**CAN THO UNIVERSITY**

**COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**🙞 🕮 🙜**

**Logo

Description automatically generated**

**GRADUATION THESIS**

**BACHELOR OF ENGINEERING IN**

**INFORMATION TECHNOLOGY**

**(HIGH-QUALITY PROGRAM)**

**MEDBOT: CHATBOT ABOUT**

**HEALTHCARE**

**Student: Nguyen Trung Tam**

**Student ID: B1910697**

**Class: 2019-2023 (K45)**

**Advisor: Dr. Lam Nhut Khang**

**Can Tho, 12/2023**

**CAN THO UNIVERSITY**

**COLLEGE OF INFORMATION AND COMMUNICATION TECHNOLOGY**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**🙞 🕮 🙜**

**Logo

Description automatically generated**

**GRADUATION THESIS**

**BACHELOR OF ENGINEERING IN**

**INFORMATION TECHNOLOGY**

**(HIGH-QUALITY PROGRAM)**

**MEDBOT: CHATBOT ABOUT**

**HEALTHCARE**

**Student: Nguyen Trung Tam**

**Student ID: B1910697**

**Class: 2019-2023 (K45)**

**Advisor: Dr. Lam Nhut Khang**

**Can Tho, 12/2023**

EVALUATION OF ADVISOR

....................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................................

Advisor

Lam Nhut Khang

ACKNOWLEDGMENTS

I wish to express my deep gratitude and sincere thanks to my professor Lam Nhut Khang– A lecturer at the College of Information and Communication Technology who gave me the golden opportunity to do this wonderful thesis on the topic “MedBot: Chatbot about healthcare”, which also helped me in doing a lot of research and I came to know about so many new things I am really thankful to them. Then I would like to thank the lecturers of Can Tho University, specifically, the lecturers of the College of Information and Communication Technology who taught me invaluable knowledge when I studied.

I am extremely grateful to my family for their love, prayers, and care for the completion of this thesis. I am very much thankful to my friends for their support when I was doing research at Can Tho University.

Sincerely,

Can Tho, 01/12/2023

Nguyen Trung Tam

TABLE OF CONTENTS

[CHAPTER 1: INTRODUCTION 1](#_Toc148190166)

[1.1. Problem 1](#_Toc148190167)

[1.2. Related works 1](#_Toc148190168)

[1.3. Purpose 3](#_Toc148190169)

[1.4. The objectives and scope 4](#_Toc148190170)

[1.5. Research Methods: 4](#_Toc148190171)

[1.6. Outline 4](#_Toc148190172)

[CHAPTER 2: THEORETICAL BACKGROUND 6](#_Toc148190173)

[2.1. Detailed description of the problem 6](#_Toc148190174)

[2.2. Background 6](#_Toc148190175)

[2.2.1. Chatbot 6](#_Toc148190176)

[2.2.2. Neuron network [9] 8](#_Toc148190177)

[2.2.3. Transformer 13](#_Toc148190178)

[2.2.4. BERT 26](#_Toc148190179)

[2.2.5. GPT [12] 27](#_Toc148190180)

[2.2.6. BART [13] 28](#_Toc148190181)

[2.2.7. Tokenizer 32](#_Toc148190182)

[2.2.8. BLEU [14] 34](#_Toc148190183)

[2.2.9. Python 34](#_Toc148190184)

[2.2.10. Tensorflow 35](#_Toc148190185)

[2.2.11. Deep learning 36](#_Toc148190186)

[CHAPTER 3: DESIGN AND IMPLEMENT OF THE CHATBOT 37](#_Toc148190187)

[3.1. Overview 37](#_Toc148190188)

[3.2. Implementation 38](#_Toc148190189)

[3.2.1. Collecting dataset 38](#_Toc148190190)

[3.2.2. Preprocessing dataset 43](#_Toc148190191)

[3.2.3. Constructing dataset 43](#_Toc148190192)

[3.2.4. Training the model 46](#_Toc148190193)

[3.2.5. Inferencing the model 47](#_Toc148190194)

[3.3. Integrate to application 48](#_Toc148190195)

[3.3.1. Facebook messager 48](#_Toc148190196)

[3.3.2. Mobile app 50](#_Toc148190197)

[3.3.3. Web application 55](#_Toc148190198)

[CHAPTER 4: EVALUATION 61](#_Toc148190199)

[4.1. Result 61](#_Toc148190200)

[4.2. Evaluate accuracy 62](#_Toc148190201)

[CHAPTER 5: CONCLUSION AND FUTURE WORK 64](#_Toc148190202)

[5.1. Conclusion 64](#_Toc148190203)

[5.2. Future work 64](#_Toc148190204)

[REFERENCES 65](#_Toc148190205)

[APPENDICES 67](#_Toc148190206)

LIST OF FIGURES

[Figure 1. Neural network 8](#_Toc148190207)

[Figure 2. Activation functions 9](#_Toc148190208)

[Figure 3. Multi-layers perception 10](#_Toc148190209)

[Figure 4. Recurrent neural networks 11](#_Toc148190210)

[Figure 5. Long short term memory network 12](#_Toc148190211)

[Figure 6. Transformer model network 14](#_Toc148190212)

[Figure 7. Vectors of self-attention encoder 16](#_Toc148190213)

[Figure 8. Multiply vector Q and K 16](#_Toc148190214)

[Figure 9. Compute softmax 17](#_Toc148190215)

[Figure 10. Calculate vector Z 18](#_Toc148190216)

[Figure 11. Calculate Q, K, V matrices 19](#_Toc148190217)

[Figure 12. Calculate self-attention 20](#_Toc148190218)

[Figure 13. Separate Q/K/V weight matrices of each head 21](#_Toc148190219)

[Figure 14. Calculate matrix Z for each head 22](#_Toc148190220)

[Figure 15. Calculate matrix Z 23](#_Toc148190221)

[Figure 16, Mask matrix 25](#_Toc148190222)

[Figure 17. Applying mask matrix 25](#_Toc148190223)

[Figure 18. Compute matrix Z with mask matrix 25](#_Toc148190224)

[Figure 19. Linear and softmax layers 26](#_Toc148190225)

[Figure 20. Bidirectional encoder 29](#_Toc148190226)

[Figure 21. Auto-regressive decoder 30](#_Toc148190227)

[Figure 22. Auto-regressive decoder 31](#_Toc148190228)

[Figure 23. BART model 32](#_Toc148190229)

[Figure 24. Processes for building the chatbot 37](#_Toc148190230)

[Figure 25. EHealthChatMini dataset samples 38](#_Toc148190231)

[Figure 26. EHealthChatMini dataset details 39](#_Toc148190232)

[Figure 27. EHealthChat dataset samples 40](#_Toc148190233)

[Figure 28. EHealthChat dataset details 41](#_Toc148190234)

[Figure 29. EHealthVNChat dataset samples 42](#_Toc148190235)

[Figure 30. EHealthVNChat dataset details 43](#_Toc148190236)

[Figure 31. Constructing dataset for traning model 44](#_Toc148190237)

[Figure 32. Overview of training the model 47](#_Toc148190238)

[Figure 33. Steps for inference the model 48](#_Toc148190239)

[Figure 34. Deploy the chatbot to Facebook messager 50](#_Toc148190240)

[Figure 35. Welcome page of the chatbot application 52](#_Toc148190241)

[Figure 36. Signup, sign-in screen mobile app 53](#_Toc148190242)

[Figure 37. Chat screen of the mobile app 54](#_Toc148190243)

[Figure 38. Signup chatbot web page 56](#_Toc148190244)

[Figure 39. Sign-in chatbot web page 56](#_Toc148190245)

[Figure 40. Chat with chatbot web page 57](#_Toc148190246)

[Figure 41. Healthcare questions page 58](#_Toc148190247)

[Figure 42. Doctors page 59](#_Toc148190248)

[Figure 43. Chat with chatbot popup 60](#_Toc148190249)

LIST OF TABLES

LIST OF ABBREVIATIONS

**ABSTRACT**

In recent years, the integration of artificial intelligence (AI) and natural language processing (NLP) technologies has ushered in a new era in healthcare with the emergence of medical chatbots. These intelligent conversational agents hold immense promise in transforming the way healthcare services are delivered, improving patient experiences, and optimizing healthcare workflows.

This abstract provides an overview of the evolving landscape of medical chatbots, their key functionalities, and their impact on the healthcare ecosystem. Medical chatbots are designed to engage in dynamic conversations with users, offering personalized health information, symptom assessment, medication reminders, and appointment scheduling. They empower patients with on-demand access to medical guidance, reducing the burden on healthcare professionals and enhancing overall patient engagement.

**OVERVIEW**

Medical Chatbot – a chatbot application for answering questions about medical health using Python and ML

Nowadays, the necessity for health care is becoming more and more important, especially in a technology world, because of the rapid development of IT and the need for fast and accurate information about healthcare. However, there are a lot of questions and problems that traditional healthcare is limited in providing fast and accurate information. The old manual approach needs humans to search for the info and then prepare the answer for the customers.

Medical chatbot was created to provide a useful tool to assist those who want to find and answer questions about medical health care in the most effective, fast, and convenient way. In addition, the chatbot also supports collecting and updating practical healthcare information from doctors to update the chatbot accordingly. Besides, the chatbot also allows users to search for medical information according to their desire to help them find answers and advice quickly.

# INTRODUCTION

## Problem

Nowadays, healthcare is becoming more and more important, especially in a fast-paced technology world where everyone is so busy with their phone and digital devices. Therefore, the necessity to have a medical chatbot to help people have fast and reliable information about healthcare is considerably important. However, the development of medical chatbots has faced several challenges and problems in the healthcare industry, which modern chatbots aim to address. Here are some key problems that have spurred the development of medical chatbots:

* Healthcare Accessibility: Many individuals face challenges in accessing healthcare services, particularly in remote or underserved areas. Medical chatbots can provide instant, round-the-clock access to basic medical information and advice, bridging the gap in healthcare accessibility.
* Appointment Scheduling: Booking appointments with healthcare providers can be cumbersome and time-consuming. Chatbots can streamline the appointment scheduling process, making it more convenient for patients and reducing the administrative burden on healthcare facilities.
* Information Overload: The internet is filled with vast amounts of health information, making it difficult for individuals to find reliable sources. Medical chatbots offer a trusted source of medical information, helping users sift through the noise to find accurate answers to their health-related questions.
* Health Awareness and Education: Many people lack access to comprehensive health education. Medical chatbots can serve as educational tools, providing users with valuable information about various health topics and promoting health awareness.

While medical chatbots offer innovative solutions to these problems, it's important to recognize that they are not a replacement for professional medical care. They should complement, not substitute, the expertise of healthcare providers. Additionally, ensuring the accuracy of medical information, maintaining data privacy, and addressing ethical concerns remain ongoing challenges in the development and deployment of medical chatbots.

## Related works

Related chatbots:

CITBot [1]: a chatbot system for students at the College of Information and Communication Technology (CICT) of Can Tho University in Vietnam using Rasa. The CITBot chatbot is constructed using a retrievalbased method in a closed domain with short conversations to answer introductory questions about CICT, including programs and staff, academic regulations, study plans and classes. The CITBot trained on the dataset includes 441 questions belonging to 19 intents, 253 entities, 133 stories and 1,336 response actions and can predict and provide relevant answers for misspelled questions. The prediction of entity labels by the CRF model achieves 95% accuracy and 92.78% relavant responses.

Woebot [2]: The conversational agent was built using Decision Tree and appropriate NLP algorithms. conversational agent consists of two platforms. First, a template-based platform that contains conversations with pre-defined options and exercises that assist participants in improving their mental distress based on CBT principles (CBT-based). Second, a generative dialogue platform that allows conversations regarding various emotional issues in an open-ended manner (i.e., without requiring the users to choose predefined conversational options) and provides emotional support. All conversational responses are stored in a database. The specific area of application is Cognitive Behavioural Therapy (CBT) for anxiety and mood disorders targeting young adults who are college students. The data for the training chatbot are collected using surveys. The data processing and analysis are carried out by qualitative analysis of the responses from the surveyors. There were several categories on which Woebot could provide therapy. Results showed that there was a significant improvement in the mean PHQ-9 score compared to controls in the mental condition of participants with 2 weeks of therapy.

**Nura Esfandiari and colleagues** [3] proposed a chatbot model with a vanilla GAN model consists of a generator and a discriminator. Generator is a full transformer model that generates fake answers in test phase. Discriminator includes only encoder part of transformer followed by a classifier. It provides the probability of real or fake answers. The experimental results of the model chatbot with BLEU-4: 0.96, ROUGE-L: 0.965, F-measure: 0.989, Meteor: 0.87 on the Chit-Chat dataset and BLEU-4: 0.71, ROUGE-L: 0.818, F-measure: 0.622, Meteor: 0.52 on the Cornell dataset.

**EduChat** [4] was proposed by Yuhao Dan and colleagues with an LLM-based chatbot system for intelligent education. The model was first pretrained on a large number of educational books and 4 million cleaned diverse instructions to learn the fundamental knowledge. After that, the model was finetuned on 500 thousand high-quality cus- tomized instructions to activate education-specific functions by aligning with the feedbacks from psychology experts and frontline teachers. EduChat achieved an avenger 40.7 score without retrieval and 49.3 score with retrieval respectively on the C-Eval benchmark, a comprehensive Chinese evaluation suite for foundation models.

**Kyle Swanson and colleagues** [5] implemented a retrieval-based chatbots with dual encoders (two neural encoders) to encode the context and the response, respectively. Each encoder takes a sequence of tokens as input. Each encoder consists of a recurrent neural network followed by a multi-headed attention layer to perform pooling. The multi-layer, bidirectional SRUs are used as the recurrent network. The model achived the AUC (area under the receiver) of 0.977, AUC@0.1 (area under the portion of the ROC curve where the false positive rate is ≤ p) of 0.885, AUC@0.05 with 0.816 and AUC@0.01 of 0.630.

**MILABOT** [6]**:** deep reinforcement learning chatbot developed by the Montreal Institute for Learning Algorithms (MILA) for the Amazon Alexa Prize competition. MILABOT is capable of conversing with humans on popular small talk topics through both speech and text. The system consists of an ensemble of natural language generation and retrieval models, including template-based models, bag-of-words models, sequence-to-sequence neural network and latent variable neural network models. The bot resulting with average user score of 3.15 with A/B testing with about eight hundred user ratings were collected after discarding returning users.

DiagGPT: DiagGPT stands for Dialogue in Diagnosis model Based on GPT-4. This is a multi-agent AI system, which has automatic topic management ability to enhance its utility in task-oriented dialogue scenarios. The system is designed to guide users towards a specific goal and assist them in accomplishing the task throughout the dialogue progression. This is achieved by advancing a sequence of predefined topics throughout the dialogue. DiagGPT has the ability to proactively pose questions based on a predefined checklist, thereby collecting necessary information from users. The system is capable of automatically managing topics throughout the dialogue, tracking topic progression, and effectively engaging in discussions centered around the current topic. It performs well in managing various topic changes in complex dialogues.

CHAI [7]: an algorithm for learning task oriented dialogue that utilizes a language model in conjunction with offline RL that leads the policy to generate goal-oriented dialogue that is both realistic and functional, and does not require training against a simulated model of human language. CHAI fine-tune on task-specific dialogue corpus using the proposal sampling approach, the target value Q is computed via a modified Bellman operator that utilizes a proposal distribution based on the language model, to generate N response proposals, and then uses the target Q-function, to score those responses and selects the highest one. CHAI achive Fluency, Coherency, Human-Likeness on the Human evaluation.

## Purpose

The application "MedicalBot: Chatbot about healthcare" is used for users around the world who have healthcare problems in particular and users across the world who want to search for information about healthcare and advice for avoiding health issues. Practical healthcare news, advice, and answers are provided to users who need fast and reliable information about health issues.

## The objectives and scope

The application provides main functions to help users find, search, and explore information related to healthcare problems and advice. This thesis will be focused on the problem of searching and providing reliable answers to the medical industry. The scope of the study is: researching the problem of searching and finding health-related information online and solving the problem by implementing a chatbot with friendly UI/UX and handy features.

## Research Methods:

- Requirements analysis: study the problem related to chatbots, especially in the medical field on the network, research papers, then analyze the function, and describe the requirements to build and train the chatbot.

- Data collection: Collect questions, answers, and relevant information about the medical problem to train the chatbot.

- Implementation: Using Python, and Google Colab for training the chatbot and then deploying the chatbot to Facebook messager with Python, and Flask using REST API.

## Outline

- Using acquired knowledge of analysis, research, and information gathering, build an application with medical chatbot Python, and Machine Learning.

Thesis outline

**Chapter 1**: Introduction

This chapter will cover the overview of chatbot, theory to implement it. Besides, the chapter also mention the objective and scope of the thesis as well as the method for researching and implementing the chatbot.

**Chapter 2**: Background

General information about the study and main functions of the system. Categorizing chatbot and neural network and it’s variant architecture that are used related to text, the Transformer model and its variant such as BART.

**Chapter 3**: Design and implementation

Introduction of UI/UX designs, models, and implementation, describing technologies that will be used in the study. The chapter also cover the process for collecting, processing and constructing the dataset, the training model process as well as the evaluation of the model.

**Chapter 4**: Results

This chapter will presenting how to assess the chatbot, evaluate the model with BLEU score and the experimental result of the model.

**Chapter 5**: Conclusion and future work

The chapter will summarize the results of the research and make recommendations development direction for the topic.

# THEORETICAL BACKGROUND

This chapter introduces the basics of chatbots, along with concepts neural networks, how neural networks work, extended models of neural networks (RNN, LSTM), the Transformer models and BART model along with the evaluation method. Besides, this chapter also provide information about the tools and framework are used to build the chatbot system.

## Detailed description of the problem

Medical chatbot provides users with functions such as answering questions related to healthcare and searching for advice about health issues in an easy, fast, and reliable way. In addition, the chatbot feature is integrated into the website to help students get answers related to actual practice accurately and quickly.

Medical chatbot application consists of 2 main parts: the chatbot and the UI application.

The application part helps users interact with the chatbot and supports collecting questions and answers to improve the effectiveness of the training chatbot. In addition, the application section also supports users to find questions, answers, and news related to practical healthcare problems. The application section will also support an administrator function to maintain and update the chatbot to improve its accuracy of the model over time.

## Background

### Chatbot

A chatbot (conversational agent (CA), dialogue system) is a computer software that acts as an interface between human users and a software application, using spoken or written natural language as the primary means of communication.

Chatbot components [8]

The functioning of a chatbot involves several key components:

Natural Language Processing (NLP): Natural language understanding (NLU) is the first core component of the conversational agents which is responding by providing a semantic representation for user utterance such as in the form of logic or class’s intent, extracting the “meaning” of an utterance. NLP is the technology that enables chatbots to understand and interpret human language.

Dialog Management: Dialogue Manager is The second core component in any chatbot and we can differentiate the chatbots through this component which has many parts that can be improved or adding some parts in the future if will be discovered that it will serve the DM.

Knowledge Base or Backend Integration: Many chatbots are connected to databases, APIs, or other systems to access information or perform actions. For example, a customer support chatbot might access a database of FAQs or connect to a CRM system to look up customer information.

User Interface: This is the medium through which users interact with the chatbot. It could be a chat window on a website, a messaging app like WhatsApp, or a voice-activated device like Amazon Echo.

Machine Learning (Optional): Some advanced chatbots incorporate machine learning algorithms. These bots can learn from user interactions to improve their responses over time. They become more effective as they gather more data and refine their understanding of user queries.

Types of chatbot

There are different types of chatbots based on their capabilities and functionalities:

Rule-Based Chatbots: These chatbots follow predefined rules and patterns. They provide responses based on keywords and phrases. Rule-based chatbots are typically used for straightforward and specific tasks.

AI-Powered Chatbots: These chatbots leverage artificial intelligence and machine learning to provide more dynamic and context-aware responses. They can handle more complex and open-ended conversations.

Voice Assistants: Voice-activated chatbots, like Siri, Google Assistant, or Alexa, are designed to respond to spoken language inputs. They use speech recognition technology to understand and fulfill user requests.

Retrieval-based: The retrieval-based Chatbots select the best matching answer for the user’s question by searching a pre-constructed conversational repository.

Learning-based: This approach is usually based on Seq2Seq learning model. In the case of long sentences and conversations (more than 20 words), all essential information of a source sentence should be compressed into a fixed-length vector which is challenging.

Generative-based: Generative-based methods learn the answer distribution using the generative models. These models use the Reinforcement Learning (RL) techniques. In the SeqGAN, The RL reward signal comes from the discriminator, which is judged on a complete sequence, and fed back to the intermediate state action steps using Monte Carlo search.

### Neuron network

The inspiration for a Neural Network [9] (NN) originates from the human brain, where biological neurons (nerve cells) respond to the activation of other neurons they are connected to. At a very simple level, neurons in the brain take electrical inputs that are then channeled to outputs.

Input: It is the set of features that are fed into the model for the learning process. For example, the input in object detection can be an array of pixel values pertaining to an image.

Weight: Its main function is to give importance to those features that contribute more towards the learning. It does so by introducing scalar multiplication between the input value and the weight matrix. For example, a negative word would impact the decision of the sentiment analysis model more than a pair of neutral words.

Transfer function: The job of the transfer function is to combine multiple inputs into one output value so that the activation function can be applied. It is done by a simple summation of all the inputs to the transfer function.

Activation Function: It introduces non-linearity in the working of perceptrons to consider varying linearity with the inputs. Without this, the output would just be a linear combination of input values and would not be able to introduce non-linearity in the network.

Bias: The role of bias is to shift the value produced by the activation function. Its role is similar to the role of a constant in a linear function.

For a single neuron/node with input u ∈ R n, a mathematical model, named the perceptron, can be described as

|  |  |
| --- | --- |
|  | (1) |

where y is the activation of the neuron/node, Wi are the weights and b is the bias.

The function σ : R → R is called activation function. Originally, it was proposed to choose the Heaviside function as activation function to model whether a neuron fires or not

|  |  |
| --- | --- |
|  | (2) |

Popular activation functions are:

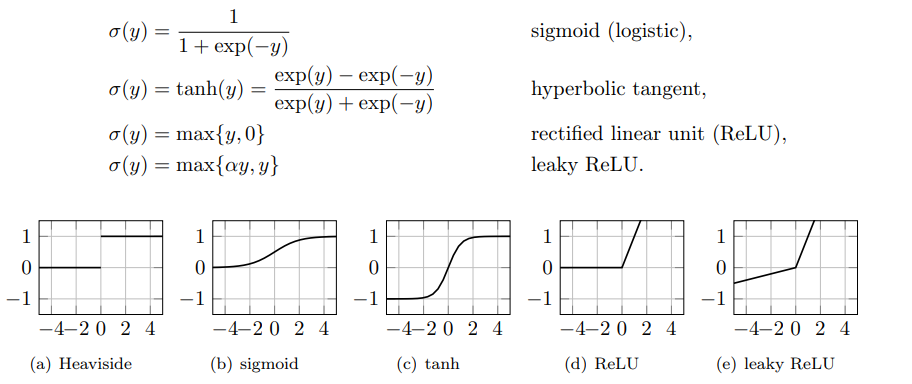


Figure 2. Activation functions

When multiple neurons are stacked together in a row, they constitute a layer, and multiple layers piled next to each other are called a multi-layer neural network.

ANNs consist of interconnected artificial neurons or nodes organized in layers, each layer serving a specific purpose in the network. Here's an overview of the architecture of artificial neural networks:

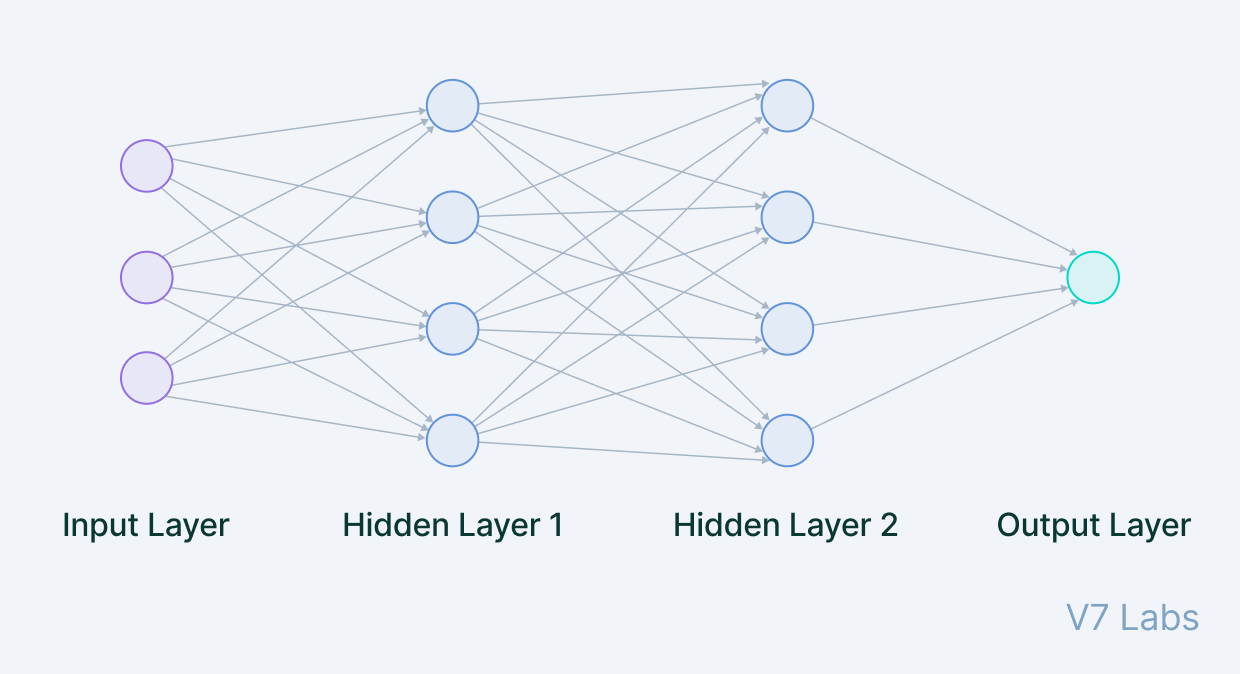


Figure 3. Multi-layers perception

Input Layer: The input layer is the first layer of the neural network. Neurons in this layer receive input data, which could be features from a dataset or raw sensor data, depending on the application. The number of neurons in the input layer is determined by the dimensionality of the input data.

Hidden Layers: Between the input and output layers, there can be one or more hidden layers. Each hidden layer consists of multiple neurons. Neurons in the hidden layers process the input data through weighted connections and activation functions. The number of hidden layers and neurons in each layer is a hyperparameter that can be adjusted to optimize network performance.

Output Layer: The output layer takes input from preceding hidden layers and comes to a final prediction based on the model’s learnings. It is the most important layer where we get the final result. In the case of classification/regression models, the output layer generally has a single node. However, it is completely problem-specific and dependent on the way the model was built.

Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a type of neural network architecture which is mainly used to detect patterns in a sequence of data. What differentiates Recurrent Neural Networks from Feedforward Neural Networks also known as Multi-Layer Perceptrons (MLPs) is how information gets passed through the network. While Feedforward Networks pass information through the network without cycles, the RNN has cycles and transmits information back into itself. This enables them to extend the functionality of Feedforward Networks to also take into account previous inputs and not only the current input. Recurrent Neural Networks have the power to remember what it has learned in the past and apply it in future predictions.

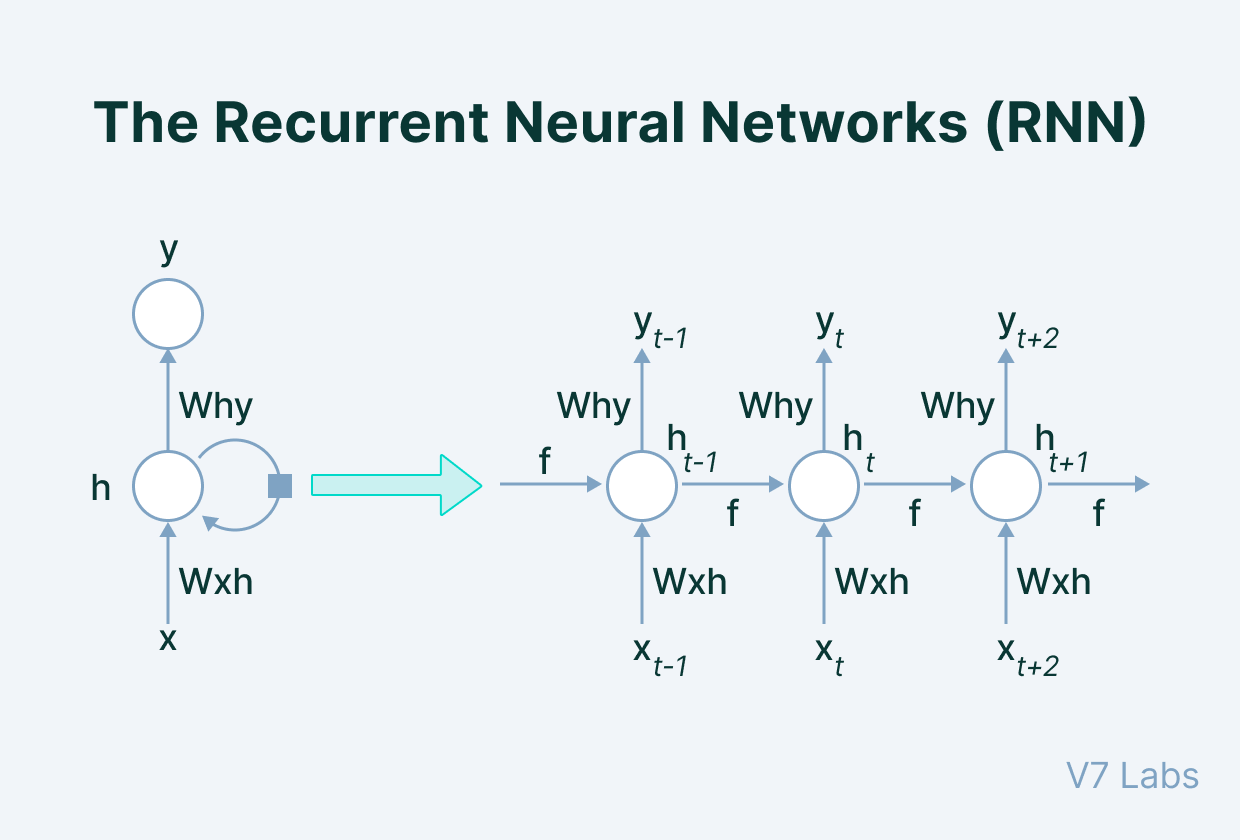


Figure 4. Recurrent neural networks

The input is in the form of sequential data that is fed into the RNN, which has a hidden internal state that gets updated every time it reads the following sequence of data in the input. The internal hidden state will be fed back to the model. The RNN produces some output at every timestamp.

We can describe this process of passing information from the previous iteration to the hidden layer as below:

|  |  |
| --- | --- |
|  | (3) |

Where:

: The hidden state at time step t

: The input at time step t

: The weight matrix

: The hidden-state-to-hidden-state matrix

: The bias

: The activation function

The Long Short Term Memory Network (LSTM)

Long Short Term Memory networks – usually just called “LSTMs” – are a special kind of RNN, capable of learning long-term dependencies. In RNN each of our predictions looked only one timestamp back, and it has a very short-term memory. It doesn't use any information from further back. To rectify this, we can take our Recurrent Neural Networks structure and expand it by adding some more pieces to it. The critical part that we add to this Recurrent Neural Networks is memory. We want it to be able to remember what happened many timestamps ago. To achieve this, we need to add extra structures called gates to the artificial neural network structure.

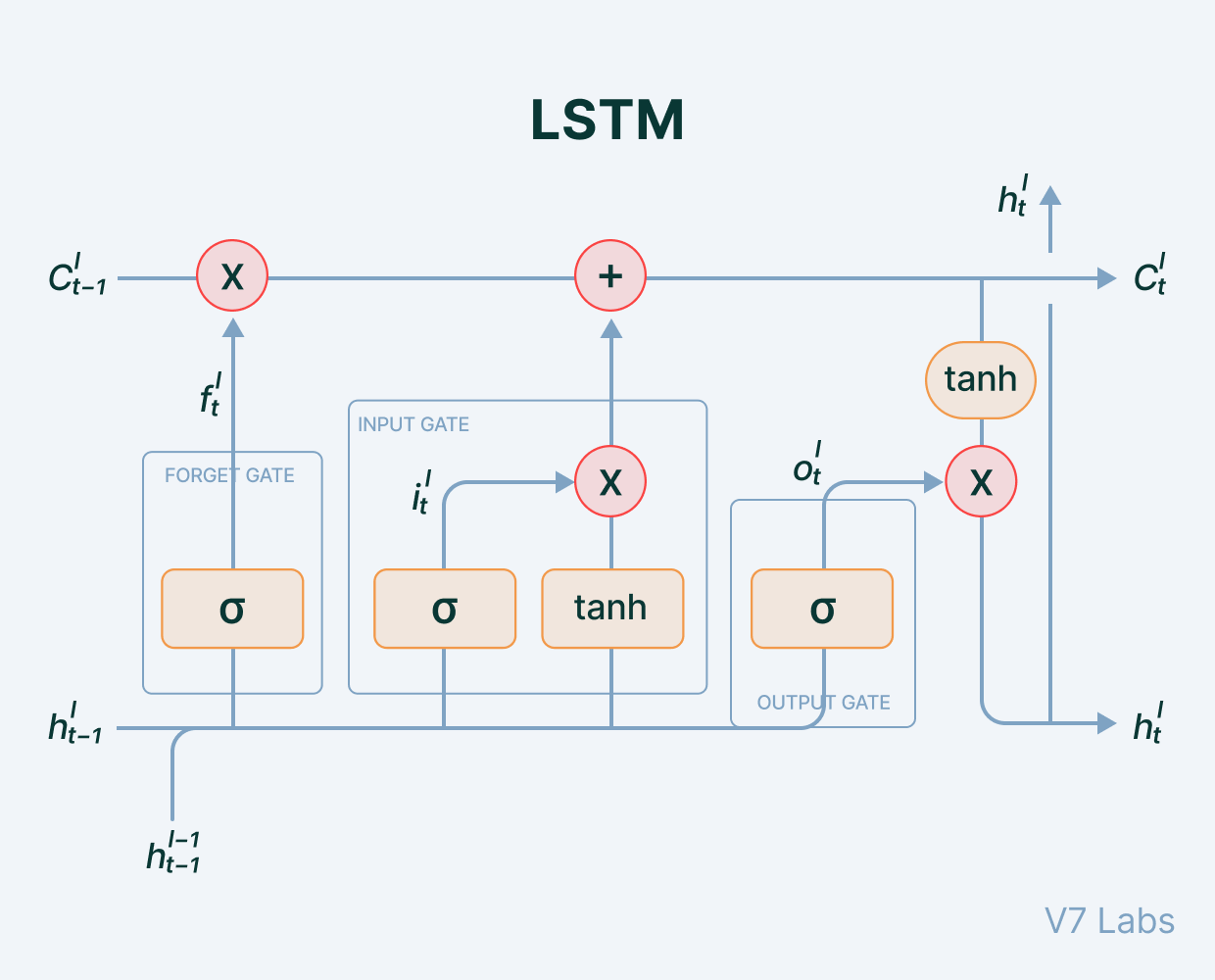


Figure 5. Long short term memory network

Cell state: the key to LSTMs, the cell state is kind of like a conveyor belt. It runs straight down the entire chain, with only some minor linear interactions. It corresponds to the long-term memory content of the network.

The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates. Gates are a way to optionally let information through. They are composed out of a sigmoid neural net layer and a pointwise multiplication operation. The sigmoid layer outputs numbers between zero and one, describing how much of each component should be let through.

Forget Gate: Some information in the cell state is no longer needed and is erased. The gate receives two inputs, x\_t (current timestamp input) and h\_t-1 (previous cell state), multiplied with the relevant weight matrices before bias is added. The result is sent into an activation function, which outputs a binary value that decides whether the information is retained or forgotten.

Input gate: It decides what piece of new information is to be added to the cell state. It is similar to the forget gate using the current timestamp input and previous cell state with the only difference of multiplying with a different set of weights.

Output gate: The output gate's job is to extract meaningful information from the current cell state and provide it as an output.

### Transformer

The Transformer was proposed in the paper “Attention is All You Need” [10]. The Transformer models have an encoder-decoder structure. The encoder maps an input sequence of symbol representations to a sequence of continuous representations. The decoder then generates an output sequence of symbols one element at a time. At each step the model is auto-regressive consuming the previously generated symbols as additional input when generating the next.

The Transformer using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in below.

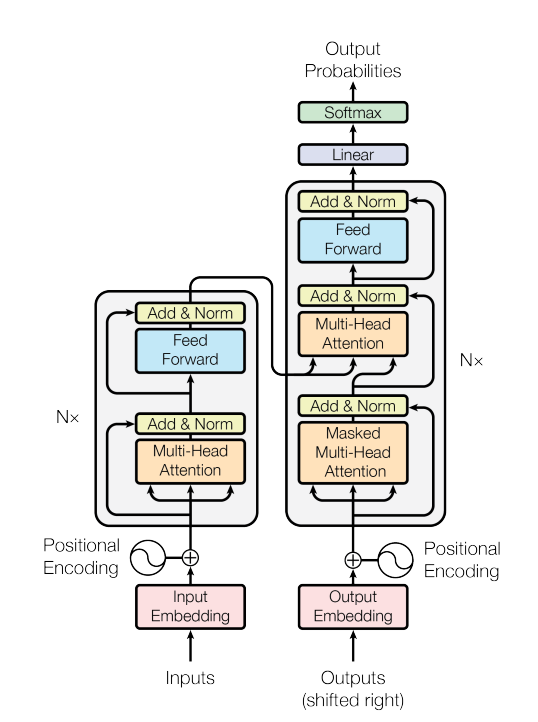


Figure 6. Transformer model network

#### Encoder

The encoder is composed of a stack of N identical layers. Each layer has two sub-layers. The first is a multi-head self-attention mechanism, and the second is a simple, positionwise fully connected feed-forward network. We employ a residual connection around each of the two sub-layers, followed by layer normalization. That is, the output of each sub-layer is LayerNorm(x + Sublayer(x)), where Sublayer(x) is the function implemented by the sub-layer itself. To facilitate these residual connections, all sub-layers in the model, as well as the embedding layers, produce outputs of dimension .

Embedding

An embedding is a mapping of a discrete, categorical variable to a vector of continuous numbers. In the context of neural networks, embeddings are low-dimensional, learned continuous vector representations of discrete variables. In the Transformer model, the embedding layer is used to represent the input into a low-dimensional vector using word embedding.

Positional Encoding

Since the model contains no recurrence and no convolution, in order for the model to make use of the order of the sequence, the model must be injected some information about the relative or absolute position of the tokens in the sequence. To this end, the "positional encodings" are added to the input embeddings at the bottoms of the encoder and decoder stacks. The positional encodings have the same dimension as the embeddings, so that the two can be summed. There are many choices of positional encodings, learned and fixed.

The sine and cosine functions are used to calculate the positional encoding of different frequencies:

|  |  |
| --- | --- |
|  | (4) |

|  |  |
| --- | --- |
|  | (5) |

where *pos* is the position and *i* is the dimension. That is, each dimension of the positional encoding corresponds to a sinusoid. The wavelengths form a geometric progression from 2π to 10000 · 2π.

Attention

Attention allows Transformer to look at other positions in the input sequence for clues that can help lead to a better encoding for this word. An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

Steps to calculate the attention:

Step 1: Creating three vectors from each of the encoder’s input vectors (in this case, the embedding of each word). So for each word, we create a Q vector (query vector), a K vector (key vector), and a V vector (value vector). These vectors are created by multiplying the embedding by three matrices that we trained during the training process.

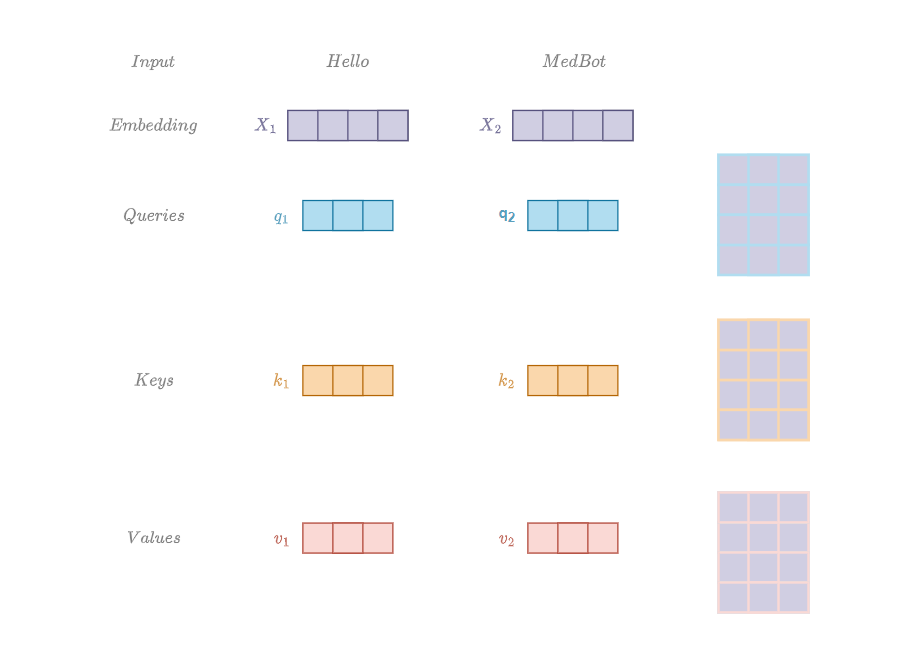


Figure 7. Vectors of self-attention encoder

Step 2: Calculating the score (the score determines how much focus to place on other parts of the input sentence as we encode a word at a certain position) by taking the dot product of the query vector with the key vector of the respective word.

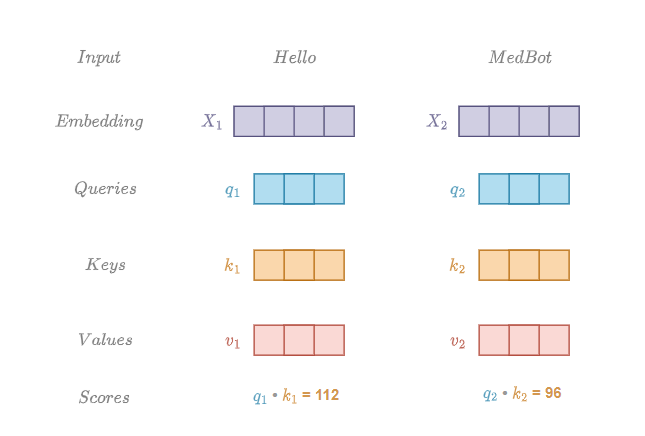


Figure 8. Multiply vector Q and K

Step 3: Divide the scores by the square root of the dimension of the key vectors (help stabling the gradients), then pass the result through a softmax operation. Softmax normalizes the scores so they’re all positive and add up to 1.

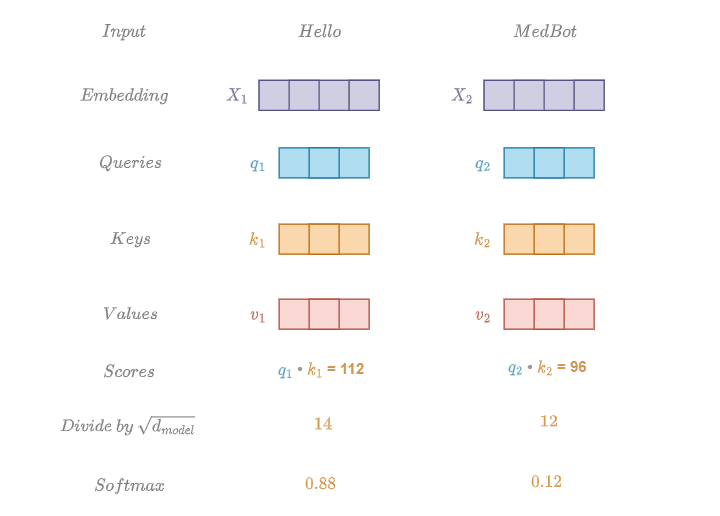


Figure 9. Compute softmax

Step 4: Multiply each value vector by the softmax score to keep intact the values of the word(s) and drown-out irrelevant words.

Step 5: Sum up the weighted value vectors. This produces the output of the self-attention layer

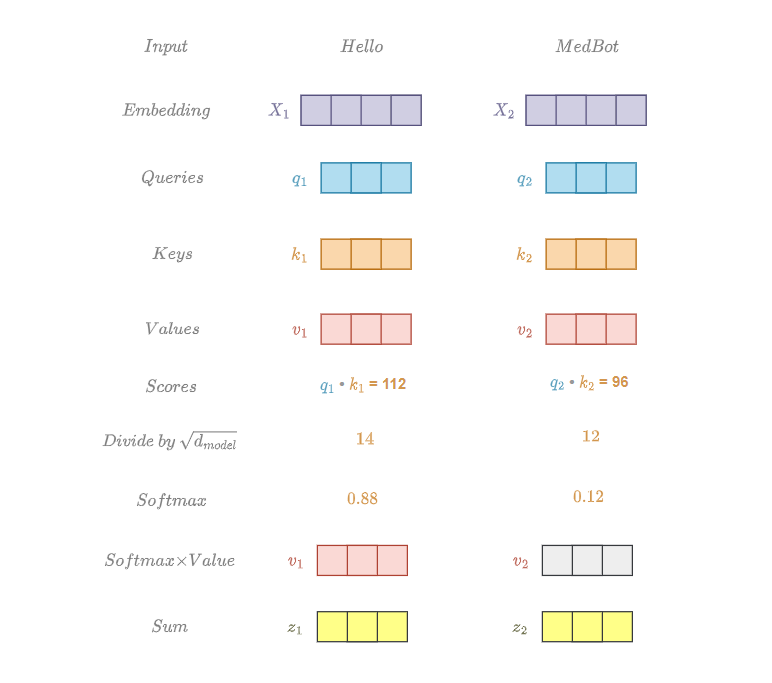


Figure 10. Calculate vector Z

Matrix Calculation of Self-Attention

In the actual implementation of the Transformer model, the attention are calculated using matrices by packing the Q, K, and V embedding matrices into a matrix X, and multiplying it by the weight matrices we’ve trained (WQ, WK, WV).

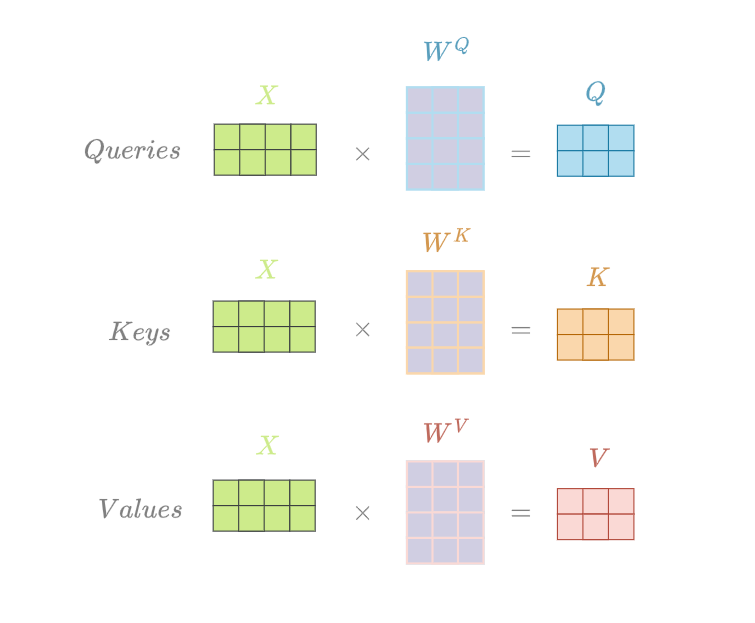


Figure 11. Calculate Q, K, V matrices

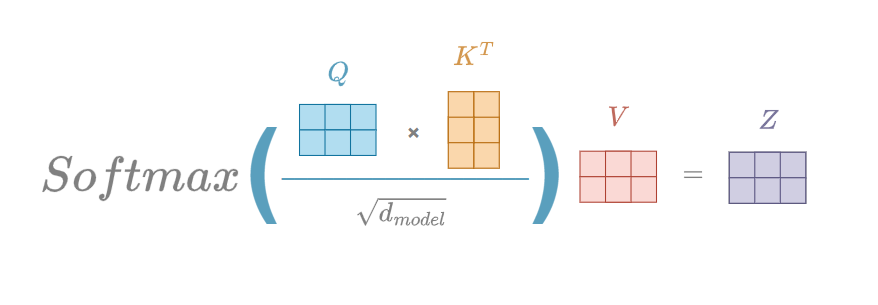


Figure 12. Calculate self-attention

Multi-Head Attention

Instead of performing a single attention function with dmodel-dimensional keys, values and queries, we found it beneficial to linearly project the queries, keys and values h times with different, learned linear projections to dk, dk and dv dimensions, respectively. On each of these projected versions of queries, keys and values we then perform the attention function in parallel, yielding dv-dimensional

output values. These are concatenated and once again projected, resulting in the final values, as depicted in Figure 2. Multi-head attention allows the model to jointly attend to information from different representation subspaces at different positions. With a single attention head, averaging inhibits this

|  |  |
| --- | --- |
| Where | (6) |

Where the projections are parameter matrices , , and. For building the chatbot, the number of head h = 8 are used.

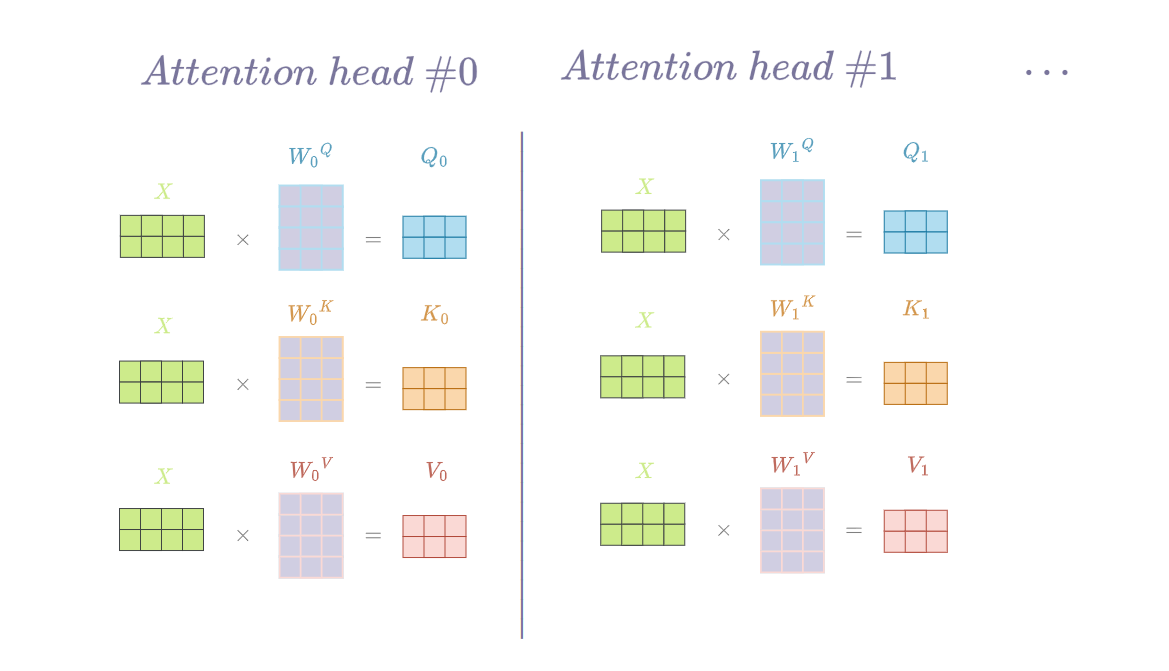


Figure 13. Separate Q/K/V weight matrices of each head

If we do the same self-attention calculation we outlined above, just eight different times with different weight matrices, we end up with eight different Z matrices

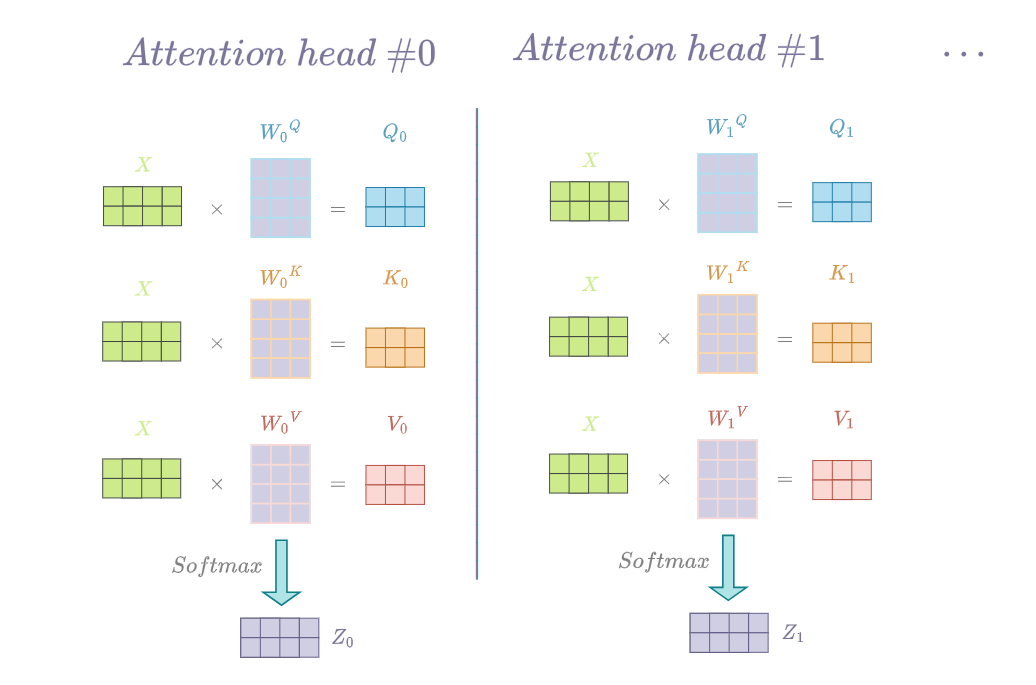


Figure 14. Calculate matrix Z for each head

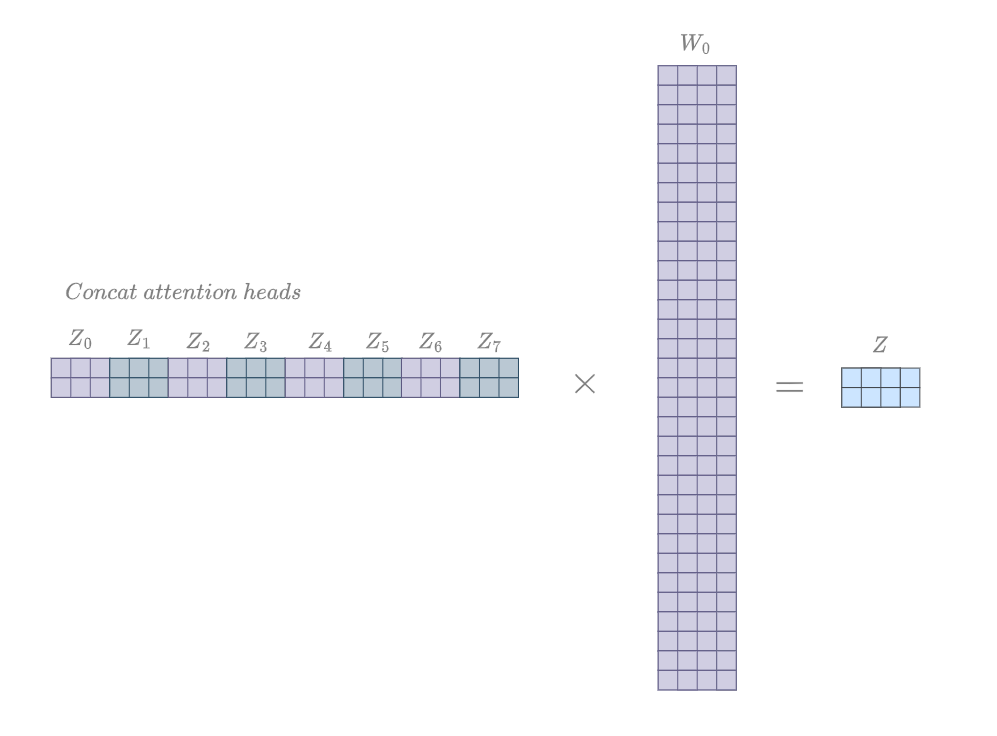


Figure 15. Calculate matrix Z

The Residuals

The residual connection is employed around each sub-layer (self-attention, ffnn) in each encoder and decoder, and is followed by a layer-normalization step.

Residual connection provides another path for data to reach latter parts of the neural network by skipping some layers. This help solving the problems such as exploding gradients and vanishing gradients when training the model.

Feed Forward

Each of the layers in our encoder and decoder contains a fully connected feed-forward network, which is applied to each position separately and identically. This consists of two linear transformations with a ReLU activation in between.

|  |  |
| --- | --- |
|  | (6) |

While the linear transformations are the same across different positions, they use different parameters from layer to layer.

Decoder

The decoder is also composed of a stack of N identical layers. In addition to the two sub-layers in each encoder layer, the decoder inserts a third sub-layer, which performs multi-head attention over the output of the encoder stack. The residual connections are also employed around each of the sub-layers, followed by layer normalization. The self-attention sub-layer in the decoder stack was modified to prevent positions from attending to subsequent positions. This masking, combined with fact that the output embeddings are offset by one position, ensures that the predictions for position *i* can depend only on the known outputs at positions less than *i*.

The components of the decoder and the encoder are pretty similar but have some different. The decoder start by processing the input sequence. The output of the top encoder is then transformed into a set of attention vectors K and V. These are to be used by each decoder in its “encoder-decoder attention” layer which helps the decoder focus on appropriate places in the input sequence.

Masked Multi-Head Attention

The Decoder Masked Multi-Head Attention works just like the Encoder Multi-Head Attention, except that it using one more mask to mask out the next word that it need to predict. For example, with the “Hi MedBot”, the word “MedBot” will be masked out.

Masking will be used to hide the next word so that, at first, it will predict the next word itself using previous results without knowing the real translated word. Only the previous word of the sentence are used for learning purposes. So, while performing parallelization with the matrix operation, the matrix will mask the words appearing later by transforming them into zeroes so that the attention network can’t use them.

A look ahead mask is a matrix of the same size as the matrix outputted by the multi-head attention layer. It contains 0s and negative infinities.

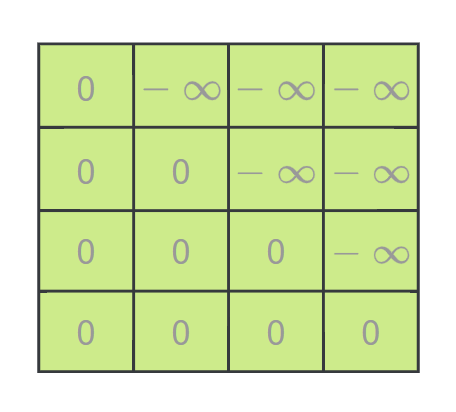


Figure 16, Mask matrix

The mask is applied to the scores after scaling but before the softmax.

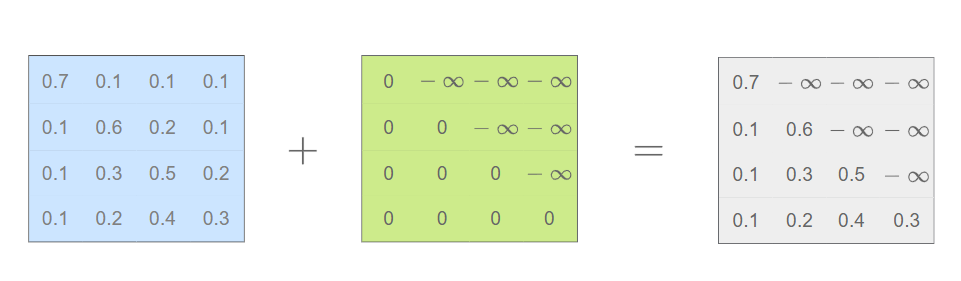


Figure 17. Applying mask matrix

The softmax then just zeros out the future words for every word in the scaled matrix, and then multiply with matrix V to produce the matrix probability Z.

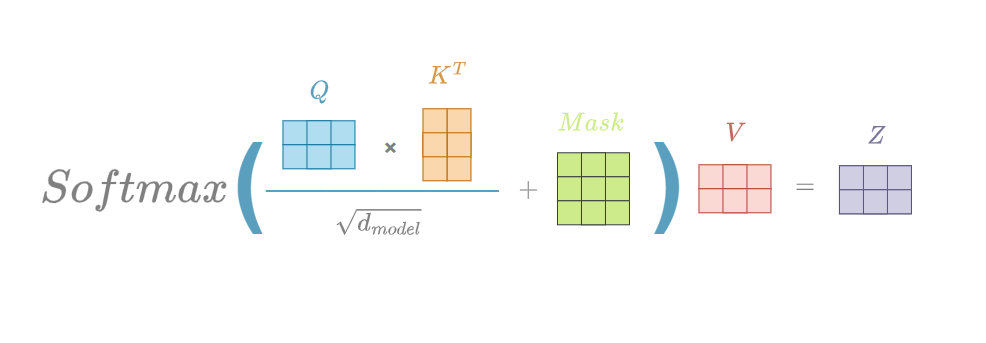


Figure 18. Compute matrix Z with mask matrix

Encoder-decoder attention block

The resulting attention vectors from the previous layer and the vectors from the encoder block are passed into another multi-head attention block. This is where the results from the encoder block also come into place. The output of this block is attention vectors for every word in the sentences. Each vector represents the relationship with other words in both sentences.

Linear and Softmax Layer

The Linear layer is a simple fully connected neural network that projects the vector produced by the stack of decoders, into a much, much larger vector called a logits vector. The softmax layer then turns those scores into probabilities (all positive, all add up to 1.0). The cell with the highest probability is chosen, and the word associated with it is produced as the output for this time step.

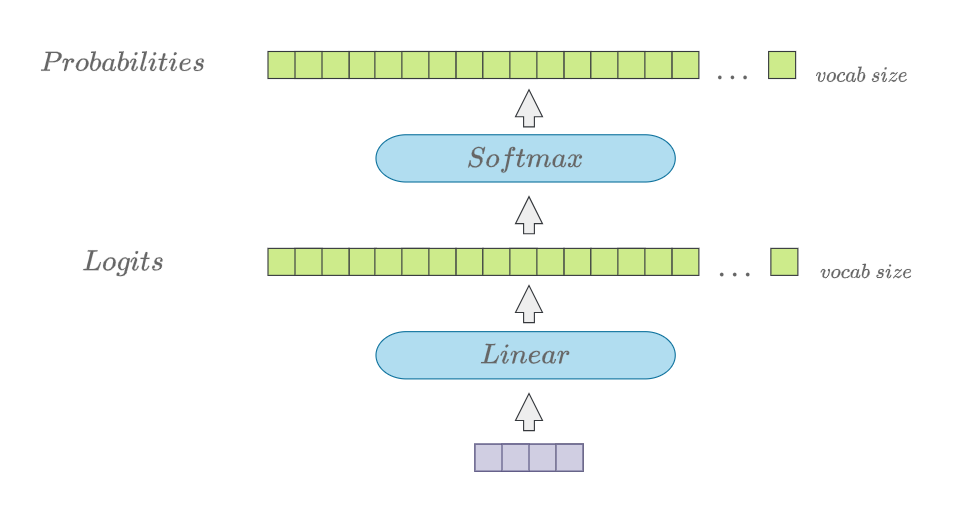


Figure 19. Linear and softmax layers

### BERT

BERT which stands for Bidirectional Encoder Representation Transformer [11], a transformer based language model published by Google Research Team with some modifications. BERT has the ability to capture the context of words in a sentence by leveraging bidirectional training.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text. Since BERT’s goal is to generate a language model, only the encoder mechanism is necessary. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is considered bidirectional, though it would be more accurate to say that it’s non-directional. This characteristic allows the model to learn the context of a word based on all of its surroundings (left and right of the word).

Masked LM (MLM)

Before feeding word sequences into BERT, 15% of the words in each sequence are replaced with a [MASK] token. The model then attempts to predict the original value of the masked words, based on the context provided by the other, non-masked, words in the sequence. In technical terms, the prediction of the output words requires:

* Adding a classification layer on top of the encoder output.
* Multiplying the output vectors by the embedding matrix, transforming them into the vocabulary dimension.
* Calculating the probability of each word in the vocabulary with softmax.

The BERT loss function takes into consideration only the prediction of the masked values and ignores the prediction of the non-masked words. As a consequence, the model converges slower than directional models, a characteristic which is offset by its increased context awareness

BERT model takes the CLS token as input first, then it is followed by a sequence of words as input. Here CLS is a classification token. It then passes the input to the above layers. Each layer applies self-attention and passes the result through a feedforward network after then it hands off to the next encoder. The model outputs a vector of hidden size. Now, this trained vector can be used to perform a number of tasks such as classification, translation, etc

### GPT

The GPT [12] model largely follows the original transformer work. The model consist of N-layer decoder-only transformer with masked self-attention heads followed by the position-wise feed-forward networks.

Unsupervised pre-training

Given an unsupervised corpus of tokens , a standard language modeling objective is to maximize the following likelihood:

|  |  |
| --- | --- |
|  | (7) |

where is the size of the context window, and the conditional probability is modeled using a neural network with parameters. These parameters are trained using stochastic gradient descent.

A multi-layer Transformer decoder for the language model, which is a variant of the transformer, applies a multi-headed self-attention operation over the input context tokens followed by position-wise feedforward layers to produce an output distribution over target tokens:

|  |  |
| --- | --- |
|  | (8) |
|  | (9) |
|  | (10) |

Where is the context vector of tokens, n is the number of layers, is the token embedding matrix, and is the position embedding matrix.

Supervised fine-tuning

After training the model with the objective in Eq. 7, the trained parameters were adapted to the supervised target task. Giving a labeled dataset C, where each instance consists of a sequence of input tokens, , along with a label . The inputs are passed through the pre-trained model to obtain the final transformer block’s activation , which is then fed into an added linear output layer with parameters to predict :

|  |  |
| --- | --- |
|  | (11) |

This gives the following objective to maximize:

|  |  |
| --- | --- |
|  | (12) |

An auxiliary objective was used to help improve the generalization of the supervised model and accelerate convergence. Specifically, need to optimize the following objective (with weight λ):

|  |  |
| --- | --- |
|  | (13) |

### BART

BART [13] is a denoising autoencoder that maps a corrupted document to the original document it was derived from. It is implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right autoregressive decoder. For pre-training, we optimize the negative log-likelihood of the original document.

BART uses the standard sequence-to-sequence Transformer architecture from the original Transformer of Google, except, following GPT, BART modifies ReLU activation functions to GeLUs and initializes parameters from N (0, 0.02). For the base model, BART uses 6 layers in the encoder and decoder, and for the large model, BART uses 12 layers in each. The architecture is closely related to that used in BERT, with the following differences:

* Each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder (as in the transformer sequence-to-sequence model)
* BERT uses an additional feed-forward network before word prediction, which BART does not. In total, BART contains roughly 10% more parameters than the equivalently sized BERT model.

For every text sequence in its input, the BERT encoder outputs an embedding vector for each token in the sequence as well as an additional vector containing sentence-level information. In this way, the decoder can learn for both token and sentence-level tasks making it a robust starting point for any future fine-tuning tasks.

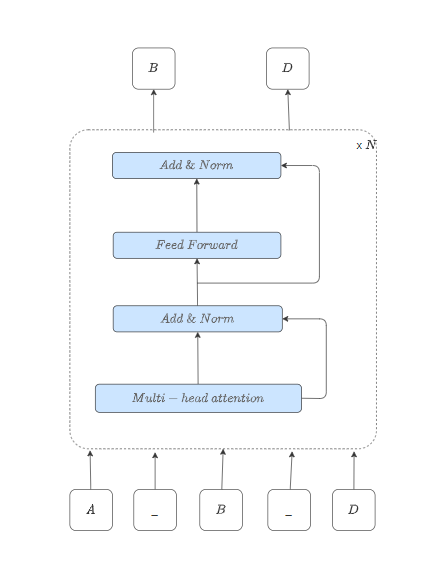


Figure 20. Bidirectional encoder

Once we get the token and sentence-level representation of an input text sequence, a decoder needs to interpret these to map with the output target. However, by using a similarly designed decoder, tasks such as next-sentence prediction or token prediction might perform poorly since the model relies on a more comprehensive input prompt. In these cases, we need model architectures that can be trained to generate the next word by only looking at the previous words in the sequence. Hence, a causal or autoregressive model that looks only at the past data to predict the future comes in handy.

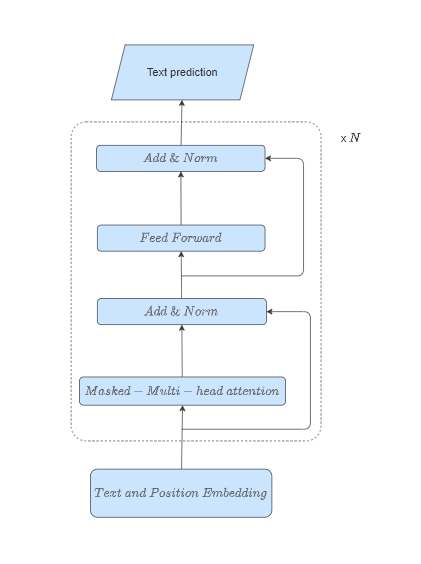


Figure 21. Auto-regressive decoder

First, the model is pre-trained on tokens “t” looking back to “k” tokens in the past to compute the current token. This is done unsupervised on a vast text corpus to allow the model to “learn the language.”

|  |  |
| --- | --- |
|  | (8) |

Next, to make the model robust on a specific task, it is fine-tuned in a supervised manner to maximize the likelihood of label “y” given feature vectors x1…xn.

|  |  |
| --- | --- |
|  | (9) |

Combining 1 and 2, we get the objective in 3. Lambda represents a learned weight parameter to control the influence of language modeling.

|  |  |
| --- | --- |
|  | (10) |

The below image shows how the autoregressive decoder processes its input.

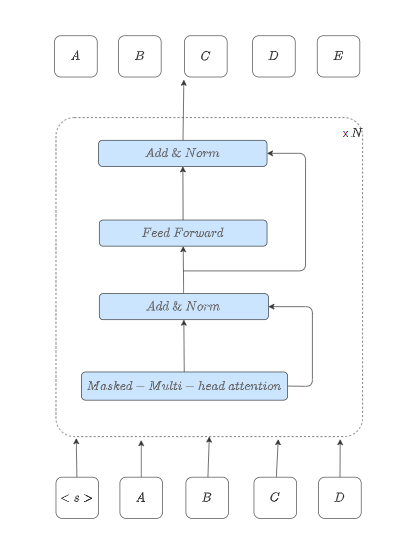


Figure 22. Auto-regressive decoder

Although we separate the decoder from an encoder, the input to the decoder would still be a learned representation (or embedding) of the original text sequence. Thus, BART attaches the bi-directional encoder to the autoregressive decoder to create a denoising auto-encoder architecture. Based on these two components, the final BART model would look something like this:

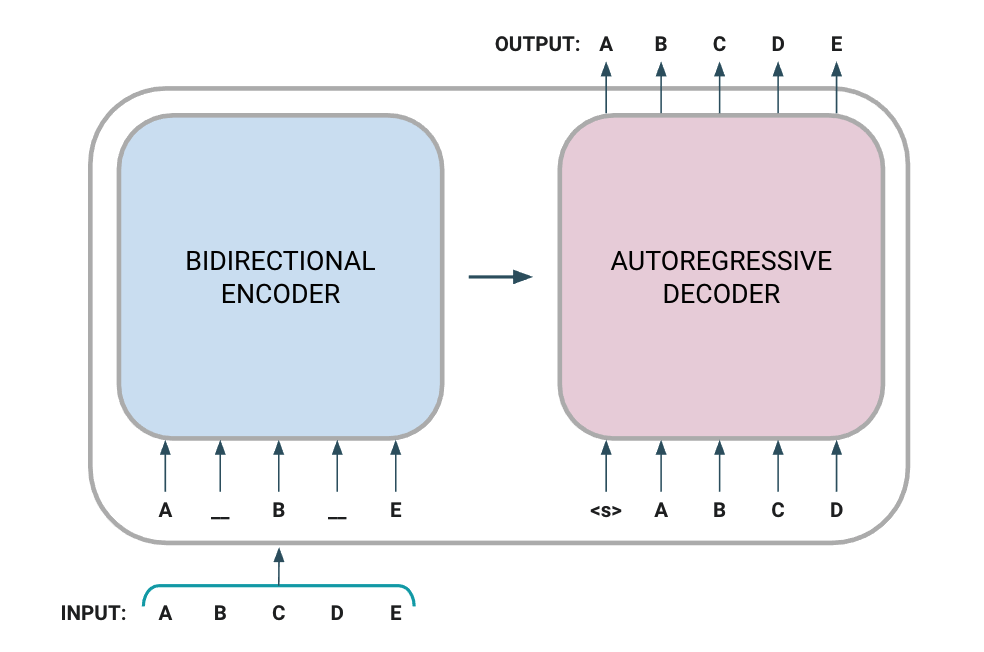


Figure 23. BART model

In the above figure, the input sequence is a masked (or noisy) version of [ABCDE] transformed into [A(MASK)B(MASK)E]. The encoder looks at the entire sequence and learns high-dimensional representations with bi-directional information. The decoder takes these thought vectors and regressively predicts the next token. Learning occurs by computing and optimizing the negative log-likelihood as mapped with the target [ABCDE].

### Tokenizer

Tokenizing a text is splitting it into words or subwords, which then are converted to ids through a look-up table. Converting words or subwords to ids is straightforward, so we will focus on splitting a text into words or subwords (i.e. tokenizing a text)

Subword-based tokenization

Subword-based tokenization is a solution between word and character-based tokenization. The main idea is to solve the issues faced by word-based tokenization (very large vocabulary size, large number of OOV tokens, and different meanings of very similar words) and character-based tokenization (very long sequences and less meaningful individual tokens).

The subword-based tokenization algorithms do not split the frequently used words into smaller subwords. It rather splits the rare words into smaller meaningful subwords. For example, “boy” is not split but “boys” is split into “boy” and “s”. This helps the model learn that the word “boys” is formed using the word “boy” with slightly different meanings but the same root word.

Some of the popular subword tokenization algorithms are WordPiece, Byte-Pair Encoding (BPE), Unigram, and SentencePiece.

Byte-Pair Encoding (BPE)

BPE is a simple form of data compression algorithm in which the most common pair of consecutive bytes of data is replaced with a byte that does not occur in that data. BPE relies on a pre-tokenizer that splits the training data into words. Pretokenization can be as simple as space tokenization.

After pre-tokenization, a set of unique words has been created and the frequency with which each word occurred in the training data has been determined. Next, BPE creates a base vocabulary consisting of all symbols that occur in the set of unique words and learns to merge rules to form a new symbol from two symbols of the base vocabulary. It does so until the vocabulary has attained the desired vocabulary size. Note that the desired vocabulary size is a hyperparameter to define before training the tokenizer.

Encoding and Decoding

Let us now see how we will decode our example. To decode, we have to simply concatenate all the tokens together to get the whole word. For example, the encoded sequence [“the</w>”, “high”, “est</w>”, “range</w>”, “in</w>”, “Seattle</w>”], we will be decoded as [“the”, “highest”, “range”, “in”, “Seattle”] and not as [“the”, “high”, “estrange”, “in”, “Seattle”]. Notice the presence of the “</w>” token in “est”.

For encoding the new data, the process is again simple. However, encoding in itself is computationally expensive. Suppose the sequence of words is [“the</w>”, “highest</w>”, “range</w>”, “in</w>”, “Seattle</w>”]. We will iterate through all the tokens we found in our corpus — longest to the shortest and try to replace substrings in our given sequence of words using these tokens. Eventually, we will iterate through all the tokens and our substrings will be replaced with tokens already present in our token list. If a few substrings are left (for words our model did not see in training), we will replace them with unknown tokens.

In general, the vocabulary size is big but still, there is a possibility of an unknown word. In practice, we save the pre-tokenized words in a dictionary. For unknown (new) words, we apply the above-stated encoding method to tokenize the new word and add the tokenization of the new word to our dictionary for future reference. This helps us build our vocabulary even stronger for the future.

### BLEU evaluation

BLEU [14], or the Bilingual Evaluation Understudy, is a metric for comparing a candidate translation to one or more reference translations. Although developed for translation, it can be used to evaluate text generated for different natural language processing tasks, such as paraphrasing and text summarization.

The BLEU score is not perfect, but it’s quick and inexpensive to calculate, language-independent, and, above all, correlates highly with human evaluation.

The BLEU score can be calculated as below:

|  |  |
| --- | --- |
|  | (11) |

With BP is the brevity penalty factor:

|  |  |
| --- | --- |
|  | (12) |

Where:

: the modified n-gram precisions measure how many n-grams in the reference sentence are reproduced by the candidate sentence.

|  |  |
| --- | --- |
|  | (13) |

: The number of unigrams, bigrams, trigrams, and four-grams (*i*=1,...,4) match their n-gram counterpart in the reference translations.

: The total number of unigrams in the candidate.

The BLEU metric ranges from 0 to 1. It is zero iff none of the n-substrings in the candidate is in reference. It is one iff every n-gram in the candidate that appears in reference, at least as many times as in the candidate.

### ROUGE evaluation

ROUGE [15] stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

ROUGE-N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed as follows:

|  |  |
| --- | --- |
|  | (14) |

Where n stands for the length of the n-gram, , and is the maximum number of n-grams co-occurring in a candidate and a set of references.

### Python

Python is a high-level, versatile programming language known for its simplicity and readability. Created by Guido van Rossum and first released in 1991, Python has gained widespread popularity in various domains, including web development, data science, artificial intelligence, scientific computing, and more.

Python is one of the most popular programming languages for machine learning and artificial intelligence. Its versatility, extensive libraries, and active community support make it an ideal choice for developing machine-learning models and conducting data analysis. Here's an overview of how Python is used in the field of machine learning:

* Libraries and Frameworks: Python has a rich ecosystem of libraries and frameworks that simplify machine learning tasks. Some of the most commonly used ones include:
* numpy: numpy is a fundamental library for numerical computing in Python. It provides support for multi-dimensional arrays and mathematical functions, making it essential for data manipulation.
* pandas: pandas is a data manipulation library that provides data structures like DataFrames and Series. It is used for data cleaning, transformation, and analysis.
* scikit-learn: scikit-learn is a popular machine learning library that offers a wide range of machine learning algorithms for classification, regression, clustering, dimensionality reduction, and more. It also provides tools for model selection and evaluation.
* TensorFlow: Developed by Google, TensorFlow is an open-source machine learning framework that's widely used for deep learning tasks. It allows you to build neural networks for various applications, including image recognition, natural language processing, and reinforcement learning.
* PyTorch: PyTorch is another deep learning framework that has gained popularity for its flexibility and dynamic computation graph. It is known for its ease of use and is commonly used in research and development of neural network models.
* Keras: Keras is a high-level neural networks API that runs on top of other deep learning frameworks like TensorFlow and Theano. It simplifies the process of building and training neural networks.

Python's role in machine learning continues to evolve, and it remains at the forefront of AI and data science. Whether you're a beginner looking to get started or an experienced practitioner, Python's ecosystem and community support make it a valuable tool for tackling a wide range of machine-learning challenges.

### Tensorflow

TensorFlow is a free and open-source software library for machine learning and artificial intelligence. It can be used across a range of tasks but has a particular focus on training and inference of deep neural networks.

TensorFlow can run on multiple CPUs and GPUs (with optional CUDA and SYCL extensions for general-purpose computing on graphics processing units). TensorFlow is available on 64-bit Linux, macOS, Windows, and mobile computing platforms including Android and iOS.

### Deep learning

Deep learning is the branch of machine learning which is based on artificial neural network architecture. An artificial neural network or ANN uses layers of interconnected nodes called neurons that work together to process and learn from the input data.

In a fully connected Deep neural network, there is an input layer and one or more hidden layers connected one after the other. Each neuron receives input from the previous layer neurons or the input layer. The output of one neuron becomes the input to other neurons in the next layer of the network, and this process continues until the final layer produces the output of the network. The layers of the neural network transform the input data through a series of nonlinear transformations, allowing the network to learn complex representations of the input data.

Deep learning can be used for supervised, unsupervised as well as reinforcement machine learning. It uses a variety of ways to process these.

# DESIGN AND IMPLEMENT OF THE CHATBOT

## Overview

The process for building the chatbot includes the following steps:

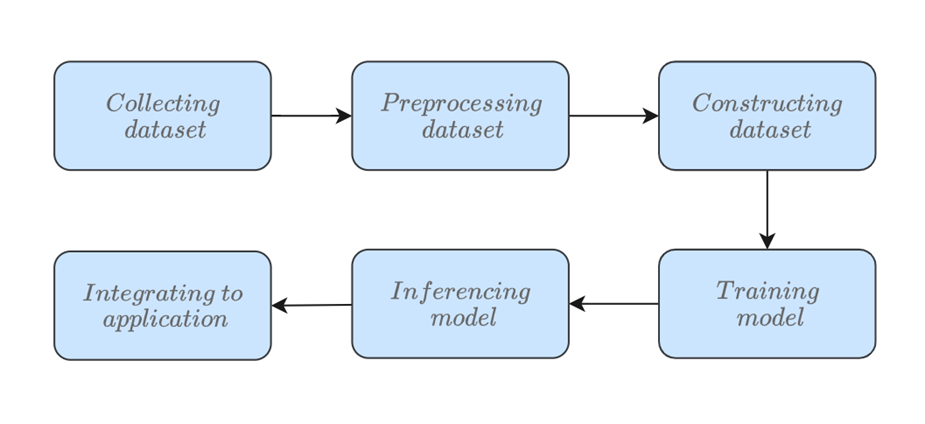


Figure 24. Processes for building the chatbot

* Collecting dataset: research, crawl, and construct the dataset for training the model.
* Preprocessing: From the dataset obtained in the above step, proceed with eliminating special characters in sentences (!@#%&\*,..), remove multiple spaces, blank lines, remove meaningless words, short sentences (hmm, ag,..), duplicate sentences, etc. Then save the results into the .csv file.
* Training the model: Read the datasets obtained in the above step, then input them into the model for training purposes.
* Inferencing: Enter input data, and proceed to process that data similarly as a preprocessing step but does not have to be saved to a file. Then use the trained model to predict the answer, and print the results to the screen.
* Integrate to applications: The trained model will be used to create an API for the applications to send the request containing the question and retrieve the answer back in the response. The answer will be extracted in the response to display on the UI for the users in various applications.

## Implementation

### Collecting dataset

To train the bot to answer questions about healthcare, a list of collected questions and answers needs to be analyzed and processed. Based on the analysis, the final data to train the chatbot is organized and defined in train, test, and validation dataset files. Once the data is finalized, the chatbot needs to be fed all the data with the corresponding responses. Then the chatbot created needs to be tested and trained to fine-tune the results.

There are 3 datasets are collected and used for training the chatbot:

* The EHealthChatMini dataset is the dataset collected from crawling on popular healthcare websites (English) with more than 299,000 question-answer pairs related to healthcare.

Below are some examples from the dataset:

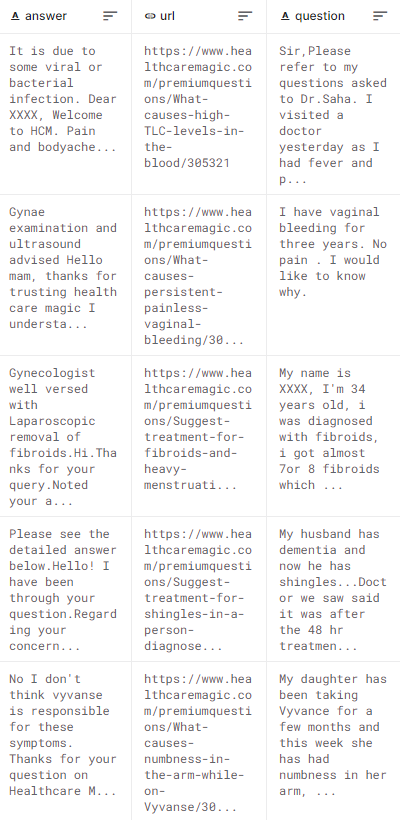


Figure 25. EHealthChatMini dataset samples

Also, details information about the dataset is shown below:

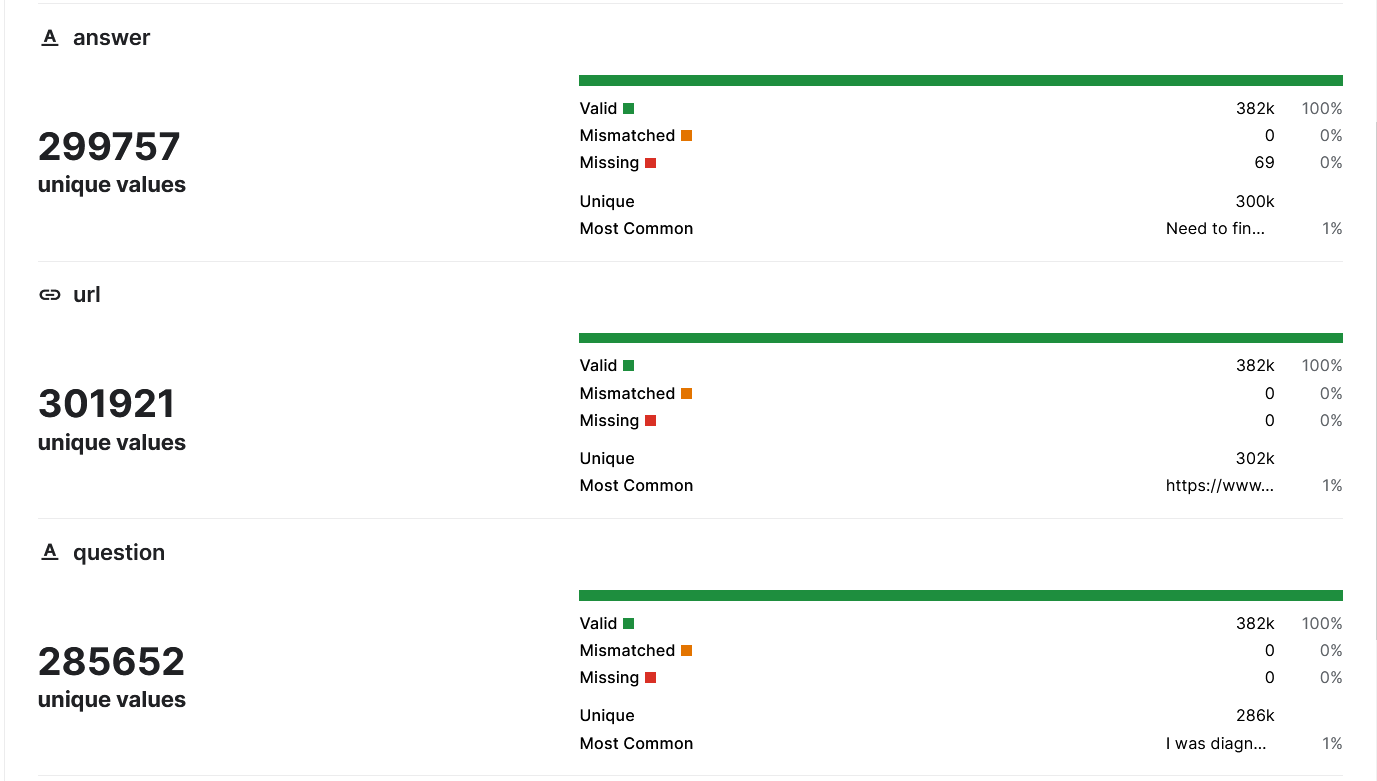


Figure 26. EHealthChatMini dataset details

* The EHealthChat dataset is the dataset provided by Kaggle (English) with more than 222,000 question-answer pairs about healthcare problems.

Below are some examples from the dataset:



Figure 27. EHealthChat dataset samples

Also, details information about the dataset is shown below:

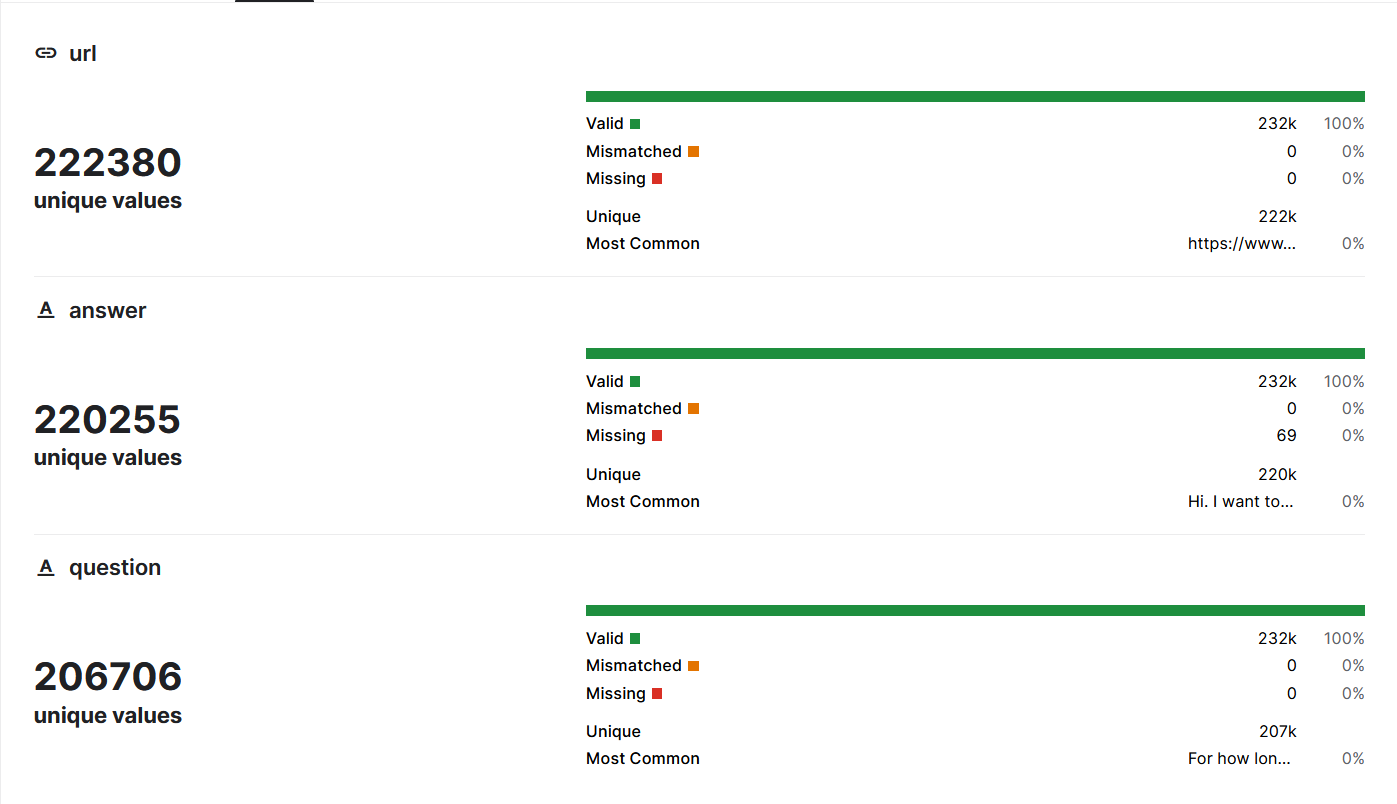


Figure 28. EHealthChat dataset details

* The EHealthVNChat dataset is the dataset collected from crawling popular healthcare websites (Vietnamese) with more than 20,000 question-answer pairs related to healthcare.
* Below are some examples from the dataset:

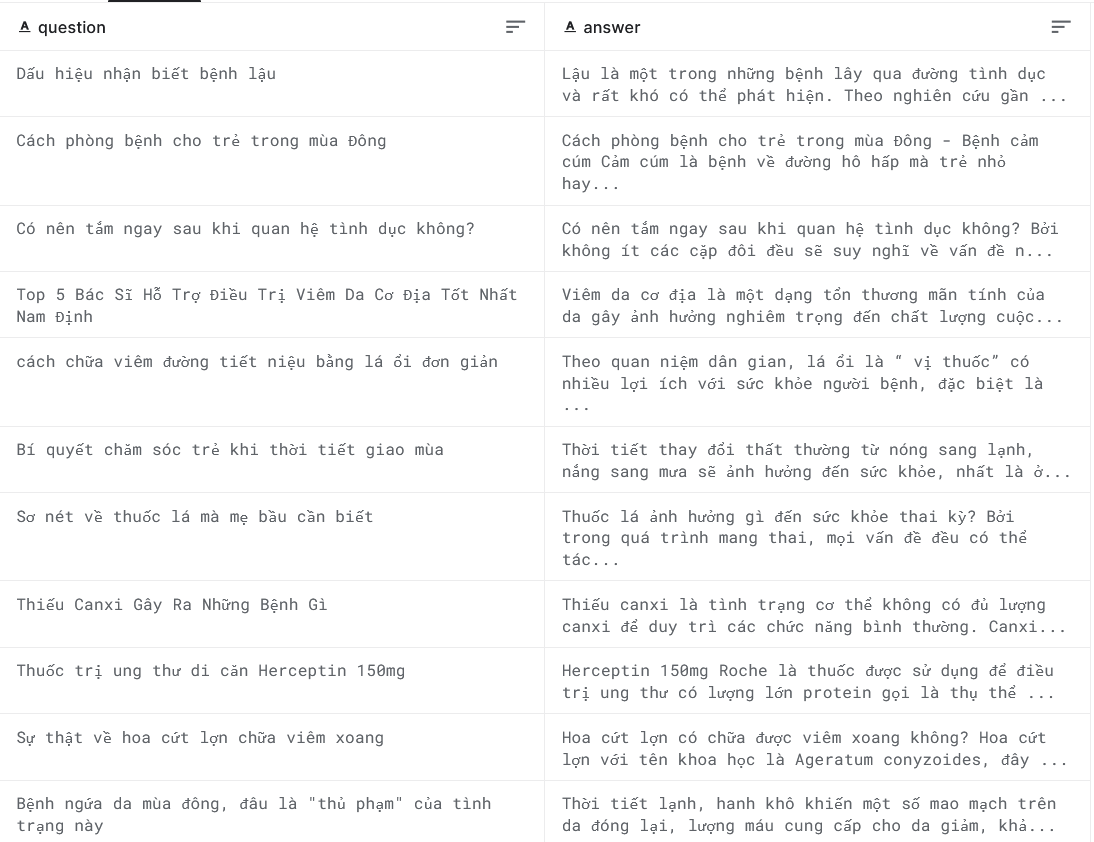


Figure 29. EHealthVNChat dataset samples

Also, details information about the dataset is shown below:

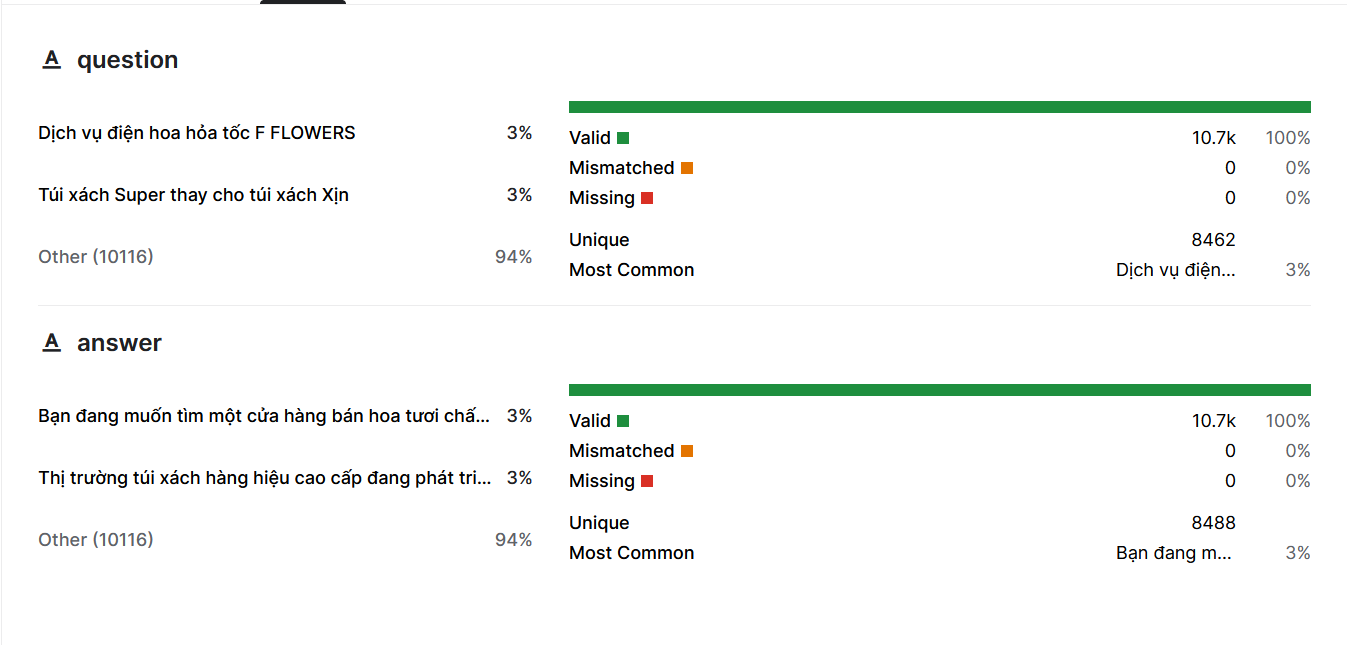


Figure 30. EHealthVNChat dataset details

### Preprocessing dataset

The data was obtained in raw form from crawling the web and text file, containing many special characters, along with many words abbreviations or words that have no meaning, so to prepare the dataset for training the model, the data cleaning steps are performed as follows:

* Remove all special characters in the sentences (@#$%^,..). For example, “How to cure headache???” to “How to cure headache?”.
* Remove all subtitle, watermarked sentences such as “answered by”, “asked by”.
* Remove multiple spaces and blank lines.
* Remove all sentences with lengths below 6.
* Remove repeated pattern such as “Question:”, “Answer:”,…

After completing the preprocessing dataset, the processed datasets with be converted into CSV files for easy loading and suitable for training of the model with datasets loader libraries.

### Constructing dataset

From the preprocessed dataset above, proceed to add start of sentent keywords (<sos>) and end of sentence (</eos>), after data collection, after that the data input sentence whose length more than max\_sequence\_length will be truncated.

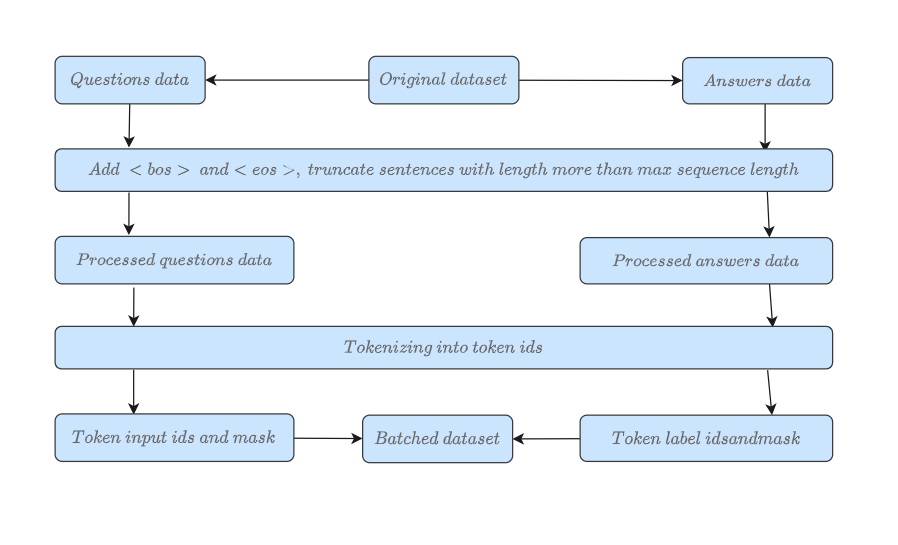


Figure 31. Constructing dataset for traning model

With the EHealthChat dataset, the chatbot will be trained with the max\_sequence\_length in the question and answer set words will be 512 tokens per sentence, because with a length of 512 tokens per sentence, more than 93% of the dataset can be used as below:

|  |  |  |
| --- | --- | --- |
| EHealthChat dataset | | |
| Sequence length | Question | Answer |
| Number of samples | 299,757 | 299,757 |
| Sample’s length below 512 | 280,680 | 280,680 |
| Sample’s length more than 512 | 19,077 | 19,077 |
| Minimum sentence length | 6 | 6 |
| Maximum sentence length | 684 | 1978 |

With the EHealthChatMini dataset, the chatbot will be trained with the max\_sequence\_length in the question and answer set words will be 512 tokens per sentence, because with a length of 512 tokens per sentence, more than 90% of the dataset can be used as below:

|  |  |  |
| --- | --- | --- |
| EHealthChatMini dataset | | |
| Sequence length | Question | Answer |
| Number of samples | 222,380 | 222,380 |
| Sample’s length below 512 | 201,200 | 201,200 |
| Sample’s length more than 512 | 11,180 | 11,180 |
| Minimum sentence length | 6 | 6 |
| Maximum sentence length | 380 | 1280 |

With the EHealthVNChat dataset, the chatbot will be trained with the max\_sequence\_length in the question and answer set words will be 512 tokens per sentence, because with a length of 512 tokens per sentence, more than 97% of the dataset can be used as below:

|  |  |  |
| --- | --- | --- |
| EHealthVNChat dataset | | |
| Sequence length | Question | Answer |
| Number of samples | 22,780 | 22,780 |
| Sample’s length below 512 | 22,128 | 22,128 |
| Sample’s length more than 512 | 652 | 652 |
| Minimum sentence length | 6 | 6 |
| Maximum sentence length | 312 | 1892 |

The datasets will be converted into tokens using the tokenizer, which tokenize the inputs (including converting the tokens to their corresponding IDs in the pretrained vocabulary) and put it in a format the model expects, as well as generate the other inputs that the model requires.

### Training the model

The training process is implemented as described below:

The English model version is trained using two processed datasets as above, and the Vietnamese version uses the VN processed dataset for training purposes.

Firstly, the dataset will be converted into the batched dataset, where each batch contains N samples (forming a matrix from multiple vector input to speed up the training process), the input matrix will be passed into the tokenizer to convert into the token IDs (numerical value) and are passed to the embedding layer to convert into the vector and adding the position encoding to capture the position info inside the input. The different length vectors are padded into a fixed length with padding value (usually 0) to optimize compute matrix multiplication. Next, the data will be passed through the encoder and decoder block respectively, and finally flow through the linear and softmax layer to produce the probability of the next word for all the words in the dictionary. The softmax output will be used to compute the loss, and then compute the gradients for updating the model weights. After processing all the batches inside the dataset, the model has finished an epoch. The model will repeat this process for a fixed amount of time (called epochs). Finally, the model will be saved for inference later and integrated into other applications.

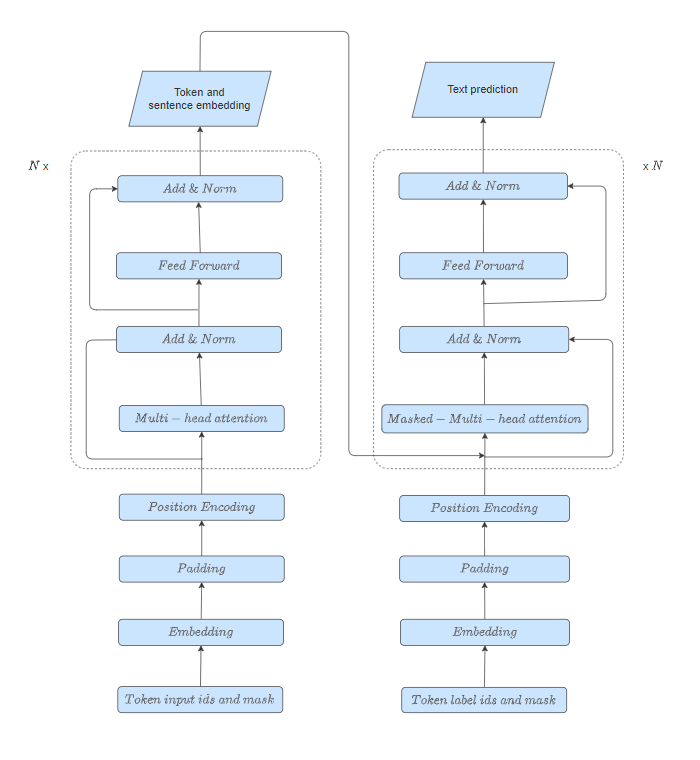


Figure 32. Overview of training the model

### Modifying the model

For the experiment with the custom model, the attention of the BART model was re-implemented using the re-attention following the DeepViT Transformer [15]. The model attention employs the attention maps from the heads as the basis and generates a new set of attention maps by dynamically aggregating them. A learnable transformation matrix is used to mix the multi-head attention maps into re-generated new ones, before being multiplied with V. Specifically, the Re-attention is implemented by:

|  |  |
| --- | --- |
|  | (11) |

where transformation matrix Θ is multiplied by the self-attention map A along the head dimension. Here Norm is a normalization function used to reduce the layer-wise variance.

The advantages of the proposed Re-attention are two-fold. First of all, re-attention exploits the interactions among different attention heads to collect their complementary information and better improves the attention map diversity. Furthermore, the re-attention is effective and easy to implement. It needs only a few lines of code and negligible computational overhead compared to the original self-attention.

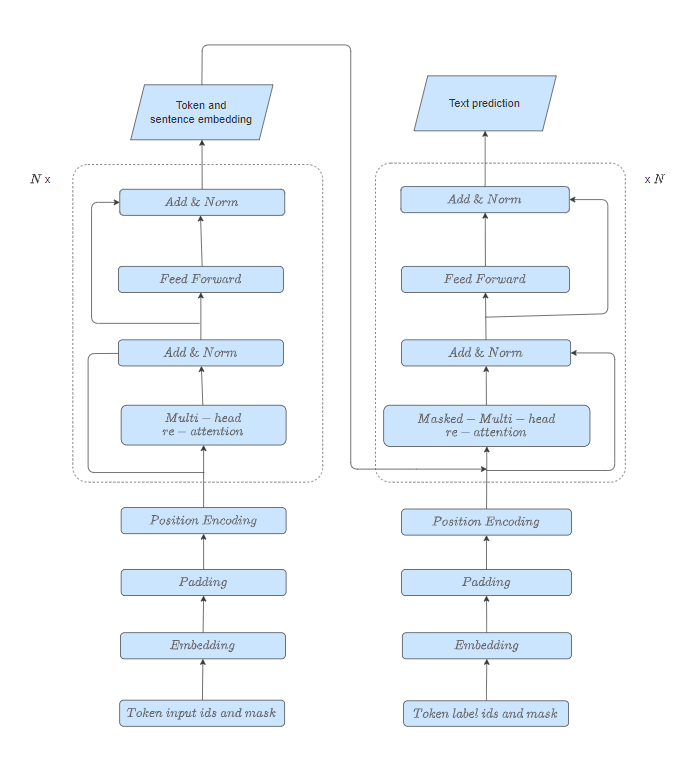


Figure 33. Modifying the model with re-attention

### Inferencing the model

After the model has been trained, it can be used to generate answers from user inputs. The steps to inferencing the model can be described below:

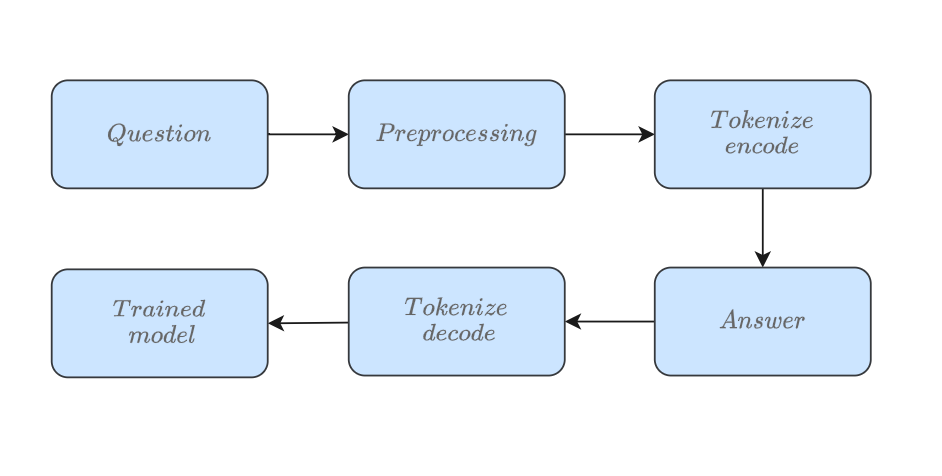


Figure 34. Steps for inference of the model

The user input will be pre-processed and tokenized into token IDs, then the token IDs will be fed into the model to generate the probability of the next token, the index of the token with the highest probability will be mapped into the token by the tokenizer decoding. The previous predicted token will be concat into the user input to form a new input and the process begins with the new input being the original input that has been concat with the predicted token. The process will continue until the max sequence length reaches or the end of string token is found.

## Integrate to application

### Facebook messager

There are 3 main steps to deploy the chatbot to the Facebook messager application as shown below:

* Create a server that listens to messages from Facebook (using Flask and Python)
* Define a function for sending messages back to users (using requests)
* Forward an HTTPS connection to the flask server (using ngrok or serveo)

#### The server

The first step is to create an HTTP server that listens to messages sent by Facebook, gets a response to that message, and eventually sends the response back to the user on Facebook. The Flask Python framework will be used to create this server. The basic idea is the following:

* Create a flask app that listens for messages sent to localhost:5000/webhook. When messages are sent on Facebook they will arrive as HTTP requests to this URL.
* The listen() function handles these HTTP requests and checks that they contain a valid Facebook message
* If the message is valid, the get\_response() function is called and the response is sent back to Facebook Messenger

In addition, the function verify\_webhook() needs to be implemented to handle the initial authentication between Facebook and the server

#### The response

A function that sends a response back to Facebook Messenger using a Python library called requests. In particular, the method will get the response from the chatbot according to the message included in the request and send it back to Facebook with a POST request.

#### Expose a HTTPS endpoint

The ngrok or serveo server will be used to set up an HTTPS endpoint to get forwarded to the Flask server.

#### Create a Facebook app and page

The next step is to create an app and a page on Facebook. Once the page has been created, go to the Token Generation settings and select this page from the drop-down menu. Copy the Page Access Token into the placeholder for PAGE\_ACCESS\_TOKEN in the Flask server. After that, the server webhook will be registered on the Facebook developer's page. Finally, just start the chatbot server and the users can chat with the chatbot via the Facebook messager. Below is an example of the chatbot after integrated to Facebook Messager.

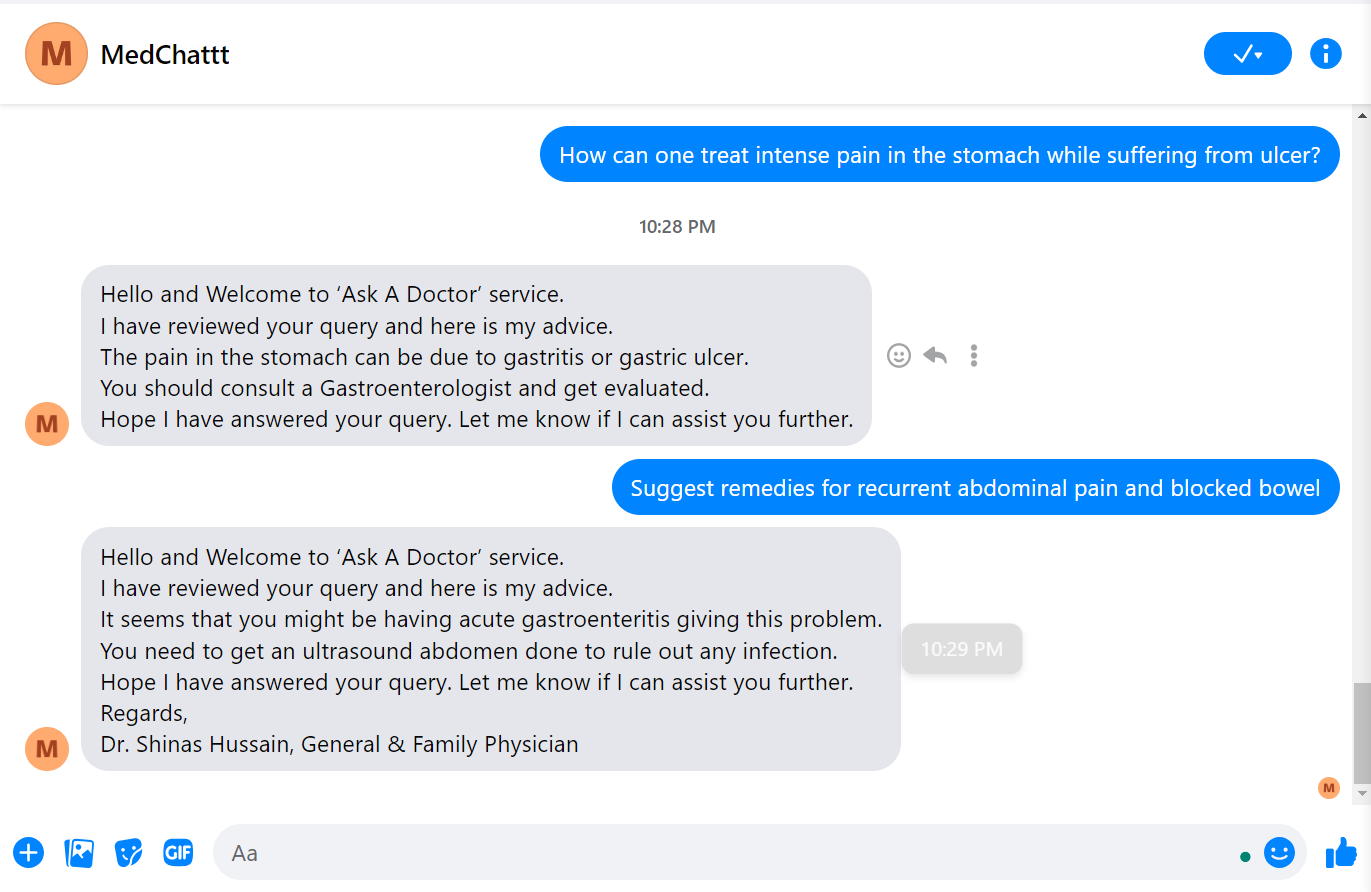


Figure 35. Deploy the chatbot to Facebook messager

### Mobile app

There are 4 main steps to deploy the chatbot to a mobile application as described below:

* Create a server that listens to messages from the application (using Flask and Python)
* Define a function for sending messages back to users (using requests)
* Create a mobile application UI to interact with the chatbot server using Flutter
* Forward an HTTPS connection to the flask server (using ngrok or serveo)

#### The server

The first step is to create an HTTP server that listens to messages sent by the mobile application, gets a response to that message, and eventually sends the response back to the user on the mobile app. The Flask Python framework will be used to create this server. The basic idea is the following:

* Create a flask app that listens for messages sent to localhost:5000/webhook. When messages are sent on the mobile they will arrive as HTTP requests to this URL.
* The chat() function handles these HTTP requests and checks that they contain a valid message.
* If the message is valid, the get\_response() function is called and the response is sent back to the application.

#### The response

A function that sends a response back to the mobile application using a Python library called requests. In particular, the method will get the response from the chatbot according to the message included in the request and send it back via a POST request.

#### Create a mobile application

The mobile application will be created using Flutter to allow users to interact with the chatbot system with a simple yet effective and beautiful UI.

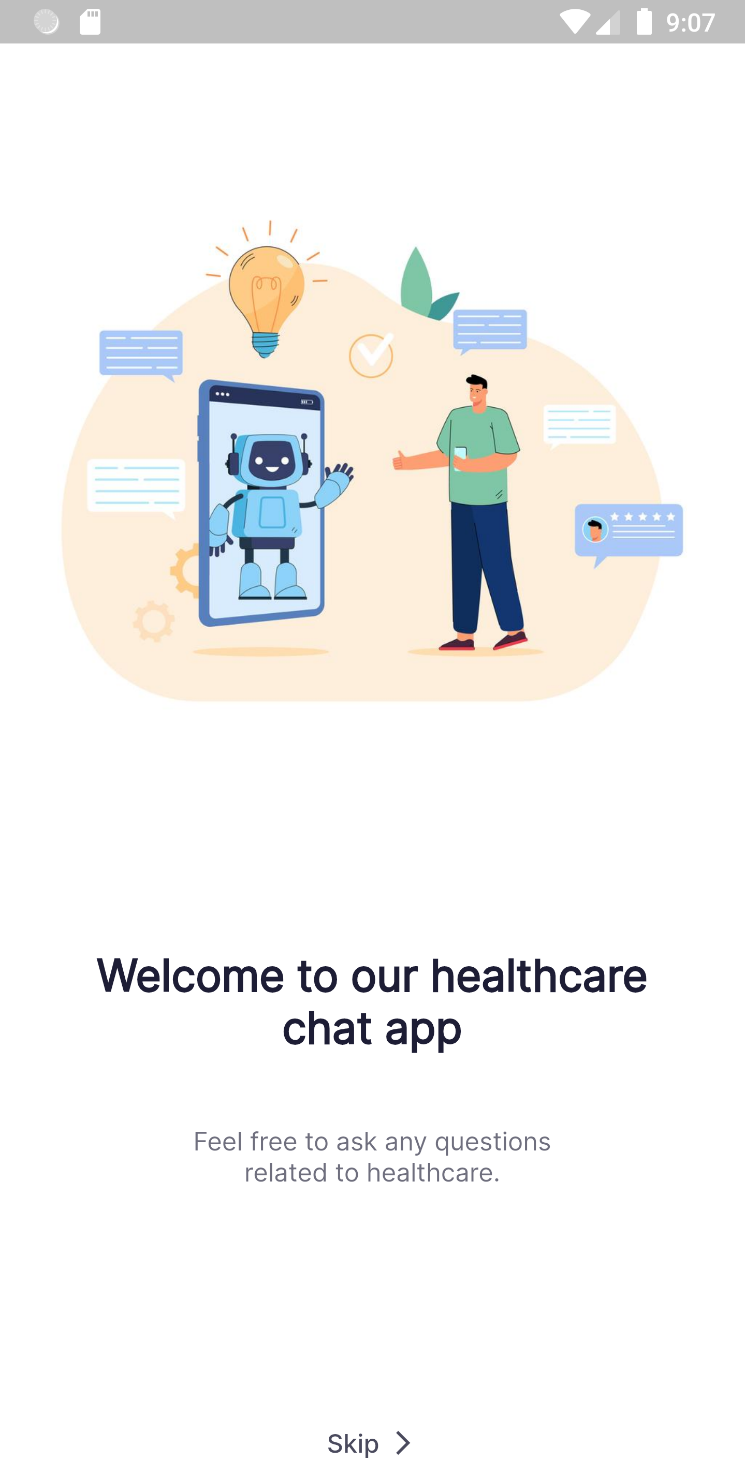


Figure 36. Welcome page of the chatbot application

1. **Sign up feature**

The application allows the user to create an account to help keep track of the data and the conversation with the chatbot. When users sign up, they have to enter their email address, username, and password. If there are some errors in the signing-up process, the system will show notifications. Otherwise, show a success notification to the user.

1. **Sign in feature**

Along with the signup feature, the users can also sign in with a signed-up email and password, and the system will authenticate, if the email and password are both valid and, in the database, the system will direct users to the conversation pages. Otherwise, the system will show an error message.



Figure 37. Signup, sign-in screen mobile app

1. **Chat feature**

The chat feature allows users to ask and receive answers related to healthcare problems from the chatbot server. When users input a question, the chatbot will generate a response from the data it had trained before and send it to the users. Users can ask as much as they want and the conversation (the questions and answers) will be saved and the user can view it later if they wish.

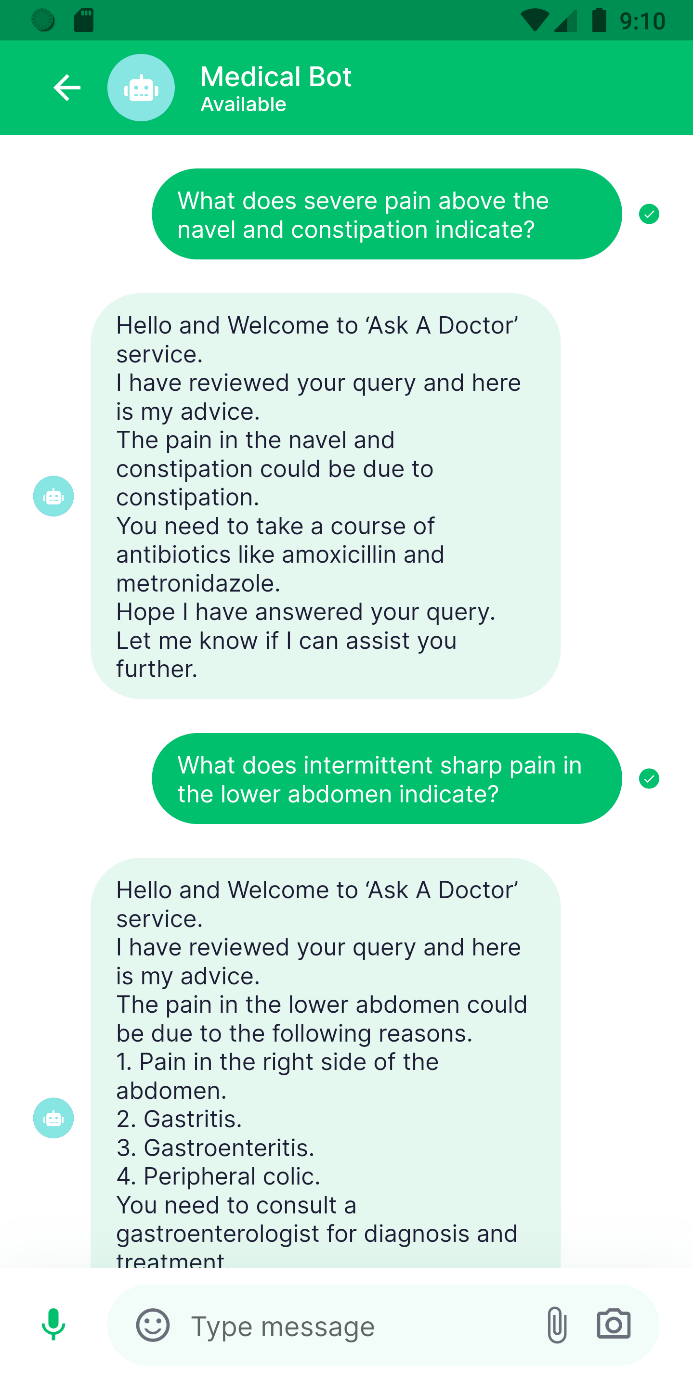


Figure 38. Chat screen of the mobile app

#### Expose a HTTPS endpoint

The ngrok or serveo server will be used to set up an HTTPS endpoint to get forward the requests of the application to the Flask server.

### Web application

There are 4 main steps to deploy the chatbot to a web application:

* Create a server that listens to messages from the application (using Flask and Python)
* Define a function for sending messages back to users (using requests)
* Create a web application UI to interact with the chatbot server using VueJs, NodeJs
* Forward an HTTPS connection to the flask server (using ngrok or serveo)

#### The server

The first step is to create an HTTP server that listens to messages sent by the mobile application, gets a response to that message, and eventually sends the response back to the user on the mobile app. The Flask Python framework will be used to create this server. The basic idea is the following:

* Create a flask app that listens for messages sent to localhost:5000/webhook. When messages are sent on the mobile they will arrive as HTTP requests to this URL.
* The chat() function handles these HTTP requests and checks that they contain a valid message.
* If the message is valid, the get\_response() function is called and the response is sent back to application.

#### The response

A function that sends a response back to the web application using a Python library called requests. In particular, the method will get the response from the chatbot according to the message included in the request and send it back via a POST request.

#### Create an web application

1. **Sign up feature**

The application allow the user to create an account to help keep track of the data and the conversation with the chatbot. When users sign up, they have to enter their email address, username, and password. If there are some errors in the signing-up process, the system will show notifications. Otherwise, show a success notification to the user.

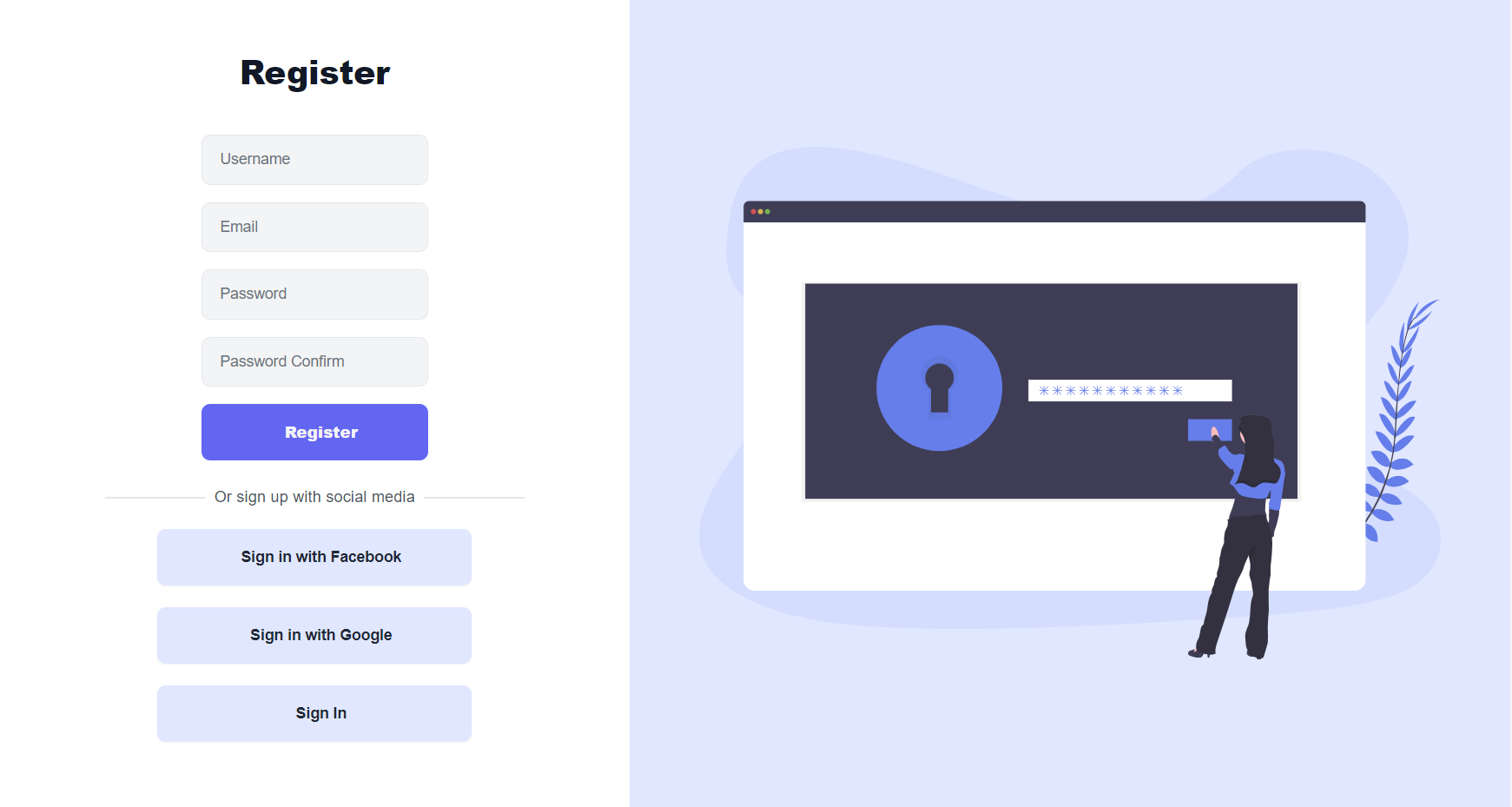


Figure 39. Signup chatbot web page

1. **Sign in feature**

Along with the sign-up feature, the users can also sign in with a signed-up email and password, and the system will authenticate, if the email and password are both valid and, in the database, the system will direct users to the conversation pages. Otherwise, the system will show an error message.

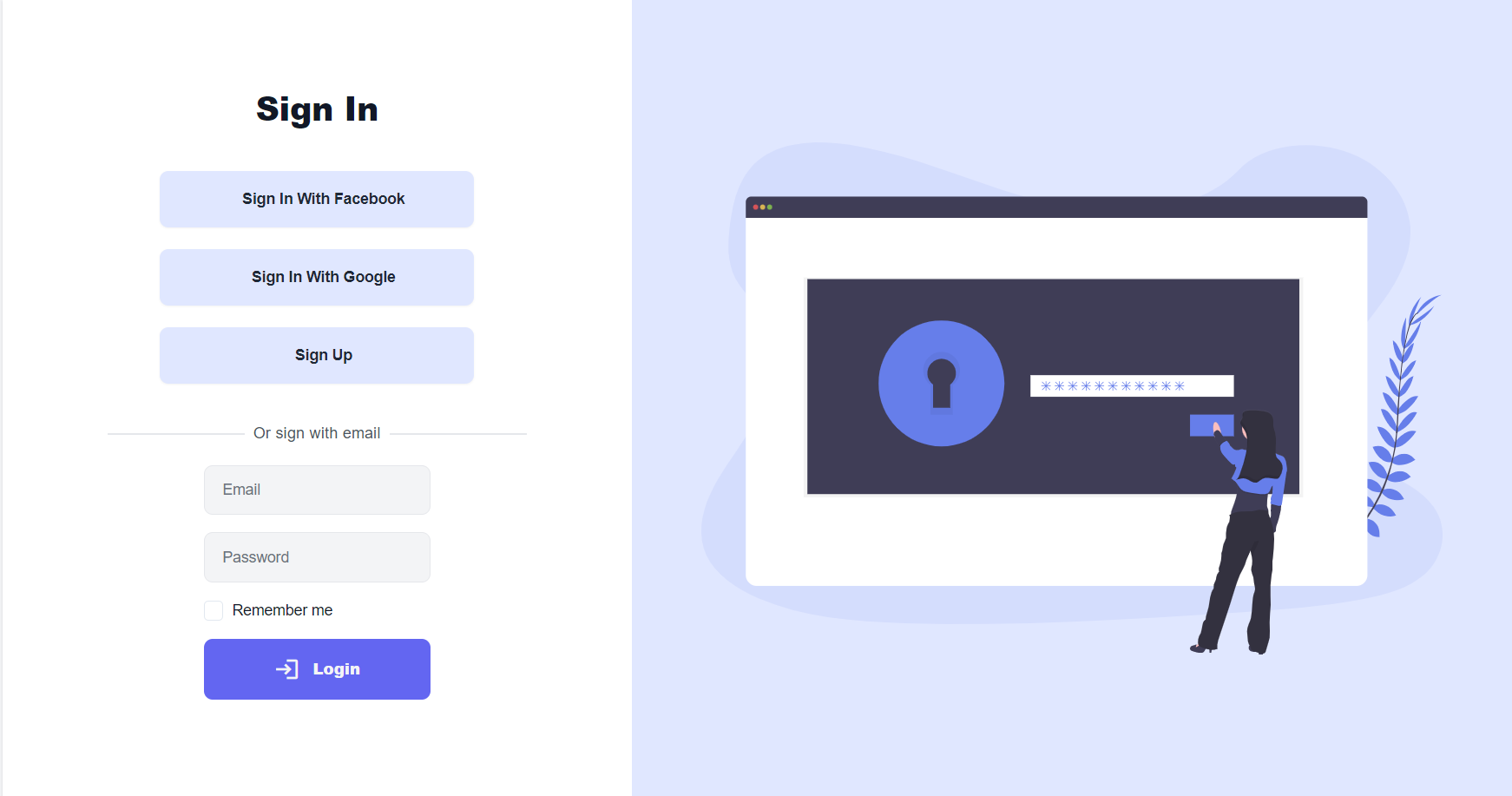


Figure 40. Sign-in chatbot web page

1. **Chat feature**

The chat feature allows users to ask and receive answers related to healthcare problems from the chatbot server. When users input a question, the chatbot will generate a response from the data it had trained before and send it to the users. Users can ask as much as they want and the conversation (the questions and answers) will be saved and the user can view it later if they wish.

On the chat page, users can chat with the chatbot to ask about the problem they want to know about healthcare. After input, the chatbot will respond to the users with the answer to the problem using the trained model. The users can create as many conversations for easy to keep track of them for preview later instead of chatting in only one conversation which leads to a long and hard-to-remember conversation.

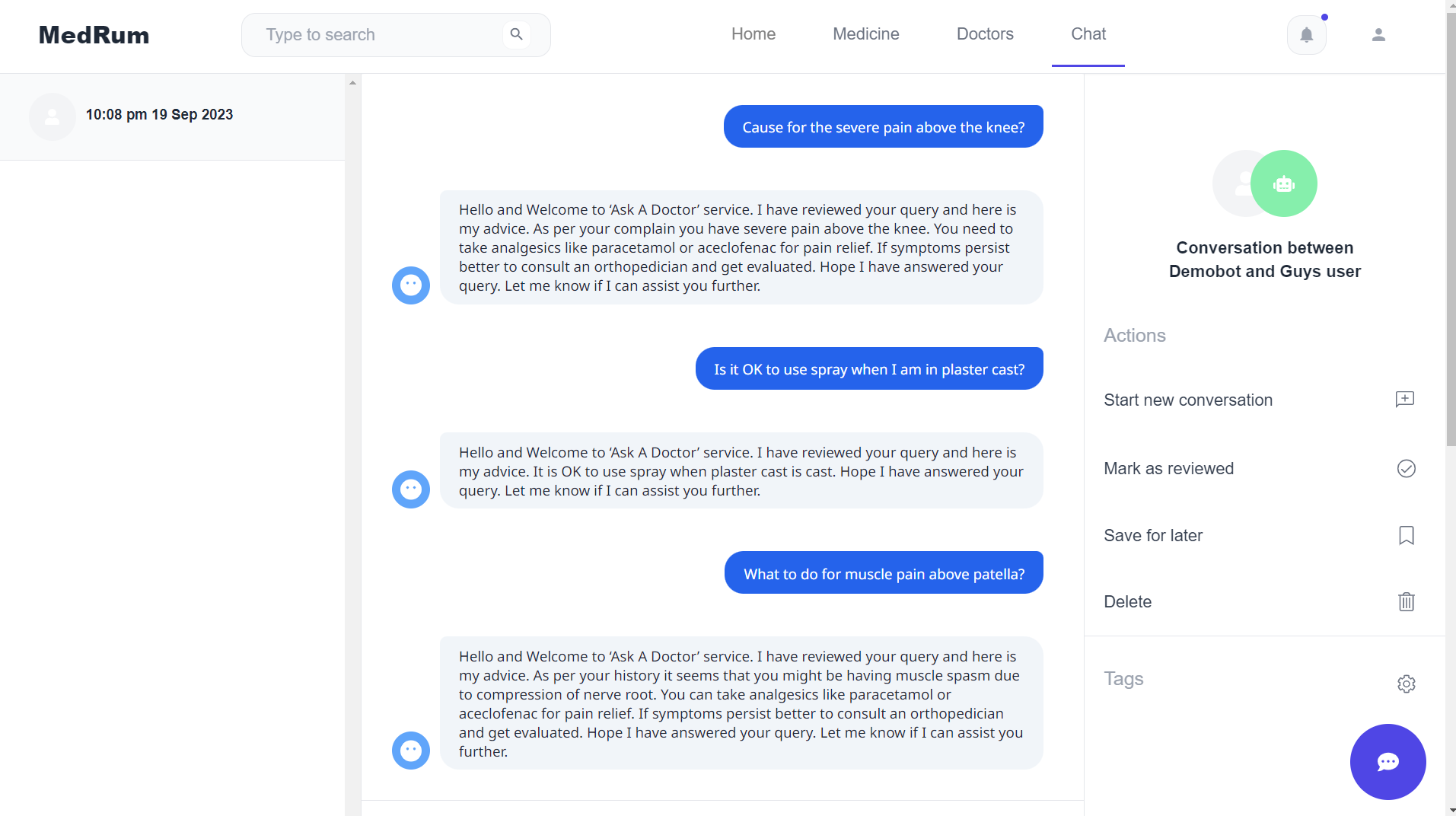


Figure 41. Chat with chatbot web page

#### Questions page

On the questions page, users will see the most popular questions to select to view detail. There is an option to allow users, who have logged in to the website, to add new questions quickly by clicking on the button “Add new”, after clicking the button, the user will be redirected to the “New question” page to create a new question.

Besides, users can also see the trending topics of questions to filter questions by topic on the left sidebar. In addition, on the right sidebar, users will see their profile and a list of top questions on the website.

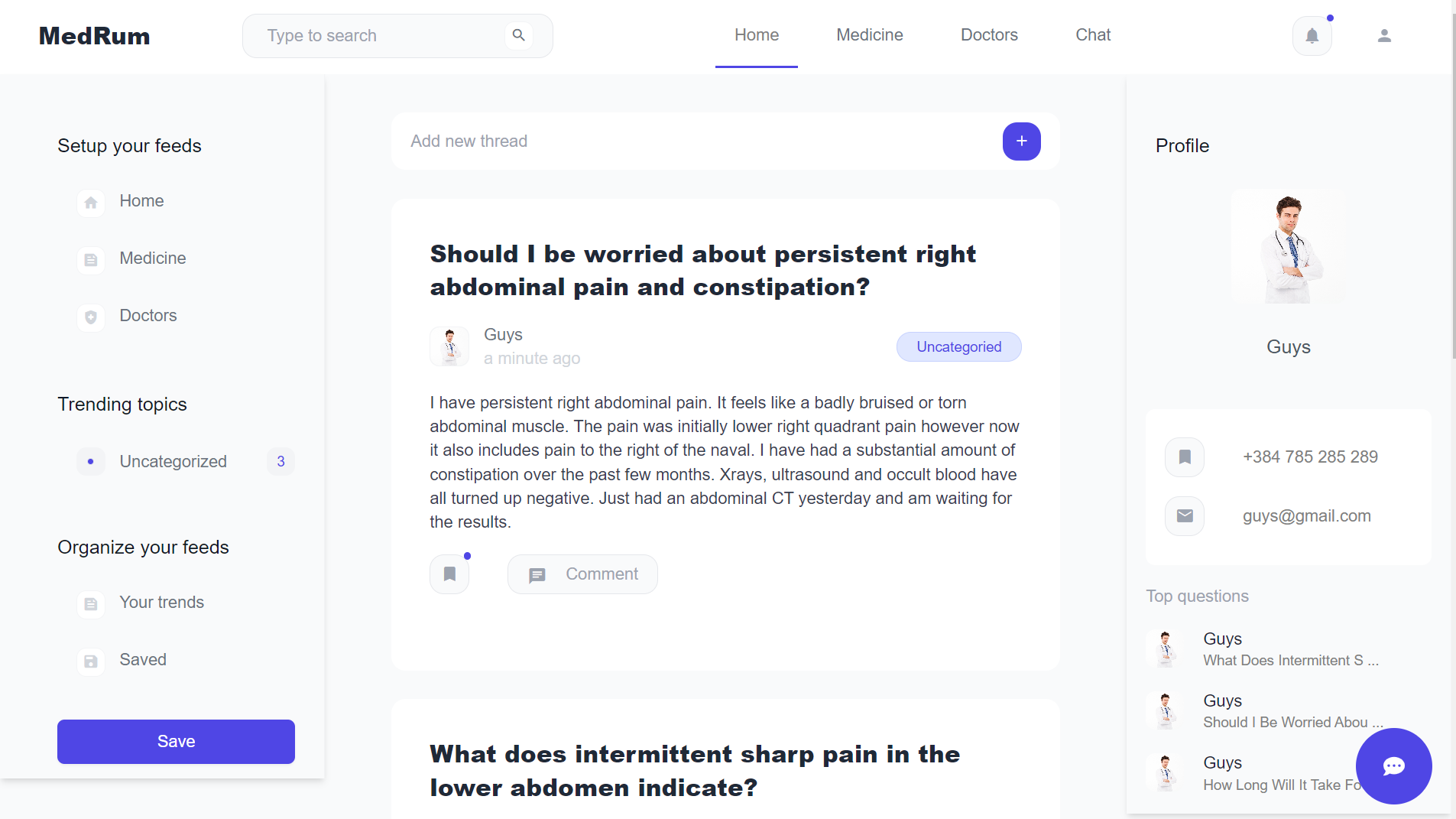


Figure 42. Healthcare questions page

#### Doctors page

On the doctors page, users will see all the doctors that have registered on the website and select to view details. Besides, users can also see the specializations of the doctors to filter doctors by their specializations on the left sidebar.

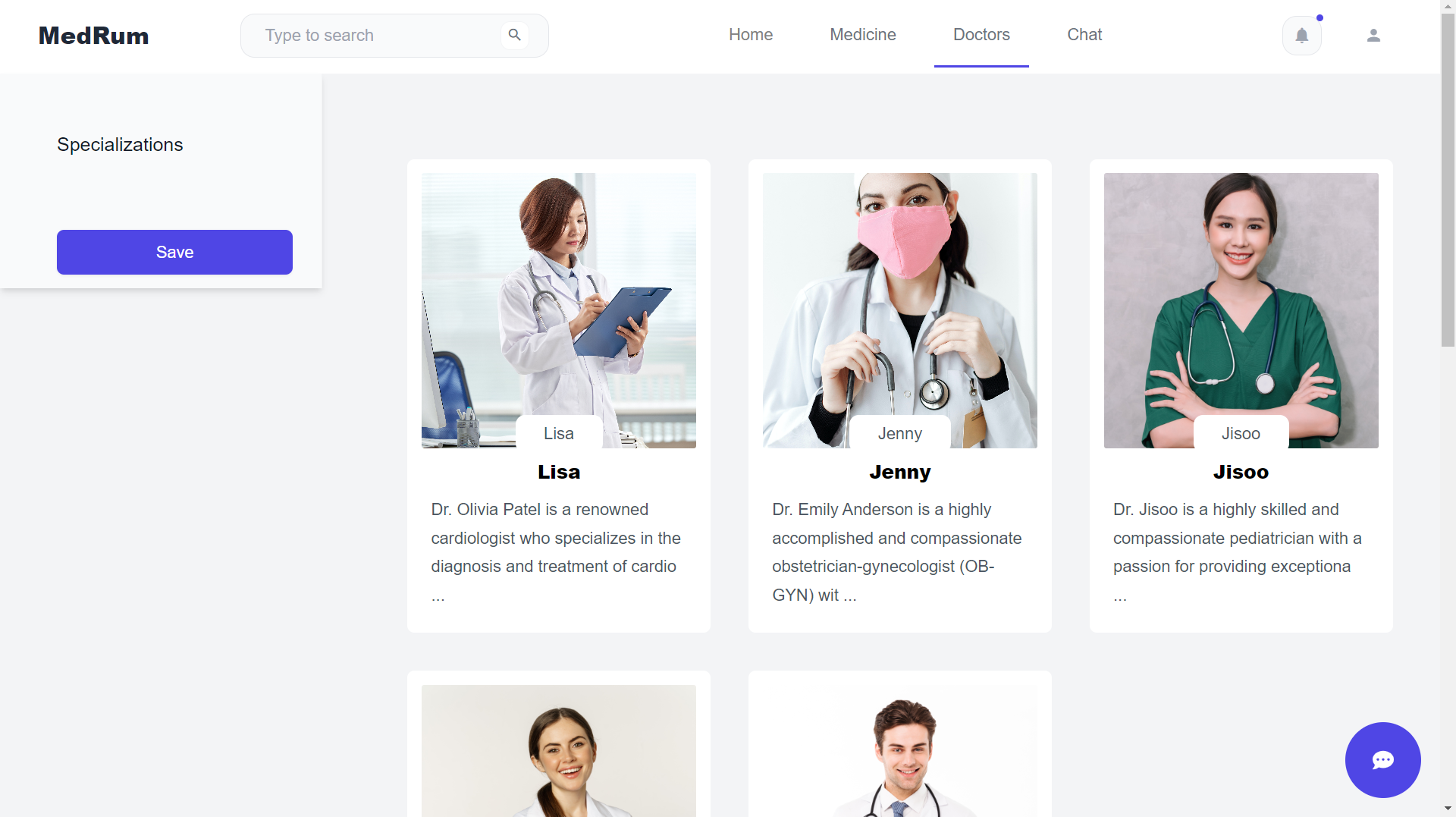


Figure 43. Doctors page

#### Chat popup

The chatbot popup window allows users to input questions and receive answers directly from the chatbox of the website. However, the conversation will not be saved when users chat using this feature.

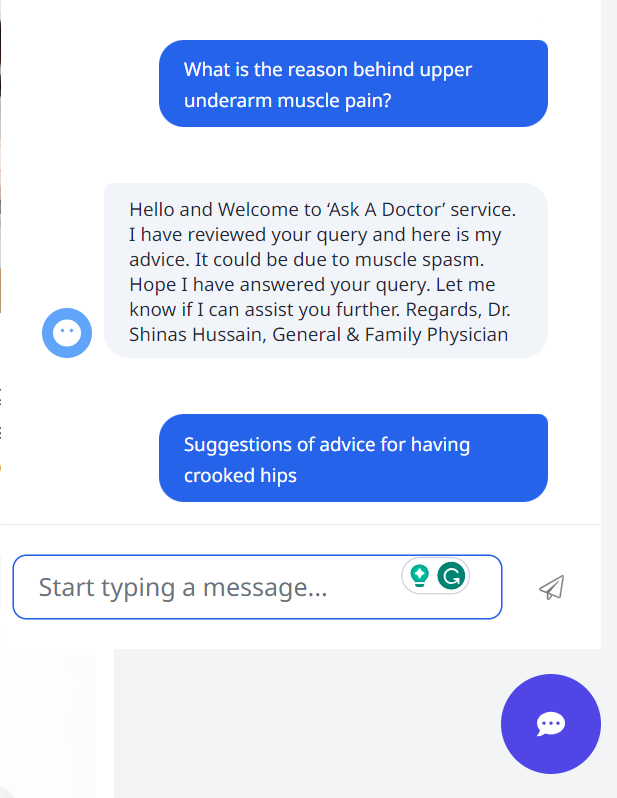


Figure 44. Chat with chatbot popup

#### Expose a HTTPS endpoint

The ngrok or serveo server will be used to set up an HTTPS endpoint to get forwarded the requests of the application to the Flask server.

# EVALUATION

## Result

The model is trained on Google Colab (with 12GB RAM and GPU T4) as well as on Kaggle (with 12GB RAM GPU P100) for 100 epochs, each epoch takes about 3 hours to complete.

Firstly, the chatbot was trained on the Medical Chat Dataset with more than 200,000 question-answer pairs about healthcare.

Next, the chatbot continues to train on the collected datasets crawled from healthcare websites with more than 100,000 question-answer pairs.

The chatbot was initially trained with different max sequence lengths (64, 128, 256, 512) on a relatively small dataset (about 10,000 question-answer pairs) to find the optimized hyper-parameters but still have good accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequence length | 64 | 128 | 256 | 512 |
| Epochs | 100 | 100 | 100 | 100 |
| Dataset | 222,380 pairs | 222,380 pairs | 222,380 pairs | 222,380 pairs |
| Token dictionary | 50,265 | 50,265 | 50,265 | 50,265 |
| Time for 1 epoch | 45 minutes | 1 hour | 1 hour 15 mins | 2 hours |

Table 1. Model training time on EHealthChat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequence length | 64 | 128 | 256 | 512 |
| Epochs | 100 | 100 | 100 | 100 |
| Dataset | 299,757 pairs | 299,757 pairs | 299,757 pairs | 299,757 pairs |
| Token dictionary | 50,265 | 50,265 | 50,265 | 50,265 |
| Time for 1 epoch | 45 minutes | 1 hour | 1 hour 15 mins | 2 hours |

Table 2. Model training time on EHealthChatMini dataset

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Sequence length | 64 | 128 | 256 | 512 |
| Epochs | 100 | 100 | 100 | 100 |
| Dataset | 22,780 pairs | 22,780 pairs | 22,780 pairs | 22,780 pairs |
| Token dictionary | 50,265 | 50,265 | 50,265 | 50,265 |
| Time for 1 epoch | 10 minutes | 12 minutes | 15 minutes | 20 minutes |

Table 3. Model training time on EHealthChatVN dataset

And below are the experimental results obtained with the internal, closed domain chatbot where “YOU” is the user input and “CITBOT” the virtual assistant.

In the chatbot experimental results, there are still questions that do not completely receive the exact answer as the questioner intended, but the rate of wrong answers is very low. Open domain chatbot results, with “BOT” as user input and “KEY” as system automatic reply.

With the above training results, the rate of correct answers from the chatbot is quite good short. But it can be seen that the chatbot has given different answers for each question, the quality of the answer will depend greatly on the learning data set. Below is an example illustrating when entering a sentence containing the wrong entities with a domain chatbot close

## Evaluate accuracy

To evaluate the accuracy of the chatbot, the test dataset need to be constructing. With each dataset, there is 10% of the data are extracted for testing purpose and not used in training process. After constructing the test dataset, the model will be evaluated using the BLEU score. The results of the BLEU scores are shown in the table below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset/Seqence length | BLEU | | | |
| BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
| English (512) | 0.078 | 0.070 | 0.068 | 0.066 |
| Vietnamese (512) | 0.078 | 0.070 | 0.068 | 0.066 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset/Seqence length | ROUGE | | | |
| ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSum |
| English (512) | 15.91 | 11.46 | 15.06 | 15.71 |
| Vietnamese (512) | 38.037900 | 18.684400 | 29.099300 | 36.344300 |

After testing the chatbot model, the experimental results of the chatbots is still not high due to the long sequence length of the answers while the length of the questions is pretty short. However, empirical observations show that chatbots do can generate answers with contextual relevance, the answer outcome depends to the initial data set built for training, because they are manageable dataset to include in chatbot training.

Compared with the results using the open domain chatbot model using Seq2Seq and LSTM in thesis by Nguyen Van Vi, with a BLEU score of 0.07, the domain chatbot model Open using Transformer gives a result of 0.07, with the result though both giving the same result BLEU score, but with our model the model training time is not yet up to 3 hours also achieved similar results, compared to training times of up to 72 hours now in Nguyen Van Vi's thesis, showing the learning speed of the model Transformer is very superior.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset/Seqence length | BLEU | | | |
| BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 |
| English (512) | 0.262 | 0.210 | 0.186 | 0.171 |
| Vietnamese (512) | X | X | X | X |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Dataset/Seqence length | ROUGE | | | |
| ROUGE-1 | ROUGE-2 | ROUGE-L | ROUGE-LSum |
| English (512) | 40.84 | 32.00 | 18.16 | 40.34 |
| Vietnamese (512) | X | X | X | X |

# CONCLUSION AND FUTURE WORK

## Conclusion

MedBot has reached the goal. We successfully created an ehealth chatbot and integrated it into others application that allows users to ask questions related to the health care problems and receive answers from it through a friendly UI/UX. Besides, the chatbot application also allows users to do other things such as: supporting users to search for questions, answers, and health care recommeded that they are interested in.

Although MedBot provides various useful functionalities, there are also some limits. At the present, MedBot does not have a very high accuracy performance and does not understand other sentences that are different from the healthcare domain. In additions, the chatbot also does not give a very details answer yet. The search functionality is limited, only supports search with name, but not its attributes such as content or topic of the question.

## Future work

Although the chatbot work pretty well with the questions on the healthcare domain, the chatbot still has some limits. For improving the chatbot further, we will expand and improve the dataset first, by filter all unreleated and all unnessesary long answers to make the dataset cleaner to help enhance the performance of the chatbot. Secondly, we will also try different parameters as well as try to modify others parts of the model to help improve the accuracy of the chatbot with long answers.

# REFERENCES

|  |  |
| --- | --- |
| [1] | Khang Nhut Lam, Nam Nhat Le, Jugal Kalita, "Building a Chatbot on a Closed Domain using RASA," 2022. |
| [2] | M. Z. N. B. Aleksandar Petrovic, "Singibot - A Student Services Chatbot". |
| [3] | K. K. R. R. Nura Esfandiari, "A Conditional Generative Chatbot using Transformer Model," 2023. |
| [4] | Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, Aimin Zhou, Ze Zhou, Qin Chen , Jie Zhou, Liang He, Xipeng Qiu, "A Large-Scale Language Model-based Chatbot System for Intelligent Education," 2023. |
| [5] | Kyle Swanson, Lili Yu, Christopher Fox, Jeremy Wohlwend, Tao Lei, "Building a Production Model for Retrieval-Based Chatbots," 2019. |
| [6] | Iulian V. Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, "A Deep Reinforcement Learning Chatbot," 2017. |
| [7] | Siddharth Verma, Justin Fu, Mengjiao Yang, Sergey Levine, "A CHatbot AI for Task-Oriented Dialogue with Offline," 2022. |
| [8] | Moataz Mohammed, Mostafa M. Aref, "Chatbot System Architecture," 2022. |
| [9] | E. Herberg, "Neural Network Architectures," 2023. |
| [10] | Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention Is All You Need," 2017. |
| [11] | Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for," 2018. |
| [12] | Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, "Improving Language Understanding," 2018. |
| [13] | Mike Lewis, Yinhan Liu, Naman Goyal\*, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural," 2019. |
| [14] | Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," 2002. |
| [15] | N. U. o. Singapore, "DeepViT: Towards Deeper Vision Transformer," 2021. |
| [16] | "Boostrap," [Online]. Available: https://www.boostrap.com. |
| [17] | "TailwindCSS," [Online]. Available: https://tailwindcss.com/. |
| [18] | "MySQL," [Online]. Available: https://www.mysql.com. |
| [19] | "Python," [Online]. Available: https://www.python.org. |
| [20] | N. W. Rohit Tamrakar, "Design and Development of CHATBOT - A Review". |
| [21] | A. W. Vitor Rocio, "Building a Chatbot for student support". |
| [22] | W. Z. X. Z. Z. W. Z. H. Sahand Sabour, "Chatbots for Mental Health Support: Exploring the Impact of Emohaa on Reducing Mental," 2022. |
| [23] | K. Ganesan, "ROUGE 2.0: Updated and Improved Measures for," 2018. |
| [24] | B. U. O. T. A. ECONOMICS, "Deep Learning Based Chatbot Models," 2019. |
| [25] | Nura Esfandiari, Kourosh Kiani, Razieh Rastgoo, "A Conditional Generative Chatbot using Transformer Model," 2023. |
| [26] | OpenAI, "Language Models are Few-Shot Learners," 2020. |

# APPENDICES