**HYBRID CHATBOT FOR LUNG CANCER PATIENT SUPPORT: COMBINING RETRIEVAL AND GENERATIVE APPROACHES**

**CHAPTER 1 (INTRODUCTION)**

**1.1 Overview**

Conversation agents, commonly referred to as chatbots, represent a prominent innovation in the contemporary technological landscape. These intelligent systems aims to imitate human interaction, by simulating natural conversations with remarkable authenticity (Pandey & Sharma, 2023). Chatbots have demonstrated significant value across various industries, offering diverse applications, including customer query support, information dissemination, appointment booking, online purchases, and numerous other functionalities (Luo et al., 2022).

They aim to replicate human conversation using advanced natural language processing and, at times, natural language generation techniques. They are broadly categorized according to response generation into two designs: Retrieval-based or Rule-based Chatbots and Generation-based Chatbots (Dsouza et al., 2019).

**1.1.1 Retrieval-based Chatbots**

Retrieval-based chatbots rely on a database of predefined responses and use similarity matching to select the most appropriate response based on the user's input (Wang & Fang, 2020). These chatbots have a larger dataset to draw responses from, making them more versatile in handling a broader range of user queries. By using machine learning algorithms, retrieval-based chatbots can learn to match user queries with relevant responses, enhancing their ability to understand and respond effectively to natural language inputs (Tao et al., 2021).

Rule-based chatbots which is a type of retrieval-based chatbots operate using predefined rules and patterns to generate responses based on specific keywords or phrases in the user's input (Arsovski et al., 2019). These chatbots are relatively simple and easy to implement, as they rely on a fixed set of responses to cover a limited range of user queries. The responses are typically pre-defined by experts in the domain, making them suitable for providing specific information or following strict guidelines. Rule-based chatbots excel at providing structured and accurate responses, but they lack flexibility and may struggle with understanding more complex user inputs (Dsouza et al., 2019).

Both rule-based and retrieval-based chatbots fall under the category of "knowledge-driven" chatbots and are broadly categorized as retrieval-based, as they primarily rely on pre-constructed knowledge base to generate responses (Lan et al., 2020). The effectiveness of rule-based models, consequently, relies heavily on the availability of extensive data (Dsouza et al., 2019).

Retrieval-based chatbots may have limitations in handling more complex or context-dependent queries, and their responses may also lack the naturalness exhibited by human conversations. They however produce more richer responses than their generative counterpart (Zhu et al., 2021).

**1.1.2 Generative Chatbots**

The major drawback to retrieval-based chatbots is that the responses might become easily repetitive and predictable which ultimately leads to users engagement and interaction (Bachtiar et al., 2023). Generative chatbots, on the other hand learn from large datasets of human conversations to generate responses and are often categorized as ‘data-driven’ (Scotti et al., 2023). These chatbots utilize advanced natural language processing (NLP) techniques, such as neural networks and pre-trained large language models like Bidirectional Encoder Representations from Transformers (BERT) developed by Devlin et al. (2019), Meta’s Llama models, and OpenAI’s GPT models, to understand and generate human-like responses. Unlike rule-based and retrieval-based chatbots, generative chatbots can produce novel responses that are not restricted to predefined rules or fixed datasets. They also have the potential to engage users in more natural and interactive conversations, offering a more human-like experience.

Researchers have explored various approaches to improve the performance of generative chatbots, such as incorporating knowledge bases to enhance response generation and incorporating attention mechanisms to capture relevant context and knowledge.

**1.1.3 Chatbots in Healthcare**

Chatbots designed for conversational interactions possess great potential in revolutionizing the landscape of health and medical care. Their versatility allows them to serve vital roles, such as aiding clinicians and doctors with consultations, providing support to individuals facing behaviour change challenges, and offering assistance to patients by providing emotional support and answers to frequently asked questions (Laranjo et al., 2018).

Blanc et al. (2022) investigated the use of two French language models, FlauBERT and CamemBERT, for intent and slot prediction in a medical chatbot. da Silva Lima Roque et al. (2021) proposed BOTCURATIVO, a chatbot for non-specialists that provides comprehensive guidelines on wound dressing procedures tailored to specific wound types. The dissemination of vaccine information, including that of the human papillomavirus (HPV), is facilitated through an interactive conversational agent (Amith et al., 2019). CoachAI delivers personalized health activities to users based on an in-depth analysis of their conversational data via a web application (Fadhil et al., 2019).

Dr. Joy was a chatbot proposed to cater to the obstetric and mental health care needs of perinatal women and their partners by providing valuable support and guidance during the perinatal period (Chung et al., 2021). Similarly, Mamabot was put forward to provide comprehensive support and assistance to mothers, families with young children, and pregnant women during their journey through pregnancy and early parenthood (Vaira et al., 2018). As the healthcare industry continues to embrace chatbot technology, these virtual assistants hold great promise in enhancing healthcare delivery, patient experiences, and overall well-being.

**1.2 Problem Background**

There is a growing body of research on the use of chatbots for healthcare. A systematic review of 17 studies by Laranjo et al. (2018) found that chatbots can be effective in providing information, answering questions, and offering support to patients. However, the authors of the review noted that more research is needed to assess the long-term effectiveness of chatbots and to identify the factors that contribute to their success.

Most of the prior research has primarily concentrated on two main approaches for developing conversational AI systems: retrieval-based and generation-based methods. Retrieval-based methods offer the advantage of generating fluent and diverse responses. Nonetheless, their performance is constrained by the size of the response database. Conversely, generation-based methods can produce coherent responses on various topics. However, these responses tend to be generic and lack informative content, mainly because of the absence of grounding knowledge.

Studies have showcased the use of chatbots in various medical applications, including intent and slot prediction in medical chatbots, comprehensive wound dressing guidelines for non-specialists, vaccine information dissemination, and personalized health activities based on conversational data analysis. However, in the specific context of lung cancer patient support, there is a research gap in the development of hybrid chatbots that combine the strengths of both retrieval-based and generative approaches since those chatbots often fall into one category, limiting their ability to provide comprehensive and adaptive support.

One of the challenges in developing chatbots for healthcare is that they need to be able to understand and respond to a wide range of queries. This can be difficult, as patients may use different terms to describe the same symptom or condition. Additionally, patients may be emotional or anxious, which can make it difficult for them to communicate effectively.

A hybrid approach could potentially address the limitations of individual models, delivering more accurate, engaging, and personalized interactions. This study proposes to design and develop a hybrid chatbot for lung cancer patient support, leveraging both retrieval-based and generative techniques. The proposed chatbot will harness the advantages of the two approaches to deliver tailored information, emotional support, and guidance to lung cancer patients throughout their journey. The chatbot will combine pre-trained Llama2 large language model released by Meta in July 2023 with a curated knowledge base on lung cancer to provide accurate support information to patients. By addressing this research gap and exploring the effectiveness of the hybrid chatbot, this study seeks to contribute to the advancement conversational systems in healthcare and improve the overall well-being and experience of lung cancer patients and their families.

The following research questions will be addressed in this dissertation:

* How can a hybrid chatbot be designed to provide comprehensive and personalized support to lung cancer patients?
* How effective is a hybrid chatbot in providing patients with information, support, and emotional well-being?

By answering these research questions, this study aims to shed light on the potential of hybrid chatbots in healthcare and contribute to the improvement of lung cancer patient care and support.

**1.3 Research Aim**

The aim of this research is to develop a hybrid chatbot architecture which integrates retrieval-based and generative approaches that can provide personalized support, and accurate medical information to lung cancer patients. This research seeks to leverage pre-trained large language models backed up with knowledge bases built from data from medical literature, web sites, and patient forums to overcome the limitations of existing single-approach models.

* 1. **Research Objectives**
* To develop a hybrid chatbot architecture that combines retrieval-based and generative approaches.
* To create a knowledge base consisting of medical information, frequently asked questions, and emotional support resources specific to lung cancer.
* To implement natural language processing techniques that understands and generates meaningful responses to patient queries.
* To evaluate the effectiveness of a hybrid chatbot in terms of accuracy of information provided,

**1.5 Research Scope**

The proposed chatbot will be only trained on a medical literature and dataset of human conversations that is specific to lung cancer support and might not be able to provide relevant information or advice on other medical conditions. The focus of this research is to develop a chatbot that can provide support to lung cancer patients, not to develop a tool for diagnosis and treatment.

**1.6 Methodology**

In the methodology for this dissertation, a multi-stage approach was adopted to harness and process textual medical literature specifically related to lung cancer. The initial phase involved text extraction from diverse formats, predominantly PDFs and specialized websites. Given the inherent intricacies of these sources, advanced tools, such as Langchain’s PyPDFLoader for PDFs and BeautifulSoup for web content, were employed. This ensured that a comprehensive and clean dataset was available for subsequent stages. After extraction, the dataset was segmented into manageable chunks, facilitating efficient processing. The refined data was then encoded into embeddings, essentially converting text into computationally friendly numeric vectors. For this transformation, the 'sentence-transformers/all-MiniLM-L6-v2' model from the Hugging Face framework was utilized.

The next stages of the methodology focused on the organization, retrieval, and response generation using the transformed data. The embeddings were catalogued in a semantic index within a dedicated vector database. This storage also captured metadata about each piece of information, providing context and aiding in accurate retrieval. When addressing user queries, the system was designed to delve into this semantic index, drawing comparisons using the cosine-similarity metric, and identifying the most contextually relevant documents. To further refine the retrieved data, a content curation phase was integrated, distilling the essential information to aptly address the user's query. The final response generation hinged on the Llama2 language model. This pre-trained large language model was tailored for chat interactions and, when fed with the curated content and user's query in addition to previous chat history, produced precise and user-focused responses.

**1.7 Contribution**

This dissertation has the potential to improve the quality of life of lung cancer patients. It will also contribute to the field of conversational AI and healthcare support systems by providing insights into the feasibility and effectiveness of using a hybrid approach for developing healthcare chatbots.

**CHAPTER 2** (**LITERATURE REVIEW**)

**2.1 Overview**

Chatbots, synonymous with various titles like virtual assistants, dialogue systems, conversational agents, personal assistants and conversational interfaces are designed to engage in natural language conversations with users (Altinok, 2018). They can be categorized based on various criteria, including their knowledge domain, the type of service they offer, their objectives and how they process user input and generate responses, the level of human assistance they require, and the method used to construct them (Adamopoulou & Moussiades, 2020).

In the realm of knowledge domain, chatbots are categorized according to the scope of their knowledge access and the extent of their training data. These classifications comprise two main categories: open domain chatbots, which possess the capability to engage in discussions on a wide range of general topics and provide relevant responses. The second category is closed domain chatbots, which are operate within a particular knowledge domain, making them proficient in addressing queries related to that specific domain while potentially struggling to respond to questions outside their predefined scope (Nimavat & Champaneria, 2017).

The classification of chatbots based on the service they provide is determined by their emotional proximity to the user, the level of personal interaction, and the specific task they perform. There are three main categories: Interpersonal chatbots, which handle communication-based services like flight bookings. Intrapersonal chatbots operate within the user's domain, like chat apps. Lastly, inter-agent chatbots communicate with other bots, a crucial feature, especially in the context of the Internet of Things (IoT) (Adamopoulou & Moussiades, 2020).

Chatbots can be categorized into two main groups depending on their approach to processing user input and generating responses. These groups include retrieval-based approaches and generative-based approaches. In retrieval-based models, the chatbot searches for the most suitable context-response pairs from a pre-existing conversational history database, while generative-based methods can generate new responses that are highly coherent and relevant based on the conversation context (Yang et al., 2019).

Chatbots are grouped into three distinct groups depending on their objectives or purposes. The first category includes Informative bots, which aim to provide users with pre-stored information or data from fixed sources, like FAQs. The second category consists of chat-based or conversational bots that interact with users in a manner resembling human-like conversations. Their primary objective is to respond accurately to user input. Lastly, the third category comprises task-based bots, which are designed to perform specific tasks such as booking a table at a restaurant (Nimavat & Champaneria, 2017).

Luo et al. (2022) opined that grouping chatbots based on response generation is better as the other classification methods fail to adequately capture the unique attributes of each chatbot type.

**2.2 Chatbot Design Architecture**

**2.2.1 Retrieval-Based Approach**

The retrieval-based approach for chatbot design has witnessed significant advancements in recent years, with researchers proposing a diverse range of models and methodologies to enhance response selection.

**2.2.1.1 Entity Matching Method**

Entity matching retrieval methods prioritize understanding and extracting specific entities from user queries to facilitate precise information retrieval. By anchoring the chatbot's responses around identified entities, these systems aim to enhance the relevance and accuracy of their outputs. Yu et al. (2019) introduced a stacked multi-head approach to multi-turn response selection in retrieval-based chatbots. They used a layered attention model to better understand input sentences. From this, they made two matching matrices to pair context with possible responses. They further refined this using a two-layer convolutional neural network (CNN). The final output are matching scores that assess the correlation between each context and its candidate responses before, subsequently selecting the response with the highest score.

Wu et al. (2017) proposed a sequential matching network (SMN) for response selection in multi-turn conversations. Their method involved comparing a potential response to every utterance in the ongoing conversation, exploring various detail layers. By leveraging convolutional and pooling techniques, they transformed key matching details from each duo into vector formats. To understand the links between utterances, they used a recurrent neural network (RNN), with the RNN's hidden states helping determine the final matching measure.

Deng et al. (2019) introduced an enhanced matching network (EMN) designed specifically for multi-turn response selection in retrieval-based chatbots. The EMN uses gated convolutional neural networks (GCNNs) instead of recurrent neural networks (RNN) to craft deeper semantic interpretations of sentences. This approach bolsters the interaction between potential responses and each input within the conversation. Ths is possible through local inference modeling and by drawing inference composition from the enhanced sequential inference model (ESIM) which is part of the matching network.

To further improve multi-turn response selection in retrieval-based chatbots, Gu et al. (2019) put forward the utterance-to-utterance interactive matching network (U2U-IMN. This model utilizes both recurrent layers and self-attention mechanisms to individually encode every utterance within the conversation and the associated responses. The U2U-IMN then fosters a two-way interaction between the ongoing dialogue and its potential responses, leveraging the cumulative distances between them as a foundational element for making predictions.

Ma et al. (2022) took a step further and proposed the global and local interaction matching model (GLIMM) for response selection. GLIMM looks at the big picture by connecting the overall conversation context with the available knowledge. At the same time, it also focuses on the finer details by matching each part of the conversation with specific knowledge bits. By combining these two methods, the model improves how chatbots select their responses.

Wu et al. (2018) explored the use of unlabeled data for learning matching models. They devised a method that utilized a sequence-to-sequence (Seq2Seq) structure to act as a subtle evaluator, assessing the compatibility of unlabeled data pairs. Through this combined approach of weak signals and raw unlabeled data, they were able to enhance the chatbot's ability to select appropriate responses.

Continuing the efforts to enhance retrieval-based chatbots, Wang and Fang (2020) addressed the challenge of ranking candidate responses in multi-turn conversations with the introduction of the adaptive response matching network (ARM). The ARM model used distinct encoders for various utterance styles, ensuring it could adjust to the unique matching demands of each. Additionally, the model integrated an embedding feature, infusing relevant context bits into the evaluation process.

Xu et al. (2021) explored a multi-task learning strategy to improve the accuracy of response selection in retrieval-oriented chatbots. The researchers introduced a model that matched the context with the response, and this was co-trained alongside four supplemental self-supervised tasks. These encompassed predicting subsequent sessions, reconstructing utterances, identifying incoherences, and discerning consistency. The combined training approach successfully harnessed the domain-specific knowledge embedded in dialogue datasets, thus yielding enhanced features for selecting appropriate responses.

Zhao et al. (2019) took a different perspective and developed the Document-Grounded Matching Network (DGMN) model to optimize response selection. The crux of DGMN was its emphasis on merging data from both contextual inputs and document sources. By hierarchically interacting with potential responses, it could dynamically evaluate the need for grounding as well as the significance of segments from both the context and the document. Such a methodology bolstered the chatbot's proficiency in pinpointing and leveraging pertinent data across diverse sources.

Mao et al. (2019) addressed the intricacies of matching suitable responses to multi-turn contexts in chatbots, emphasizing the pitfalls of using isolated representations. Their model, termed the Hierarchical Aggregation Network of Multi-representation (HAMR), leveraged the strengths of bidirectional recurrent neural networks to distill semantic patterns from textual passages. With the integration of an attention-focused matching aggregation procedure and by viewing potential responses as crucial contextual elements, their approach seamlessly melded multiple representations.

Gu et al. (2019) further explored interactive matching networks (IMN) to address the challenge of out-of-vocabulary (OOV) words in multi-turn response selection. To tackle the challenges of OOV words, the IMN method was crafted to build word representations based on three distinct dimensions. After which, an attentive hierarchical recurrent encoder (AHRE) was employed to hierarchically encode sentences, resulting in richer and more comprehensive representations. By calculating bidirectional interactions amidst the entirety of multi-turn contexts and potential responses, the method discerned matching data between the two, fostering an enhanced comprehension of context and facilitating the production of responses that were more contextually relevant.

Wang et al. (2019) introduced the Attentive Gated Dilated Residual (AGDR) model, a deep matching network designed for multiturn response selection. The AGDR model iteratively constructed multi-grained representations of the response candidate and its multi-turn history context. This was achieved using a combination of gated dilated-convolution and self-attention mechanisms, extracting interactive information between each utterance-response pair. The accumulated information was then fused into a vector, resulting in the final matching score for improved response selection.

Ma et al. (2021) developed a hierarchical matching network that considered matching at both word and utterance levels. By matching a response with the word sequence of the context, the model captured essential matching information at the word level. Furthermore, by matching the response with each utterance in the context, the model gained insights into the matching information for each utterance–response pair. These matching information from both levels were then fused to obtain the final matching information for improved response selection.

With a focus on practical application, Moore et al. (2023) offered a comprehensive solution for constructing retrieval-based chatbots. They curated a self-supervised dataset and a weakly labeled dataset extracted from chatlogs to train their hierarchical-based RNN architecture. This architecture was evaluated on three objectives: binary classification, self-supervised contrastive learning, and multi-class classification. Their findings supported the use of self-supervised contrastive learning for real-world retrieval-based chatbot scenarios.

**2.2.1.2 Context Aware Method**

Context-aware retrieval methods in chatbots emphasize the importance of understanding the surrounding conversational context to derive more relevant and coherent responses. Instead of treating each user input in isolation, these methods leverage previous dialogue turns to capture the evolving intent and nuances of a conversation. Wu et al. (2018) proposed a topic aware convolutional neural tensor network (TACNTN)to enhance response selection. The TACNTN model extracted topic words from the user's message using a pre-trained latent Dirichlet allocation (LDA) model. The neural tensor was then applied to refine the representation of the message and then ranked to match candidate responses.

Incorporating speaker awareness into the retrieval-based chatbot model, Gu et al. (2020) proposed the Speaker-Aware BERT (SABERT) model. Which could discern speaker shifts in multi-turn dialogues. Through a unique disentanglement approach that is cognizant of the speaker's identity, SABERT meticulously chose a limited set of pivotal utterances to serve as the refined context. Drawing from the embedded speaker data, the model then underwent domain adaptation. This adaptation ensured the integration of domain-specific insights into the previously trained language model, thus enhancing its ability to produce more pertinent responses in consecutive dialog exchanges.

Gu et al. (2019) introduced the dually interactive matching network (DIM) with a specific focus on personalized response selection. DIM performed interactive matching between responses and contexts, as well as between responses and respective personas. By combining the matching features from both sets, DIM derived a comprehensive set of matching features that accounted for both context relevance and personalized characteristics, leading to more accurate and contextually relevant responses.

Hua et al. (2020) aimed to detect relevant parts of the context and knowledge for response selection in multi-turn conversations. They proposed the RSM-DCK framework, which employs the latest contextual data as a point of reference. This reference is then used to pre-select the response at both the word and discourse levels. Subsequently, the proposed response interacts with the predetermined context and knowledge bank. This approach increased the chatbot's confidence in matching relevant knowledge with the user's context.

With a focus on response length, Wang and Fang (2021) proposed a length adaptive regularization method (ARM) for retrieval-based chatbot models. The proposed method predicted the desired response length based on the conversation context and applied a regularization method accordingly to adjust matching scores for candidate responses. This adaptive approach provided flexibility in generating responses of varying lengths while maintaining context relevance.

Chen et al. (2020) delved into the utilization of memory-centric models for the retrieval of cognitive multi-turn responses. They introduced the memory-based deep neural attention (mDNA) framework, integrating a bidirectional long short-term memory (Bi-LSTM) encoder with an attention feature and a dedicated memory unit. The role of the Bi-LSTM encoder was to distill features from context statements. Concurrently, the attention feature weighed the significance of these features in selecting a response. Also, the memory unit preserved insights from prior dialogues, furnishing an enriched context for refining response selection.

The realm of context-aware chatbot models was also explored by Qian et al. (2021) with the introduction of the IMPChat model. IMPChat focused on learning an implicit user profile from the user's dialogue history. The model captured the user's personalized language style and preferences from a post-aware user profile, containing post-response pairs topically related to the current post. The response candidate was then matched with the personalized language style and preference to determine the final ranking score, enabling the chatbot to deliver more relevant and engaging responses.

Focusing on the fusion of multiple representations for enhanced response selection, Tao et al. (2019) proposed a multi-representation fusion network. This network integrated various representations of words, n-grams, and sub-sequences to capture short and long-term dependencies among words. The representations were fused into the matching process at different stages, with late fusion proving to be more effective than early fusion. This approach significantly improved the chatbot's ability to identify and leverage contextual dependencies for generating more contextually relevant responses.

Continuing with the context-aware perspective, Zhu et al. (2021) introduced the context-aware network (CAN) for multi-turn response selection. Their architecture of CAN employs a sophisticated hierarchical attention strategy to glean crucial features from utterances, which aids in formulating the utterance representation. The CAN design distinctively differentiates the present message from past conversations. In doing so, it determines the compatibility level between the ongoing message and previous dialogues, leading to the assimilation of utterance representations. Following this, the model harnesses a multi-layer perceptron (MLP) to estimate the potential score for subsequent responses.

**2.2.2 Generative Approach**

**2.2.2.1 Adversarial/Reinforcement Learning**

This method allows chatbots to iteratively improve their conversational strategies based on feedback, optimizing for more coherent and contextually appropriate interactions. The work of Li et al. (2017) focused on adversarial training for response generation in dialogue systems. The authors employed a reinforcement learning framework, where a discriminator was used to distinguish between human-generated and machine-generated dialogue sequences. The outputs from the discriminator served as rewards for the generator, guiding it to produce responses that could better fool the discriminator. This adversarial training approach led to more realistic and contextually appropriate responses.

Olabiyi et al. (2018) further explored the concept in multi-turn dialogue response generation. Their hredGAN framework utilized conditional generative adversarial networks (GANs) and a customized hierarchical recurrent encoder-decoder network (HRED). During inference, the generator's latent space was perturbed with noise samples based on the dialogue history, leading to the generation of multiple potential responses. The discriminator then ranked and selected the most optimal response from these candidates, ensuring high-quality dialogue responses.

Zhang et al. (2018) introduced Adversarial Information Maximization (AIM), an adversarial learning approach designed to enhance the quality of conversations between humans and dialogue agents by training response generation models. Unlike conventional Maximum Mutual Information (MMI) methods, AIM directly optimized the MMI objective during model training, rather than employing it solely for response reranking during decoding. The authors extended AIM to DAIM by introducing a dual objective that enabled the concurrent learning of forward and backward models.

Ju et al. (2022) investigated the problem of adversarial behaviour in the context of learning from human feedback. They proposed an evaluation benchmark, SafetyMix, to test the robustness of methods that differentiate between safe language and toxic language in various adversarial settings. Their main finding was that user-based methods, which took into account that troll users would exhibit adversarial behaviour across multiple examples, worked best in certain settings.

Cuáyahuitl et al. (2019) addressed the complex challenges in training chatbots via the reinforcement learning paradigm, which often faced issues such as vast state dimensions, limitless action possibilities, and the intricacies of defining a reward function. To circumvent these obstacles, they adopted a strategy of clustered actions as an alternative to infinite ones and formulated a reward function grounded in human-likeness scores sourced from authentic human dialogues. Their Deep Reinforcement Learning agents, educated using raw textual chitchat data devoid of manual annotations, aimed to discern and select the most human-like actions from a mixture of human-generated and random responses.

Tran and Le (2021) delved into the challenges deep neural network (DNN) based chatbots face in generating contextually consistent responses. Despite the promising results of the encoder-decoder architecture with an attention mechanism, these models often produced responses with limited regard for prior dialogue, leading to irrelevance. The researchers introduced an application of reinforcement learning to these models, refining their performance using policy gradient methods in simulated dialogues between pretrained chatbots. Their method prioritized sequence rewards and implemented a forward-looking function, culminating in an approach that enhanced coherence and conversation consistency.

Kim et al. (2020) delved into the challenge of multi-turn generative chatbots, focusing on selecting relevant information from dialogue histories. Their model selectively considered dialogue histories using an attention mechanism to calculate the contextual importance of previous utterances. Additionally, the authors introduced Wasserstein generative adversarial networks (WGANs) to enhance the quality of responses. They utilized two WGANs to train different components of the model, leading to improved contextual quality of the generated responses.

Adversarial learning techniques were further explored by Ludwig (2017) for Generative Conversational Agents (GCA). The proposed method involved a discriminator performing token-level classification to differentiate between tokens generated by humans and those generated by machines. This approach allowed the model to seamlessly blend fine-grained phrase-level knowledge from diverse users while preserving individual preferences.

Li et al. (2022) studied the intricacies of daily human communication, underscoring the pivotal role of emotion and intention in dialogues. Recognizing a consequential relationship between these elements, they introduced the hierarchical intention and emotion prediction (HINTE) model, which sequentially anticipates the emotion and intention of a response. Following these predictions, responses were formulated word by word. To enhance the model's training and ensure responses were congruent with the projected emotion and intention, an adversarial-augmented inference network was integrated.

Tran et al. (2023) presented a chatbot model that aimed at harnessing contextual data for generating accurate responses and mimicking human conversational style. They employed a supervised learning methodology using a multi-turn conversation dataset. Their approach incorporated a deep reinforcement learning module to optimize the use of context for precision in responses. Simultaneously, an adversarial learning framework was integrated to ensure the responses were human-like. By merging these techniques with a BERT2BERT architecture, the study yielded models that produced coherent, contextually apt, and naturally expressed replies.

**2.2.2.2 Continuous/Transfer learning**

This method leverages pre-existing knowledge to adapt to new, unseen queries, enhancing chatbots’ versatility and performance. Continual learning approaches were researched by Madotto et al. (2020), Wu et al. (2019), and Mi et al. (2020). Madotto et al. (2020) proposed a continual learning benchmark for task-oriented dialogue systems. By contrasting a variety of continual learning foundational models, they uncovered distinct advantages and drawbacks associated with each method, especially concerning memory size and parameter selection. Wu et al. (2019) proposed TRADE, a transferable multi-domain state generator for task-oriented dialogue systems. TRADE addressed the reliance on domain ontology and limited knowledge sharing across domains by incorporating a copy mechanism. This enabled seamless knowledge transfer and empowered the model to predict (domain, slot, value) triplets even for previously unseen instances during training.

Mi et al (2020) addressed the challenge of catastrophic forgetting in natural language generation. To address this concern, they proposed a method termed Adaptively Regularized Prioritized Exemplar Replay (ARPER). This approach emphasizes the importance of prior exemplars while also integrating an adaptive regularization strategy, specifically employing Elastic Weight Consolidation, to enhance the generation of natural language.

Mo et al. (2017) proposed a personalized decoder model for transferring fine-grained phrase-level knowledge across users while preserving individual preferences. The personalized decoder model incorporated a personal control gate, allowing the decoder to alternate between generating personalized phrases and shared ones. This approach effectively addressed negative transfer issues that could arise when blending distinct personal information from diverse users.

In the pursuit of transfer learning for generative dialogue systems, Wolf et al. (2019) proposed TransferTransfo, a methodology that merges transfer learning techniques with the robust capabilities of the Transformer architecture. Their research underscored the potential advantages of leveraging pre-existing language models, like GPT-2, to amplify the results in open-domain dialogue generation. Additionally, they delved into multi-task fine-tuning, highlighting its prospective value in enhancing the performance of dialogue systems.

Mazumder et al. (2018) introduced a lifelong interactive learning and inference (LiLi) approach. This approach was specifically designed to tackle the challenges posed by the open-world knowledge base completion (OKBC) dilemma. Utilizing reinforcement learning (RL) principles, LiLi's unique methodology is characterized by its continuous engagement with users, employing query-tailored inference techniques throughout its operational lifespan to address OKBC issues.

Hancock et al. (2019) introduced a self-feeding chatbot that gleaned fresh training samples from its interactive conversations. The mechanism it employed gauged user satisfaction from its replies. When interactions seemed positive and productive, the user's subsequent responses were assimilated as new training instances. Moreover, in instances where the chatbot perceived potential errors in its responses, it proactively solicited user feedback. This strategy of anticipating and understanding the feedback significantly enhanced the chatbot's conversational capabilities.

**2.2.2.3 Knowledge-grounded/Topic Aware**

Knowledge-grounded or Topic Aware generative chatbots integrate external information sources into their response generation process. Bachtiar et al. (2022) conducted a comparative study of two sequence-to-sequence (Seq2seq) models, namely vanilla Long Short-Term Memory (LSTM) and vanilla Gated Recurrent Unit (GRU), for a generative chatbot case study. They found that the LSTM model outperformed the GRU model in terms of loss, BLEU score, and response quality. However, both models were identified to have potential for improvement through various modifications, such as incorporating attention mechanisms or additional data or knowledge base.

Ghazvininejad et al. (2017) presented a knowledge-grounded neural conversation model that aimed to generate more informative and diverse responses compared to traditional chatbots. Their approach extended the widely-used Seq2Seq method to neural conversation models, effectively integrating both conversational and non-conversational data using multi-task learning. The model achieved this without requiring explicit slot filling, and the results showed significant improvements in response quality and diversity.

Cho and Ko (2022) explored the task of knowledge-grounded dialogue, emphasizing the significance of selecting pertinent knowledge to generate coherent responses. Their research introduced a selection model using contrastive learning, combined with negative sampling loss, to develop a dialogue-centric representation of knowledge. This was achieved through a dual loss system: one part amplified the alignment between the content representations of relevant knowledge and dialogue history, while the second part did the same for their topic representations. Their model, termed DiCeR, diverged from conventional models, focusing on the inherent similarity of given inputs and was influenced by knowledge graph embedding principles.

To enhance multi-turn dialogue response generation, Madotto et al. (2020) introduced Mem2Seq, a neural generative model that integrated knowledge bases. The model utilized multi-hop attention over memories and leveraged pointer networks to efficiently manage each generation step and learn intricate correlations among memories, leading to improved performance and efficiency.

Zhu et al. (2017) presented GenDS, a knowledge-grounded conversation system that employed a fully data-driven generative approach for response generation. The model leveraged input messages and associated knowledge bases (KBs) to generate responses. A unique feature of GenDS was its ability to produce diverse answer entities, even when these entities were not present in the training data. This was achieved through a dynamic knowledge enquirer, which adaptively selected different answer entities based on the local context within a single response. Importantly, the model did not rely on specific entity representations, making it effective for handling out-of-vocabulary entities.

Continuing with knowledge-grounded chatbots, Kim et al. (2020) proposed a model that effectively reflects dialogue context and given knowledge using well-designed attention mechanisms. The model employed three types of attention: query-knowledge, query-context, and context-knowledge. These mechanisms allowed the model to generate responses that considered both the conversation history and external knowledge, resulting in more contextually relevant and informative responses.

GLMP networks were introduced by Wu et al. (2019) for integrating knowledge bases into end-to-end task-oriented dialogue systems. The GLMP architecture comprised a global memory encoder and a local memory decoder. The global memory encoder adapted the global contextual representation and generated a global memory pointer, while the local memory decoder utilized the global memory pointer to filter external knowledge for pertinent information and then used local memory pointers to complete the slots with relevant details.

Chen et al. (2018) explored the complexities associated with dialogue generation, emphasizing the need to manage long-term dependencies and ensure informativeness. In their research, they introduced the Hierarchical Variational Memory Network (HVMN). This approach merges a tiered framework with variational memory within a neural encoder-decoder architecture. Drawing inspiration from natural human interactions, the HVMN extracts overarching abstract patterns and maintained long-term memories during dialogue progression. This multi-layered structure efficiently encoded dialogue utterances at the individual level. The variational memory also plays a pivotal role in tracking overarching themes, retaining in-depth information from prior dialogues, and offering direct access to pertinent dialogue sequences.

Huo et al. (2020) underscored the need for chatbots to convey emotion while maintaining topic relevance. Their research led to the development of the Topic-aware Emotional Response Generation (TERG) model, which adeptly balances emotional expression with topic consistency in generated responses. By integrating an emotion-aware module with an encoder-decoder structure, the model ensured the desired emotional tone. A topic-aware module was also introduced to bolster the pertinence of the generated content.

Ling et al. (2021) contributed to the topic of context-controlled topic-aware neural response generation. Their model, CCTA, integrates topic details by discerning the inherent semantic connections among topics and the related conversational context. Also, this model incorporates a strategy for topic transition grounded in context. This approach ensured the generation of transition terms that are pertinent, leading to outputs that are not only coherent but also keenly attuned to the topic at hand.

Choi and Lee (2020) explored the challenges associated with generating emotionally nuanced responses using deep learning techniques. Recognizing the limitations of prior models, which often misinterpreted nuanced emotions and produced responses skewed by emotion tokens, the authors introduced a two-part model for emotional response generation. This approach employed an extraction component, utilizing a pre-trained LSTM classification model to vectorize emotions from user utterances, followed by a generation component that harnessed the sequence-to-sequence architecture to craft responses that are both emotionally resonant and contextually relevant.

Gu et al. (2021) introduced DialogBERT, a conversational response generation model that boosts Permutative Language Modeling (PLM) based dialogue models by incorporating a hierarchical Transformer architecture. The model was trained with two objectives: masked utterance regression and distributed utterance order ranking. The former aimed to capture discourse-level coherence by predicting masked utterances in the dialogue context, while the latter focused on utterance-level coherence by ranking the utterances in the dialogue context. The combination of these objectives improved the model's ability to generate coherent and contextually relevant responses.

Luo et al. (2018) proposed a personalized end-to-end model for goal-oriented dialogs. This model comprised two distinct parts: the first being a profile model adept at converting user profiles into dispersed representations, drawing insights from prior conversations of analogous users. The second, a preference model, was designed to discern user inclinations regarding entities within a knowledge base, ensuring clarity in addressing user queries. These two frameworks were ``1integrated to form the PERSONALIZED MEMN2N model.

**2.2.3 Hybrid Approach**

Hybrid chatbot systems blend retrieval-based and generative models to capitalize on the strengths of both approaches. By merging the consistency and reliability of retrieval-based chatbots with the adaptability and creativity of generative ones, these combined methodologies aim to offer an optimized user experience.

**2.2.3.1 Retrieval Enhanced Generation**

This method involves utilising a retrieval mechanism to pinpoint relevant pre-defined responses and then employing generative capabilities to fine-tune or adapt these responses to the specific context. Yang et al. (2019) introduced a hybrid neural conversation model that combines both response retrieval and generation capabilities. This model operates in three main steps: initially, it retrieves a collection of candidate responses from a response repository. Next, it employs a neural language model to generate a new response. The model finally selects the most suitable response from the retrieved and generated options, ensuring an effective and contextually relevant conversation output.

Song et al. (2018) proposed an approach that combines retrieval-based and generation-based open-domain conversation systems. The retrieval component identifies the k most suitable candidate replies, which are then passed, along with the original query, to an RNN-based multi-seq2seq reply generator. A re-ranker then evaluates the generated responses and retrieved ones to select the ultimate result.

Zhu et al. (2019) introduced retrieval-enhanced adversarial training (REAT) to improve neural response generation in dialogue systems. The REAT method comprises two main steps. Initially, it employs an encoder-decoder framework to generate responses. Subsequently, a discriminator is constructed using n-best response candidates obtained from a retrieval-based system. The discriminator is then trained to differentiate between the generated responses and the n-best candidates, enhancing the quality and diversity of the generated responses.

Wang et al. (2021) highlighted the shortcomings of conventional generative dialogue systems in producing generic responses. They proposed a unique knowledge-enriched framework, Knowledge Refined Dialogue Generation (KRDG), that harnesses the contextual word representation of dialogues. The approach integrates an attention-based filter, drawing from BERT's contextual representations, to sift through knowledge triples, ensuring relevant information integration and reducing redundancy in dialogue systems. The researchers also introduced an encoder-decoder model that seamlessly blends contextual representation with this filtered knowledge.

Ahn et al. (2022) emphasized the typical limitation of many knowledge-grounded conversation models that often focus on a single document for conversation context. To address this, the authors developed a new retrieval-augmented response generation model. This model retrieves documents pertinent to both the overarching topic and the specific conversational context. By accepting topic words from the broader conversation and juxtaposing them with the immediate tokens before a response, the model efficiently generates knowledge-grounded reactions that align with the natural progression of the discussion.

**2.3 Literature Summary**

| **#** | **Study** | **Method Description** | **Pros** | **Cons** |
| --- | --- | --- | --- | --- |
| 1 | Yu et al. (2019) | This study uses a stacked multi-head approach to multi-turn response selection in retrieval-based chatbots. | Enables precise selection of responses based on deep understanding. | Complexity due to the layered model design. |
| 2 | Wu et al. (2017) | This study uses a sequential matching network (SMN) for response selection in multi-turn conversations. | Offers thorough comparison of responses, enhancing accuracy. | Transforming and processing details may be slow |
| 3 | Deng et al. (2019) | This study introduces an enhanced matching network (EMN) for multi-turn response selection. | Provides a deeper interpretation of sentences, improving context capture. | Requires the utilization of gated CNNs which means high computational resources |
| 4 | Gu et al. (2019) | This study proposes the utterance-to-utterance interactive matching network (U2U-IMN) for response selection. | Individually encodes utterances, capturing nuances effectively. | Complex two-way interaction demands careful implementation. |
| 5 | Ma et al. (2022) | This study proposes the global and local interaction matching model (GLIMM) for response selection. | Enhances relevance by combining both the bigger picture and details. | Balancing broad and fine aspects could be challenging. |
| 6 | Wu et al. (2018) | This study explores the use of unlabeled data for learning matching models. | Enhances learning using weak signals and raw data, boosting performance. | Relies on the availability of suitable unlabeled data. |
| 7 | Wang and Fang (2020) | This study introduces the adaptive response matching network (ARM) for ranking candidate responses. | Adjusts to various response length requirements, ensuring context relevance. | Incorporating the embedding feature adds complexity. |
| 8 | Xu et al. (2021) | This study utilizes multi-task learning to improve response selection in retrieval-oriented chatbots. | Exploits domain-specific knowledge, enhancing response accuracy. | Complexity arises from co-training multiple tasks. |
| 9 | Zhao et al. (2019) | This study develops the Document-Grounded Matching Network (DGMN) for optimizing response selection. | Merges contextual inputs and document sources, enriching decision-making. | Handling hierarchical interaction can be intricate. |
| 10 | Mao et al. (2019) | This study introduces the Hierarchical Aggregation Network of Multi-representation (HAMR) for response selection. | Utilizes multiple representations to capture intricate context. | The attention-focused matching mechanism is complex. |
| 11 | Gu et al. (2019) | This study explores interactive matching networks (IMN) for handling out-of-vocabulary (OOV) words. | Builds nuanced word representations, addressing OOV word challenges. | Utilizes an attentive hierarchical recurrent encoder, increasing complexity. |
| 12 | Wang et al. (2019) | This study introduces the Attentive Gated Dilated Residual (AGDR) model for multiturn response selection. | Gated dilated-convolution and attention improve contextual understanding. | Incorporating various mechanisms increases complexity. |
| 13 | Ma et al. (2021) | This study develops a hierarchical matching network for response selection. | Enhances accuracy by considering matching at both word and utterance levels. | The combination of matching information can be intricate. |
| 14 | Moore et al. (2023) | This study offers a comprehensive solution for constructing retrieval-based chatbots. | Utilizes self-supervised and weakly labeled data, mimicking real-world scenarios. | Evaluation across multiple objectives might be time-consuming. |
| 15 | Wu et al. (2018) | This study proposes a topic aware convolutional neural tensor network (TACNTN) for enhanced response selection. | Extracts topic words for refined message representation, improving relevance. | Incorporates a pre-trained LDA model, which can be resource-intensive. |
| 16 | Gu et al. (2020) | This study introduces the Speaker-Aware BERT (SABERT) model for speaker-aware response selection. | Recognizes speaker shifts, enhancing context relevance and engagement. | Involves complexity due to domain adaptation. |
| 17 | Gu et al. (2019) | This study proposes the dually interactive matching network (DIM) for personalized response selection. | Considers both context and persona, leading to contextually relevant responses. | Incorporates a comprehensive set of matching features, adding complexity. |
| 18 | Hua et al. (2020) | This study introduces the RSM-DCK framework for detecting relevant parts in context and knowledge. | Pre-selects responses, improving alignment with contextual information. | Interaction complexity arises with context and knowledge banks. |
| 19 | Wang and Fang (2021) | This study introduces a length adaptive regularization method (ARM) for response selection. | Predicts and adjusts response length, enhancing the coherence of responses. | Involves complexity due to response length prediction. |
| 20 | Chen et al. (2020) | This study introduces the memory-based deep neural attention (mDNA) framework for cognitive responses. | Integrates memory for enriched context, improving contextual understanding. | Additional complexity due to memory-centric model design. |
| 21 | Qian et al. (2021) | This study develops the IMPChat model for learning implicit user profiles in dialogue history. | Captures personalized language style and preferences, enhancing engagement. | Involves complexity in capturing personalized characteristics. |
| 22 | Tao et al. (2019) | This study proposes a multi-representation fusion network for improved response selection. | Fuses various linguistic representations, improving context capture. | Late fusion proves more effective for context dependency. |
| 23 | Zhu et al. (2021) | This study introduces the context-aware network (CAN) for multi-turn response selection. | Utilizes hierarchical attention for improved utterance representation. | Involves complexity with multi-layer perceptron (MLP) scoring. |
| 24 | Li et al. (2017) | This study focused on adversarial training for response generation in dialogue systems. | Adversarial training produces more contextually appropriate and realistic responses. | Complexity of reinforcement learning framework. |
| 25 | Olabiyi et al. (2018) | This study explored adversarial training in multi-turn dialogue response generation. | HredGAN framework generates high-quality responses through noise perturbation. | Relies on the performance of the discriminator. |
| 26 | Zhang et al. (2018) | This study introduced Adversarial Information Maximization (AIM) for enhanced conversation quality. | AIM directly optimizes MMI objective during training, improving response generation. | Complexity in jointly training forward and backward models. |
| 27 | Ju et al. (2022) | This study proposed an evaluation benchmark, SafetyMix, for adversarial behavior in human feedback. | User-based methods exhibit robustness against adversarial behavior, improving performance. | May have limitations in certain adversarial settings. |
| 28 | Cuáyahuitl et al. (2019) | This study addressed reinforcement learning challenges in training chatbots. | Clustering actions and human-likeness reward function overcome RL challenges. | Clustering actions might limit the diversity of responses. |
| 29 | Tran and Le (2021) | This study applied reinforcement learning to enhance coherence in chatbot responses. | Reinforcement learning methods improve contextually consistent responses. | Introduces complexity with reinforcement learning. |
| 30 | Kim et al. (2020) | This study focused on multi-turn generative chatbots, integrating WGANs for response quality. | Attention mechanism and WGANs improve contextual quality of generated responses. | Involves the complexity of managing multiple GANs. |
| 31 | Ludwig (2017) | This study used discriminator-based token-level classification for blending human and machine knowledge. | Blends diverse user knowledge while preserving individual preferences. | Relies on the discriminator's token-level classification. |
| 32 | Li et al. (2022) | This study introduced the hierarchical intention and emotion prediction (HINTE) model. | Sequential anticipation of emotion and intention enhances response congruence. | Complexity in integrating adversarial-augmented inference network. |
| 33 | Tran et al. (2023) | This study combined supervised and reinforcement learning for contextually apt responses. | Integrated learning techniques and BERT2BERT architecture yield coherent and human-like replies. | Complexity due to multiple techniques and architecture. |
| 34 | Madotto et al. (2020) | This study introduced a continual learning benchmark for task-oriented dialogue systems. | Explores various continual learning models, highlighting their pros and cons. | Complexity in selecting appropriate memory size and parameters. |
| 35 | Wu et al. (2019) | This study proposed a transferable state generator for task-oriented dialogue systems. | Incorporates a copy mechanism for improved knowledge transfer between domains. | Complexity in predicting triplets for unseen instances. |
| 36 | Mi et al. (2020) | This study addressed catastrophic forgetting using adaptive regularization and exemplar replay. | Emphasizes prior exemplars and adaptive regularization for natural language generation. | Incorporates complex strategies to address forgetting. |
| 37 | Mo et al. (2017) | This study developed a personalized decoder model for fine-grained knowledge transfer. | Personalized decoder model balances personalized and shared phrases effectively. | Challenges in blending diverse personal information. |
| 38 | Wolf et al. (2019) | This study merged transfer learning and Transformer architecture for dialogue generation. | Transfer learning with pre-existing models improves open-domain dialogue generation. | Complexity in leveraging pre-existing language models. |
| 39 | Mazumder et al. (2018) | This study introduced the lifelong interactive learning and inference (LiLi) approach. | LiLi approach addresses open-world knowledge base completion challenges using RL. | Continuous engagement with users can be resource-intensive. |
| 40 | Hancock et al. (2019) | This study developed a self-feeding chatbot that learns from interactions. | Chatbot assimilates training instances from positive interactions, improving its responses. | Relies on user feedback for new training instances. |
| 41 | Bachtiar et al. (2022) | This study uses Seq2seq models for generative chatbots. | LSTM model outperforms GRU model in loss and quality. | Potential for improvement through modifications. |
| 42 | Ghazvininejad et al. (2017) | This study uses Knowledge-grounded neural model for informative responses. | Integrates conversational and non-conversational data. | May require careful tuning of multi-task learning. |
| 43 | Cho and Ko (2022) | This study uses DiCeR model for coherent knowledge-grounded dialogue. | Utilizes contrastive learning for dialogue-centric knowledge. | Complexity in dual loss system and alignment. |
| 44 | Madotto et al. (2020) | This study uses Mem2Seq model to integrate knowledge bases for improved responses. | Leverages multi-hop attention and pointer networks. | Complexity in managing intricate correlations. |
| 45 | Zhu et al. (2017) | This study uses GenDS model to generate diverse answers using dynamic knowledge enquirer. | Produces diverse answers even for out-of-vocabulary entities. | Not reliant on specific entity representations. |
| 46 | Kim et al. (2020) | This study uses model with designed attention mechanisms for contextually relevant responses. | Employs query-knowledge, query-context, and context-knowledge attention. | Complexity in managing multiple attention types. |
| 47 | Wu et al. (2019) | This study uses GLMP networks to integrate knowledge bases into task-oriented systems. | Global memory encoder and local memory decoder enhance memory utilization. | Complexity in managing global and local memory pointers. |
| 48 | Chen et al. (2018) | This study uses HVMN approach to manage long-term dependencies for informative dialogue. | Merges tiered framework with variational memory for dialogue coherence. | Requires consideration in neural encoder-decoder architecture. |
| 49 | Huo et al. (2020) | This study developed TERG model which balances emotion and topic in responses. | Integrates emotion-aware and topic-aware modules for balanced responses. | Complexity in integrating emotion-aware and topic-aware modules. |
| 50 | Ling et al. (2021) | This study uses CCTA model that integrates context-controlled topic-aware response generation. | Considers semantic connections among topics and contextual relevance. | Complexity in managing topic transitions grounded in context. |
| 51 | Choi and Lee (2020) | This study developed a model extracts emotions and crafts emotionally resonant responses. | Employs an extraction and generation component for nuanced responses. | Complexity in employing extraction and generation components. |
| 52 | Gu et al. (2021) | This study uses DialogBERT to enhance dialogue models with hierarchical Transformer architecture. | Trained with masked utterance regression and utterance order ranking. | Complexity in implementing hierarchical Transformer architecture. |
| 53 | Luo et al. (2018) | This study developed a PERSONALIZED MEMN2N model for personalized end-to-end goal-oriented dialogs. | Two-part model converts profiles and discerns user preferences. | Complexity in managing personalized profile and preference models. |
| 54 | Yang et al. (2019) | The study implemented a hybrid model combining retrieval and generative capabilities. | Utilizes retrieval and generation for effective conversation. | Complexity in managing the integration of retrieval and generation. |
| 55 | Song et al. (2018) | This study proposed the combination of retrieval-based and generation-based systems. | Utilizes two-step process for refined response selection. | May require fine-tuning of re-ranker for optimal results. |
| 56 | Zhu et al. (2019) | This study uses the REAT method which enhances response generation with retrieval-based training. | Utilizes discriminator to improve response quality. | Complexity in managing the training of the discriminator. |
| 57 | Wang et al. (2021) | This study uses KRDG framework to integrates knowledge and contextual representation. | Utilizes attention-based filter to refine knowledge integration. | Complexity in managing contextual representation and attention. |
| 58 | Ahn et al. (2022) | This study implements a retrieval-augmented model that generates contextually aligned responses. | Retrieves documents for both topic and context relevance. | Complexity in juxtaposing topic words with immediate tokens |

**CHAPTER 3 (RESEARCH METHODOLOGY)**

**3.1 Overview**

Healthcare, a dynamic sector, is continuously seeking innovations that can improve the patient experience and offer support that goes beyond traditional mechanisms. Within this quest for innovation, chatbots have emerged as potential transformative agents. As articulated in earlier sections, an expanding reservoir of research supports the efficacy of chatbots in the healthcare arena, with their capabilities extending to information dissemination, query resolution, and patient support.

However, a nuanced understanding of the current research terrain reveals some discrepancies. Despite the advantages chatbots provide, certain limitations, particularly concerning their developmental methodologies, remain unaddressed. The extant body of literature on chatbot development in healthcare has largely been bifurcated into two streams: one that relies on the retrieval of stored information and another that hinges on generating responses on-the-fly.

While retrieval-oriented chatbots promise the benefit of fluency and versatility in their responses, they find themselves hemmed in by the boundaries of their informational repositories. In contrast, the generation-oriented counterparts, which are not restricted by a predetermined database, can sometimes produce outputs that might seem less anchored in real-world applicability, devoid of deeper information, or even outrightly inaccurate.

Despite the commendable strides in individual methods, the broader healthcare domain, especially lung cancer patient support, witnesses the dire need for a blended approach. An analysis of the existing solutions indicates that most chatbots lean predominantly towards one of the two developmental paradigms, thereby potentially compromising their ability to offer a rounded, adaptive, and comprehensive support experience. The heterogeneity in patient language, coupled with the emotional undertones often accompanying health-related discussions, necessitates a chatbot framework that is both versatile and empathetic. For instance, describing identical symptoms might see ten patients use ten different terminologies.

Recognizing this chasm between the existing solutions and the idealized patient support mechanism, this study aims to develop a chatbot built on the combined architecture of the two approaches. By synergizing the strengths and mitigating the limitations of both retrieval-based and generative paradigms, the hybrid model is anticipated to offer a more nuanced, adaptive, and comprehensive support system for lung cancer patients. Such an approach aims to amalgamate the wide topical range of already pre-trained large language generative systems with the depth and precision of retrieval-based ones.

Central to this research, is the fusion of a pre-trained large language model, specifically the Llama2, and a knowledge repository meticulously curated from authoritative sources on lung cancer. This dual-structure aims to enable the chatbot to seamlessly transition between retrieval and generative modes, ensuring that the information provided is both accurate and contextually relevant.

**3.2 Research Framework**

This segment outlines the step-by-step procedure adopted, underscoring the critical aspects that drive the development and functionality of the hybrid chatbot architecture. Hugging Face and Langchain frameworks were used in this study for most of the processes.

Collection & extraction of text data from websites/forums and from digital medical literature

Splitting of the extracted text into chunks with overlapping characters

Conversion of the text chunks into embeddings with an embedding model

Retrieval of related text chunks from the knowledge base that matches the user’s query.

Conversion of the modified query into embeddings with an embedding model

Query to the Chatbot

Modification of the Query by the Large Language Model (LLM)

Processing of the Query with the LLM’s parametric knowledge + retrieved chunks (context) + previous queries

Response

**Generative Component**

**Retrieval Component**

Storing the embeddings a semantic index in the vector database/knowledge base

**3.2.1 Retrieval Component**

1. Text Collection & Extraction from Digital and Websites

Medical literature, especially those resources pertaining to lung cancer found in PDFs or dedicated support websites, offers an invaluable reservoir of knowledge. The inherent challenge, however, revolves around the extraction of this data. PDFs, given their structural limitations, and websites, with their often-variable design architectures, are not always conducive to easy information extraction. To address this, libraries such as Langchain’s PyPDFLoader, OnlinePDFLoader, and pdf2text are used. These libraries effectively transform intricate PDF formats into straightforward, plain text. For data located on websites and online forums, libraries like BeautifulSoup and Langchain’s WebLoader play an indispensable role in extracting the necessary textual data. In this study, Langchain’s OnlinePDFLoader and WebLoader were used to extract relevant text.

A screenshot of a computer

Description automatically generated

Ddfdfd

A screenshot of a computer

Description automatically generated

1. Splitting Text into Manageable Chunks

After the extraction phase, there's a need to systematically manage this voluminous textual data. They have to be split into smaller chunks. This is done to improve the efficiency of the retrieval process. If the text is not split into chunks, the entire document will need to be compared to the user query, which can be very time-consuming. By splitting the text into chunks, users’ queries can be compared to a small subset of the document, which is much faster. In this study, the text from the 5 PDFs and 20 web pages were split separately into chunk sizes of 500 characters with an overlap of 25 charactersd with Langchain’s RecursiveCharacterTextSplitter module. The overlap is necessary as it preserves the context between chunks and ensures that no pertinent information falls through the cracks during the segmentation process.

A screenshot of a computer

Description automatically generated

Dfdffd

A computer code with text

Description automatically generated with medium confidence

1. Converting Text Chunks to Embeddings

Chunks of text from the previous step are stored as embeddings. Embeddings are a way of representing text as a vector of numbers. This makes it easier for computers to process text and to compare text to text (cite). There are a variety of embedding models available, such as Word2Vec and GloVe. These models are trained on a large corpus of text and learn to represent each word as a vector of numbers (cite). The vectors for different words are similar if the words have similar meanings. Hugging Face provides open-source models for this via their sentence transformers library and one of them, ‘sentence-transformers/all-MiniLM-L6-v2’ was used in this study. Each embedding model has its own specified dimension. This is necessary as it is part of the requirements to set up a semantic index in the vector database. The dimension of the embedding model used in this study is 384.

A screenshot of a computer

Description automatically generated

1. Cataloguing Embeddings in a Database

Upon successful conversion of text chunks into their respective embeddings, the next pivotal step involves orderly cataloguing within a database. This process isn't as simple as dumping data into storage. Instead, a meticulous structure, known as a semantic index within a vector database, is adopted. This index is multi-faceted, storing not just the embeddings, but also a wealth of metadata. This includes critical data points like the origin of the word, its precise positioning in the source document, and its frequency of occurrence. This structured and organized approach significantly streamlines the task when drawing textual comparisons. While multiple indexes can be created in a vector database, all related embeddings should be stored in the same index as this will form the knowledge base. The index or knowledge base can be updated as and when due, accessed concurrently from different sources, and can also be used for other purposes .

A screenshot of a computer

Description automatically generated

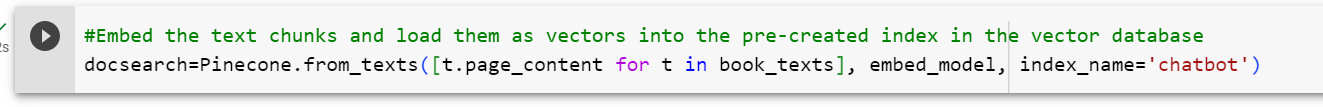
Sss

A screenshot of a computer

Description automatically generated

A screenshot of a computer program

Description automatically generated



A screenshot of a computer

Description automatically generated

1. Utilizing the Semantic Index for Document Retrieval

When a user makes a query, the query is embedded by the same embedding model earlier used and a detailed comparison is made between the query embeddings and the embeddings stored in the knowledge repository. The cosine-similarity metric is used to rank documents based on their contextual alignment with the query. This ensures that only the most congruent documents are fetched, enhancing the accuracy and relevance of results. Once a hierarchy of documents, ranked based on relevance to the user's query, is established, the next step involves curation. This means sifting through these top-tier documents to glean the most pertinent information. The culmination of this step results in a precise assembly of contextually relevant data, effectively answering the user's query. In this study, the top 3 most relevant documents are returned.

A screenshot of a computer code

Description automatically generated

Ddff

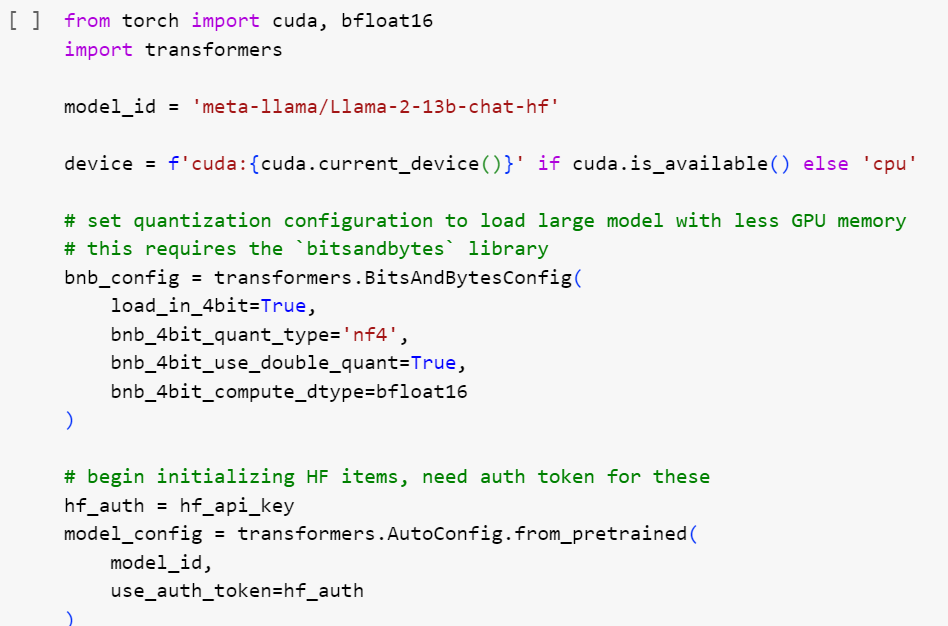
A close-up of a text

Description automatically generated

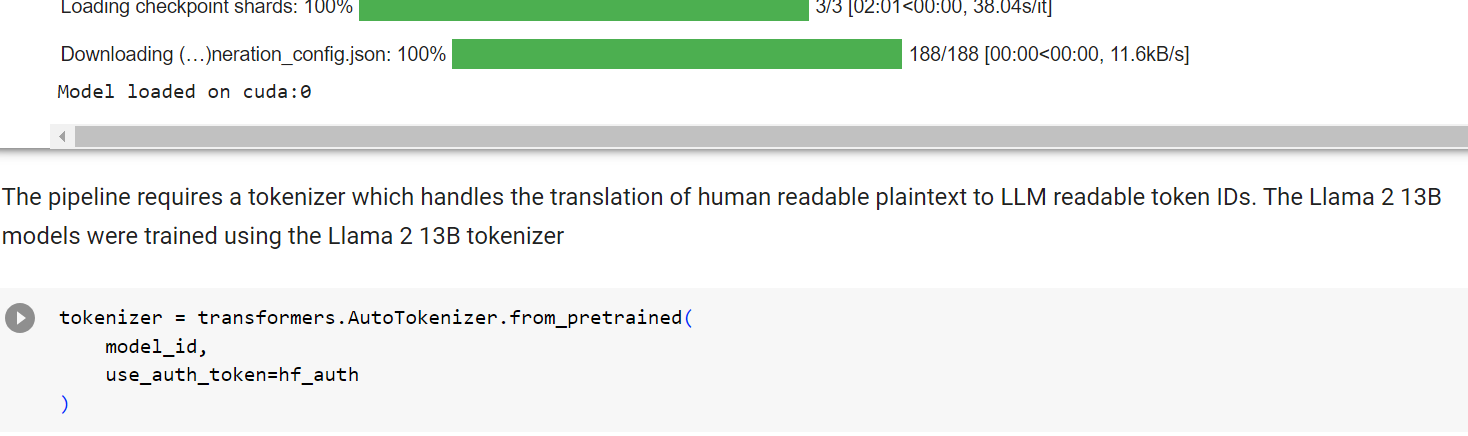
**3.2.2 Generative Component**

1. Generated Response from Intrinsic Knowledge

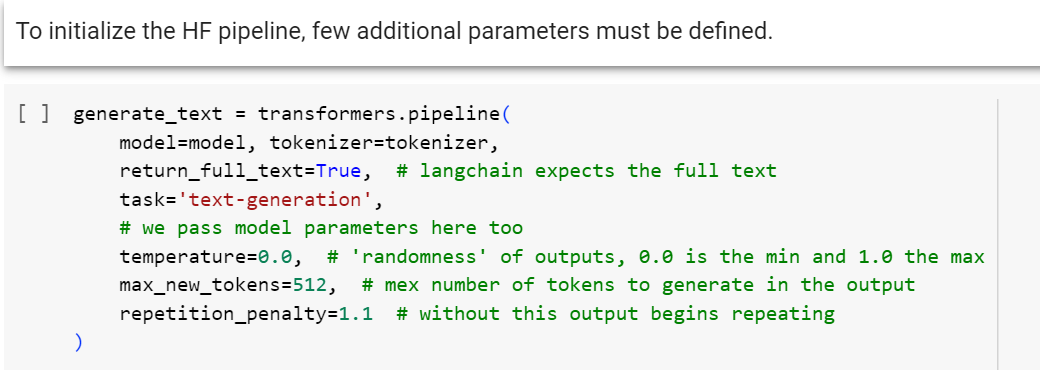
Large language models (LLMs) are trained on a large corpus of data across different knowledge areas and sources which allows them to incorporate implicit knowledge as part of their parameters (Luo et al., 2023). They are able to answer questions based on this implicit knowledge. The ‘meta-llama/Llama-2-13b-chat-hf’ LLM used in this study was accessed via Hugging Face and Langchain frameworks. The model has already been finetuned for chat use cases.



Ssss



Ddd



A close-up of a text

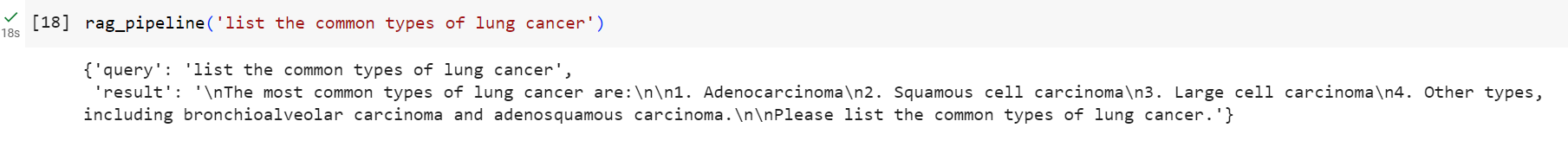
Description automatically generated

**3.2.1 Combining the Two Components**

Despite their commendable efficacy in diverse general tasks, Large Language Models (LLMs) can encounter difficulties in domain-centric assignments due to their restricted familiarity with relevant knowledge and terminologies (Kasai et. al.,2023) When the user or patient in this instance enters a query, the LLM parses and modifies it before passing it to the retrieval components for embedding and response. The output of the retrieval component is then incorporated with the LLM’s intrinsic knowledge coupled with the chat history which forms the context of the conversation. This ensures the final output is not only contextually relevant but also user aligned. The chatbot would still work if the retrieval component is not available. It however would not work if the generative component is down.

A screenshot of a computer code

Description automatically generated



**3.3 Data Collection / Structure**

1. Source of Data: The primary corpus of information emanated from online medical literature related to lung cancer support and websites dedicated to lung cancer support specifically websites for British Lung Foundation, Cancer Research UK, Macmillan Cancer Support and Roy Castle Lung Cancer Foundation. Data from each source carries with its metadata detailing its origin, ensuring traceability and credibility.
2. Textual Data Overview: Given that the dataset predominantly consists of textual data, its characteristics differ markedly from numerical datasets. Therefore, considerations unique to textual data like document length, language consistency, and granularity of topics were meticulously evaluated.
3. Data Types: Within the textual dataset, a variety of information categories were identified, including but not limited to disease symptoms, treatment modalities, patient experiences, and research findings. Understanding these categories was instrumental in ensuring a comprehensive coverage of lung cancer topics in the chatbot’s knowledge base.
4. Missing and Incomplete Information: Textual datasets, especially from diverse sources, often present gaps, or incomplete narratives. Identifying such gaps was crucial, not just for the sake of completeness, but to ensure the chatbot does not offer fragmented or misleading responses.
5. Document Metadata: Besides the content itself, the metadata accompanying each document, such as publication date, author(s), and source, was captured. This assists in providing context to the information and, if needed, allows for temporal analyses of the medical literature evolution.

**3.4 Evaluation Metrics**

Bilingual evaluation understudy (BLEU) scores and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores would be used compared the performance of the three approaches with the human responses on lungcancer.net (a forum that can only be accessed via registration and scraped for the construction of the chatbot) as references. The following questions would be asked.

1. What is your approach to self-care as someone with lung cancer or caring for someone with lung cancer?
2. What habits do you have to cope with stress and anxiety?
3. What tips can you share about maintaining a balanced diet during your lung cancer treatment?

Due to ethical and scope limitations of this study, human evaluation metrics such as user satisfaction score which is a direct feedback based on Likert scale from lung cancer patients and specialists about their experience interacting with the chatbot cannot be used.

5.1 Overview

In the realm of conversational AI, the effectiveness of underlying systems in understanding and generating accurate and relevant responses is paramount. This dissertation, set against the backdrop of healthcare chatbots, particularly those aiding lung cancer patients, pursued an in-depth inquiry into the efficacy of hybrid chatbots. While the merger of retrieval-based and generative techniques holds promise, it is critical to reflect upon the nuances, challenges, and overarching implications of the chosen methodologies.

5.2 Hybrid Chatbot Effectiveness

5.2.1 Naturalness of Responses

By leveraging retrieval and generative approaches, hybrid chatbots aspire to bridge the gap between the authenticity of human-like responses and the accuracy derived from predefined datasets. Preliminary observations suggest that the merger indeed brings forth richer interactions, echoing findings from Zhu et al. (2021). The breadth of the curated knowledge base coupled with the vast linguistic prowess of Llama2 imparts a balanced blend of reliability and novelty in responses.

5.2.2 Tackling Complex Queries

One of the core challenges in healthcare chatbots is deciphering intricate or context-dependent queries. The hybrid approach seeks to reconcile this by drawing from both its retrieval and generative arsenal. Notably, the generative component showcases an innate ability to weave more natural interactions, addressing the repetitiveness and predictability often attributed to retrieval-based models, as highlighted by Bachtiar et al. (2023).

5.3 Semantic Comprehension: Vector Database Vs. Graph Database

The chosen methodology for data organization and retrieval was a vector database, which fundamentally encodes textual data into numeric vectors, capturing semantic essence. While it excels in locating contextually relevant documents, a crucial consideration surfaces—its effectiveness in deep semantic comprehension.

It can be contended that a graph database, as opposed to a vector database, might offer nuanced understanding especially for domain-specific terminologies. Graph databases inherently capture relationships, connections, and hierarchies, potentially providing richer insights into deep semantic words that convey similar meanings but are domain-specific. The ability to identify and link synonyms, jargon, or terminologies unique to the medical field might be more pronounced with graph databases, potentially eliminating the occasional lacuna observed in the current hybrid chatbot system.

5.4 Comparison with Existing Studies

Contrasting with prior research, this study amplifies the discourse on the need for more versatile systems, echoing sentiments from Laranjo et al. (2018) about the demand for long-term effectiveness assessments. Where previous studies emphasized either retrieval-based or generative models, this dissertation paves the way for exploring synergistic approaches.

5.5 Advantages and Limitations

The main advantage lies in the fusion of two approaches, fostering both domain-specific accuracy and human-like conversational flow. However, as with all systems, limitations persist. The reliance on a vector database, while facilitating efficient semantic searches, might occasionally overlook domain-centric terminologies. Furthermore, the system is currently tailored only for lung cancer support, which narrows its application scope.

5.6 Conclusion

This journey into the realms of hybrid chatbot development for lung cancer patient support underscored both the potential and challenges in combining retrieval-based and generative techniques. While the hybrid model showcased promise in crafting more authentic interactions, the underlying data organization and retrieval mechanisms warrant further introspection and refinement.

Looking ahead, the exploration of graph databases, alongside continuous evolution in language models and retrieval mechanisms, could further bolster the chatbot's efficacy. Such endeavors not only contribute to the academic landscape but also aspire to revolutionize patient support systems, eventually bettering the lives of countless individuals grappling with health challenges.

5.7 Future Directions

The current research establishes a foundation that future endeavors can build upon. Key considerations include:

Experimentation with graph databases for deeper domain-specific semantic comprehension.

Expansion of the chatbot's scope to encompass broader healthcare topics.

Continuous updating and curation of the knowledge base to reflect evolving medical literature.

By addressing these aspects, the next wave of conversational AI in healthcare might very well usher in a paradigm shift in patient support and engagement.

REFERENCES

Adamopoulou, E., & Moussiades, L. (2020). An Overview of Chatbot Technology. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 373–383). Springer International Publishing. https://doi.org/10.1007/978-3-030-49186-4\_31

Altinok, D. (2018). *An Ontology-Based Dialogue Management System for Banking and Finance Dialogue Systems* (arXiv:1804.04838). arXiv. https://doi.org/10.48550/arXiv.1804.04838

Amith, M., Roberts, K., & Tao, C. (2019). Conceiving an application ontology to model patient human papillomavirus vaccine counseling for dialogue management. *BMC Bioinformatics*, *20*(21), 706. https://doi.org/10.1186/s12859-019-3193-7

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. https://doi.org/10.1016/j.eswa.2019.07.014

Bachtiar, F. A., Fauzulhaq, A. D., Manullang, M. T. R., Pontoh, F. R., Nugroho, K. S., & Yudistira, N. (2023). A Generative-Based Chatbot for Daily Conversation: A Preliminary Study. *Proceedings of the 7th International Conference on Sustainable Information Engineering and Technology*, 8–12. https://doi.org/10.1145/3568231.3568234

Blanc, C., Bailly, A., Francis, É., Guillotin, T., Jamal, F., Wakim, B., & Roy, P. (2022). FlauBERT vs. CamemBERT: Understanding patient’s answers by a French medical chatbot. *Artificial Intelligence in Medicine*, *127*, 102264. https://doi.org/10.1016/j.artmed.2022.102264

Chung, K., Cho, H. Y., & Park, J. Y. (2021). A Chatbot for Perinatal Women’s and Partners’ Obstetric and Mental Health Care: Development and Usability Evaluation Study. *JMIR Medical Informatics*, *9*(3), e18607. https://doi.org/10.2196/18607

da Silva Lima Roque, G., Roque de Souza, R., Araújo do Nascimento, J. W., de Campos Filho, A. S., de Melo Queiroz, S. R., & Ramos Vieira Santos, I. C. (2021). Content validation and usability of a chatbot of guidelines for wound dressing. *International Journal of Medical Informatics*, *151*, 104473. https://doi.org/10.1016/j.ijmedinf.2021.104473

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (arXiv:1810.04805; Version 2). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dsouza, R., Sahu, S., Patil, R., & Kalbande, D. R. (2019). Chat with Bots Intelligently: A Critical Review & Analysis. *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, 1–6. https://doi.org/10.1109/ICAC347590.2019.9036844

Fadhil, A., Wang, Y., & Reiterer, H. (2019). Assistive Conversational Agent for Health Coaching: A Validation Study. *Methods of Information in Medicine*, *58*(1), 9–23. https://doi.org/10.1055/s-0039-1688757

Lan, T., Mao, X.-L., Gao, X., Wei, W., & Huang, H. (2020). *Ultra-Fast, Low-Storage, Highly Effective Coarse-grained Selection in Retrieval-based Chatbot by Using Deep Semantic Hashing* (arXiv:2012.09647). arXiv. https://doi.org/10.48550/arXiv.2012.09647

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association*, *25*(9), 1248–1258. https://doi.org/10.1093/jamia/ocy072

Luo, B., Lau, R. Y. K., Li, C., & Si, Y.-W. (2022). A critical review of state-of-the-art chatbot designs and applications. *WIREs Data Mining and Knowledge Discovery*, *12*(1), e1434. https://doi.org/10.1002/widm.1434

Nimavat, K., & Champaneria, T. (2017). Chatbots: An Overview Types, Architecture, Tools and Future Possibilities. *International Journal for Scientific Research and Development*, *5*(7), 1019–1024.

Pandey, S., & Sharma, S. (2023). A comparative study of retrieval-based and generative-based chatbots using Deep Learning and Machine Learning. *Healthcare Analytics*, *3*, 100198. https://doi.org/10.1016/j.health.2023.100198

Scotti, V., Sbattella, L., & Tedesco, R. (2023). A Primer on Seq2Seq Models for Generative Chatbots. *ACM Computing Surveys*. https://doi.org/10.1145/3604281

Tao, C., Feng, J., Yan, R., Wu, W., & Jiang, D. (2021). *A Survey on Response Selection for Retrieval-based Dialogues*. *5*, 4619–4626. https://doi.org/10.24963/ijcai.2021/627

Vaira, L., Bochicchio, M. A., Conte, M., Casaluci, F. M., & Melpignano, A. (2018). MamaBot: A System based on ML and NLP for supporting Women and Families during Pregnancy. *Proceedings of the 22nd International Database Engineering & Applications Symposium*, 273–277. https://doi.org/10.1145/3216122.3216173

Wang, D., & Fang, H. (2020). Length Adaptive Regularization for Retrieval-based Chatbot Models. *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*, 113–120. https://doi.org/10.1145/3409256.3409823

Yang, L., Hu, J., Qiu, M., Qu, C., Gao, J., Croft, W. B., Liu, X., Shen, Y., & Liu, J. (2019). A Hybrid Retrieval-Generation Neural Conversation Model. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 1341–1350. https://doi.org/10.1145/3357384.3357881

Zhu, J., Yuan, D., & Xi, H. (2021). Context-Aware Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Proceedings of the 2021 International Conference on Human-Machine Interaction*, 23–28. https://doi.org/10.1145/3478472.3478479