Symptoms-Disease Detecting Conversation Agent using Knowledge Graphs

Ila Ananta
Sonia Khetarpaul
Dolly Sharma
ia813@snu.edu.in
sonia.khetarpaul@snu.edu.in
dolly.sharma@snu.edu.in
Shiv Nadar University
Gautam Buddha Nagar, Uttar Pradesh, India

ABSTRACT

Conversational agents have become extraordinarily popular over the last few years, with accelerated adoption due to COVID-19. Even though a lot of work has been done to devise a real-time agent very few of them focus on dynamic responses. The challenges for automatic medical diagnosis not only include issues for topic transition coherency and question understanding but also issues regarding the context of medical knowledge and symptoms of disease relations. In this paper, we propose a conversational agent that not only generates answers to specific medical questions but also makes more natural and human-like conversations and can adapt to the context and evolve over time. We propose an End-to-End knowledge-routed Relational Dialogue System that would incorporate a rich medical knowledge graph into the topic transition in dialogue management, and make it accommodative with NLU (Natural Language Understanding) and NLG (Natural Language Generation). A knowledge-routed graph for topic decision-making is used, which helps to identify relationships between symptoms and symptom-disease pairs. However, there are constraints on the extent of questions that knowledge graphs can address independently. To overcome these, we have used a fine-tuned GPT-3 model. While knowledge graphs organize data as interconnected entities, GPT-3 generates human-like text using learned patterns from large datasets. This approach enhances responses to intricate queries.

ACM Reference Format:

Ila Ananta, Sonia Khetarpaul, and Dolly Sharma. 2024. Symptoms-Disease Detecting Conversation Agent using Knowledge Graphs. In *2024 Australasian Computer Science Week (ACSW 2024), January 29–February 02, 2024, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3641142.3641165

1 INTRODUCTION

Conversational agents are computer systems designed to converse with humans. They are dialogue systems that use Natural Language Processing (NLP) to respond to queries in human languages. The motive of this work is to design a natural, real-time conversational agent which can be employed to cater to some of the real-world problems specifically that of disease classification based on symptoms.

Disease classification is central to the practice of medicine; it systematizes clinical knowledge and experience. Classification is essential for diagnosis and effective treatment of human disease. To get new insights into disease taxonomy, etiology, and pathogenesis, it is important to understand how diseases are related to each other. Conversational agents are getting used to imitating individuals from a miscellany of fields of work. These include counselors, copy marketers, salespeople, and other mediators. They have shown to be exceedingly functional in domains such as healthcare, business, customer service, marketing, and others ([9]). Studies on various human-bot conversations disclose that, in practice, conversations can be multi-turn and multi-intent ([17]). In multi-turn conversations, there might be a need to extract some missing information from the user's utterances. For instance, if the user asks what the weather is like today, missing particulars like location should be satisfied by the agent before an actual API is called. Furthermore, various studies have shown that humans have a natural tendency to switch between different intentions while conversing. Hence, there is a need for more dynamic and rich abstractions to represent and reason about multi-turn and multi-intent conversational patterns.

Healthcare conversation agents are highly useful in reducing the strain on healthcare workers and patients. They cut down the need for constant hospital visits as the patients are aware of their symptoms and treatment requirements. From the patient's perspective, there is less money spent on different procedures and tests and they get quick access to the doctor. A dialogue-based system for medical diagnosis converses with patients to obtain additional symptoms and make a diagnosis automatically, which has significant potential to simplify the diagnostic procedure and reduce the cost of collecting information from patients. Further, patient condition reports and preliminary diagnosis reports generated by the dialogue system may assist doctors to make a diagnosis more effectively.

Our effort intends to accomplish three key goals and is in line with the major themes of healthcare conversation agents. The primary goal is to minimize user resources and time waste while using chatbot technology to access information. With a human-like interface, we hope to assist users in finding the needed information

ACSW 2024, January 29–February 02, 2024, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1730-7/24/01
https://doi.org/10.1145/3641142.3641165

as soon as possible. The second goal is to give common users with less topic knowledge more accurate responses. With AI technology, we anticipate that the system will be able to comprehend the meaning of natural language and respond with high-quality feedback. Making it simpler to administer and expand the features and databases is the third goal. This would help in creating a system that is scalable and versatile.

The novelty of our work lies in the utilization of a hybrid QA model that combines a knowledge graph database and a fine-tuned GPT3 model. A user's question firstly will be queried from the knowledge graph. If it cannot find any result, the question would be given as input to the GPT-3 model which is fine-tuned on our data.

2 LITERATURE REVIEW

This section focuses on various state-of-the-art works done by researchers in the Healthcare question-answering domain.

Healthcare Helper system with a Hybrid QA model ([7]) is an online web-based medical chatbot system based on Knowledge Graph and Hierarchical Bi-Directional attention. In this paper, the authors have proposed a framework that comprises two modules. The first module is the user interface i.e. the GUI and a back-end to handle the database. The second module is a hybrid QA model which uses a mixture of knowledge graphs and a deep learning-based text representation and similarity comparison model: the HBAM, to respond to user queries.

A Graph-based Chatbot for Cancer Patients ([8]) proposes a chatbot specifically to cater to the needs of cancer patients or people who want to know about cancer. In the proposed framework the collected data is preprocessed using different NLP techniques. This data is then converted to a graph model using a graph database-Neo4j. The engine will shortlist cancer by considering the data provided by the user. Then it suggests remedies and related information to the user.

Task-oriented Dialogue System for Automatic Diagnosis ([16]) proposes a dialogue System that typically contains three components, namely Natural Language Understanding(NLU), Dialogue Manager (DM), and Natural Language Generation (NLG). The DM for automatic diagnosis consists of two sub-modules, namely, dialogue state tracker (DST) and policy learning. It is a reinforcement learning-based framework for medical DS. They have used Markov Decision Process Formulation for Automatic Diagnosis.

DialoGPT (dialogue generative pre-trained transformer) ([18]) is a transformer model trained on 147M conversation-like exchanges extracted from Reddit comment chains over a period spanning from 2005 through 2017. DialoGPT extends the Hugging Face PyTorch transformer to attain performance close to humans both in terms of automatic and human evaluation in single-turn dialogue settings.

DistilBERT base model (uncased) ([14]) fine-tuned with the SQuAD dataset ([5]) can also be used for interactive question-answering. This model is a distilled version of the BERT base model. It is a smaller and faster model as compared to BERT, which was pre-trained on the same corpus in a self-supervised fashion, using the BERT base model as a teacher. To generate the answer you need to provide the context and the question to the model and it generates answers from that context.

Customizing GPT-3 for your application ([3]) highlights how to fine-tune GPT-3 on your own data. Developers may fine-tune a GPT-3 model by harnessing the potential of their own data, a transformative capability that empowers them to craft a tailor-made version ideally suited to their specific application.

Chatbots meet eHealth: automatizing healthcare [6] proposes a chat-bot program that has been skillfully created and trained to behave and communicate with patients like a human being with the aim of offering helpful advice for a number of disease preventive routes. Text-based Healthcare Chatbots Supporting Patients and Health Professional Teams: Preliminary Results of a Randomized Controlled Trial on Childhood Obesity [13] proposes a text-based healthcare chatbot (THCB) to assist patients and health care professionals in therapeutic settings beyond on-site consultation. The design purpose of this chatbot was childhood obesity intervention. Best Practices for Designing Chatbots in Mental Healthcare - A Case Study on iHelpr [10] proposes iHelpr, a chatbot for mental healthcare that offers well-being and self-help material through a conversational interface.

Moreover, several investigations have been undertaken to examine the realm of conversational agents in the healthcare domain. Chatbots as Conversational Healthcare Services [12] analyze how the new chatbots approach design issues pertinent to providing healthcare services, putting a focus on human-AI interaction aspects and the transparency of AI automation and decision-making.

Can Chatbots Help Support a Person's Mental Health? Perceptions and Views from Mental Healthcare Professionals and Experts [15] examine how mental health experts feel about using chatbots and other conversational user interfaces to promote users' mental health and wellness. An online survey has been included in this study to gauge participants' knowledge and opinions toward mental healthcare.

Many other chatbots exist in the medical and healthcare field like Sensely [4] is a virtual nurse app, Babylon [1], a symptom Checker to analyze your symptoms, and SafeDrugBot to provide information about the drugs used during breast-feeding

Drawing inspiration from various pioneering approaches, we present a novel model that embodies the culmination of these efforts. Our chatbot prototype stands out for its ability to engage users in natural and contextually relevant interactions, akin to human-like conversations. Most of the work done till now focuses on a simple question-answer system where the agent answers one question at a time based on the symptom-disease relationship. However, we wish to design an interactive real-time agent which not only answers questions based on different relationships but also generates a follow-up question based on the answer received from the previous question. Therefore, our proposed model tends to be more suitable and useful in real-world scenarios. By delving into the amalgamation of cutting-edge NLP techniques and sophisticated AI algorithms, we demonstrate the potential of our model to users in a new era of chatbot communication, offering practical and adaptable solutions across a diverse range of domains.

3 PROPOSED MODEL AND METHODOLOGY

The healthcare industry is one of the largest and most rapidly expanding sectors globally, catering to a vast customer base with diverse healthcare needs. We introduce a pioneering methodology for assessing disease probabilities based on symptoms, utilizing a knowledge graph as the fundamental framework. By leveraging knowledge graphs, we effectively organize and structure data in a coherent manner, enhancing the accessibility and analyzability of information.

However, there are limitations to what can be answered using knowledge graphs alone. In recent years, there has been a growing interest in using artificial intelligence (AI) for answering complex questions. To address the shortcomings of knowledge graphs, we utilized a fine-tuned version of the GPT-3 language model. While knowledge graphs represent data as a graph of interconnected entities and relationships, GPT-3 is a language model that can generate human-like text based on the patterns it learns from large datasets.

Our focus lay on the approach adopted for training GPT-3 to address medical questions and generate appropriate answers. To achieve this, we utilized a combination of medical question answers sourced from ChatGPT and our Kaggle dataset using prompt engineering, which resulted in 126 question-answer pairs. Prior to fine-tuning the model, we transformed these question-answer pairs into the prompt-completion format, which conforms to GPT-3's requirements. This format serves as a structured means of providing a starting point or context for the language model, enabling it to generate text responses. By implementing this strategy, we aimed to enhance the accuracy and relevance of GPT-3's medical question-answering capabilities, ultimately contributing to medical conversational AI systems.

3.1 Dataset Description

The initial step involved creating a comprehensive compilation of diseases and their common symptoms. To achieve this, we utilized a symptom disease dataset, as depicted in Figure 1. This dataset was obtained from Kaggle [11]. It comprises 4920 rows and 18 columns. Every row gives the symptoms of a particular disease. It is important to note that these rows are not exclusive, i.e., multiple rows of the same disease can exist. Further, the first column gives the name of the disease, and the remaining 17 columns provide the symptoms. A particular disease can have less than 18 symptoms. The dataset served as a valuable foundation, providing information about various diseases and their commonly observed symptoms.

For streamlining the data and enhancing its usability, data preprocessing was performed. This preprocessing involved consolidating the symptoms of each specific disease, resulting in the creation of a combined dataset. As illustrated in Figure 2, this combined dataset presented all the symptoms associated with a particular disease in a cohesive and structured manner. Grouping our diseases reduced the number of rows to 41. Further, all the symptoms for that disease was clubbed under one column. This organization allowed easier navigation and analysis of the data, enabling efficient data retrieval and information extraction. Subsequently, we harnessed the power of the combined dataset to construct a knowledge graph.

Overall, the generated dataset and development of the knowledge graph played a pivotal role in our research, providing a sophisticated and efficient framework for investigating disease-symptom associations and contributing to advancements in medical data analysis.

Disease	Symptom_1	Symptom_2	Symptom_3	Symptom_4
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	
Fungal infection	itching	skin_rash	dischromic _patches	
Fungal infection	itching	skin_rash	nodal_skin_eruptions	
Fungal infection	skin_rash	nodal_skin_eruptions	dischromic _patches	
Fungal infection	itching	nodal_skin_eruptions	dischromic _patches	
Fungal infection	itching	skin_rash	dischromic _patches	
Fungal infection	itching	skin_rash	nodal_skin_eruptions	
Fungal infection	itching	skin_rash	nodal_skin_eruptions	dischromic _patches
Allergy	continuous_sneezing	shivering	chills	watering_from_eyes
Allergy	shivering	chills	watering_from_eyes	
Allergy	continuous_sneezing	chills	watering_from_eyes	
Allergy	continuous_sneezing	shivering	watering_from_eyes	
Allergy	continuous_sneezing	shivering	chills	
Allergy	shivering	chills	watering_from_eyes	
Allergy	continuous_sneezing	chills	watering_from_eyes	
Allergy	continuous_sneezing	shivering	watering_from_eyes	
Allergy	continuous_sneezing	shivering	chills	

Figure 1: Sample original dataset of diseases and their symptoms

3.2 Proposed Model

The model proposed comprises two modules: the frontend to handle the user interface and the backend. Further, the backend is a hybrid QA model that uses a mixture of knowledge graphs and a fine-tuned Generative Pre-trained Transformer 3(GPT-3) to generate answers to users' queries and/or produce a follow-up question.

3.2.1 The Hybrid Q-A Model. The bot is centrally built based on a knowledge graph. Hence, the primary aim is to build a database of selected diseases and their symptoms. Further, the relationship *has_symptom* between different diseases and symptoms is stored in the graph-based database from where we can fetch the results. We perform NLP techniques like lemmatization and stemming on our database. This process reduces words to their base forms, facilitating better search results and improving the efficiency of the knowledge graph's search operations.

The core novelty lies in the design of the knowledge graph, wherein nodes represent diseases in two distinctive formats: individual diseases and groups of diseases with common symptoms. This organization enables the system to swiftly identify sets of diseases that share similar symptomatology, thereby expediting the disease prediction process. Figure 3 shows a node in the knowledge graph representing a single disease. In contrast, Figure 4 showcases nodes representing groups of diseases along with their shared symptoms.

For initiating the disease prediction process, the system employs a variety of NLP techniques to analyze the users' queries. It begins with the tokenization of the query terms. Subsequently, stemming and lemmatization are applied to reduce the tokens to their basic forms. This preprocessing step ensures that the query is adequately prepared for further analysis, thereby increasing the accuracy of symptom extraction from the user input. The user query is segmented, and symptoms are extracted either one phrase at a time or in bi-word phrases. The objective is to identify symptoms that match those stored in the knowledge graph's database. This step

Name	Symptoms
Fungal infection	['itching', ' skin_rash', ' nodal_skin_eruptions', ' dischromic _patches']
Allergy	[' continuous_sneezing', ' shivering', ' chills', ' watering_from_eyes']
GERD	[' stomach_pain', ' acidity', ' ulcers_on_tongue', ' vomiting', ' cough', ' chest_pain']
Chronic cholestasis	['itching', ' vomiting', ' yellowish_skin', ' nausea', ' loss_of_appetite', ' abdominal_pain', ' yellowing_of_eyes']
Drug Reaction	['itching', ' skin_rash', ' stomach_pain', ' burning_micturition', ' spotting_ urination']
Peptic ulcer diseae	[' vomiting', ' loss_of_appetite', ' abdominal_pain', ' passage_of_gases', ' internal_itching', ' indigestion']
AIDS	[' muscle_wasting', ' patches_in_throat', ' high_fever', ' extra_marital_contacts']
Diabetes	[' fatigue', ' weight_loss', ' restlessness', ' lethargy', ' irregular_sugar_level', ' blurred_and_distorted_vision', ' obesity', ' excessive_hunger', ' increased_appetite', ' polyuria']
Gastroenteritis	[' vomiting', ' sunken_eyes', ' dehydration', ' diarrhoea']
Bronchial Asthma	[' fatigue', ' cough', ' high_fever', ' breathlessness', ' family_history', ' mucoid_sputum']
Hypertension	[' headache', ' chest_pain', ' dizziness', ' loss_of_balance', ' lack_of_concentration']
Migraine	[' acidity', ' indigestion', ' headache', ' blurred_and_distorted_vision', ' excessive_hunger', ' stiff_neck', ' depression', ' irritability', ' visual_disturbances']
Cervical spondylosis	[' back_pain', ' weakness_in_limbs', ' neck_pain', ' dizziness', ' loss_of_balance']
Paralysis (brain hemorrhage)	[' vomiting', ' headache', ' weakness_of_one_body_side', ' altered_sensorium']

Figure 2: Sample combined dataset after clubbing the symptoms

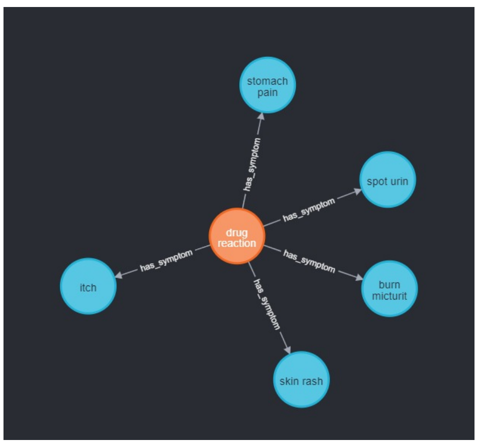

Figure 3: Node of Drug Reaction

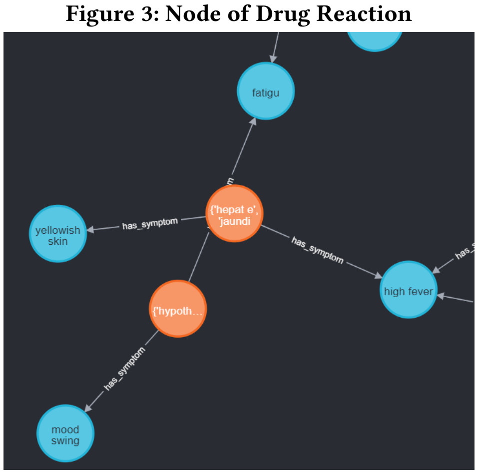

Figure 4: Grouped Node of hepatitis E and jaundice

is critical as it determines whether the query contains relevant symptom information that can aid in predicting potential diseases.

If the symptom/s are found in the database, they are queried to the knowledge graph which generates a set of set of diseases. The system then performs a set intersection operation on the resulting diseases. This results in a grouped set of diseases that possess all the mentioned symptoms in the user query. Through this step, the model gains insight into diseases that align closely with the provided symptoms. The process of grouping diseases with shared symptoms significantly speeds up the prediction process and enhances the accuracy of potential diagnoses. Subsequently, the diseases within the group undergo a symptom subtraction step. Common symptoms among the diseases in the group are eliminated, leaving behind the remaining symptoms. These remaining symptoms are then presented to the user, and the model initiates an iterative interaction with the user to obtain responses regarding the presence or absence of these symptoms. The user's responses play a crucial role in generating probabilities for all possible diseases in the refined group. This probabilistic approach enhances the accuracy of the disease predictions, as the system considers the user's input in assessing the likelihood of each disease in the group. Figure 5 highlights this flow of code through the knowledge graph.

If the system is unable to retrieve any results from the knowledge graph, a fine-tuned version of the DaVinci model of GPT-3 model is employed to respond to users' queries. GPT-3 is an autoregressive language model released in 2020 that uses deep learning to produce human-like text. When given a prompt, it will generate text in continuation of that prompt.

Using the combined dataset mentioned above, a set of question-answer pairs specific to each disease was generated, resulting in 126 pairs that would serve as training data for the fine-tuning process.

Before fine-tuning the GPT-3 model, we transformed these question-answer pairs into the prompt-completion format compatible with GPT-3's requirements. This format allows the model to better understand and generate responses in a prompt-based manner. The conversion to prompt-completion pairs ensured that the model would effectively learn to respond to questions about diseases and their corresponding symptoms. Figure 6 shows this prompt-completion structure [2].

The prompt-completion pairs were structured in a jsonl format, enabling smooth integration with the fine-tuning process. The DaVinci model of GPT-3 was selected for this fine-tuning. This model in GPT-3 is a variant of the language model that uses a

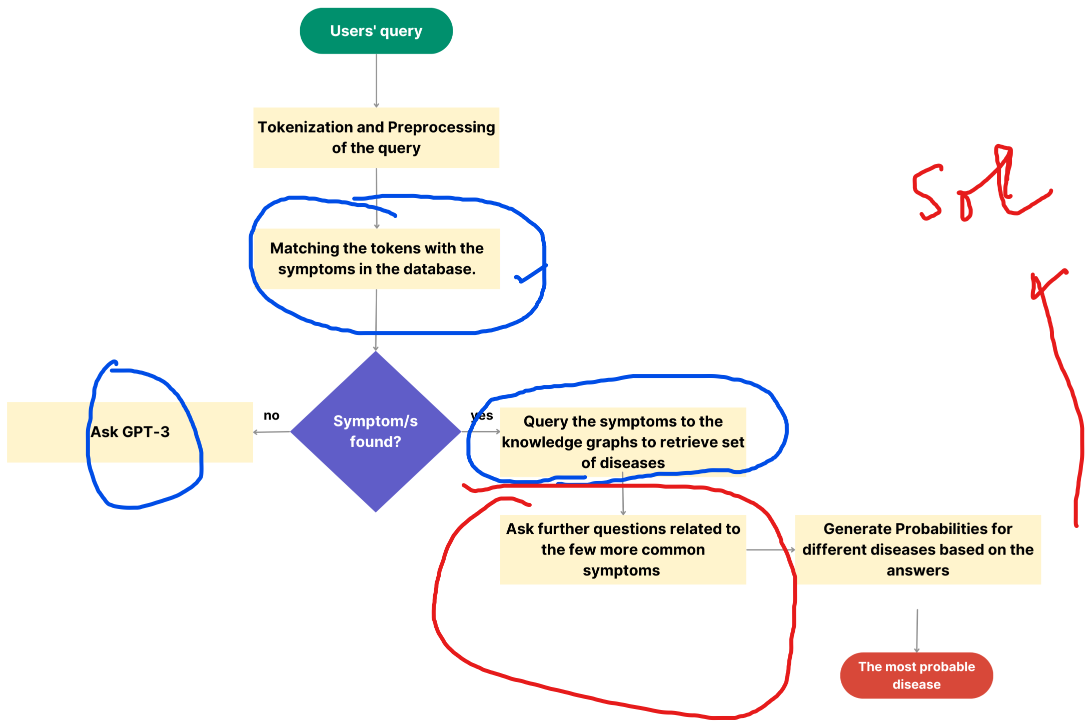

Figure 5: Flow of Code through Knowledge Graphs for disease diagnosis

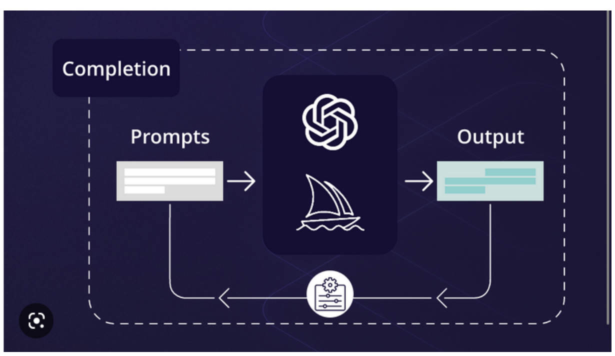

Figure 6: Prompt-Completion Form of GPT-3 [2]

larger training dataset and more advanced training techniques to achieve improved performance on various language tasks. It is capable of generating more human-like text and can produce coherent and informative responses to a wide range of prompts, making it a valuable tool for applications such as natural language understanding, text completion, and conversation generation.

With the completion of the fine-tuning of the GPT-3 model, we can equip the symptom to handle user queries concerning diseases or any other questions that yield no results when queried to the knowledge graph. In such cases, the user's question is passed to the fine-tuned GPT-3 model, which then generates an appropriate and contextually relevant response. Figure 7 highlights the fine-tuning process of GPT-3. By incorporating fine-tuned GPT-3 capabilities into the system, we have extended its ability to effectively respond to a wider range of user queries beyond those found in the initial knowledge graph. This enhancement ensures a more comprehensive and informative user experience when seeking information about diseases or other related topics.

3.2.2 The User Interface. In the development of our web application's frontend, we have leveraged the immense capabilities of the widely acclaimed JavaScript framework, REACT. With its robust and flexible architecture, REACT has proven to be an ideal choice for building dynamic and interactive user interfaces. To optimize our development workflow and enhance performance, we seamlessly integrated Vite, a cutting-edge build tool, and development server for JavaScript applications. The heart of our frontend architecture lies in the thoughtfully designed UI component. This all-encompassing component encapsulates the entire user interface, providing users with a seamless and engaging experience.

To establish a robust connection between the frontend and backend systems, we employed APIs and Websockets. This combination created a highly effective communication channel, enabling real-time data transfer and updates between the client-side and the server-side.

The methodology employed in our work leverages a combined dataset of diseases and their common symptoms to generate a knowledge graph and to fine-tune the GPT-3 model using prompt-completion pair. The preprocessing of the database using lemmatization and stemming ensures optimal information retrieval, while the novel organization of diseases within the knowledge graph accelerates the prediction process. The integration of the fine-tuned GPT-3 model with the system allows for improved response generation, ensuring accurate and relevant answers to user queries,

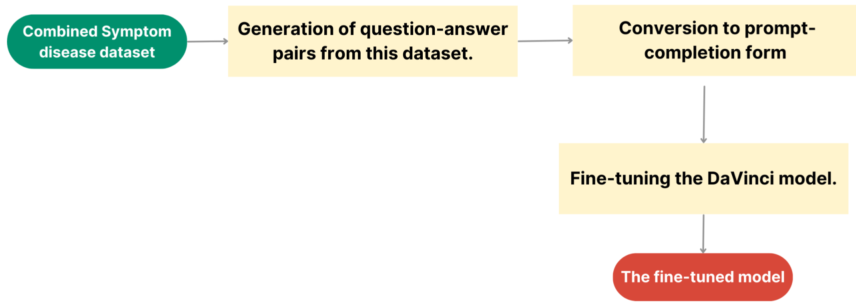

Figure 7: Fine-tuning GPT-3 for medical question-answering

particularly when the initial knowledge graph query yields no results. This approach enhances the system's overall performance and provides users with a more comprehensive and satisfying information retrieval experience. Figure 8 showcases the architecture of our hybrid model.

4 EXPERIMENTATION AND RESULTS

The evaluation of our hybrid QA model yielded promising results, showcasing its effectiveness in addressing user queries and generating coherent follow-up questions. The key factor contributing to its success was the strategic grouping of diseases in the knowledge graph. By organizing related diseases together, the model demonstrated a remarkable ability to comprehend user inputs more accurately and tailor follow-up questions accordingly. This approach proved pivotal in enhancing the overall user experience and encouraging meaningful interactions with the web application.

To comprehensively assess the hybrid model's performance, we encountered the following cases. In the first case, when users provided symptoms already present in our database, the model generated appropriate follow-up questions using the mentioned methodology. These follow-up questions delved deeper into the user's reported symptoms, offering more personalized and relevant information. Moving on to the second scenario, when users inquired about diseases that were already part of our database, the hybrid model deployed the fine-tuned GPT-3 to provide precise and contextually accurate responses. Lastly, in situations where users asked about symptoms or diseases not present in our database, the hybrid model dynamically adapted by employing GPT-3 to generate suitable responses. The integration of GPT-3 as a supplementary resource bolstered the model's understanding of specific disease-related queries, resulting in responses that were comprehensive and tailored to the user's inquiry.

In the first experiment, our innovative approach of utilizing a knowledge graph to generate probabilities for diseases based on reported symptoms demonstrated considerable promise. By tapping into the wealth of information available within the knowledge graph, we could efficiently retrieve data on diseases and their associated symptoms, enabling the generation of probabilities for all potential diseases based on user-provided symptoms. This novel approach holds the potential to significantly enhance the accuracy and speed of disease diagnosis, particularly in cases where patients might struggle to express their symptoms accurately or when multiple diseases share similar manifestations. The ability of our chatbot to consider a wide array of potential diseases based on reported symptoms can be valuable in guiding healthcare professionals toward more precise and timely diagnoses, leading to improved patient outcomes.

For the second experiment, we pursued a fine-tuned version of the GPT-3 language model, specifically trained on a dataset comprising 126 medical question-answer pairs generated from ChatGPT. The results obtained from this experiment demonstrated the promising capabilities of fine-tuned GPT-3 models as a complementary tool for answering complex medical questions. The fine-tuned model provided satisfactory answers to the questions posed, showcasing its ability to grasp medical concepts and decipher intricate language patterns. One noteworthy aspect of the fine-tuned GPT-3 model's performance was its capacity to generate plausible answers for rare medical conditions, even in cases where such conditions were not explicitly represented in the database. This flexibility highlighted the model's potential as a valuable resource for addressing more obscure or less common medical queries, broadening the scope of information available to healthcare professionals and users alike.

Nevertheless, it is crucial to acknowledge that the accuracy and effectiveness of the fine-tuned GPT-3 model heavily depend on the quality and diversity of the training data as well as the prompt completion format used during the fine-tuning process. As such, future endeavors should focus on enhancing the training dataset by incorporating a wider range of medical question-answer pairs and refining the fine-tuning process to optimize the model's performance further.

In our research, we conducted thorough testing of our model, considering various inputs to evaluate its performance and effectiveness in providing diagnoses and symptom information. The results obtained from these tests shed light on the model's capabilities and potential applications in different scenarios.

1) The first type of input we examined involved a patient directly entering the disease they believe they have and seeking a preliminary diagnosis to decide whether immediate medical treatment is necessary or if self-medication is sufficient. We analyzed this scenario in Figure 9, which demonstrates the interaction between the patient and the chatbot in such cases. The results showed that the model was successful in providing relevant and accurate diagnoses based on the patient's input, enabling them to make informed decisions about their health.

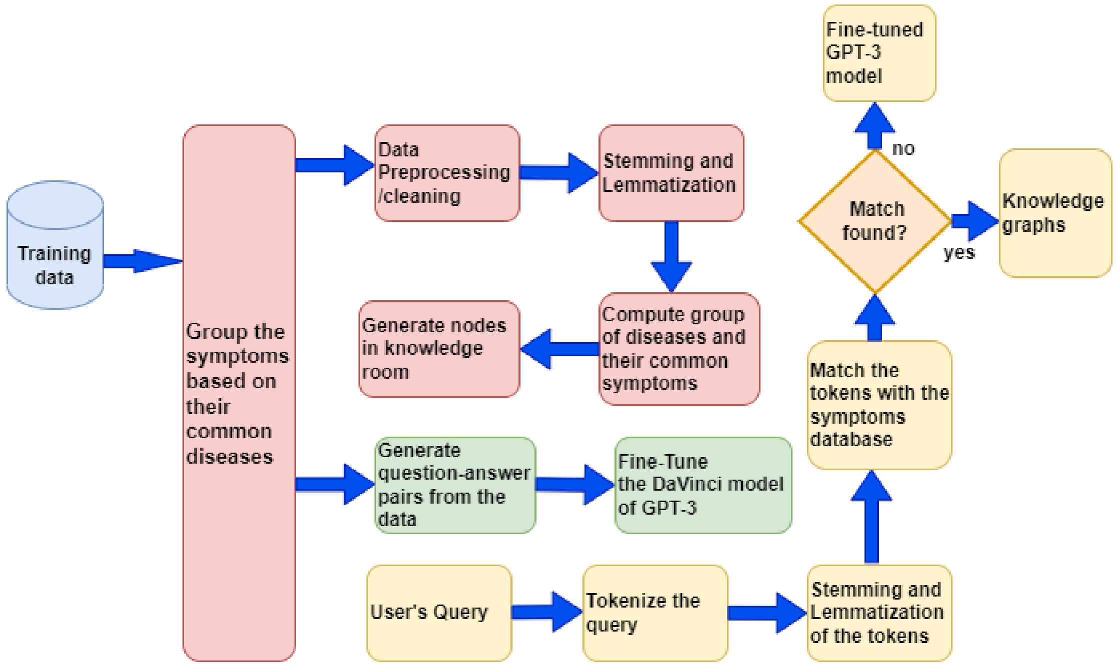

Figure 8: Architecture of the proposed model

```
Hello How can I help you? :I have malaria. what should I do?
Contact a physician immediately for treatment and start taking anti-malaria medication as prescribed.
Do you have more questions:no
```

Figure 9: User's input for inquiring about Malaria

2) The second type of input focused on users who suspected they had a particular disease and sought to retrieve the most common symptoms associated with it for self-assessment. Figure 10 illustrates the chatbot's response in these situations. The model efficiently presented the most relevant symptoms related to the queried disease, assisting users in gaining a better understanding of their health condition and enabling them to seek appropriate medical attention if necessary.

```
Hello How can I help you? :What are the symptoms of typhoid
The symptoms of typhoid include chills, vomiting, fatigue, high fever, constipation, diarrhea, sweats, headache, nausea, loss o
f appetite, mild fever, yellowish skin, abdominal pain, and swelled spleen.
Do you have more questions:no
```

Figure 10: User's input for inquiring about the symptoms of Typhoid

3) The third type of input involved users reporting specific symptoms they were experiencing, prompting the chatbot to interactively diagnose the potential issue. Two separate cases were examined to cater to user preferences. In Figure 11, we explored the scenario where the user requested a general diagnosis, and the chatbot provided a comprehensive assessment based on the reported symptoms. Additionally, in Figure 12, we investigated cases where users explicitly asked for more accurate results. To manage the number of symptoms considered in the diagnosis, we kept a counter that was incremented if the user sought more precise assessments. The results indicated that the model adeptly adjusted its response based on user preferences, delivering appropriate and personalized diagnostic insights.

```
Hello How can I help you? :I have high fever and cough
Do you experience congestion:no
Do you experience yellowing of eyes:yes
Do you experience loss of appetite:yes
Do you experience chest pain:yes
Do you experience phlegm:no
Do you experience rusty sputum:no
Do you want more accurate resultsno

You most probably have Tuberculosis
```

Figure 11: User's input specifying the symptoms they have been experiencing

4) The output, as shown in Figure 13, depicts the knowledge graph generated which has nodes for disease and symptoms. Further, it has nodes with a group of diseases and their shared common symptoms. The edge depicts the relationship *has_symptom* i.e. if the disease node is connected to the symptom node then it is a symptom of that disease or the group of diseases and vice-versa. Our graph comprises of 149 diseases and 129 symptoms with 474 *has_symptom* edges between them.

5 EVALUATION

The evaluation of our model's performance necessitates a dual assessment of its two integral components: the GPT-3 model and the Knowledge Graph integration. A detailed breakdown of these evaluations follows.

```
Hello How can I help you? :I have cough
Do you experience high fever:yes
Do you experience throat irritation:yes
Do you experience sinus pressure:yes
Do you experience runny nose:yes
Do you experience swelled lymph nodes:no
Do you experience mucoid sputum:no
Do you want more accurate resultsyes
Do you experience chest pain:yes
Do you experience fatigue:yes
Do you experience chills:no
Do you experience yellowing of eyes:no
Do you experience vomiting:no
Do you want more accurate resultsno

You most probably have Pneumonia
```

Figure 12: Third input when the user asks for more accurate results

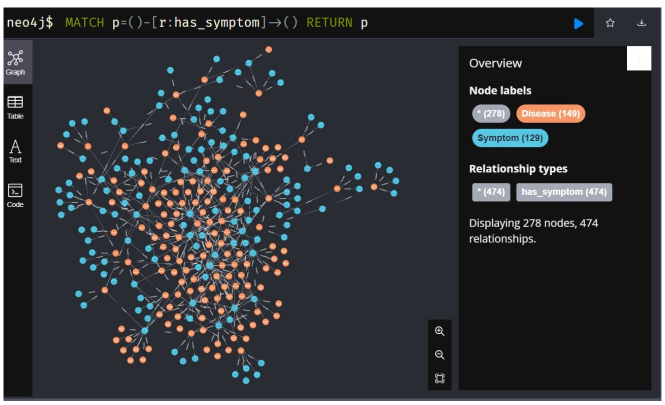

Figure 13: Pictorial representation of knowledge graph

5.0.1 Evaluation of GPT-3 Generated Responses. To rigorously assess the effectiveness of the GPT-3 component, we constructed a dedicated dataset comprising sample questions and corresponding answers. The line of questioning revolved around the symptoms associated with a particular disease, while the answers encompassed all the relevant symptoms present in our dataset. This structured evaluation aimed to quantitatively estimate the degree of semantic similarity between the responses generated by GPT-3 and our designed answers. The evaluation procedure employed the cosine similarity metric, a widely accepted technique for quantifying the similarity between vectors. By subjecting the responses from GPT-3 and our tailored answers to this metric, we derived a numerical measure reflecting their semantic correspondence. Among the 41 question-answer pairs generated, an average cosine similarity of 67.99% was calculated, reaching a remarkable maximum similarity of 98.8%.

Nevertheless, a closer examination is imperative, given the overall accuracy observed at 67.99%. This disparity stems from the fine-tuning process GPT-3 underwent on our specific dataset, absorbing a certain level of inherent knowledge. Consequently, the generated responses reflect the fine-tuned data and encompass the broader pre-existing knowledge of the model. This combination occasionally deviates from our intended answers, leading to lower cosine similarity scores for specific questions.

The amalgamation of fine-tuned data and the model's inherent knowledge is responsible for some questions exhibiting reduced cosine similarity. Figure 14 illustrates the distribution of cosine similarity values for different questions, highlighting these variations. Although we experimented with other matrices, such as the BLEU score, we encountered unsatisfactory results due to the same factor of pre-existing knowledge. The cosine score offered the most valuable insights into our model's performance.

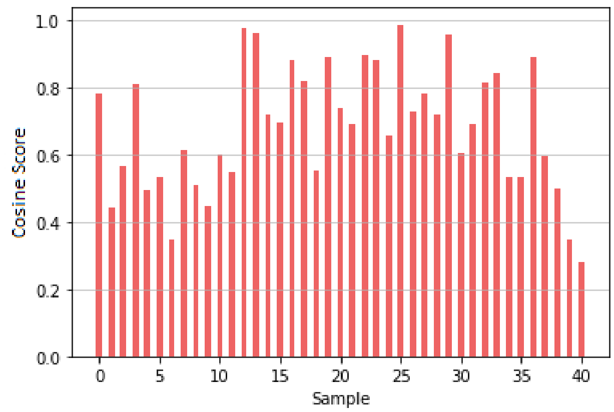

Figure 14: The cosine similarity for different questions generated

Further, drawing comparisons with state-of-the-art models like BERT and HBAB [7], our GPT-3 model gives added functionalities. Firstly, we have fine-tuned the already-existing GPT-3 model. Hence, the answer generated is not solely biased towards our training data. Instead, it contains pre-existing knowledge of GPT-3 as well. Also, compared to the HBAB model, which finds the most similar question from a large QA dataset, our model is not restricted to fixed question-answer pairs and generates a large variety of answers for the given questions.

5.0.2 Evaluation of Knowledge Graph Predictive Performance. The second phase of our evaluation strategy focuses on appraising the predictive capacity of the Knowledge Graph integration. This evaluation was executed manually by providing the Knowledge Graph with symptom-related information concerning a specific disease. The system then harnessed this input to predict the most probable disease while generating pertinent follow-up questions to enhance diagnostic accuracy.

In the course of this evaluation, one distinctive edge cases emerged, warranting thorough consideration. In certain scenarios, if the system inquired about a symptom of the target disease and the user denied experiencing that symptom, the system prematurely halted the process of eliciting further symptoms relevant to that disease. This premature termination led to incorrect predictions due to insufficient symptom data. An illustrative instance of this situation is showcased in Figure 15. Here, the objective was to detect migraines. However, when the user denied having "depression", a symptom of migraines, the system ceased further symptom inquiries, subsequently predicting hypertension as the disease.

```
Hello How can I help you? :I have headache
Do you experience depression:no
Do you experience slurred speech:no
Do you experience fatigue:yes
Do you experience high fever:no
Do you experience sweating:no
Do you experience muscle pain:no
Do you want more accurate resultsyes
Do you experience chest pain:yes
Do you experience loss of balance:no
Do you experience altered sensorium:no
Do you experience spinning movements:yes

You most probably have Hypertension
```

Figure 15: Model detecting Hypertension instead of Migraine

In summation, beyond this identified edge case, the model consistently demonstrated its ability to generate appropriate responses across diverse scenarios. This thorough evaluation has furnished us with a profound understanding of the model's capabilities, unveiling its strengths and areas that warrant refinement. Further, as compared to the other knowledge graph models like the Healthcare Helper system [7], our model uses the knowledge database not only to answer questions but also to generate follow-up questions from the retrieved information. This makes the diagnosis process more interactive and efficient.

It is important to note that these results represent a significant milestone in the development of our model. However, further validation and testing on larger datasets and diverse user populations are essential to ensure its robustness. The performance of the model also highlights the importance of continually updating and refining the underlying data and algorithms to enhance its diagnostic accuracy and user-friendliness.

6 CONCLUSION AND LIMITATIONS

In conclusion, the research conducted in this study represents a significant advancement in the field of healthcare informatics, particularly in the domain of chatbot-based medical diagnosis and symptom assessment. Our novel approach of generating probabilities for diseases based on symptoms using a knowledge graph has shown remarkable promise. Furthermore, the integration of the fine-tuned GPT-3 language model in our approach proved to be highly advantageous. By training GPT-3 on medical question-answer pairs, the model demonstrated its ability to provide satisfactory answers to complex medical inquiries. The fine-tuned GPT-3 complemented our knowledge graph-based approach, offering valuable insights into rare medical conditions and enhancing the accuracy and depth of our chatbot's responses.

The results from our experiments highlight the potential impact of our approach on healthcare accessibility and early disease detection. The chatbot prototype exhibited a capacity to engage users in meaningful interactions and provide informative and personalized diagnostic assessments. Patients and individuals seeking medical advice can benefit significantly from the chatbot's assistance in self-screening for health issues, thereby empowering them to take proactive measures in safeguarding their well-being.

However, we acknowledge that there are still challenges and opportunities for improvement in our research. Although our approach for generating disease probabilities based on symptoms using a knowledge graph shows promise, there are limitations to our current implementation.

One significant limitation is that our current implementation relies heavily on the accuracy of the symptom-disease dataset used to generate the knowledge graph and fine-tune GPT-3. If the dataset is incomplete or contains errors, this could negatively impact the accuracy of our results. To mitigate this limitation, we could consider using a more robust dataset or implementing additional data cleaning and validation procedures.

It is crucial to emphasize that while our chatbot represents a valuable tool in healthcare information retrieval, it should not be regarded as a substitute for professional healthcare workers. As with any AI-driven system, there exists the potential for misleading advice or inaccurate information, which could have detrimental consequences for patients. The chatbot's capabilities are limited to preliminary diagnoses and providing general information, and it should never replace the expertise and personalized care that healthcare professionals offer. Furthermore, one important aspect of future work would be to use a more dense and verified dataset.

Looking ahead, we envision the potential of our chatbot model to be integrated into healthcare systems, local pharmacies, and telemedicine platforms, expanding its reach and impact on healthcare accessibility.

REFERENCES

[1] [n. d.]. babylon-Symptom Checker. https://www.babylonhealth.com/en-us/what-we-offer/chatbot. Accessed: 2023-03-21.
[2] [n. d.]. Best practices for prompt engineering. https://medium.com/@saipragna.kancheti/best-practices-for-prompt-engineering-bf4117833371. Accessed: 2023-02-24.
[3] [n. d.]. Customizing GPT-3 for your application. https://openai.com/blog/customizing-gpt-3#exmple. Accessed: 2023-01-04.
[4] [n. d.]. Sensely an empathy-driven conversational platform. https://sensely.com/. Accessed: 2023-04-20.
[5] [n. d.]. The squad dataset. https://huggingface.co/datasets/squad. Accessed: 2023-01-14.
[6] Flora Amato, Stefano Marrone, Vincenzo Moscato, Gabriele Piantadosi, Antonio Picariello, Carlo Sansone, et al. 2017. Chatbots Meet eHealth: Automatizing Healthcare.. In *WAIAH@ AI* IA*. 40–49.
[7] Qiming Bao, Lin Ni, and Jiamou Liu. 2020. HHH: an online medical chatbot system based on knowledge graph and hierarchical bi-directional attention. (2020), 1–10.
[8] RV Belfin, AJ Shobana, Megha Manilal, Ashly Ann Mathew, and Blessy Babu. 2019. A graph based chatbot for cancer patients. (2019), 717–721.
[9] Nivedita Bhirud, Subhash Tataale, Sayali Randive, and Shubham Nahar. 2019. A literature review on chatbots in healthcare domain. *International journal of scientific & technology research* 8, 7 (2019), 225–231.
[10] Gillian Cameron, David Cameron, Gavin Megaw, Raymond R Bond, Maurice Mulvenna, Siobhan O'Neill, Cherie Armour, and Michael McTear. 2018. Best practices for designing chatbots in mental healthcare–A case study on iHelpr. In *British HCI Conference 2018*. BCS Learning & Development Ltd.
[11] Ebrahim Elgazar. [n. d.]. Doctor's Specialty Recommendation. https://www.kaggle.com/datasets/ebrahimelgazar/doctor-specialist-recommendation-system/versions/1?resource=download&select=Original_Dataset.csv. Accessed: 2023-02-4.
[12] Mlađan Jovanović, Marcos Baez, and Fabio Casati. 2020. Chatbots as conversational healthcare services. *IEEE Internet Computing* 25, 3 (2020), 44–51.
[13] Tobias Kowatsch, Marcia Nißen, Chen-Hsuan I Shih, Dominik Rüegger, Dirk Volland, Andreas Filler, Florian Künzler, Filipe Barata, Sandy Hung, Dirk Büchter, et al. 2017. Text-based healthcare chatbots supporting patient and health professional teams: preliminary results of a randomized controlled trial on childhood obesity. *Persuasive Embodied Agents for Behavior Change (PEACH2017)* (2017).
[14] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108* (2019).

[15] Colm Sweeney, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D Mulvenna, Siobhan O'neill, Martin Malcolm, Lauri Kuosmanen, Catrine Kostenius, Alex Vakaloudis, et al. 2021. Can Chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts. *ACM Transactions on Computing for Healthcare* 2, 3 (2021), 1–15.

[16] Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuan-Jing Huang, Kam-Fai Wong, and Xiang Dai. 2018. Task-oriented dialogue system for automatic diagnosis. (2018), 201–207.

[17] Shayan Zamanirad, Boualem Benatallah, Carlos Rodriguez, Mohammadali Yaghoubzadehfard, Sara Bouguelia, and Hayet Brabra. 2020. State machine based human-bot conversation model and services. In *International Conference on Advanced Information Systems Engineering*. Springer, 199–214.

[18] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536* (2019).