. Pillare, M. Chaoudhari, V. Hiwrale, S. Ade, P. Dudhe, and S. Chinte, “A SURVEY PAPER ON CHATBOT,” no. 10, pp. 2915–2919, 2023.

[21] A. Toprak, M. Turan, and F. S. Toprak, “INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING Deep Learning Based Chatbot Adapted to the Electronic Funds Transfer Process of Turkish Banking,” vol. 11, no. 1, pp. 356–362, 2023.

[22] H. Soffar, “Apple Siri features, use, advantages, disadvantages and using of Siri For Learning,” Online Sciences. [Online]. Available: https://www.online-sciences.com/technology/apple-siri-features-use-advantages-disadvantages-using-of-siri-for-learning/.

[23] S. I. Hussain, S. Deepalakshmi, R. J. Benilla, and V. C. Nivetha, “Automation of Smart Home using Smart Phone via Google Assistant,” Proc. - 5th Int. Conf. Smart Syst. Inven. Technol. ICSSIT 2023, no. Icssit, pp. 499–505, 2023, doi: 10.1109/ICSSIT55814.2023.10060979.

[24] “Google assistant, your own personal google,” google. [Online]. Available: https://assistant.google.com/.

[25] Z. M. Paul, “Cortana-Intelligent Personal Digital Assistant: a Review,” Int. J. Adv. Res. Comput. Sci., vol. 8, no. 7, pp. 55–57, 2017, doi: 10.26483/ijarcs.v8i7.4225.

[26] J. O. Sarfo, “Artificial Intelligence Chatbot – ChatGPT and High-Tech Plagiarism Concerns in a Digital Age: Is Detection Possible?,” J. Advocacy, Res. Educ., vol. 10, no. 2, pp. 55–58, 2023, doi: 10.13187/jare.2023.2.55.

[27] N. S. Reddy, N. P. Chaitanya, P. S. Varshitha, R. C. Varma, P. K. Reddy, and J. Dandu, “Intelligent Chatbot for Educational Institutions,” Proc. Int. Conf. Circuit Power Comput. Technol. ICCPCT 2024, vol. 1, pp. 1337–1343, 2024, doi: 10.1109/ICCPCT61902.2024.10673214.

[28] M. S. Abdul Razak, M. V. Manoj Kumar, C. R. Nirmala, R. Naseer, B. S. Prashanth, and H. R. Sneha, “Enhancing Student Support with AI: A College Assistance Chatbot Using NLP and ANN,” 2nd IEEE Int. Conf. Networks, Multimed. Inf. Technol. NMITCON 2024, pp. 1–8, 2024, doi: 10.1109/NMITCON62075.2024.10699279.

[29] S.-V. Fulgencio, “Developing Effective Educational Chatbots with GPT: Insights from a Pilot Study in a University Subject,” Trends High. Educ., vol. 3, no. 1, pp. 155–168, 2024, doi: 10.3390/higheredu3010009.

[30] M. Burkhard, S. Seufert, M. Cetto, and S. Handschuh, “Educational Chatbots for Collaborative Learning: Results of a Design Experiment in a Middle School,” Proc. 19th Int. Conf. Cogn. Explor. Learn. Digit. Age, CELDA 2022, no. Celda, pp. 209–218, 2022, doi: 10.33965/celda2022_202207l027.

[31] DeepSeek-AI et al., “DeepSeek-V2: A Strong, Economical, and Efficient Mixture of-Experts Language Model,” vol. 50000, pp. 1–52, 2024, [Online]. Available: http://arxiv.org/abs/2405.04434.

[32] J. P. Dhivvya and S. B. Karnati, “BuddyBot: AI Powered Chatbot for Enhancing English Language Learning,” 2024 IEEE Int. Conf. Interdiscip. Approaches Technol. Manag. Soc. Innov. IATMSI 2024, vol. 2, pp. 1–6, 2024, doi: 10.1109/IATMSI60426.2024.10502595.

[33] O. Baker, A. Rehman, O. Basas, and J. Liu, “Deep Learning for Intelligent Customer Service Automation: Development of GRU, LSTM, and Recurrent Neural Network Architectures for Chatbot Applications,” Proceeding - 2024 Int. Conf. Inf. Technol. Res. Innov. ICITRI 2024, pp. 111–117, 2024, doi: 10.1109/ICITRI62858.2024.10699245.

[34] D. Al-Ghadhban and N. Al-Twairesh, “Nabiha: An Arabic dialect chatbot,” Int. J. Adv. Comput. Sci. Appl., vol. 11, no. 3, pp. 452–459, 2020, doi: 10.14569/ijacsa.2020.0110357.

[35] R. Alotaibi, A. Ali, H. Alharthi, and R. Almehamadi, “AI Chatbot for Tourism Recommendations A Case Study in the City of Jeddah, Saudi Arabia,” Int. J. Interact. Mob. Technol., vol. 14, no. 19, pp. 18–30, 2020, doi: 10.3991/ijim.v14i19.17201.

[36] N. A. Al-Madi, K. A. Maria, M. A. Al-Madi, M. A. Alia, and E. A. Maria, “An Intelligent Arabic Chatbot System Proposed Framework,” 2021 Int. Conf. Inf. Technol. ICIT 2021 - Proc., pp. 592–597, 2021, doi: 10.1109/ICIT52682.2021.9491699.

[37] M. R. Douglas, “Large Language Models,” pp. 1–47, 2023.

[38] J. Gehrke, P. Lee, and H. Palangi, “Sparks of Artificial General Intelligence : Early experiments with GPT-4,” arxiv, 2023.