Computer Science NEA – Chatbot

By Kazik Sobotnicki

Contents

[Analysis 3](#_Toc120700477)

[Problem identification 3](#_Toc120700478)

[Stakeholders 3](#_Toc120700479)

[Interview 4](#_Toc120700480)

[Interview: Tasnim 5](#_Toc120700481)

[Why should it have a computational solution 6](#_Toc120700482)

[Possible computational methods 6](#_Toc120700483)

[Decomposition 6](#_Toc120700484)

[Abstraction 7](#_Toc120700485)

[Research 8](#_Toc120700486)

[Chirpy Cardinal 8](#_Toc120700487)

[Kuki AI 10](#_Toc120700488)

[Features of my solution 13](#_Toc120700489)

[Limitations 13](#_Toc120700490)

[Developer requirements 14](#_Toc120700491)

[Hardware 14](#_Toc120700492)

[Software 14](#_Toc120700493)

[Stakeholder requirements 15](#_Toc120700494)

[Design 15](#_Toc120700495)

[Functionality 15](#_Toc120700496)

[Hardware 15](#_Toc120700497)

[Software 16](#_Toc120700498)

[Design 17](#_Toc120700499)

[Create Venv 17](#_Toc120700500)

[ML model 17](#_Toc120700501)

[Choosing model 18](#_Toc120700502)

[Implementing model 19](#_Toc120700503)

[User interface 21](#_Toc120700504)

[Create HTML web pages 22](#_Toc120700505)

[Use CSS styling for the web pages 22](#_Toc120700506)

[Create HTML forms for users to input text 22](#_Toc120700507)

[Link to python backend using Flask 23](#_Toc120700508)

[User interaction 24](#_Toc120700509)

[Text from user should be read by the model 24](#_Toc120700510)

[Response from model outputted to user 24](#_Toc120700511)

[Allow user to enter message using microphone 24](#_Toc120700512)

[Record Audio 24](#_Toc120700513)

[Decide on model to convert text to speech 25](#_Toc120700514)

[Fine-tune model 25](#_Toc120700515)

[Implement model 29](#_Toc120700516)

[Allow users to report messages 29](#_Toc120700517)

[Reported messages and the report reason should be logged 30](#_Toc120700518)

[Testing 30](#_Toc120700519)

# Analysis

## Problem identification

These days, people often go out and talk less, spending more time in doors and therefore meeting less people and engaging in conversation less with people. As a result of this, individual’s social skills, and their confidence in those, are worse off. This means that individuals create less meaningful connections, and conversate with less new people. This shows that people need a way to practice these skills in order to build up confidence. A way where there is no risk of embarrassment or awkwardness, and no one to make fun of you. People need an environment they are familiar with, to allow them to build up those skills, becoming more familiar with casual conversations, in particular with people they do not know.

This can also be an educational tool, to show users the uses and abilities of AI. But more simply, it can be a fun piece of software for the users to play around and experiment with.

## Stakeholders

The potential clients and users for this product are wide and varied. From people that want to develop their conversational skills, to people that just want someone to talk to when they are bored, many people can make use of this chatbot. A lot of people may also feel that they cannot talk to other real people and want to talk about private matters that they would want to keep themselves.

People with mental health issues or terrible illnesses may not feel like they can talk to others. Therefore, an online chatbot would allow them to talk about their problems and express themselves, in a safe environment where they should have none of the worries about expressing themselves that they usually would when conversating with real people. This should hopefully help them to become comfortably with expressing themselves and their problems and insecurities to real people.

## Interview

My interview questions are:

1. Have you ever used an online chatbot before?

* This question should show whether they have experience with chatbots, as previous experience may influence their opinions, either in a more positive or negative light.

1. If you have used online chatbots, or other used other language-centred applications that use AI, what are your opinions on how comprehensible, understandable, and natural current AIs are?

* This question is can only really be answered if they answered yes to question one, but this gives me a good view of the general opinion on current AI language generation, what could be improved and what has worked in the past. It gives me ideas for what techniques my bot should use, and what problems other AIs have struggled with.

1. What can make discussing personal things with people uncomfortable?

* This question can show what they may have to gain from the chatbot, and how it can help them. This will help give me a better idea of what the program will be used for, allowing me to have a more precise set of aims and requirements for the chatbot, so I can tailor the chatbot to the exact uses of the user

1. How often do you message people online or through text?

* This shows how comfortable they are with a messaging-based interface, hopefully backing up the previous evidence about how frequent online messaging is. It may also be an indicator for how much they could benefit from the chatbot.

1. Is it easier to talk over text than over phone calls or in person?

* This shows how much they can benefit from the chatbot and supports my previous knowledge that many people prefer to text over in person or on the phone conversation. This, combined with some of the other answers will give me a good idea on what some potential users are like, and what they would get out of using my chatbot.

1. What features (current features or ones you wish for) can make messaging easier

* This will give me a good idea of what features users would want to message with. For example, some users (particularly those with disabilities) may find speech-to-text (entering a message by speaking into the microphone) very useful. Others may put a lot of emphasis on privacy features or different themes.

1. If you wanted to talk about yourself, would you find it easier for the other person to lead the conversation, asking you questions, or if they just let you talk and react to what you say.

* This helps decide how the bot should communicate. Whether the bot should be starting off and controlling the conversation in a way, or if the users want to be able to enter things about any topic, and have the bot react to what I am saying. Each method would require a different approach to creating the artificial intelligence, so it would be useful to get a second opinion on which one any potential users find the best.

### Interview: Tasnim

1. Have you ever used an online chatbot before?

* Yes, a few years ago. Mostly to play around with.

1. If you have used online chatbots, or other used other language-centred applications that use AI, what are your opinions on how comprehensible, understandable, and natural current AIs are?

* Whilst some were understandable and comprehensible, none felt at all natural. It was easy to tell that I was talking to an AI and not an actual human.

1. What can make discussing personal things with people uncomfortable?

* When you are talking face to face, you can see immediate reactions to what you are saying, through tone of voice (of any response), facial expressions and vocal expressions, which might make you feel uncomfortable if those are not the reactions you intended (even if the person does care and is trying to help).

1. How often do you message people online or through text?

* Very regularly, every day.

1. Is it easier to talk over text than over phone calls or in person?

* It depends on the situation, but it can be easier to talk over phone calls as it is easier to get across the meaning of what you are saying.

1. What features (current features or ones you wish for) can make messaging easier

* Reactions, emojis, emotes and direct replies. A search bar to search through older conversations. Maybe customisable interface (such as light and dark mode)

1. If you wanted to talk about yourself, would you find it easier for the other person to lead the conversation, asking you questions, or if they just let you talk and react to what you say.

* I would find it easier to lead the conversation when talking about myself, as it allows me to talk about what I want to talk about, and not rely on hoping the other person is talking about something relevant to me, or that I want to discuss

#### Analysis

This interview shows that there would be interest in an AI chatbot, however that interest is strongly dependent on the quality of the chatbot. Poor, unintelligible chatbots would quickly lose the user’s interest.

The suggestion to have the user lead the conversation is interesting and would allow more interesting conversations. Conversations lead by the bot would have a risk of all being very similar, so this would help to add some variety and make the conversations feel more natural, and less repetitive. This would however mean the AI model would have to be more complex, as it would have to be able to understand a wider variety of topics and user inputs, more of which could be unexpected.

It was also pointed out that talking can feel more natural and take less effort than typing. This could mean an option of speech-to-text input may be useful, as it would allow the user to feel more like they are having a normal conversation, whilst still be understood by the bot.

## Why should it have a computational solution

Text conversations are becoming more common these days. Statistics have shown that young adults think text conversations are just as meaningful as phone calls. Furthermore, a not insignificant portion of the adult population prefers texts over all other forms of conversation. This shows that online messaging and texting are a very familiar way of communicating, one that people are comfortable with. It would therefore be best if the way for people to practice and build up their social skills should be through a text like interface, on a phone or computer. They would already know how to use this and would not need to spend any time becoming familiar with the interface, so they could spend all their time focusing on the conversations.

Furthermore, as I have already discussed, people do not always feel comfortable discussing freely with another person, so the anonymity of a computer screen, and the knowledge you are not talking to a real person can, in some ways, be comforting. The use of an AI allows this to be achieved, and on a large scale, where the number of users would make it near impossible to have dedicated staff or moderators watching over or even controlling the conversations. When purely using software for this, artificial intelligence is the only way to truly accomplish this complex natural language processing task.

I could just have a dataset of conversational responses and use basic comparison techniques to find the most similar one to the user’s input, and output that. However, these responses may feel unnatural as there is no way to have a response for every possible user input, so some of the responses may be off topic. Also, if the program is just selecting responses from a list, there is a risk that the same response will be printed twice, which would feel very unnatural to the user, and make the program feel more automated then human.

## Possible computational methods

### Decomposition

Decomposition is the process of breaking down the problem into smaller parts, each one can be solved individually. Then, when all the solutions to the smaller problems are combined, you have the solution to the main problem. This process makes the problem less daunting, and uncomplicates it. As my program will be very complex, being comprised of multiple features, each one having a host of problems, my program needs to be decomposed into many subroutines.

The project can be broken down into a set of smaller steps, such as:

1. Train an AI model for conversational intelligence
2. Build algorithms for interacting with the model
3. Have a simple, easily accessible interface
4. Take in user input, either through them typing it into a box or perhaps saying it into their microphone
5. Have the AI read the input and output a text response
6. Print out this response onto the interface

This should happen smoothly, and as quickly as possible so it could imitate a normal online conversation between real people in an as natural way as possible. As well as inappropriate responses, long response time and a clunky UI can take away from the immersion, and make the conversation feel less natural. This would take away from the familiarity and comfortability I am trying to make the users feel when using the program. This means that the program would no longer be able to achieve its aims of imitating a natural conversation and would not help them build up their confidence in real conversations.

Also, the AI model would only have to be trained once, and could then be stored in the software files, to be accessed by the program later on. This would decrease the size of the program and prevent long loading times when it is first run (loading times which would be heavily dependent on the speed of the user’s computer).

For the most part, decomposing my problem is not necessary, and I could just use minimal subroutines and just have long blocks of code, attempting to solve most problems at once. However, this makes the development process more complicated and time consuming. Trying to solve every problem at once also makes the code more convoluted and harder to understand, which would make it harder to go back and update the program.

### Abstraction

Once the problem has been decomposed into subroutines, it becomes easier to see what parts of the problem can be abstracted.

Abstraction is the process of removing unnecessary details to create a representation of reality. It can help to remove unnecessary complexity from the program, such as:

* As the conversational AI model would not be able to take images or emojis as input, only text input needs to be considered and coded for. The only way images could be taken as input is if another AI model was created, trained, to be used for image classification. However, this is a whole other complex program in and of itself, which is way beyond the scope of this program. I could ask users to give captions to images they want to send to the model, however, at that point the images are useless, as the user has to explain them, and the model only pays attention to the explanation.
* Any reactions to the AIs messages can be ignored, as these would be unlikely to give any extra information that the AI could not already gain from the user’s messages. The user should still be able to send reactions, but these are purely for the user, as the AI would not be able to understand these and there is nothing it could gain from it.

## Research

### Chirpy Cardinal

#### Link: https://stanfordnlp.github.io/chirpycardinal/

#### Overview:

Graphical user interface, text, application, chat or text message

Description automatically generatedChirpy Cardinal is a chat bot developed by a natural language processing research group at Stanford university, with the aim to research and advance conversational AI. It is seen very highly and won 2nd place in the Alexa prize. This is good, as it means it is a great example of the best of current conversational AI. It has a simple, easy to use interface, which although bare bones, allows the user to simply chat with the bot without needing to navigate through clutter.

Figure

Graphical user interface, text, application, chat or text message

Description automatically generatedIn terms of the AI and other algorithms behind it, it seems to control the conversation, asking the user questions and then reacting to their answers, continuing on the conversation. It is quite good at continuing the conversation, for instance, going from talking about cooking, to being vegetarian, to having pets.

It is generally relatively good at reacting to the user’s inputs, though this is partially because it is good at using less precise responses. For example, in Figure 2, its response to “I’m not vegetarian”, is “Oh okay, I’m actually vegetarian”. This is a clever response, as the “Oh, okay” gives the impression of a reaction to what I said, but could in fact be a perfectly understandable response whether the user said that they are vegetarian or not. It also has a pattern of never directly referencing what the user said. So when it asked for food reccomendations and I gave one, it didn’t directly include the name of that food in the response, it just said “That sounds really good”, which would make sense whatever I put in.

Figure

#### What am I going to implement in my own Chatbot?:

I like the clean, simple user interface, simply showing the text boxes and a bar to enter your response, with a button to submit the response. This simplifies the user experience, allowing them to easily interface with the product, therefore making them less stressed out by the interface and other things distracting them, and focusing entirely on the bot. I want to do something similar in mine, though I may add one or two other features, but I will try to make those as less overbearing and distracting as possible.

I also like the text message style of interface, with the send icon being common in messaging apps, and the text bubbles being familiar. This adds a familiarity to the program, as the users are already comfortable with the interface and how it works, so they feel more at home. This also gives the chat app a more casual feel, as the users likely use other messaging apps regularly, so using my chatbot will feel similar to just texting a friend.

Text

Description automatically generatedFurthermore, I like how casual the conversation is. Although it is led by the bot, who is the one asking questions, it feels much more like a casual conversation with a friend and less like an interrogation.

Moreover, I liked how it could occasionally refer back to something you said earlier in the conversation, for example in Figure 3, when it refers to the user mentioning school earlier and starts a conversation.

Figure

#### What am I not going to implement:

Although I like the simple, clutter-free interface Chirpy Cardinal uses, I would like to add one or two extra features. These could go a long way to help the user interact with the chatbot, and more varied ways the user and chatbot can communicate. These may also make it easier for the user to use, as they may be accessibility features such as speech-to-text.

Also, a lot of the bot’s conversations are similar, and I have found only three different starting conversation prompts once I entered my name, all asking me what food I like. I feel like a bit more freedom, and the user having more control over the conversation, should help them express themselves a bit more.

### Kuki AI

#### Link: https://www.kuki.ai/

#### Graphical user interface, text, application, chat or text message Description automatically generatedOverview:

Figure

Graphical user interface, text, application

Description automatically generatedKuki is a conversational AI that users can chat with over a text messaging like platform. Kuki has been given an avatar and has used this to expand to other media such as video games. This bot is also seen very highly, winning 5 Turing competitions, and is widely used with 25 million people chatting with it. Kuki’s interface has slightly more features to it, whilst still keeping the text message like appearance. It has features like mini games and quizzes, and a whole currency that can be earned by chatting with it and spent on gifts for it.

Figure

The quizzes can be personality quizzes or pop quizzes and are presented through the same text message-like format as the rest. The coins slowly build up over time, with the user getting small amounts when they talk with Kuki. Users also have the option to buy coins to supplement their funds. These coins can be used to purchase gifts like food items to give to Kuki. There are also other games, like small, simply games that work just with text and emojis, like tic-tac-toe and connect 4.

The chatbot itself also controls the flow of the conversation, asking questions for the user to answer, though unlike Chirpy cardinal, Kuki’s questions are a lot more open ended allowing for a wider range of user responses, which the bot is not too bad at doing. It has more direct references to what the user said, for example in Figure 4, where the user mentions liking bike riding and finding it refreshing, Kuki, comments on that, directly talking about bike riding and talking about it being refreshing in a different. It is also quite good at call backs, as shown in Figure 6.

Figure

This is a useful feature as it makes this conversation feel less random. It allows the bot to switch topics, but still not in a completely unexpected way. It also gives the feeling that the bot is truly listening and knowing what you say, rather than just giving off premade responses.

The bot also has further features, like video calls. These, whilst in theory make the bot more personable and human like, in reality often draw the user out, as reinforces to them that the bot is not real and can often lead to an uncanny valley-like effect, as the conversation may be realistic and human-like, but the bot is obviously not human. However, it can be useful as it allows users to talk using a microphone and hear responses as the bot says them, rather than only being able to write and read responses. This gives the user more ways to communicate with the bot and is more accessible.

#### What am I going to implement in my own Chatbot?:

Similar to Chirpy Cardinal, I like the familiar feeling of the interface. It draws the user in and makes them feel comfortable. It also means that they don’t have to spend time becoming familiar with the interface, as it is already similar to services that they use every day. However, I also like the visual flair added to separate it from other chat apps: the different chat bubbles and the avatar next to the bot’s messages.

I like how the bot directly references what the user says, either in direct responses or call backs to topics that were discussed earlier in the session. These makes the user feel heard, and make it think the bot is smarter as it is not just giving off generic responses that could work for anything the user inputs but is tailoring the user’s responses to exactly what was inputted. This also makes the user feel like the bot is genuinely listening and caring about what is said, instead of just sending out a response seemingly randomly picked from a list of responses.

#### What am I not going to implement:

Graphical user interface, application

Description automatically generatedI am not going to use the quizzes. I feel like they take away from the casual feel of the conversation that the bot has been good at creating, and instead feels more like the user is taking some internet quiz. It doesn’t help that the actual questions feel clunky. From the “Type START to begin the quiz”, to the “Please answer either A, B or C for question 1 or Quit to finish the quiz” (as shown in Figure 5), It feels less like a causal conversation and more like entering commands into a computer (which yes, is technically kind of what the user is doing anyway, but the bot is better at masking that at other times).

Figure

I am also not going to use the currency feature. I don’t want to introduce a monetary or economic part to my program. I like the feeling of just having a casual conversation with the bot, either to pass time or build up conversational skills. I think introducing a monetary aspect to it can make the user feel more stressed on what to spend the money on, and to earn as much as possible. This becomes more like a game, and less like a conversation with a friend. Also, as you earn coins by chatting, users may feel pressured to talk with the bot as much as possible, perhaps giving shorter responses, with less thought put into them, so the responses will be quicker to enter, allowing them to earn more money. They may care less about having a good conversation, and more about gaining money, which takes away from the purpose of the bot.

Furthermore, I will not implement a video chat feature. This is for a few reasons. Firstly, it is a very large addition, which would end up taking way too much time to implement, especially compared to the algorithm behind the chatbot, which should be the main focus. But secondly, it introduces an uncanny valley like effect, drawing attention to the fact that the bot is not real, which would be easier to avoid with just the messaging conversation. However, I will take some parts from it. The ability for the user to interact with, and enter messages, using their voice is a good feature, both to add more ways to interact with the bot, but also to cater for people that may use more accessability features.

## Features of my solution

My solution will be a chatbot that you can communicate with through a text message like interface. It will use a conversational AI instead of a rule-based approach, as this generates more natural feeling messages. It will use a sequence-to-sequence model with a separate encoder (to analyse the users input and find the semantic information about the user’s input) and decoder (which takes in the semantic information from the user’s input, and returns the next word in the sequence, in order to generate the response).

It will use a messaging like interface, with a bar to enter a message, and text bubbles to show previous messages. It will also have a speech-to-text option, which will allow users to enter a response by talking into their microphone and use automated speech recognition to convert that into a text input.

The interface would have the option for the user to react to messages with emoji-style reaction bubbles, as well as report messages that they believe to be incorrect (to allow for better debugging and improvement)

If possible, I may include the option to have different personalities for the user to choose from, which would allow the user to choose a bot that would be best for them, as well as have some variety if they get bored of the bot they are talking to.

Finally, users will have the option to report messages, adding a reason onto the report. This allows for easier bug reporting and checking, making future improvements to the program easier.

### Limitations

Part of the limitations will stem from the conversational AI model. If this cannot generate very natural responses, the conversation will feel unnatural and less human-like. Also, having multiple pretrained models stored as part of the program will dramatically increase the disk space it uses. The option to have multiple personalities would be heavily dependent on finding a good enough dataset, preferably one with multiple different people’s conversations.

## Developer requirements

### Hardware

* **A fast enough computer.** This must be able to train machine learning algorithms, in a reasonable time, which requires a reasonably fast processor. The standard peripherals of a mouse, keyboard and monitor are also needed for using the software.
* **Microphone.** As the program will have a speech-to-text feature, A microphone is needed, although the standard microphone built into most computers and headphones is good enough.

### Software

* **Windows, Mac, or Linux operating system.** These are supported by both python and PyTorch (the python library used for creating and training machine learning models).
* **Python interpreter.** This is the programming language my chatbot will be developed in, as it is good for machine learning, and support many useful pieces of machine learning software, such as PyTorch.
* **Pip.** This is python’s package manager/installer, which will be used to install all the packages required to develop the chatbot. By default, this usually comes installed with python.
* **PyTorch.** This is a library for python that is the most widely used for developing and training machine learning algorithms, with support for many different types of algorithms. This is important for creating and training the machine learning model used for the chatbot. Whilst TensorFlow, the other popular machine learning library, is an option, I am going to use PyTorch as it is more pythonic. This means that common debugging tools, even simply print statements, work with PyTorch, whereas TensorFlow requires you to use its own debugging tool alongside normal python ones.
* **Transformers.** This library has pre-trained models for many different typed of transformers, and will be used to access the model, and fine-tune it.
* **Pandas.** This is a library for python which can be used for data analysis and data manipulation. I will need it for reading in and filtering the data from the dataset.
* **NumPy.** This is a python library for doing complex mathematical operations and is also a pre-requisite for all of the other libraries used.
* **External Datasets.** Appropriate datasets of conversations are needed to train the model.
* **Flask.** A web application framework for python that can be used to create the web app for the chatbot. It allows the html files for the website to interact with the python code for the chatbot, sending data in between. This will allow users to enter messages into a html form, which will be processed by the python code, the output from which will be displayed on a new html page.
* **Any other possible packages.** This may not be all the necessary packages needed to develop the chatbot, more may be discovered during development, and they will need to be installed.
* **A web browser.** A web browser is needed to open the web app, as it will be run out of html files. This can be any mainstream web browser, such as Google Chrome or Firefox.

## Stakeholder requirements

### Design

* **Intuitive, simple interface.** An easy-to-understand interface should be used to allow the user to quickly understand how to use the program without the need for complex instructions. This interface should be similar to a messaging app, as this is a type of interface users are already familiar with.This reduces the barrier to entry for users to access and use the program.
* **Customisable display.** Users should be able to customise their display, to make it suitable for their needs and preferences. This should at least consist of light mode or dark mode options.

### Functionality

* **A working AI model.** The model should be able to produce understandable text, that makes sense in the wider context of the conversation currently going on between the user and the chatbot.
* **Reactions.** Users should be able to react to the AI’s messages with icons/emojis. This adds to the user’s interaction, and makes the conversation feel more organic, and more like a conversation the user may have on a chat application like Discord with a real person.
* **Reporting.** Users should be able to report messages the bot has sent that make little sense or are inappropriate for the context of the conversation. This could allow the users to add a message explaining why they reported it. This should allow for me to test and debug the program easier.
* **Speech to text.** Users should be able to enter messages into the model using their voice, with the program recording them through their microphone and converting the audio to text. This feature allows more ease-of-use, as users do not have to be directly sitting at their computer to talk to the chatbot. It is also better for those who find it hard to use a keyboard and would prefer to simply talk to the bot instead. Therefore, whilst I could only let the user enter text through typing into a text box, allowing the user to enter text through their microphone makes the program more accessible.

### Hardware

* **A decent computer.** A computer with decent specifications will be needed to run the chatbot, as the models will take some computing power to use. The chatbot technically can be run on lower powered computers but will be incredibly slow. The standard peripherals of a mouse, keyboard and monitor are also needed for using the software.
* **Microphone.** As the program will have a speech-to-text feature, A microphone is needed, although the standard microphone built into most computers and headphones is good enough.

### Software

* **Windows, Mac, or Linux operating system.** These are supported by both python and PyTorch (the python library used for creating and training machine learning models).
* **Python interpreter.** This is the programming language my chatbot will be developed in. It will need to be installed on the user’s computer to allow them to run python programs.
* **All packages used in the program.** All the packages used in the program will be listed in a requirements text file, allowing them to be easily downloaded using pip (pythons package manager and installer).
* **Pip.** This is python’s package manager/installer, which will be used to install all the packages listed in the requirements text file. By default, this usually comes installed with python.
* **A web browser.** A web browser is needed to open the web app, as it will be run out of html files. This can be any mainstream web browser, such as Google Chrome or Firefox.

# Design

## Create Venv

A virtual environment is a folder structure that allows the developer to use a lightweight, but isolated environment. This means that the current python version, plus all modules and libraries used in the development of the program will be stored with the program, independently of all other python modules installed globally.

I could just store all packages globally, however there are a number of issues with this approach, including dependency conflicts. When programming for this chatbot, I would be using a certain version of each of the packages, whilst the packages may be updated in this time – updates which could have conflicts with the versions I am currently using, potentially breaking the program if used. For example, if I were to install these updated libraries for use in another project and hadn’t used a virtual environment for the chatbot, these updated libraries (which would overwrite the old ones) could cause conflicts with the chatbot, which had been programmed for earlier versions of the package.

This is not only an issue that would affect me, but also other users who install the program, as these users might have other versions (older or newer) of the packages installed. Users will need to be able to reproduce the environment on their own computers, which will need to not conflict with any possible packages they already have installed.

There is also the possible case that users may be using a computer that does not have administrator privileges, and so won’t be able to install packages into the python directory. Therefore, these users should also make use of a Venv.

## ML model

### Choosing model

#### Why transformers?

An important part of creating a good chatbot is to choose the right machine learning model. One of the best performing types of machine learning models for natural language processing tasks is the Transformer, introduced in 2017.

Transformers are a type of neural network architecture that has seen wide use in the field of natural language processing, performing better much better than other types of machine learning models like recurrent neural networks. RNN’s loop through the language data inputted, finding relationships between data points that it can connect together to gain a better understanding of the input data.

Whilst I could create my own model using TensorFlow and Scikit-learn, Studies have shown that Transformers perform much better across a wide range of natural language tasks. Furthermore, models available from Hugging Face are pre-trained by the big companies and researchers that created them, a process that requires incredibly expensive hardware and trains them to a much better level than I ever could. Therefore, using transformer models from Hugging Face allows me to use models that are a lot more refined and accurate than a more basic model that may make use of a neural network. However, this becomes much more computationally expensive as the input sequences get longer, or more sequences are used, as the RNN tries to compress the sequence into a fixed length vector.

This is where the principle of attention is useful. This means that the model only pays attention to the parts of the input with the most information, therefore allowing the model to maximise the information gained from the input, whilst also being able to handle longer inputs

Transformers, which use the principle of attention, have shown to have much higher accuracy scores whilst having lower training costs than non-attention-based models across a number of language tasks, therefore showing it to be the best type of model for natural language processing tasks.

#### What transformer?

Whilst there are a number of different Transformer models, many of them are created and pre-trained for set purposes. For example, Facebook’s BART model is best used for summarisation, and Google’s T5 is best used for translation.

I have decided to use Meta’s BlenderBot, which is created specifically for Conversational AI, the task I need the model for. All iterations of BlenderBot have been shown to outperform other competing models – like Microsoft’s DialoGPT – in human evaluation tests. And the latest iteration, BlenderBot 3, performs even better than its predecessors. It is factually incorrect 47% less of the time more up to date (on topical questions) 82% of the time. However, as this is a developing field, it is still not perfect – with 1.2% of users reporting responses as incorrect or nonsensical, and 0.12% reporting messages as off-topic. Overall, whilst it is not perfect it is still by far the best transformer model for conversational intelligence. As it is created and trained specifically for conversational AI, it will produce better results in that specific task that I need it for, as opposed to using a more general-purpose model like Google’s BERT. Using a standard, pre-trained version of the model is a lot more practical for my purposes, as Meta states that the model requires a 32GB V100 graphics card to fine tune, which makes it impractical for me.

### Implementing model

Transformer models are publicly available through the Hugging Face library, which allows you to access the models and their tokenizers.

The BlenderBot model has different sizes, with the larger size giving higher quality outputs, but being more computationally expensive.

#### Tokenizer pseudocode

Tokenizers are required to convert the input data, in this case a string, into a type of data understandable by the model, in this case tensors. Hugging Face provides access to the BlenderBot tokenizer, which has been pre-trained by Meta. Meta provides a number of versions of BlenderBot, including a small 90M model, a medium sized 400M model and larger 1B, 3B and 9B models. I am going to use the 400M model as it strikes a balance between good performance and not being too slow. Whilst I could use a larger model, this would make the program run even slower, which would make the programme a lot harder to use for users with lower powered computers. So therefore, as I want to maximise the number of possible users that can use my program, I will use a smaller model.

Text

Description automatically generatedThe name of the model needs to be specified, and the tokenizer should be called, and used to encode the input message.

#### Model pseudocode

Text

Description automatically generatedThe model needs to be called and then is used to generate the ids for the reply tokens

#### Decoding pseudocode

The reply tokens then need to be decoded into a string of text that is readable for humans.

#### Overall pseudocode

This can be combined into one function which can be called whenever a message needs to be generated. As this is done as a reusable component, it means less code needs to be written, allowing for quicker development time and simpler looking code. Text

Description automatically generated

## A picture containing graphical user interface Description automatically generatedUser interface

Figure

For my user interface, I am going to use multiple web pages, linked to a python backend using the Flask package. Figure 11 shows a mock up what the light mode version of the GUI will look like.

For my design, I took inspiration from popular chat apps like WhatsApp and Discord. This means that users will already be familiar with the user interface, therefore reducing the time it would take for them to get comfortable using the app. I have kept the interface somewhat basic, in order to make it less cluttered and easier to look at. Apps many elements in their UI, and many things happening on the screen at once, can make the user feel overwhelmed, and too distracted to use the app for its original function.

Some of the blank space left is also useful if I were to add more features to the app in later updates. For example, there is a lot of space in the character list for extra characters, so if I decided to update the app to add more characters the user can talk to, I can just add them on the end. If I had left very little space in the UI, then if I come to add more features, I may have to drastically rearrange the UI, making this choice not very practical.

I have used bright high contrast colours to easily differentiate the different elements in the UI, making everything clear and standing out, as well as creating a design that is more aesthetically pleasing than a more basic design. Whilst I could have just used black text and black borders on a plain white background, not only would this not have looked as good, but the more colourful design is also better for accessibility, as it makes everything stand out from the background, and from other parts of the UI. This makes it easier to see and understand for those with difficulties with sight. I tried to avoid using certain colours, like using both red and green together, to prevent issues for colour blind users. Whilst this does only cater for those with the most common types of colour blindness, I think this is the more practical as it caters for most colour-blind people whilst still allowing me to make a colourful design.

### Create HTML web pages

### Use CSS styling for the web pages

Icon

Description automatically generatedCSS will be used to style the web pages: adding colours and other appearance tweaks, as well as creating grids.

Figure

Icon

Description automatically generatedCSS grids are a useful way to arrange and order elements on a web page and can be used to make the design more consistent. I would make the page into a grid, can then create each element to fill a certain number of cells in the grid. An alternative to using CSS grids is to use a flexbox, however flexboxes only work for one dimension, either in rows or columns. As you can see in Figure 13, the flexbox is only ordering items in rows, with the rows wrapping around to the next line if needed. This is alot less organised than the grid shown in Figure 12, which has neatly ordered items by row and column, organising the elements in a much more structured way.

Figure

I will use external CSS, keeping all the CSS in one separate file as I will have a constant style across all pages. If I use internal CSS, I will have to make edits to the CSS in every HTML file, often making the same edits each time, whilst external CSS means I just have to make the edits once. It also prevents repetition, as due to keeping a relatively constant styling, using internal CSS would mean I would be adding the same CSS code to every HTML page. This not only means it is harder to test and refactor, but also increases development time and disk space used (although the reductions in disk space would be minimal due to the small size of CSS files in comparison to other parts of the program).

### Create HTML forms for users to input text

There will be two elements where the user can enter text. The first one is the message box, where users can enter text that the model will read and output a response to. The second one is the search box, where users can search through previous messages. Both of these require the HTML page to take the user input and send it to the python code that can take it as an input, and the HTML page will update to show the output.

### Link to python backend using Flask

Flask is a python package, which functions as a framework to connect html.

Flask organises the main Python script into separate functions, one for each HTML page, that it runs when the page is loaded. It can use the GET and POST methods to take the input from the HTML forms and store it as a variable in the function. This allows it to be run through the rest of the algorithm. The function can then return the next web page to be loaded and transfer variables from the function to the new HTML page. Flask then allows you to use the Jinja template engine (built into Flask) to output the variables onto the web page. Jinja also allows you to use basic programming tools like for loops in the HTML templates, allowing you, for example, to create a div for every variable in a list.

A popular alternative to Flask is the web framework Django, however Flask has a number of advantages over Django. Firstly, Flask is more lightweight, requiring a smaller codebase, speeding up development time and reducing disk usage. Flask also works better with external packages as it functions as a light, modular framework that creates the web framework and leaves everything else open to the developer. Flask also allows much easier access to the request object – used for access data from HTML forms, whereas Django requires the Request object to be explicitly passed around.

#### Text Description automatically generatedText Description automatically generatedBase Flask pseudocode

Here I have specified functions to be run when the user access the home page and the message page. The home page function will just display the home page for the web app. The message page function will take the inputted message from the html page, use a separate function (to be written later) to generate the reply and then renders the message page, sending the generated reply onto the page.

## User interaction

### Text from user should be read by the model

Flask’s request class provides a number of ways to request data for the flask app, for example ‘request.json’ parses JSON data, ‘request.args’ parses data from the URL query and ‘request.files’ parses in uploaded files. As my html files will contain input forms for the user to enter their message into, ‘request.form’ is the best way to do this. It requires you to specify the name of the form you are getting the data from, allowing multiple forms to be used. When using this, the type of requests used need to be put into the methods parameter in ‘@app.route’. ‘GET’ is the default method and is used to request the html web page with the form. When the user submits data into the form, this sends a ‘POST’ request, which the function handles.

The data is requested (using the ‘POST’ request) from the messages HTML form and stored as a variable to be used later in the program.

#### Form request pseudocodeA screenshot of a computer Description automatically generated with medium confidence

### Response from model outputted to user

The page should be updated to show the message the user sent, as well as the message the model outputted in message bubbles, moving the others up.

Flask uses the ‘render\_template’ method to render HTML pages stored in the templates folder, which is usually used as the return of a function for the original page. The ‘render\_template’ method supports inputting other variables alongside. These variables will be used in the template as Jinja variables, that can be used in the HTML code of the page

#### Returning html page pseudocodeText Description automatically generated

### Allow user to enter message using microphone

Users should have the option to submit a message to the user using a microphone. This should be done by having the python program record input from their microphone in the background whilst the user is looking at the html page, and then converting this input into a text string.

### Record Audio

When the user selects the microphone option, the python script should record audio and save it as a temporary file in the program folder. The recorded audio will have a set length of 5 seconds. Whilst in theory I could have the user choose to end the recording themselves, it would be much simpler to code for the recording to automatically stop after a set time, and it is unlikely any audio messages would be longer than this (I will test to make sure this is true).

I will use the Sound Device and Sound File libraries to record and save the audio. First, I declare the duration and sample rate of the recording (the sample rate will be 16kHz to match the training used for the Wav2Vec model which will transcribe the recordings). Then I use Sound Device’s record function to record the audio, specifying the sample rate and the use of two audio channels. Finally, I will use Sound File to save the audio file, using Sound File’s write function.

#### Recording audio pseudocodeGraphical user interface, text Description automatically generated

### Decide on model to convert text to speech

To convert the audio file to a readable string that can be used to generate a reply to, another transformer will be used. In this case, Meta’s Wav2Vec transformer, a state-of-the-art model for automatic speech recognition. The model uses self-supervised training, being trained on unlabelled data to allow it to achieve the best speech representation possible. It then uses supervised fine-tuning, using labelled data to get better at predicting certain words or phenomes. These approaches, particularly the self-supervised training, give it a massive advantage over other suitable models. Allowing the model to learn good speech representation allows it to achieve accurate results on a small amount of labelled data.

Chart, scatter chart

Description automatically generatedWhilst I could use a different model, if that model does not use the same self-supervised training, it would require a much larger amount of training data to reach similar accuracy levels. Whereas Wav2Vec reaches very good accuracy rates with a dataset 10 times smaller than what is usually used. As you can see from the graph above, the latest version of Wav2Vec (2.0) outperforms every competing transformer model, making it the best model for me to use.

### Fine-tune model

Fine-tuning a model is the process of training the model on a dataset specific to the task it will be used for, making it perform better at that task.

Whilst I could just use the default, pre-trained version of Wav2Vec 2.0 that is available on Hugging Face, it has been shown that fine-tuning a pre-trained Wave2Vec 2.0 model on even a very little amount of data yields a very small word error rate.

#### Dataset pseudocode

Fine-tuning a model requires a dataset to train the model on, and Hugging face provides suitable datasets for this purpose. I will use the TIMIT dataset, a set of recordings of 630 speakers, speaking in 8 different English dialects, all reading the same 10 sentences. This makes the dataset particularly useful for me as training it on a wide variety of English dialects means it should be more likely to understand a wide variety of English speakers. It has a version specifically for ASR (automatic speech recognition – the task the model will be used for) that I will use, that is especially good for this task as all the audio files are in the same 16kHz format, allowing me to use a more appropriate, more customised training algorithm.

As well as a standard train-test split of the dataset (which the dataset comes with already), I will also need to split the chunks of speech in the dataset into letters and use this to create a dictionary of the letters used and how many times they occur – using this as the vocabulary the transformer will use.

Text

Description automatically generated

Before this is done, the dataset will also need to be reformatted to get rid of the special characters found in the text, like punctuation. This is done as these special characters don’t have a specific sound linked to them, so the model would struggle to link them to the audio.



#### Creating data collator pseudocode

The first step to creating the training algorithm is to create a data collator. Data collators are used to combine many samples from a dataset into one batch, which can then be processed by the processor I created earlier and used to train the model. The data collator used to fine-tune Wav2Vec is different from other data collators in that it will need to dynamically pad the samples, ensuring they are padded to match the longest sample in their batch, instead of the overall longest sample. This is particularly necessary due to the differences in length of Wav2Vec’s input samples in comparison to its output samples.

To do this, I need to apply separate padding functions to the labels and the input values, as input speech and output speech are of different modalities.

The following class will be created for the data collator:

Text

Description automatically generated

This diagram shows the data collator class will have a number of attributes. Firstly, padding will specify to the collator whether or not the returned sequences should be padded. Max\_length sets the maximum length of the input values of the returned data, and max\_length\_labels sets the maximum length of the labels of the returned data. Finally, pad\_to\_multiple will pad the sequences to a multiple of the given value.

The call method will be defined as:Text

Description automatically generated

Firstly, the inputs and labels are split as, as stated above, they have to be of different length and will therefore be padded differently. I then use the processor’s pad method to pad first the inputs and then the labels, using their respective maximum length and padding arguments. As I am using PyTorch, I will also specify that the padding process should return PyTorch tensors.

#### Evaluation metrics pseudocode

During training, the model should be evaluated using the word error rate. Loading the word error rate will return a sequence of logit vectors, which will contain the log-odds for each word in the vocabulary dictionary I created earlier. As I want the most likely prediction, I will take the argmax of the logits. I will also decode the prediction ids and label ids. Text

Description automatically generated

#### Loading modelText Description automatically generated

Here I load the pre-trained Wav2Vec model from Hugging Face.

#### Training arguments pseudocode

Here I define all the parameters related to training the model.

Text

Description automatically generatedGroup by length groups samples of similar lengths together into one batch, speeding up training by reducing the number of padding tokens that are passed through the model. Per device train batch size specifies the size of the batch that will be run through each CPU core. Setting the evaluation strategy to ‘steps’ means that the evaluation is done at a set rate (set by the eval\_steps) parameter. Gradient checkpointing is used to save memory, which is useful due to the lower computing power I have access to. However, this does mean the backward pass will be slower. The use of fp16 mixed precision training reduces the also saves memory, with the trade-off of having a lower precision. This is a trade off I feel is necessary to ensure the algorithm can run smoothly on the lower computing power.

Any arguments not explicitly specified will use their default value, and the values I have used are simply estimates, and will be tweaked based on trial and error when developing the training algorithm

#### Creating trainer pseudocode

Now all of these aspects can be passed into a trainer, which can then be used to train the model. A screenshot of a computer

Description automatically generated with medium confidence

This trainer can then be run using trainer.train() to run the training algorithm on the model

### Implement model

Now that the model has been trained, it should be used to transcribe the audio message that has been stored as a .wav file.

Firstly, the audio file should be loaded, and for this I will use the Sound File library that I used earlier to save the recorded audio files.

Then I will use the fine-tuned processor to process the speech, getting the input values the model requires to transcribe the audio. Then the input values will be run through the model, extracting the logits of the predicted values. As I want the most likely prediction, I take the argmax of the logits. Finally, I use the processor to decode the predicted logits, returning a transcription that should be in readable English and match the audio fed into the model. Text

Description automatically generated

### Allow users to report messages

There will be an option next to every message from the bot to report the message. When clicked, a popup will appear with a HTML form in. There will be a text box for the user to input their reason for reporting the message.

The popup will be created using a Bootstrap modal. This requires a trigger – a button that the user must click to open the modal. The modal part contains a header, main content, and a footer (which will contain a button to close the modal). The main content will contain the HTML input form, which will have a textbox, allowing the user to enter a report reason, and a submit button.

### Reported messages and the report reason should be logged

The program will then take in the reported message, and find the previous message sent from the user. The reported message, the user’s previous message and the report reason will be stored in a CSV file.

This allows for easier error reporting, as I would simply need the CSV file to see how the model made an error and could use that information to change the model’s training if needed.

## Testing

# Development