The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM)
ISSN: 2602-3199

The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 2024

Volume 27, Pages 246-262

IConTech 2024: International Conference on Technology

Conversational Agents: An Exploration into Chatbot Evolution, Architecture, and Important Techniques

Wassem Saad Nsaif
University of Diyala

Huda M. Salih
University of Diyala

Hassan Hadi Saleh
University of Diyala

Bashar Talib Al-Nuaimi
University of Diyala

Abstract: A chatbot is a computer software that mimics a user interaction. It is often referred to as a dialogue system or conversational agent. Developers and academics are increasingly focused in designing and implementing Chatbots. This in-depth look focuses on the ever-changing realm of chatbot technology, concentrating on chatbot evolution, architecture, and techniques that drive the most recent breakthroughs. We begin with a brief history and then follow the progress of Chatbots, emphasizing major milestones. The review focuses on the many architectures used in chatbot creation, ranging from classic rule-based systems to cutting-edge machine learning (ML) and natural language processing (NLP) approaches. We analyze the present status of chatbot technology and its breakthroughs, which include advances in NLP and interpersonal interactions. While demonstrating effective industry practices, we also discuss the architecture of software. Recommendations are made for academics, developers, and enterprises, identifying possible areas for future investigation and development in this quickly changing industry. The paper finishes by projecting future trends and developments in chatbot development.

Keywords: Chatbot, Artificial Intelligence (AI), Natural language processing (NLP), Machine learning (ML)

Introduction

A chatbot is a computer program that mimics and interprets spoken or written human communication, enabling people to engage with digital gadgets in the same way they would with a real person (Berry, 2023) . Chatbots can be simply defined as one-line programs that respond to basic questions, or they can be as complex as digital assistants that learn and develop over time to provide ever-more-personalized services as they collect and analyze data (Kalla, 2023). Chatbots are being used in a range of industries and applications, ranging from entertainment to education, e-commerce, and healthcare. chatbots, like jessie humani, and mitsuku can offer both support and fun to users (Joseph, 2023) . these "small talk" chatbots may encourage a sense of social connection. chatbots appear to be more engaging than static frequently asked questions (faq) pages on websites. chatbots are more efficient and cost-effective than human customer assistance since they can help several users at once. chatbots can provide customer service, entertainment, and connection to the end user (Lin, 2023).However, users' engagement and confidence in Chatbots are influenced by their amount of embodiment

- Selection and peer-review under responsibility of the Organizing Committee of the Conference

(human-likeness) and disclosure (how and when the bot's presence becomes apparent to the user) (Alabed, 2023; Suhaili, 2021).

Chatbots have become more popular in recent years due to increased processing power and open source technology. Advancements in Artificial Intelligence (AI) and NLP have made Chatbots more user-brotherly (Kar, 2016) , adaptable, and capable of mimicking human conversation (Suhaili, 2021) . Chatbots have become more popular in recent years due to higher computational capacity and open-source technological advances. Chatbots use conversation systems to efficiently manage tasks including gathering information, directing requests to the appropriate channels (Gandon, 2018) , and delivering customer service (Dhola, 2021). Some Chatbots utilize sophisticated NLP (Arora, 2016) and word classification algorithms to understand and interpret user input Amodu et al. (2023) . These Chatbots comprehension the context and complexities of the conversation, leading in more accurate and detailed responses. Some Chatbots, on the other hand, employ a simpler method of searching for generic keywords and constructing responses from pre-defined terms in a library or database Cho Wang et al. (2023). Virtual assistants and website popups are the primary methods for contacting Chatbots online Ashari et al. (2014). Virtual assistants, such as voice-activated Chatbots, provide engaging conversational experiences on smartlaptop s and smart speakers Shamir and Gilad (2023).

Website popups, on the other hand, are chatbot interfaces that display on websites and allow users to have text-based discussions Nirala, et al. (2019).These two communication strategies have a wide range of applications, including business (for example, e-commerce help), education, entertainment, finance, health, news, and productivity Grudin et al. (2019) . Chatbots can be classed in several ways depending on certain characteristics. Here are some popular chatbot categories. *Rule-based* Chatbots are those that obey a set of rules or scripts and are hence referred to as "rule driven" Chatbots. They generate predefined replies based on phrases or patterns found in user inputs. Rule-based Chatbots are basic and have limited functionality because they can only answer to a specific set of preset questions Caldarini et al. (2022) . Chatbots based on machine learning evaluate and respond to human input using AI and NLP techniques. They can refine their replies over time by observing user interactions. *ML-based Chatbots* are more complex and can answer a wider variety of questions than rule-based Chatbots Maher et al. (2020) . *Retrieval-based Chatbots* answer using predetermined responses stored in a database. They do not generate new replies, but instead select the most appropriate one depending on user input. Retrieval-based Chatbots are commonly used for certain activities or places where predetermined replies are available Horacek, Helmut. "Natural language processing, 1990 . *Generative Chatbots* generate their own replies rather than depending on preprogrammed ones. They may exhibit innovative, interactive behaviors in response to human stimulation. Generative Chatbots give more flexible and unique replies than retrieval-based Chatbots, but they are more complex to design and train Miller (2024). *Task-oriented Chatbots* are Chatbots that are trained to do certain tasks, such as organizing appointments, providing customer service, or making bookings Al-Amin et al. (2024). . They have a limited functional range and are focused on completing a certain task. Conversational Chatbots are meant to converse with humans in an open-ended manner. They may not have specific jobs or responsibilities because their primary goal is to encourage conversational interactions. Conversational Chatbots are commonly used for amusement, customer engagement, and social interaction Dsouza et al. (2019) . The rest of this article is organized as follows: Section 2 provides some background on Chatbots and their evolution through time, Section 3 describes the Chatbots architecture, Section 4 presents an analysis of the state of the art in terms of Chatbots techniques, in Section 5 we will discuss the challenges and we conclude the paper in Section 6.

Evolution of Chatbots

The term "chatbot" may have gained popularity recently, but the idea has been around since people began creating computer-mediated communication methods. Chatbots have a history dating back to the 1960s, when the first such technology assistant was developed. Joseph Weizenbaum created ELIZA in 1966 1 . It could recognize certain phrases and reply accordingly, simulating a conversation with a person. Since 2000 more and smarter bots have been created. They use natural language processing to 'understand' human speech and respond appropriately. Conversational systems have grown alongside advancements in computing and NLP techniques. The first chatbot, ELIZA, was developed in 1966 and used linguistic rules and pattern-matching algorithms (Berry, 2023; Nirala & Kumar, 2022). It can converse with users using a keyword-matching system. The algorithm looks for a suitable transformation rule to reformulate the input and return an answer to the user. Eliza was a milestone system that encouraged additional research in the field. Elıza's knowledge was limited due to its reliance on basic context identification and the inflexibility of pattern-matching rules for implementing new domains Seering, Joseph et al, 2019 . Chatbots are a type of AI that interpret and converse with users by using NLP Tulshan et al. (2018). They are often used in a variety of platforms and apps to interact with users, respond to inquiries, and offer automated support. Over the years, Chatbots have had a remarkable transformation, progressing from simple pattern-matching algorithms to complex conversational agents utilizing cutting-edge technology. Their past reveals an amazing development in NLP and AI. Fig. 1, shows the major stages of chatbot development.

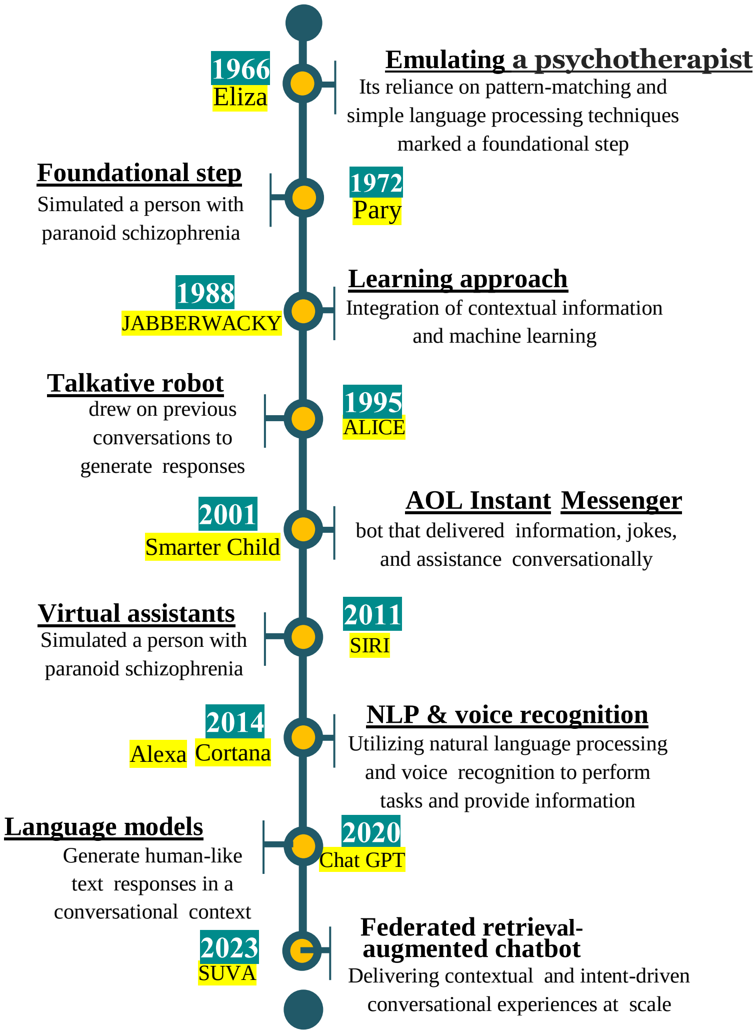

Figure 1. Chatbots through time

In the mid-1960s, the first bot ELIZA appeared as an early NLP software that mimicked a psychiatrist. Its focus on pattern-matching and basic language processing techniques was a critical advance Hadi, Muhammad Usman, et al, 2023 . In 1972, the PARRY software emulated a person with paranoid schizophrenia, providing insight on the difficulty of replicating complicated human cognitive processes Yogesh et al, 2021 . The learning methodology gained prominence in 1988 with JabberWacky, which used previous talks to create replies, demonstrating the combination of contextual knowledge with ML Han et al. (2021) . In the middle of 1990, A.L.I.C.E., a talking robot, used pattern-matching and rule-based approaches to simulate conversation Luo et al. (2022) . In 2001, CHILD, an AOL Instant Messenger bot that gave knowledge, jokes, and assistance conversationally, made its debut Horacek, Helmut, 1990 . Then, the following years introduced virtual assistants like SIRI (2011), Cortana (2014), and ALEXA (2014), each built by big tech firms, using NLP and voice recognition to do tasks and offer information Miller, James R, 1981 . ChatGPT made its public release in 2020, marking a big step forward in language models. OpenAI built it on the GPT-3.5 architecture with the goal of producing human-like text answers in a conversational setting Reiter (2007) . By 2023, ChatGPT had advanced significantly, leveraging the capabilities of retrieval augmented generation, ML, NLP, natural language question answering (NLQA), generative AI, and an insights engine. This enabled it to address customer and employee support inquiries 24/7, demonstrating a comprehensive integration of several AI technologies Firdaus et al. (2020) .On June 15, 2023, SearchUnify introduced SUVA. The world's first federated retrieval-augmented chatbot, SearchUnify Virtual Assistant (SUVA), provides contextual and intent-driven conversational interactions at scale. SUVA, which is powered by large language models (LLMs), uses

ML, NLP, NLQA, and retrieval-enhanced generation to provide secure and tailored customer and staff assistance queries Skantze (2007) .

Chatbot Architecture

To improve system design, it's important to first understand its structure and architecture. The architecture is often represented by its components and their interactions Lee et al. (2010) . We did a thorough review of survey papers, publications, and journals. We discovered several conversational agents, each with its unique architecture. For example, the Amazon Alexa bot uses user data, such as speech, to generate Automatic Speech Recognition (ASR) about the Amazon ASR service. Then, process the supplied data about various Amazon Web Services (AWS) and use Amazon DynamoDB to store the conversations and their states Adamopoulou and Moussiades (2020). Google Assistant is the most successful bot of all. Which receives recordings from users and sends them to Google's servers for processing. It breaks down the voices into component sounds and tries to match each sound with the most comparable word's pronunciation. A standard chatbot architecture is made up of five major components, see fig.2. A user interface, a natural language understanding (NLU) component, a dialogue management (DM) component, a backend component, and a response generation (RG) component Khanna and Pandey (2019), as shown in the image below:

- **User Interface** The user interface enables users to converse and engage with Chatbots using messaging apps such as Facebook Messenger, Cortana, and Slack. The operation of a chatbot begins with a user request Khanna, and Pandey (2019) .

- **Natural Language Understanding:** When the system gets a user request, the NLU component extracts information and creates a representation of its meaning for subsequent use Abdul-Kader and Woods (2015) . NLU focuses on three tasks: conversation act categorization, intent classification, and slot filling Arsovski and Osipyan (2019) . Conversational agents reply by giving a semantic representation for user utterances Hahm et al. (2018). , such as logic or class purpose, and extracting the "meaning" of an utterance de Melo and Hose (2013) . Parsing is the primary function of an NLU, which takes a string of words and provides a linguistic framework for the speech. The mechanism used by an NLU to parse input is implementation-dependent, and it can use context-free grammars, pattern matching, or data-driven methods. NLU outputs must be manageable by a conversation manager Bollacker et al. (2008) .

- **Dialogue Management (DM)**: The Dialogue Management component manages information from other components, updates discussion context, and governs the chatbot's activities Sean (2010).

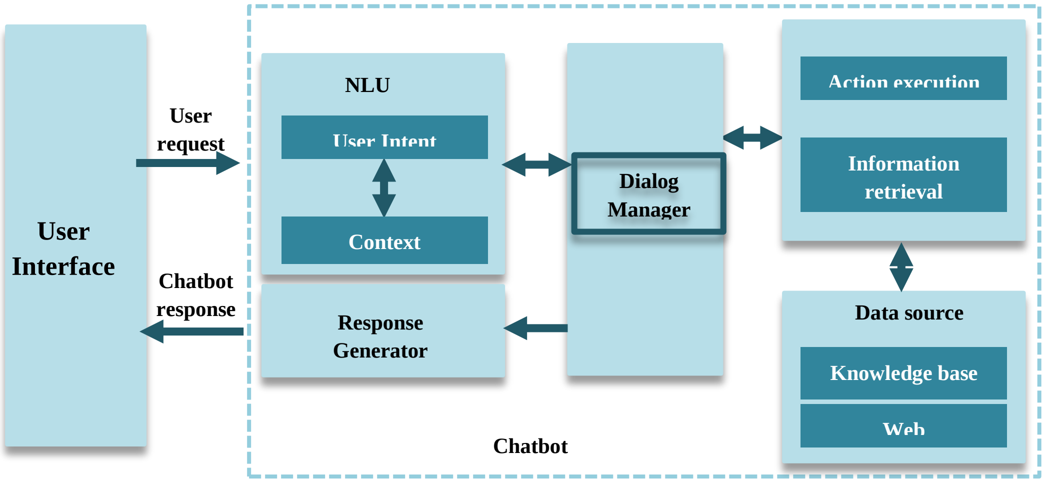

Figure 2. Operational mechanics and architectural components of chatbots

Dialogue Manager is the second key component of every chatbot, and we can distinguish between Chatbots using this component, which has various aspects that may be modified or added in the future if it is discovered to serve the DM. DM accepts user input from the NLU and generates system replies at the concept level for the natural language generation (NLG). The reaction that the DM will pick is determined by the approach that was chosen. Strategies involve preserving a conversational state and modeling the discourse structure beyond a single statement Khanna and Pandey (2019) . The methodologies include rule-based, knowledge-based, retrieval-based, and generative. The rule-based techniques include background, intent templates, and entity-based templates, which are arranged in order of priority. Because rule-based techniques encode human knowledge into templates, they produce the most accurate results. If one of these tactics recognizes the input, the system will respond with a template. If there is no matching template for the input, the system may attempt to obtain a response from a knowledge-based question answering (Q/A) engine Adamopoulou and Moussiades (2020). Otherwise, the input is processed by a collection of neural network models and information retrieval modules to provide a generic conversation output.

- **Backend:** Chatbots gather information from the backend and send messages to the Dialogue Management and Response Generation components . Rule-based Chatbots require a Knowledge Base (KB) to hold their own rules. To ensure the chatbot's resilience, the Knowledge Base rules should be broad and comprehensive Abdul-Kader and Woods (2015). .A chatbot can utilize a Relationship Data Base (RDB) to retrieve previous interactions. Using past knowledge improves the chatbot's consistency, precision, and reliability Arsovski S., Osipyan H., 2019 . Developing the knowledge base (KB) can be time-consuming and labor-intensive due to manual effort. To address this challenge, engineers created a method for automatically creating a new KB from an existing chatbot Hahm et al. (2018). A software may convert a corpus into an AIML knowledge base Bollacker et al. (2008) . Rule-based Chatbots often employ user replies to direct and fulfill knowledge base questions Motta (2010) . Information is increasingly being saved digitally on the World Wide Web and other online sources. Large repositories contain information in machine-readable and accessible formats Pool, Steven, and Brian Pool, 2007 . Examples of knowledge bases are Google's Knowledge Graph et al. (1955) and DBpedia , Freebase, and Wolfram Alpha . Jacobs and Bean (1963). Knowledge Graphs were launched in 2012. Structured information on a topic or summary is provided by crowd-sourcing and manual duration of data. DBpedia uses structured data from Wikipedia and is accessible online. These repositories often have subject-predicate-object triples and a graphical framework to display them. The nodes represent entities, such as topics and objects, whereas edges represent their relationships.

- **Response Generation:** Once the appropriate information has been retrieved, the next step for the dialogue system is to determine the content of the response and the best way to express it. The Response Generation component is responding for generating responses in user understandable format. The final key component of any chatbot. It gets a communicative act from the DM and produces a corresponding textual representation. The NLG must execute two functions: content planning (Content Filter | Engagement Ranking) and language generation (using only text or speech via Text to Speech). After going through one or more of the DM's techniques, the pipeline moves on to the reply generator Yorozu et al. (1987), This generator will first apply a content filter to exclude incoherent or problematic candidates. If there are several legitimate responses, a ranking procedure is utilized to sort candidate utterances, first by priority, and then by engagement ranking. Finally, the chatbot sends the selected utterance as a text in the final output.

Chatbot Techniques

- Domain Adaptive Pre-training (DAPT): is a simple technique. Pre-training the model on an extremely small corpus or particular task can yield substantial advantages. In addition to working on ever-larger LMs, it can be beneficial to find and employ domain- and task-relevant corpora in tandem to specialize models Baevski and Edunov (2019) . This model can investigate additional adaptability by extending the pre-training of big LM into two types of unlabeled data: (i) huge domain-specific text corpora Beltagy et al .(2019) and (ii) readily available unlabeled data related to a specific task Lewis et al. (2020) .

- Retrieval Augmented Generation (RAG): pre-trained endow, parametric-memory generation models with a non-parametric memory through a general-purpose fine-tuning technique which we refer to as retrieval-augmented generation (RAG) Payal Bajaj, Daniel Campos, 2016 . Using the same retrieved document to generate the entire sequence, technically, the model treats the retrieved document as a single latent variable that is pooled to obtain a probability seq2seq p(yx) using a top-K approximation Devlin and Chang (2019) . Specifically, the top-K documents are retrieved using the retriever and the generator produces the resulting sequence probability for each document, which is then marginalized Hussain et al. (2019) .

- Parsing: In order to ascertain the semantic structure of the text or the relationships between its terms, it transforms the text into a meaningful representation string. Lexical parsing is one type of parsing approach that can be used to facilitate manipulation and extract information from text by breaking it down into simpler atomic words. Following the application of lexical parsing, syntactical and semantic parsing can be used. By translating text into a machine-understandable representation of its meaning, these two parsing approaches ascertain the grammatical structure of a phrase and extract a particular meaning Cooper (2020) . "Set your eyes on my brother" and "could you see my brother," for instance, would both produce the identical parsed version of "see my brother." Additionally, this method assists in locating the ambiguity so that a user can be asked to restate his input Ramesh et al. (2017). As an illustration, there are two ways you could understand)the line "I saw my brother with my laptop ": 1) Did I see my brother with my laptop ? ; 2) Did I see my brother with my laptop in his hand?

- Matching patterns: Using this method, Chatbots generate responses with patterns when they are created manually, which is a laborious procedure. Even though it speeds up response times, the responses could be monotonous and repetitive, which makes for boring conversations devoid of spontaneity and the human touch Motgeret al. (2022) . The chatbot, for instance, can identify terms such as "books" and "office supplies" in response to the input "Where the stickers are?" It can also identify other inputs like "different types of supplies in the office supplies aisle".

- Artificial Intelligence Markup Language (AIML): Chatbots use a variety of technologies, such as AIML, to use a pattern matching approach syntax to find the best possible response Adamopoulou and Moussiades (2020). Derived from the Extensible Mark-up Language (XML), AIML is an open standard language. Topics and categories are the two components that make up AIML data objects. A category is a rule that matches a template for input and an optional pattern for response. A subject is an optional top-level element that has a group of related categories. AIML files are used to sort the objects. It must offer a pattern for every possible response and update it regularly, which cannot be done automatically, despite its readability, usefulness, and efficient use of response time Sutskever et al.

- RNNs. The chatbot can handle sequential data and take into account the input from the present users thanks to the RNNs. It retains the input from prior users because of the inherent limited memory. Put another way, an RNN allows data to stay, in contrast to a standard neural network. RNNs work on the basic principle of storing an output from one layer and using it as an input for the subsequent layer in order to predict the outcome S. Hochreiter (1993). For certain applications, the unaltered version of RNN is inappropriate because of the vanishing or expanding gradient problem Chung et al. (2022) . Two distinct approaches addressing this issue are Gated Recurrent Units (GRU) Cho et al. (2014) and Ma et al. (2023) , LSTM Jozefowicz et al. (2015).

- LSTM. An unique variety of RNN is the LSTM Jozefowicz et al. LSTM addresses the vanishing or ballooning gradient issue in RNN and is built to manage long-term dependency. Consequently, the LSTM introduces gates. The main element of an LSTM is a gate, which determines which information is remembered. The gates also output a value between zero and one, where zero denotes memorizing nothing and one denotes allowing everything to go to the next stage. In addition, LSTM provides three different types of gates to regulate the information flow: forget gates, output gates, and input gates. The forget gates choose which data should be committed to memory, whereas the input gates handle the state update procedure. The output from the hidden layer is also determined by the output gate. These three gates make up the LSTM memory cell. Gender is among the things the LSTM can remember because it was designed as a short-term memory solution. Thus, the chatbot can use "his/her" based on the previously recalled input. An alternative LSTM architecture called BiLSTM takes into account input coming from the opposite direction as well Vinyals and Le (1964) . In addition, the LSTM and RNN's primary rival is the GRU Cho et al. (2014) and Ma et al. (2023) . Its architecture makes it less complex and more popular than LSTM. One "update gate" is created by combining the input and forget gate.

- Sequence to Sequence: The first design to be developed to tackle translation concerns is the Seq2Seq structure, and its success is encouraging for NLG. Diverse datasets and domains are used to train the seq2seq end-to-end. Additionally, seq2seq is the industry standard structure because of its versatility, ease of use, and generality in solving many NLP tasks Weizenbaum (1966) . In theory, seq2seq is made up of an encoder and a decoder, two RNNs. Word by word, the Encoder interprets the user's input, and word by word, the Decoder creates the response based on previously had discussions. When creating Chatbots, the challenge was to translate user input to the chatbot's response rather than translating across languages. One benefit of seq2seq structure over others is that the lengths of the input and response sequences can also vary. Depending on table 1 below, Papers Makatchev et al. (2010) and Liu et al. (1971) discussed Rules and patterns matching technique as a set of predefined human made rules with pros Easy and less expensive implementation Fast deployment. No overtime to understand the intent of the user. Cons is unable to learn on their own. Unable to react outside its preconceived notion.

- To map words to actual umber vectors, a language modelling and feature extraction technique was used. Kim and Kwon (2020) Combination of generation and retrieval-based approaches which "Xiaolce: more popular example from Microsoft". In Combination of generation and retrieval-based approaches, Tran and Nguyen (2020) "Proposed PS, GP. And PRFDevlin et al. (2022) and Toutanova (2022) "Develop matching method based on the seq2seq, Radford et al. (2018). "Develop a model using the Twitter LDA model and attention mechanism Song and Wang (2022) "Integrate AIML technique with a SNC model", Albeladi et al. (2023) "Multi-strategy process including LSTM with an attention mechanism beside rule-based technique". Lewis et al. (2020).Domain Adaptive Pre-training (DAPT) technique where text modeling for upcoming tasks inside the domain is enhanced by domain adaptive pre-training, which is the ongoing unsupervised pre-training of a language model on text unique to the domain. Finaly, Retrieval Augmented Generation (RAG) technique where models that blend non- parametric and parametric memory that has been trained beforehand to generate language Lin et al. (2023)

- Also in Rule –based Jia (2009) and Moubaiddin et al. (2015) and Belgaumwala (2019) discussed the Parsing technique is Converting a text to be less complicated terms. The pros is Providing the text's semantic structure or the dependence relationships between terms. The cons is the same guidelines and trends that correspond with drawbacks. The Corpus based the articles Candra (2017) and Noori et al. (2014), describe Pattern matching technique which predefined structures of responses, pros is adequate for basic jobs, Pick insightful answers from the list of candidate answers. Greater adaptability compared to rule-based, cons is providing Chatbots devoid of intelligence and relativity, Repetitive responses and restricted capacities.

Mavridis et al. (2019) describe AIML technique Represents the knowledge as objects which derived from XML, pros is benefits of matching patterns strong in creating a conversational flow, cons are developing every potential motif by hand Very challenging to scale. Generation –based has linear support vector machine (LSTM) technique which the papers Zhang et al. (2022) and Shawar and Atwell (2004) describe the CNN that usually used for learning features automatically by utilizing convolution and pooling processes. Shawar and Atwell (2004), depict GRU is a type of RNN technique related with LSTM. Shawar (2011) , RGDA based on gradient reinforcement learning Stacked LSTM method that teaches computers to make decisions in a way that produces the best outcomes. It simulates the process of learning by making mistakes. Kadeed (2014) , characterize Stacked LSTM that a deeper and more abstract model is produced by stacking many hidden LSTM layers on top of one another in a LSTM. Two concealed LSTM layers in the opposite direction make up a bidirectional LSTM, which uses information from both sides. Ali and Habash (2020) characterize Stacked LSTM and BILSTM which is the input is followed in both directions. Al-Ghadhban and Al-Twairesh (2020) HRED generates context and response.

In hybrid technique Palasundram et al. (2019) paper expressing the “Attention” which is attention method, in which the encoder assigns attention weights to each concealed state. The computation of these weights establishes the relative importance of an encoder state to a decoder state in producing the subsequent state based on the energy associated with each weight. In seq2seq learning based on encoder- decoder architecture, Al-Madi et al. (2021) based on RNNs cell and improved technique of seq2seq learning. Zhang et al. (2019) Mapping a sequence of input words to another representation of response sequence. TNaous et al. (2019) GRU is in compression to LSTM. GRU required fewer parameters training and it not being required for an additional cell state. Hu et al. (2022) paper expresses RNN. Attention mechanisms Improved technique of seq2seq learning based on RNNs cell. Resolve the problem of systems incapability to remember a longer sequence. Prassanna (2020) "Enhancement of RNN-GRU" which "This technique is improved by adding three additional cells which are Refinement, Adjustment, and output cells Zhang and Dinan (2022), Boussakssou and Ezzikouri (2022). Pre-trained GPT-2, DIALOGPT, BoB, aubmindlab, CakeChat, asafaya is

Chabot Challenges and Limitations

Chatbot creation faces numerous challenges, including linguistic, technical, and user experience issues. These include NLU, handling ambiguity, generating responses, preventing user dissatisfaction, supporting multiple languages, managing dynamic contexts, ensuring data security and privacy, enhancing personalization, integrating chatbots with systems, identifying user intent, and managing memory retention Chaves et al. (2021) . Efficient assessment, transplatform coherence, and ethical considerations like bias and fairness are also crucial. Adapting to new information requires robust learning techniques Chiang et al. (2024) . A multidisciplinary approach involving domain-specific knowledge, machine learning, NLP, and user experience design is needed to overcome these challenges. Continuous research and developments in AI technology can help reduce these challenges and enhance chatbot performance. Overall, overcoming these challenges is crucial for the development of chatbots Sudalairaj et al. (2024) . Limited flexibility, inflexible responses, complex scalability, managing ambiguity, dependence on rule quality, absence of learning capabilities, expensive and resource-intensive maintenance, difficulty managing variability, restricted generalization, scalability problems, user experience issues, reliance on expertise, and limited adaptability are some of the challenges that rule-based chatbots must overcome Naık et al. (2023). Because of these difficulties, rule-based systems are best suited for certain applications with clearly defined domains and predictable user behavior.

Table 1. Approaches and techniques for Chatbot

Approach		Techniques	Description	Advantages	Disadvantages	Articles
Rule –based		Rules and patterns matching	Set of predefined human -made rules	Easy and less expensive implementationFast deployment. No overtime to understand the intent of the user	Unable to learn on their own. Unable to react outside its preconceived notion	Makatchev et al. (2010),Weizenbaum (1966) , Colby(1971)
Rule –based		Parsing	Converting a text to be less complicated terms	Providing the text's semantic structure or the dependence relationships between terms	The same guidelines and trends that correspond with drawbacks	Jia (2009) , Moubaiddin(1962) , Belgaumwala, (2019)
	(Retrievalbased)	Pattern matching	predefined structures of responses	adequate for basic jobs, Pick insightful answers from the list of candidate answers.Greater adaptability compared to rule-based	Providing chatbots devoid of intelligence and reativity.Repetitive responses and restricted capacities.	Candra and Noori(2014)
	(Retrievalbased)	AIML	Represents the knowledge as objects which derived from XML	Benefits of matching patterns Strong in creating a conversational flow	Developing every potential motif by hand Very challenging to scale	Mavridis and AlDhaheri, (2011), Roca and Sancho (2020)
		CNN	CNN usually used for learning features automatically by utilizing convolution and pooling processes	LSTMs are beneficial for managing long sequences in time series analysis and natural language processing due to their ability to retain information from previous time steps. They also help solve the vanishing gradient problem by regulating information flow across the network.	Long Short-Term Memory (LSTM) models have drawbacks such as computational complexity, overfitting, hyperparameter tuning, and limited interpretability. They require more computation than feedforward networks or basic RNNs, and are prone to overfitting when there is insufficient training data.	Zhang and Chen (2021), Zhang and Qin (2022), Shawar and Atwell (2004)
		CNN and GRU	GRU is a type of RNN technique related with LSTM .	CNN offers high accuracy in machine learning tasks, while GRU networks are faster and less computationally expensive. They manage long-term dependencies in sequential data selectively.	CNN design is time-consuming and labor-intensive, and GRU models face low learning efficiency and slow convergence rate issues.	Shawar and Atwell (2011), Wang (2022).
Corpus based	LSTM	RGDA based on gradient reinforce ment learning	method that teaches computers to make decisions in a way that produces the	Reinforcement learning has the following benefits: Performance maximization. Maintain Change for an extended length of time. An excess	It is preferable to use reinforcement learning to solve difficult problems rather than simple ones. It demands a great deal of work and a large amount	Shawar and Atwell (2011) ,Wu et al. (2020), Yu et al. (2019).

Stacked LSTM	best outcomes. It simulates the process of learning by making mistakes.	of states resulting from reinforcement can taint the outcomes.	of data. The cost of maintenance is substantial.	
Stacked LSTM	A deeper and more abstract model is produced by stacking many hidden LSTM layers on top of one another in a stacked linear support vector machine (LSTM). Two concealed LSTM layers in the opposite direction make up a bidirectional LSTM, which uses information from both sides.	To effectively capture intricate patterns and long-term dependencies in sequential data, numerous LSTM layers may be used.	In order to guarantee efficient training and avoid overfitting, it also adds more complexity and calls for cautious tuning and regularization.	T. Kadeed, 2014 , Ma, M., Liu, C., Wei, R., Liang, B., & Dai, J. (2022) , Jørgensen, R. K., Hartmann, M., Dai, X., & Elliott, D. (2021).
Stacked LSTM and BILSTM	BiLSTM: The input is followed in both directions.	Reliable on a bigger dataset and useful for recalling lengthier sequences	makes use of a lot of parameters, more memory size was needed, Extended execution duration and intricate complexity	Ali and Habash (2020)
HRED	HRED generates context and response			Al-Ghadhban and Al-Twairesh (2020)
Attention	using the attention method, in which the encoder assigns attention weights to each concealed state. The computation of these weights establishes the relative importance of an encoder state to a decoder state in producing the subsequent state based on the energy associated with each weight.	Using a BiLSTM network with attention, they map each user statement to an EPA (Evaluation Potency Activity) vector. They then provide a corresponding EPA response vector that is used to condition the response generation.	BiLSTM is a far slower model that takes longer to train. As a result, they advise against utilizing it unless absolutely necessary.	Palasundram et al.(2019)
seq2seq Seq2seq learning	based on RNNs cell and improved	"Same advantages of tachniques that seq2seq	"Same disadvantages of tachniques that seq2seq depending on it"	Al-Madi et al. (2021)

based on encoder-decoder architecture	technique of seq2seq learning Mapping a sequence of input words to another representation of response sequence.	depending on it Support variable-length size of input and response"		Zhang et al . Wang and Koji (2018)
GRU	In compression to LSTM. GRU required fewer parameters training and it not being required for an additional cell state	Uses less training parameter Uses less memory Take less time in execution Less complex structure"	"Not suitable for large Dataset Not suitable for long- distance relations"	Naous et al. (2019)
RNN	"Attention mechanisms Improved technique of seq2seq learning based on RNNs cell. Resolve the problem of systems incapability to remember a longer sequence"	Uses less training parameter Uses less memoryTake less time in execution - Less complex structure"	Suffer from gradient exploding and vanishing problems. Difficult to process very longer sequences"Not suitable forparallelizing or stacking up"	Hu et al. (2022)
"Enhancement of RNN-GRU"	"This technique is improved by adding three additional cells which are Refinement, Adjustment, and output cells"	Uses less training parameter Uses less memoryTake less time in execution - Less complex structure"	Suffer from gradient exploding and vanishing problems. Difficult to process very longer sequences"Not suitable forparallelizing or stacking up"	Prassanna (2020)
Pre-trained GPT-2, DIALOGPT , BoB, aubmindlab, CakeChat, asafaya,	To map words to actual umber vectors, a language modeling and feature extraction technique was used.			Zhang et al. (2018) , Boussakssou, (2020)
Combination of generation and retrieval-based approaches	"Xiaolce: more popular example from Microsoft"	"Easy to select the attributes (relevance) from ranked features list."	If the hybridization technique is not complementary to each other, the performance quality may decrease	Kim and Kwon (2020)
Combination of generation and retrieval-based approaches	"Proposed PS, GP. And PRF" "Develop matching method based on the seq2seq"	"Easy to select the attributes (relevance) from ranked features list."	If the hybridization technique is not complementary to each other, the performance quality may decrease	Tran and Nguyen Devlin et al. (**2022**

	"Develop a model using the Twitter LDA model and attention mechanism"			Radford et al.(2018)
	"Integrate AIML technique with a SNC model"			Song and Wang (2022)
	"Multi-strategy process including LSTM with an attention mechanism beside rule-based technique"			Albeladi et al. (2023)
Domain Adaptive Pre-training (DAPT)	Text modeling for upcoming tasks inside the domain is enhanced by domain adaptive pretraining, which is the ongoing unsupervised pretraining of a language model on text unique to the domain.	focus on adapting to multiple languages within a specific domain. they propose different techniques to compose pretraining corpora that enable a language model to both become domain-specific and multilingual.	MDAPT is a complex, resource-intensive, and multilingual model that requires careful consideration of linguistic and domain-specific characteristics. It also faces challenges in fine-tuning and requires sufficient data for effective performance across multiple languages and domains.	Lewis et al. (2020)
Retrieval Augmented Generation (RAG)	models that blend non-parametric and parametric memory that has been trained beforehand to generate language	They present RAG models, in which a pre-trained neural retriever accesses a Wikipedia dense vector index as the non-parametric memory and a pre-trained seq2seq model serves as the parametric memory.	RAG is a resource-intensive model that relies on a large dataset for information retrieval. Its effectiveness depends on the quality of the information retrieved, as missing or erroneous data can lead to generation mistakes. RAG is also dependent on a pre-existing set of fixed retrievals, making it challenging to manage dynamic data.	Lin et al. (2023)

Furthermore, they have trouble with a variety of unexpected or diverse user inputs, which makes it challenging to update and maintain the rule base. Moreover, rule-based systems are incapable of learning, which makes it challenging for them to adjust to changing user requirements or shifting linguistic trend Sanders et al. (2023) . To get around these restrictions, hybrid systems that combine learning-based and rule-based strategies have been developed. The challenges faced by generative chatbots, which employ NLG techniques, include handling open-ended questions, avoiding biases and inappropriate content, producing creative but inaccurate responses, handling domain-specific knowledge, and handling lack of context. In addition, they encounter constraints on data efficiency and training sets because of enormous training sets, response time management, unclear user intent, and challenges in responding to user feedback. Another difficulty is achieving real-time responsiveness since some models could take longer to generate. Determining trustworthy measures to assess the caliber of produced answers is still a challenge Spivack et al. (2024) . In order to improve the overall performance and dependability of generative chatbots, advances in NLP, model designs, and ethical considerations are required

Conclusion

This research paper examines the techniques of chatbots, a type of computer software that mimics user interactions. Chatbots are used in various industries, such as entertainment, education, e-commerce, and healthcare. They can be one-line programs or complex digital assistants that learn and develop over time. Advancements in AI and NLP have made chatbots more user-friendly and adaptable. They use conversation systems to efficiently manage tasks, such as gathering information and delivering customer service. Chatbots can be classified into rule-based, machine learning-based, retrieval-based, generative, task-oriented, and conversational chatbots. The evolution of chatbots dates back to the 1960s, with more sophisticated bots created since 2000. Techniques used include parsing, pattern matching, AIML, and RNNs, DAPT where text modeling for upcoming tasks inside the domain is enhanced by domain adaptive pretraining, which is the ongoing unsupervised pretraining of a language model on text unique to the domain, and RAG where models that blend non-parametric and parametric memory that has been trained beforehand to generate language.

Scientific Ethics Declaration

The authors declare that the scientific ethical and legal responsibility of this article published in EPSTEM Journal belongs to the authors.

Acknowledgements or Notes

* This article was presented as an oral presentation at the International Conference on Technology (www.icontechno.net) held in Alanya/Turkey on May 02-05, 2024.

References

Abdul-Kader, S. A., & Woods, J. C. (2015). Survey on chatbot design techniques in speech conversation systems. *International Journal of Advanced Computer Science and Applications*, *6*(7), 72-80.

Abras, S., ThawraKadeReed, S. P., & Alani, L. (2014). *Construction of arabic interactive tool between humans and intelligent agents.* Retrieved from https://www.researchgate.net/profile/Thawra-Kadeed/publication.pdf

Adamopoulou, E., & Moussiades, L. (2020). Chatbots: History, technology, and applications. *Machine Learning with Applications*, *2*, 100006.

Alabed, A., Javornik, A., Gregory-Smith, D., & Casey, R. (2024). More than just a chat: A taxonomy of consumers' relationships with conversational AI agents and their well-being implications. *European Journal of Marketing*, *58*(2), 373-409.

Al-Amin, M., Ali, M. S., Salam, A., Khan, A., Ali, A., Ullah, A., ... & Chowdhury, S. K. (2024). History of generative Artificial Intelligence (AI) chatbots: Past, present, and future development. *arXiv preprint arXiv:2402.05122*.

Albeladi, K., Zafar, B., & Mueen, A. (2023). Time series forecasting using LSTM and ARIMA. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *14*(1).

Aleedy, M., Shaiba, H., & Bezbradica, M. (2019). Generating and analyzing chatbot responses using natural language processing. *International Journal of Advanced Computer Science and Applications*, *10*(9).

Al-Ghadhban, D., & Al-Twairesh, N. (2020). Nabiha: An Arabic dialect chatbot. *International Journal of Advanced Computer Science and Applications*, *11*(3), 0110357.

Ali, D. A., & Habash, N. (2016). Botta: An arabic dialect chatbot. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations* (pp. 208-212).

Al-Madi, N. A., Maria, K. A., Al-Madi, M. A., Alia, M. A., & Maria, E. A. (2021). An intelligent Arabic chatbot system proposed framework. In *2021 International Conference on Information Technology (ICIT)* (pp. 592-597). IEEE.

Almurtadha, Y. (2019). Labeeb: Intelligent conversational agent approach to enhance course teaching and allied learning outcomes attainment. *Journal of Applied Computer Science & Mathematics*, *13*(27), 9–12.

Alshammari, N. O., & Alharbi, F. D. (2022). Combining a novel scoring approach with arabic stemming techniques for Arabic chatbots conversation engine. *Transactions on Asian and Low-Resource Language Information Processing*, *21*(4), 1-21.

Alshareef, T., & Siddiqui, M. A. (2020, November). A seq2seq neural network based conversational agent for gulf arabic dialect. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-7). IEEE.

Amodu, O. A., Bukar, U. A., Mahmood, R. A. R., Jarray, C., & Othman, M. (2023). Age of Information minimization in UAV-aided data collection for WSN and IoT applications: A systematic review. *Journal of Network and Computer Applications*, *216*, 103652.

Arora, C., Sabetzadeh, M., Briand, L., & Zimmer, F. (2016). Automated extraction and clustering of requirements glossary terms. *IEEE Transactions on Software Engineering*, *43*(10), 918-945.

Arsovski, S., Osipyan, H., Oladele, M. I., & Cheok, A. D. (2019). Automatic knowledge extraction of any Chatbot from conversation. *Expert Systems with Applications*, *137*, 343-348.

Ashari, H. A., Heidari, M., & Parvaresh, S. (2014). Improving SMTEs' business performance through strategic use of information communication technology: ICT and tourism challenges and opportunities. *International Journal of Academic Research in Accounting, Finance and Management Sciences*, *4*(3), 1-20.

Baevski, A., Edunov, S., Liu, Y., Zettlemoyer, L., & Auli, M. (2019). Cloze-driven pretraining of self-attention networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp.5360–5369). Association for Computational Linguistics.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., ... & Wang, T. (2016). Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*.

Belgaumwala, N. (2019). Chatbot: a virtual medical assistant. *International Journal for Research in Applied Science and Engineering Technology*, *7*, 1042–1050.

Beltagy, I., Lo, K., & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp.3615–3620). Association for Computational Linguistics.

Berry, D. M. (2023). The limits of computation: Joseph Weizenbaum and the ELIZA chatbot. *Weizenbaum Journal of the Digital Society*, *3*(3).

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data* (pp. 1247-1250).

Boussakssou, M., Ezzikouri, H., & Erritali, M. (2022). Chatbot in Arabic language using seq to seq model. *Multimedia Tools and Applications*, *81*(2), 2859-2871.

Caldarini, G., Jaf, S., & McGarry, K. (2022). A literature survey of recent advances in chatbots. *Information*, *13*(1), 41.

Candra, H., & Riki, R. (2019). Designing a Chatbot application for student information centers on Telegram Messenger using fulltext search Boolean. *Bit-Tech: Binary Digital-Technology*, *2*(2), 71-80.

Chaves, A. P., & Gerosa, M. A. (2021). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, *37*(8), 729-758.

Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., ... & Stoica, I. (2024). Chatbot arena: An open platform for evaluating LLMs by human preference. arXiv preprint arXiv:2403.04132.

Cho, E., Molina, M. D., & Wang, J. (2019). The effects of modality, device, and task differences on perceived human likeness of voice-activated virtual assistants. *Cyberpsychology, Behavior, and Social Networking*, *22*(8), 515-520.

Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Colby, K. M., Weber, S., & Hilf, F. D. (1971). Artificial paranoia. *Artificial Intelligence*, *2*(1), 1-25.

De Melo, G., & Hose, K. (2013). Searching the web of data. In *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013* (pp. 869-873). Springer.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *1*, 4171–4186.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhola, K., & Saradva, M. (2021, January). A comparative evaluation of traditional machine learning and deep learning classification techniques for sentiment analysis. In *2021 11th international conference on cloud computing, data science & engineering (Confluence)* (pp. 932-936). IEEE.

Dsouza, R., Sahu, S., Patil, R., & Kalbande, D. R. (2019). Chat with bots intelligently: A critical review & analysis. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)* (pp. 1-6). IEEE.

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Eason, G., Noble, B., & Sneddon, I. N. (1955). On certain integrals of Lipschitz-Hankel type involving products of Bessel functions. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, *247*(935), 529-551.

ElGibreen, H., Almazyad, S., Shuail, S. B., Al Qahtani, M., & ALhwiseen, L. (2020). Robot framework for anti-bullying in Saudi schools. In *2020 Fourth IEEE International Conference on Robotic Computing (IRC)* (pp. 166-171). IEEE.

Firdaus, M., Golchha, H., Ekbal, A., & Bhattacharyya, P. (2021). A deep multi-task model for dialogue act classification, intent detection and slot filling. *Cognitive Computation*, *13*, 626-645.

Gandon, F. (2018). A survey of the first 20 years of research on semantic Web and linked data. *Revue des Sciences et Technologies de l'Information-Série ISI: Ingénierie des Systèmes d'information*. Retrieved from https://inria.hal.science/hal-01935898/document

Grudin, J., & Jacques, R. (2019). Chatbots, humbots, and the quest for artificial general intelligence. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-11).

Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks *.In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistic*s (pp.8342–8360).

Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Hahm, Y., Kim, J., An, S., Lee, M., & Choi, K. S. (2018). Chatbot who wants to learn the knowledge: KB-Agent. *Semdeep/NLIWoD@ ISWC*, *4*.

Han, J. M., Rute, J., Wu, Y., Ayers, E. W., & Polu, S. (2021). Proof artifact co-training for theorem proving with language models. *arXiv preprint arXiv:2102.06203*.

Hochreiter, S. (1998). The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *6*(02), 107-116.

Horacek, H. (1990).Natural language processing. *Computer Physics Communications, 61*(1-2), 76-92.

Hu, T., Xu, A., Liu, Z., You, Q., Guo, Y., Sinha, V., ... & Akkiraju, R. (2018, April). Touch your heart: A tone-aware chatbot for customer care on social media. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Hussain, S., Ameri Sianaki, O., & Ababneh, N. (2019). A survey on conversational agents/chatbots classification and design techniques. In *Web, Artificial Intelligence and Network Applications: Proceedings of the Workshops of the 33rd International Conference on Advanced Information Networking and Applications (WAINA-2019) 33* (pp. 946-956). Springer International Publishing.

Jacobs, I. S., & Bean, C. P. (1963). Fine particles, thin films and exchange anisotropy (effects of finite dimensions and interfaces on the basic properties of ferromagnets). *Spin Arrangements and Crystal Structure, Domains, and Micromagnetics*, *3*, 271-350.

Jia, J. (2009). CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, *22*(4), 249-255.

Jørgensen, R. K., Hartmann, M., Dai, X., & Elliott, D. (2021). mDAPT: Multilingual domain adaptive pretraining in a single model. *arXiv preprint arXiv:2109.06605.*

Joseph, A., & Eapen, N. G. Conversational agents and Chatbots: Current trends. In *Machine Learning and Deep Learning in Natural Language Processing* (pp. 17-52). CRC Press.

Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning* (pp. 2342-2350). PMLR.

Kalla, D., Smith, N., Samaah, F., & Kuraku, S. (2023). Study and analysis of chat GPT and its impact on different fields of study. *International Journal of Innovative science and Research Technology*, *8*(3).

Kapočiūtė-Dzikienė, J. (2020). A domain-specific generative chatbot trained from little data. *Applied Sciences*, *10*(7), 2221.

Kar, R., & Haldar, R. (2016). Applying chatbots to the internet of things: Opportunities and architectural elements. *arXiv preprint arXiv:1611.03799*.

Khanna, A., Pandey, B., Vashishta, K., Kalia, K., Pradeepkumar, B., & Das, T. (2015). A study of today's AI through chatbots and rediscovery of machine intelligence. *International Journal of u-and e-Service, Science and Technology*, *8*(7), 277-284.

Kim, S., Kwon, O. W., & Kim, H. (2020). Knowledge-grounded chatbot based on dual wasserstein generative adversarial networks with effective attention mechanisms. *Applied Sciences*, *10*(9), 3335.

Lee, C. J., Jung, S. K., Kim, K. D., Lee, D. H., & Lee, G. G. B. (2010). Recent approaches to dialog management for spoken dialog systems. *Journal of Computing Science and Engineering*, *4*(1), 1-22.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474.

Lin, B. (2023). Reinforcement learning and bandits for speech and language processing: Tutorial, review and outlook. *Expert Systems with Applications*, 122254.

Lin, C. C., Huang, A. Y., & Yang, S. J. (2023). A review of ai-driven conversational chatbots implementation methodologies and challenges (1999–2022). *Sustainability*, *15*(5), 4012.

Liu, B., & Mei, C. (2020). Lifelong knowledge learning in rule-based dialogue systems. 5.

Luo, B., Lau, R.Y., Li, C. and Si, Y.W., 2022. A critical review of state-of-the-art chatbot designs and applications. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(1), e1434.

Ma, M., Liu, C., Wei, R., Liang, B., & Dai, J. (2022). Predicting machine's performance record using the stacked long short-term memory (LSTM) neural networks. *Journal of Applied Clinical Medical Physics*, *23*(3), e13558.

Ma, N., Yin, H., & Wang, K. (2023). Prediction of the remaining useful life of supercapacitors at different temperatures based on improved long short-term memory. *Energies*, *16*(14), 5240.

Maher, S., Kayte, S., & Nimbhore, S. (2020). Chatbots & its techniques using AI: an review. *International Journal for rRsearch in Applied Science and Engineering Technology*, *8*(12), 503-508.

Makatchev, M., Fanaswala, I., Abdulsalam, A., Browning, B., Ghazzawi, W., Sakr, M., & Simmons, R. (2010). Dialogue patterns of an Arabic robot receptionist. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 167-168). IEEE.

Mavridis, N., AlDhaheri, A., AlDhaheri, L., Khanii, M., & AlDarmaki, N. (2011, February). Transforming IbnSina into an advanced multilingual interactive android robot. In *2011 IEEE GCC Conference and Exhibition (GCC)* (pp. 120-123). IEEE.

Miller, J. R. (1981). PARRY and the evaluation of cognitive models. *Behavioral and Brain Sciences*, *4*(4), 543-544.

Moubaiddin, A., Shalbak, O., Hammo, B., & Obeid, N. (2015). Arabic dialogue system for hotel reservation based on natural language processing techniques. *Computación y Sistemas*, *19*(1), 119-134.

Naik, A., Alvarado, C. S., Wang, L. L., & Chen, I. (2023). NLP for Maternal Healthcare: Perspectives and Guiding Principles in the Age of LLMs. *arXiv preprint arXiv:2312.11803.*

Naous, T., Hokayem, C., & Hajj, H. (2020). Empathy-driven Arabic conversational chatbot. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop* (pp. 58-68).

Nirala, K. K., Singh, N. K., & Purani, V. S. (2022). A survey on providing customer and public administration based services using AI: chatbot. *Multimedia Tools and Applications*, *81*(16), 22215-22246.

Noori, Z., Bandarl, Z., & Crockett, K. (2014). Arabic goal-oriented conversational agent based on pattern matching and knowledge trees. *World Congress on Engineering*, 7.

Octavany, O., & Wicaksana, A. (2020). Cleveree: an artificially intelligent web service for Jacob voice chatbot. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *18*(3), 1422-1432.

Palasundram, K., Sharef, N. M., Nasharuddin, N., Kasmiran, K., & Azman, A. (2019). Sequence to sequence model performance for education chatbot. *International Journal of Emerging Technologies in Learning (iJET)*, *14*(24), 56-68.

Patel, R., Bhagora, N., Singh, P., & Namdev, M. K. (2020). Cloud based student information Chatbot. *International Research Journal of Modernization in Engineering Technology and Science*, *2*(4), 928.

Peng, D., Zhou, M., Liu, C., & Ai, J. (2020). Human–machine dialogue modelling with the fusion of word-and sentence-level emotions. *Knowledge-Based Systems*, *192*, 105319.

Pool, S., & Pool, B. (2007). A management development model: Measuring organizational commitment and its impact on job satisfaction among executives in a learning organization. *Journal of Management Development*, *26*(4), 353-369.

Prassanna, J., Nawas, K. K., Jackson, J. C., Prabakaran, R., & Ramanathan, S. (2020). Towards building a neural conversation chatbot through seq2seq model. *International Journal of Scientific and Technology Research*, *9*(3), 1219-1222.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (p.12). Retrieved from https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

Ramesh, K., Ravishankaran, S., Joshi, A., & Chandrasekaran, K. (2017). A survey of design techniques for conversational agents. In *International Conference on Information, Communication and Computing Technology* (pp. 336-350). Singapore: Springer Singapore.

Reiter, E. (2007, June). An architecture for data-to-text systems. In *Proceedings of the Eleventh European Workshop on Natural Language Generation (ENLG 07)* (pp. 97-104).

Reshmi, S., & Balakrishnan, K. (2018). empowering chatbots with business intelligence by big data integration. *International Journal of Advanced Research in Computer Science*, *9*(1), 627–631.

Riek, L. D., Mavridis, N., Antali, S., Darmaki, N., Ahmed, Z., Al-Neyadi, M., & Alketheri, A. (2010). Ibn sina steps out: Exploring arabic attitudes toward humanoid robots. In *Proceedings of the 2nd international symposium on new frontiers in human–robot interaction, AISB, Leicester* (Vol. 1, p.8).

Roca, S., Sancho, J., García, J., & Alesanco, Á. (2020). Microservice chatbot architecture for chronic patient support. *Journal of Biomedical Informatics*, *102*, 103305.

Sanders, N. E., Ulinich, A., & Schneier, B. (2023). Demonstrations of the potential of AI-based political issue polling. *arXiv preprint arXiv:2307.04781*.

Sean, M. M. (2010). *Design of a comprehensive student information system (SIS) and user interface for the honors college at USF*. (Department of Mechanical Engineering, USF College of Engineering USF Honors College).

Seering, J., Luria, M., Kaufman, G., & Hammer, J. (2019). Beyond dyadic interactions: Considering chatbots as community members. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13).

Shamir, G. (2023). *Introducing machine learning in K-12 education for the advancement of computational thinking.* (Doctoral Disertation, Tel Aviv University).

Shawar, A., & Atwell, E. S. (2004). An Arabic chatbot giving answers from the Qur'an. In *Proceedings of TALN04: XI Conference sur le Traitement Automatique des Langues Naturelles* (Vol. 2, pp. 197-202). ATALA.

Shawar, B. A. (2011). A Chatbot as a natural web Interface to Arabic web QA. *International Journal of Emerging Technologies in Learning (iJET)*, *6*(1), 37-43.

Shawar, B. A., & Atwell, E. (2004). Accessing an information system by chatting. In *International Conference on Application of Natural Language to Information Systems* (pp. 407-412). Berlin, Heidelberg: Springer Berlin Heidelberg.

Skantze, G. (2007). *Error handling in spoken dialogue systems-managing uncertainty, grounding and miscommunication*. (Doctoral dissertation, TH Royal Institute of Technology).

Song, H., Wang, Y., Zhang, K., Zhang, W. N., & Liu, T. (2021). BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. *arXiv preprint arXiv:2106.06169*.

Spivack, N., Douglas, S., Crames, M., & Connors, T. (2024). Cognition is all you need-the next layer of ai above large language models. *arXiv preprint arXiv:2403.02164.*

Sudalairaj, S., Bhandwaldar, A., Pareja, A., Xu, K., Cox, D. D., & Srivastava, A. (2024). LAB: Large-Scale Alignment for ChatBots. *arXiv preprint arXiv:2403.01081.*

Suhaili, S. M., Salim, N., & Jambli, M. N. (2021). Service chatbots: A systematic review. *Expert Systems with Applications*, *184*, 115461.

Sutskever, I., Martens, J., & Hinton, G. E. (2011). Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* (pp. 1017-1024).

Sweidan, S. Z., Abu Laban, S. S., Alnaimat, N. A., & Darabkh, K. A. (2021). SIAAA-C: a student interactive assistant Android application with chatbot during COVID-19 pandemic. *Computer Applications in Engineering Education*, *29*(6), 1718-1742.

Sweidan, S. Z., Laban, S. S. A., Alnaimat, N. A., & Darabkh, K. A. (2021, March). SEG-COVID: a student electronic guide within Covid-19 pandemic. In *2021 9th International Conference on Information and Education Technology (ICIET)* (pp. 139-144). IEEE.

Tran, V. K., & Nguyen, L. M. (2019). Gating mechanism based natural language generation for spoken dialogue systems. *Neurocomputing*, *325*, 48-58.

Tulshan, A. S., & Dhage, S. N. (2019). Survey on virtual assistant: Google assistant, siri, cortana, alexa. In *Advances in Signal Processing and Intelligent Recognition Systems: 4th International Symposium SIRS 2018, Bangalore, India, September 19–22, 2018, Revised Selected Papers 4* (pp. 190-201). Springer.

Vijayaraghavan, V., & Cooper, J. B. (2020). Algorithm inspection for chatbot performance evaluation. *Procedia Computer Science*, *171*, 2267-2274.

Vinyals, O., & Le, Q. (2015). A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Wang, Q., Li, Y., Wang, Y., & Ren, J. (2022). An automatic algorithm for software vulnerability classification based on CNN and GRU. *Multimedia Tools and Applications, 81*(5), 7103-7124.

Weizenbaum, J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36-45.

Wu, L., Kong, C., Hao, X., & Chen, W. (2020). A short-term load forecasting method based on GRU-CNN hybrid neural network model. *Mathematical Problems in Engineering*, 1-10.

Yorozu, T., Hirano, M., Oka, K., & Tagawa, Y. (1987). Electron spectroscopy studies on magneto-optical media and plastic substrate interface. *IEEE Translation Journal on Magnetics in Japan*, *2*(8), 740-741.

Yu, L., Qu, J., Gao, F., & Tian, Y. (2019). A novel hierarchical algorithm for bearing fault diagnosis based on stacked LSTM. *Shock and Vibration*, *2019*(1), 2756284.

Zhang, D., Chen, X., Zhang, Y., & Qin, S. (2021). Template-based Chatbot for agriculture related FAQs. *arXiv preprint arXiv:2107.12595*.

Zhang, Q., Zhang, J., Zou, J., & Fan, S. (2020). A novel fault diagnosis method based on stacked lstm. *IFAC-PapersOnLine*, *53*(2), 790-795 .

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too?. *arXiv preprint arXiv:1801.07243*.

Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., & Dolan, B. (2018). Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, *31*,11.

Author Information

Wassem Saad Nsaif
Department of Computer Science, Sciences College, University of Diyala, Iraq
Contact e mail: Wassem.saad@uodiyala.edu.iq

Huda Mohammed Salih
Department of Computer Engineering, College of Engineering, University of Diyala, Iraq

Hassan Hadi Saleh
Department of Computer Science, College of Education for Pure Science, University of Diyala, Iraq

Bashar Talib Al-Nuaimi
Department of Computer Science, Sciences College, University of Diyala, Iraq

To cite this article:

Nsaif, W.S., Salih, H.M., Saleh, H. H. & Al-Nuaimi, B. T. (2024). Conversational agents: An exploration into Chatbot evolution, architecture, and important techniques. *The Eurasia Proceedings of Science, Technology, Engineering & Mathematics (EPSTEM), 27*, 246-262.