NANYANG TECHNOLOGICAL UNIVERSITY
SINGAPORE

CZ4079 Final Year Project

SCSE23-0848

Title: Empowering Communication For Mute Individuals Through AI-Assisted Conversations

FYP Final Report

19 October 2024

Supervisor: A/P Goh Wooi Boon

Done By: Au Yew Rong Roydon (U2021424J)

Contents

[2]: SpeechPathology. (2022, May 11). *Mutism: SLP graduate programs and the study of mutism*. https://www.speechpathologygraduateprograms.org/mutism/ 60

[3]: Ability Central, A. (2024, February 7). *What keeps someone from talking? information you should know about muteness*. https://abilitycentral.org/article/what-keeps-someone-talking-information-you-should-know-about-muteness#:~:text=Muteness%20is%20a%20rare%20condition%20that%20can%20be%20either%20temporary,permanent%2C%20but%20many%20are%20misunderstood. 60

[4]: Masrur Sobhan. (n.d.). (PDF) a communication aid system for deaf and mute using vibrotactile and visual feedback. https://www.researchgate.net/publication/336877461_A_Communication_Aid_System_for_Deaf_and_Mute_using_Vibrotactile_and_Visual_Feedback 60

[5]: Munir, M. B., Alam, F. R., Ishrak, S., Hussain, S., Shalahuddin, Md., Islam, M. N., Muhaimin Bin MunirMilitary Institute of Science and Technology, D., Fariha Raisa AlamMilitary Institute of Science and Technology, D., Shadman IshrakMilitary Institute of Science and Technology, D., Sonaila HussainMilitary Institute of Science and Technology, D., Md. ShalahuddinMilitary Institute of Science and Technology, D., & Muhammad Nazrul IslamMilitary Institute of Science and Technology, D. (2021, September 22). A machine learning based Sign Language Interpretation System for communication with deaf-mute people: Proceedings of the XXI international conference on human computer interaction. ACM Other conferences. https://dl.acm.org/doi/10.1145/3471391.3471422 60

[6]: *What are the different types of sign language?*. Sign Solutions. (2024, June 5). https://www.signsolutions.uk.com/what-are-the-different-types-of-sign-language/ 60

[7]: *Proloquo4Text*. AssistiveWare. (n.d.). https://www.assistiveware.com/products/proloquo4text 60

[8]: Limited, T. B. (2011, January 6). *Predictable*. App Store. 60

https://apps.apple.com/us/app/predictable/id404445007 60

[9]: *CHATGPT and disability: Benefits, concerns, and future potential*. Rocky Mountain ADA. (n.d.). https://rockymountainada.org/resources/research/chatgpt-and-disability-benefits-concerns-and-future-potential 60

[10]: Exploring ai chatbot for spontaneous word retrieval in aphasia. (n.d.). https://aphasia.talkbank.org/publications/2023/Purohit23.pdf 60

[11]: Farid, A. (2024, February 20). *Building apps with Expo & React native: Pros & cons*. Upstack Studio. https://upstackstudio.com/blog/expo-react-native/ 60

[12]: Agastya, A. (2024, January 4). *Decoding LLM performance: A guide to evaluating LLM applications*. Medium. https://amagastya.medium.com/decoding-llm-performance-a-guide-to-evaluating-llm-applications-e8d7939cafce 61

[13]: Google-Research. (n.d.). Google-Research/Bleurt: Bleurt is a metric for natural language generation based on transfer learning. GitHub. https://github.com/google-research/bleurt 61

[14]: Cleary, D. (2023, November 30). *Can you use llms as evaluators? an LLM evaluation framework*. Medium. https://medium.com/@dan_43009/can-you-use-llms-as-evaluators-an-llm-evaluation-framework-8681b400b110 61

[15]: *List of available metrics*. Ragas. (2024, October 3). https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/ 61

[16]: Anello, E. (2024, April 12). *How to improve rag performance: 5 key techniques with examples*. DataCamp. https://www.datacamp.com/tutorial/how-to-improve-rag-performance-5-key-techniques-with-examples 61

[17]: Lamers, R. (2023, August 9). *OpenAI's embedding model dethroned in MTEB by Baai General Embedding Model*. OpenAI's embedding model dethroned in MTEB by BAAI general embedding model. https://codingwithintelligence.com/p/openais-embedding-model-dethroned ... 61

[18]: Splore. (2023, November 3). *Different types of search explained: AI, keyword, Hybrid Search & More*. Medium. https://medium.com/@sploredotcom/different-types-of-search-explained-ai-keyword-hybrid-search-more-d5e24f74ef4d 61

[19]: Improve rag performance using cohere Rerank | AWS machine learning blog. (n.d.-b). https://aws.amazon.com/blogs/machine-learning/improve-rag-performance-using-cohere-rerank/ 61

[20]: Sharma, H. (n.d.). *Techniques to enhance retrieval augmented generation (RAG)*. Community.aws. https://community.aws/content/2gp2m3BJcl9mSMWT6njCIQNiz0e/techniques-to-enhance-retrieval-augmented-generation-rag?lang=en 61

[21]: Santhosh, S. (2024, February 4). *How to improve rag(retrieval augmented generation) performance*. Medium. https://medium.com/@sthanikamsanthosh1994/how-to-improve-rag-retrieval-augmented-generation-performance-2a42303117f8 61

[22]: *Getting started*. ElevenLabs. (n.d.). https://elevenlabs.io/docs/api-reference/getting-started. 61

[23]: Taivo Pungas. (2024, January 23). *Making GPT API responses faster*. https://www.taivo.ai/__making-gpt-api-responses-faster/#:~:text=There%20is%20a%20way%20to,can%20get%20faster%20answers%2C%20guaranteed. 61

[24]: gfbane23. (2023, November 9). *How can I improve response times from the openai API while generating responses based on our knowledge base?*. OpenAI Developer Forum. https://community.openai.com/t/how-can-i-improve-response-times-from-the-openai-api-while-generating-responses-based-on-our-knowledge-base/237169/5 62

Abstract

With the **rising proliferation** of **generative ai applications** it would be useful to explore its application to **enhance communication** for mute individuals. Being **mute** has a significant impact on one's daily life since it affects **social interactions**, **personal relationships** and **even professional opportunities** .

This project is focussed on delivering a **generative ai** mobile application that would serve as a companion **for mute individuals** to assist them in their conversations with a normal speaker. The application improves the **convenience** of mute users by **generating personalized responses** at real time for them to pick to reply **to a normal speaker.** Overtime, the app grows with the users by learning their **habits** and **past experiences** therefore improving the quality of **generated responses** to become more personal. Lastly, to enhance greater **flexibility** for the mute user, they are also able to edit the generated responses directly and **provide feedback** to the ai model through **liking** or **disliking** certain responses.

Since this project was built from scratch, this report highlights **all features implemented, experiments attempted, areas of improvements** and **next steps**. The final application has a fairly high **G-Eval score** of **0.714** (**High relevancy** / **personalized** responses), high context precision score of **0.9 (Low noise** of extracted contexts from the vector database) and fairly high context recall score of **0.76** (**High relevancy** of extracted contexts to the **ground truth**). **All scores are out of 1**. Refer to Section 5.2.3 for the explanation of the scores. Participants from the user studies also ranked the mobile application highly useful for mute users.

1. Introduction

Mutism is defined as speech output that is **minimal or always absent**[1]**.** The extent of mutism varies across different patients and can be classified into different categories[2].

(1) Neurogenic mutism: Usually a manifestation of extreme forms of speech disorders. It is a **lack of speech** due to **underlying damage to the brain**. It can be short or long term, static or progressive depending on the region of the brain affected and level of damage sustained.

(2) Selective mutism: Patients with this type of mutism often have the **ability to speak** but feel **unable to** due to **social anxiety.** However, in **certain situations** they can **speak normally**.

(3) Total mutism: Psychogenic mutism that causes a person to **not speak under any circumstance.**

Despite mutism/muteness being rare[3], it has a significant impact on one's daily life since it affects **social interactions**, **personal relationships** and even **professional opportunities**. Without being able to speak, it brings about great inconvenience. To address these challenges, technology plays a crucial role in **enhancing communication capabilities** for these mute individuals.

1.1 Project Overview

With the rising proliferation of generative ai applications, it has paved the way for innovative applications across various domains. Generative Ai has gained popularity in its ability to understand, learn and generate creative "human-like" text. By harnessing the capabilities of generative ai, it would be useful to explore its application to enhance communication for mute individuals.

The primary goal of this project would be to build a mobile application that would serve as a companion **for mute individuals** to assist them in their conversations with a normal speaker. The app does **real-time speech recognition** and uses the **context** of the **conversation** to generate pre-typed conversational sentences for the mute user to select in order to maintain a conversation. These generated **conversational sentences** are **personalized** to the user's past experiences. The app would further learn the **users' habits**, **past experiences** and grow with the user the more it is used. Furthermore, to enhance greater flexibility for the user, they are also able to **edit the generated responses directly** without having to retype and can provide feedback to the ai model through **liking** or **disliking** certain responses.

The goal of the app is to speed up communication and enhance convenience for mute individuals without compromising the quality of the conversation (making responses generated personalized).

1.2 Project Motivation and Target Audience

The project aims to target **mute individuals** who are incapable of any form of speech including impaired speech. These individuals usually communicate through sign languages, writing or texting[3]. They include individuals who can't speak due to underlying damage to the brain (Neurogenic Mutism)[1].

2. Related Works

With the rapid evolving landscape of technologies used to help the mute communicate, it would be useful to explore **existing solutions** and how to improve on them.

1. **Solutions that interpret visual and vibrotactile feedback**[4, 5] (EG: Sign language, gestures)

Currently, most solutions revolve around interpretation of visual and vibrotactile feedback to convert it into speech. These solutions are often in the form of assistive technologies that serves as interpretation tools. These tools can be further broken down into 3 categories:

a. Sensor-Based Technologies

The sensor-based assisted technologies mainly use some external devices like handheld or finger worn devices to detect various gestures. Examples include a magic ring that can translate a predefined gesture into information for communication or a micro-controller device with flex sensors to detect various finger movement.

Challenge: Besides the **accuracy** of the interpreted gestures, a big challenge would be the **inconvenience** of bringing along these devices for daily communication.

b. Vision-Based Technologies:

Most of these technologies use various image processing techniques to **detect sign languages or gestures** from **input images or videos**. Some examples include using the Microsoft Kinect camera, a full-duplex communication system where hand gesture is processed from input video and converted to text or speech and even a device called Digital Dactylology Convertor developed to process sign language from input images converting them to voice signal and text messages.

For example: (1) Technology that interpretates and produces sentences based on vibrations/impaired speech. (2) Technology that interprets sign language.

Challenge: With so many different types of sign language[6] (more than 300), there is immense complexity in **Sign Language interpretation**.

c. Smartphone Based Technologies:

The popularity and usability of smart devices like smart- phones and smartwatches have led to the development of many assisted technologies. A good example is a device that is used to help the deaf-mute communicate[4]. It receives voice input, produces images and vibrations for the deaf mute to view. The deaf mute then provide gesture or vibrotactile input to the phone which maps this **input into context** and **commands** before providing an output to the user.

Challenge: These solutions currently fall short in terms of **flexibility** (limited functionalities) and **texts generated often feel machine like**. With very limited number of commands and contexts, the solution is **quite restrictive** and would often **fail in bringing about a smooth human-like conversation**.

2. **Solutions that use predictive text:**

There are a few solutions that look towards predictive text that assists mute users in a faster response time. These solutions are viable since they provide greater convenience for the mute while being easily available on the app store. This includes apps such as Proloquo4Text[7] and predictable[8] which predicts what the user would tend to type based on typing habits.

Proloquo4Text predicts what mute users would say next based on what they type out so far. They can also create their own phrases for quick access.

Limitations: The app is still limited in terms of their prediction capabilities. The predicted text is often very general, **non-personalized** and hence can only be used for **common situations.**

Predictable is similar to Proloquo4Text. It uses smart word prediction technology to learn what you would type next making typing easier and more efficient. Similarly, you can store phrases for quick access,

Limitations: Similarly, predicted text feels **very general** for **common situations.**

Despite both applications speeding up and improving the convenience of communication, there are 2 issues that could be addressed:

a. **Making predicted responses** more **human-like and personal**. To ensure conversations are natural, there is a need to look for a way to incorporate past experiences into predicted responses.
b. **Improve personalization** at a greater convenience. The apps above allow some sort of personalization by allowing mute users to **store their own phrases.** But it would be a time-consuming process for them to build up these phrases. There is a need to look for a way to allow the app to grow with the users **without compromising convenience.**

With the **growth of generative ai**, it would be useful to observe how generative ai could play a role in further **enhancing** these solutions. As of now, **most generative ai solutions for the disabled** revolves more around those with **visual disabilities**[9]**.** There are limited solutions **for the mute** in which most solutions revolve around speech therapy.

However, one research that could be leveraged on is a paper on exploring AI Chatbots for Spontaneous **Word Retrieval** In Aphasia[10] (Speech disorder). People with Aphasia have difficulties in forming complete sentences. They often have trouble understanding, speaking reading or writing. With the use of ChatGPT, these words could be predicted / extracted based on the context, allowing those with Aphasia to better complete their sentences.

This research proves that **generative ai like ChatGPT** could be used to **properly identify context** and the use case **could be further extended towards generating complete sentences.**

With **creative human-like generative capabilities,** it would be useful to explore how gen ai could be used to enhance the **above predictive applications** by **predicting** how the **mute** would **reply** so as to provide a **convenient** way for them to converse with a normal person. Combining it with a smart phone solution, we can learn from the examples above to produce a robust and convenient solution for **mute users.**

3. Proposed Solution

3.1 Overview of Proposed Solution

The solution would be an application that the mute user could use to conveniently communicate with a normal person. Generated responses are human-like and personal. Refer to video attached for a demonstration. The flow is as follows:

(1) When the mute enters the application, they can see **all their stored conversations**. These are conversations that **they had with others** using the application. (**Figure 1A**)
(2) They can then choose to **create a new** conversation (with a new person) or to click on one of the **existing chats**. (**Figure 1B**)
(3) To start the conversation, the **microphone button is clicked** and the normal person speaks into the **phone application**.
(4) The speech is converted to text to be passed into the **Retrieval-Augmented Generation (RAG) System**.
(5) **Previous conversations** made by the mute user are stored in **the vector store**. Based on the query, **similar conversations are extracted** to become context/information for the response generation. This allows the **response generated** to be more **personalized** based on the mute user's **personality** and **past experiences**.
(6) The **instruction**, **top 3 most relevant conversations** extracted from RAG, and the **recent 5 replies** are **passed as context** to the ChatGPT model. The model would then **generate 3 responses** for the **mute user** to **pick,** replying to the normal person. (**Figure 1C**)
(7) If the mute user is **unsatisfied** with the **generated responses**, they have a few options. Firstly, **they can refresh and regenerate 3 new options** by clicking on the refresh button.
(8) Alternatively, they can click on the **thumbs up** or **thumbs down** button for the responses generated. By clicking on the **thumbs up**, a **high score** for that response is stored in the system so that it would **recognise that it is a good response**. By clicking on the thumbs down, the response would be **regenerated**, providing the user **with a new response** based on the scores of the **existing responses**. Hence, the good responses would influence the **regenerated response** to be **similar** to it. (**Figure 1D**)
(9) Lastly, if the generated responses are **similar to** (not exact) what the user **would like to say,** they have the option to **edit the generated responses directly**. This provides **greater convenience** since the **user can now remove or add in information directly** to the response **transforming it** into the **exact message** they would like to **reply with**. (**Figure 1E,F**)
(10) The mute user could then either **pick one of these responses** or **type in their own response**. This response would then be stored into the vector store. (**Figure 1F)**
(11) The selected response would then be converted to speech and automatically be played out to the normal person. This makes the conversation **more natural.**

The application is meant to grow with the user. As the mute user continues to use it, the application would understand the **users' past experiences**, **habits**, and **personalities** in various **situations**.

The Json file is used to store the recent 5 replies so that ChatGPT is aware of the topic at hand.

The vector store is used to store all conversations so that the response generated would be more personalized towards the mute users in various contexts.

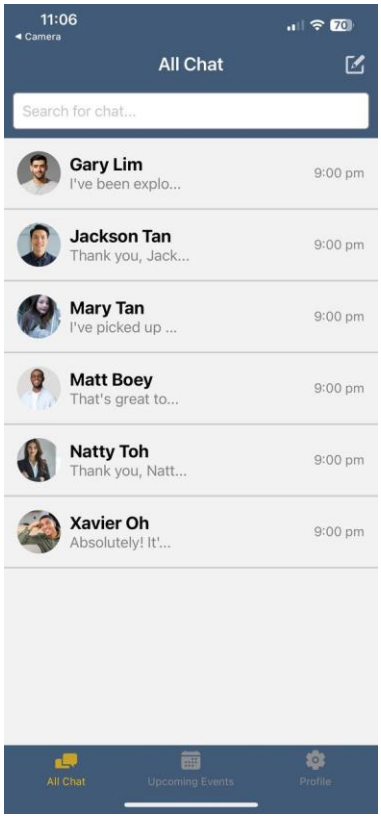


Fig 1A: All Chats

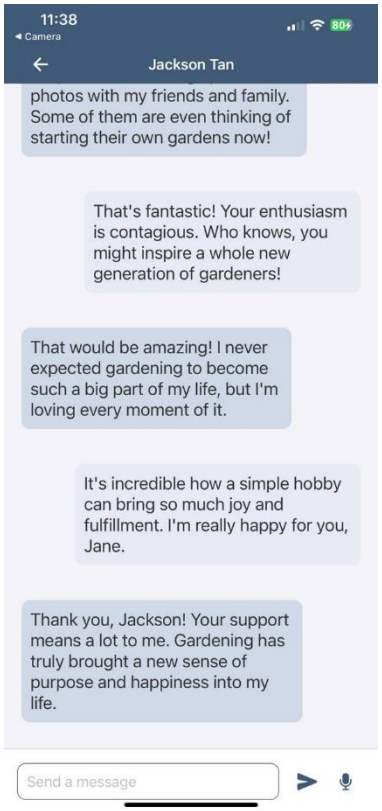


Fig 1B: Conversation History

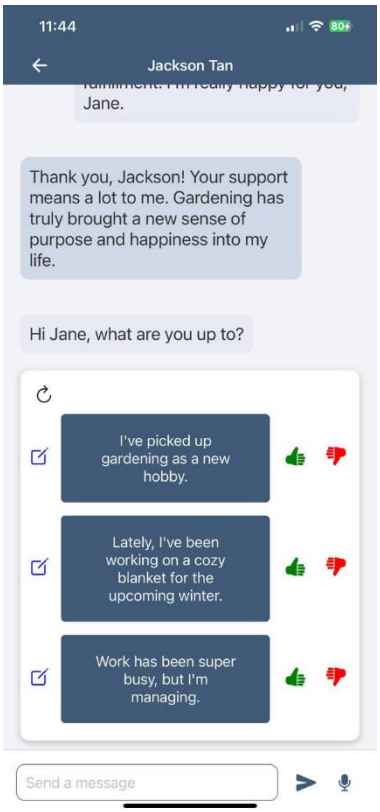


Fig 1C: Responses Generated

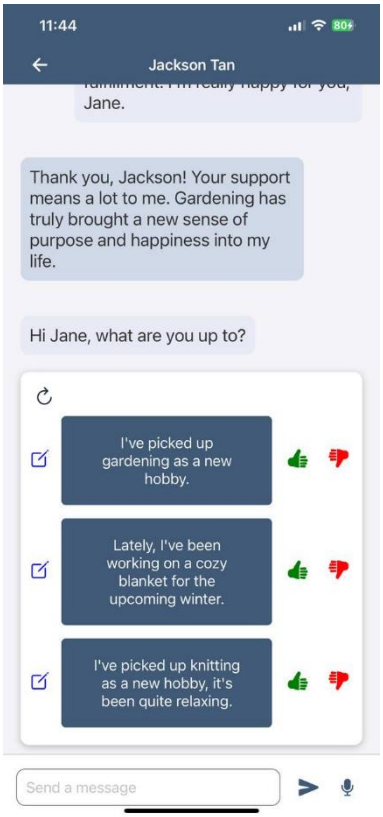


Fig 1D: After Thumbs Down On 3rd Response

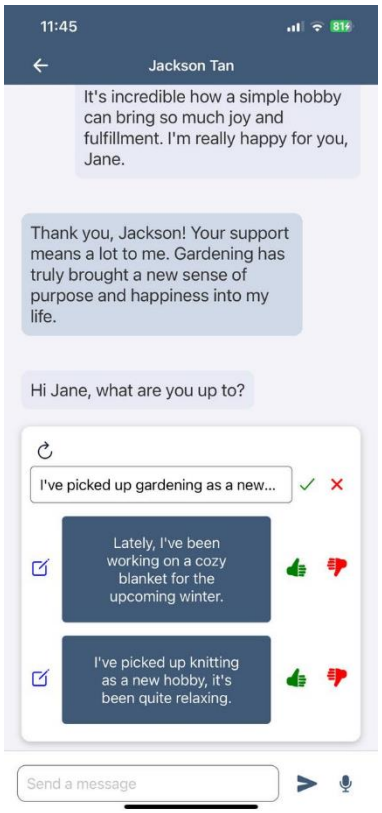


Fig 1E: Editing 1st Response

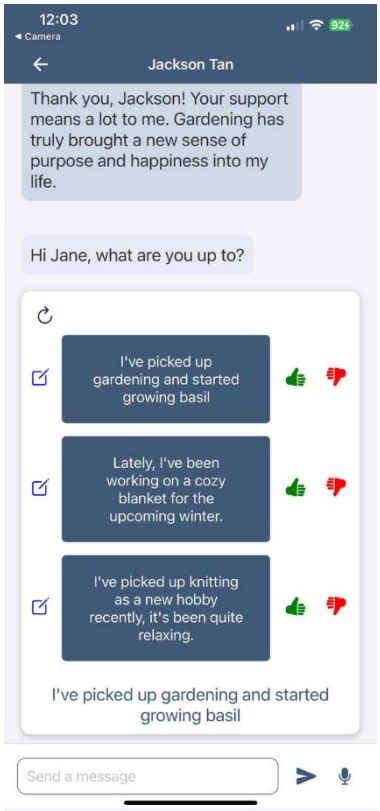


Fig 1F: Edited 1st Response And Picked It

3.2 Technical Overview of Response Generation

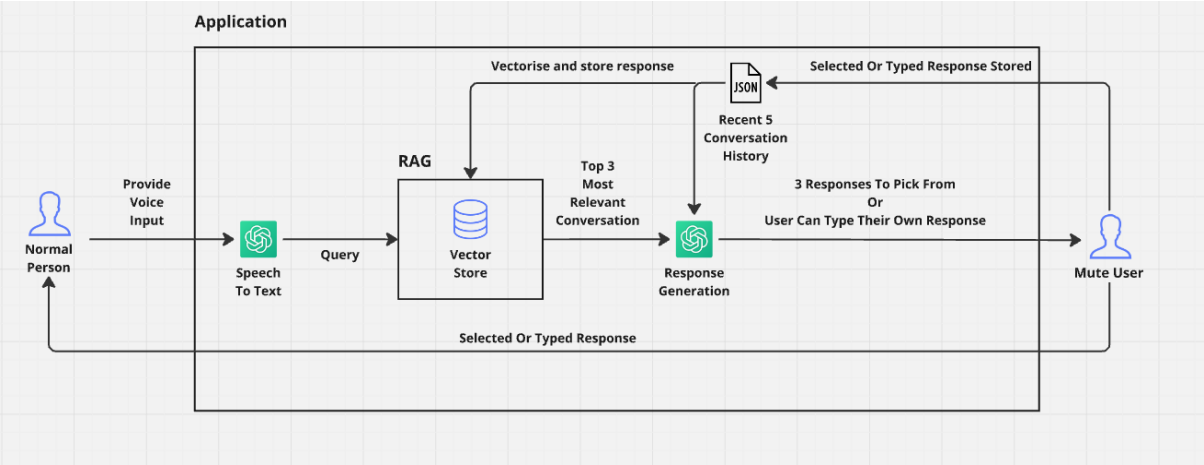


Fig 3A: Final Architecture Diagram

The above diagram shows how the 3 responses are generated. In the report below, more will be shared on the experiments done but after many experiments, this is the final architecture/set up for response generation. Refer to Appendix 5.7A for code on **context retrieval**.

There are a few core parts to the application and the flow is as follows:

(1) OpenAI's audio transcriptions api converts **normal person's speech** into **text**.
(2) Text is passed into **process_query** function to break it down in the case that the **text** is **complicated** and **contains multiple questions** joined together. (Section 5.6)
(3) Each broken down component is passed into getTopic function to obtain topic of conversation (Section 5.5.1)
(4) Each broken down component is also passed into filter_list function which will conduct both the **hybrid search** (Section 5.4.1, Section 5.4.3) and the **meta data filtering** of topics (Section 5.5.1) to obtain **the top 3 most relevant conversations** to be used as context **to generate personalized responses**.
(5) The **instruction**, **top 3 most relevant conversations** extracted from RAG vector store, and the **recent 5 replies** are **passed as context** to the ChatGPT model. The model would then **generate 3 responses** for the **mute user** to **pick**. (Section 5.3)
(6) Picked or typed out responses are then stored together with the conversation into the **Json file** and the **vector store.** These can **later be used for contexts**. Hence the more the **user uses the app**, the more the app **grows to understand the user's interests, habits** and **personalities** through his/her **past experiences**.
(7) The picked response would then be passed through **text_to_speech api** which would automatically play the audio out for the normal person to listen to. (Section 3.3.4)

Important Components Include:

RAG (Vector Store): Contains vectorised **previous conversations** made by the **mute user**. Essential in ensuring responses generated **are non-general** and contain **past experiences** that the mute user has.

Search Algorithm To Retrieve Top 3 Most Relevant Conversation: More will be shared on the experiments done in Section 5. However, the final algorithm is a combination of **meta data topic filtering** and **hybrid search.**

Json File: A Json file is also used to store the **most recent 5 responses** generated by ChatGPT so that it is aware of the topic at hand.

ChatGPT: Lastly, ChatGPT has been prompt engineered in a way such that its output would be **3 responses in a single string.** This would later be **processed and passed as 3 different options for the mute user to pick from**.

3.3 Tools / Software Used

3.3.1 Frontend

The frontend was implemented using **React Native** and **Expo** for mobile development. React native allows cross mobile development hence allowing both iPhone and android development.

Expo allows testing directly on smartphone devices, providing greater ease for rapid development[11].

3.3.2 Backend

The backend was implemented using **fastapi** and **uvicorn**. Fastapi was used due to its **ease in testing** of **apis** as well as its **high performance**.

Below are the apis used and their functionalities:

API	Purpose
health	Checking the health status of the server
post-audio	Transcribes audio into text.
post-audio-response	Transcribes audio into text. **Text** is passed into ChatGPT together with **instruction**, **top 3 most relevant conversations** extracted from RAG, and the **recent 5 replies / conversations** to generate a single string containing 3 responses for the mute user. (EG: "Response 1: … Response 2: … Response 3: …")
text-to-speech	Converts the chosen text into an mp3 audio file to be stored in the server
store-response	Stores the response selected and current conversation into vector store.
reset	Reset the history stored. (Deletes the recent 5 conversations from the Json file)
regenerate-responses	Refreshes and regenerates 3 new responses. Also clears the history deleting the 5 recent conversations from the Json file.
rank-responses	Regenerates a single bad response into a new response based on the scores of existing responses.

3.3.3 ChatGPT

To improve the quality of **responses** generated by ChatGPT, different **prompt engineering techniques were employed**.

- **Specifying of persona:** Instructions on the persona to be taken on by ChatGPT was given to better allow ChatGPT to **understand the task at hand**.
- **Giving of examples:** Providing of examples to improve the understanding of ChatGPT
- **Adding of personality and personalization:** To better generate natural responses that would accurately reflect how the person would reply, past conversations are provided to ChatGPT through RAG. This would allow understanding of the mute person's past history and type of replies in various contexts.
- **Answer formatting:** Specifying the output to be expected from ChatGPT in a certain format. In this case, 3 responses generated should be in 1 string. You might be wondering, why not just call ChatGPT 3 times to get 3 responses. That is because the 3 responses generated would be **exactly the same** or very similar. Without changing the **inputs,** the **outputs** from ChatGPT would be **the same** or very similar. Hence to bypass this and be able to provide 3 **different responses** for the users, the instruction for **generating 3 responses** into **1 string** must be made known to ChatGPT for **easier post processing (**Splitting of the string into the 3 responses**) through standardisation** of output format from ChatGPT. (Refer to Section 5.3 for prompt used)

3.3.4 Text To Speech Using Eleven Labs API

In order to convert the chosen reply into speech, I decided to use Eleven Labs API which could easily convert any text into a wide array of voices[22]. After converting the text into an audio file, it would be automatically played for the normal person to listen to. **This makes the conversation more natural** as if 2 people are speaking to one another.

To automatically play the file, **since expo is unable to access locally stored files in the drive**, I had to **mount** the **audio files** onto the server. Once the reply has been selected, **eleven labs api converts it into an audio file** and stores it in the **mounted location** at the **server**. To save space, the **new reply chosen** would replace the old file, hence there would only be **one response audio file** at a time in the **server**. Refer to Appendix 3A for the code.

3.3.5 Other Tools

- **Github:** Github and git was used for version control allowing me to keep track of different changes I made to the project ensuring a systematic way to backtrack if needed.

4. Software System Architecture

Before showing all experiments conducted, it would be useful to show what was explored in terms of the system architecture, before finally arriving at the architecture above.

4.1 Version 1 Architecture:

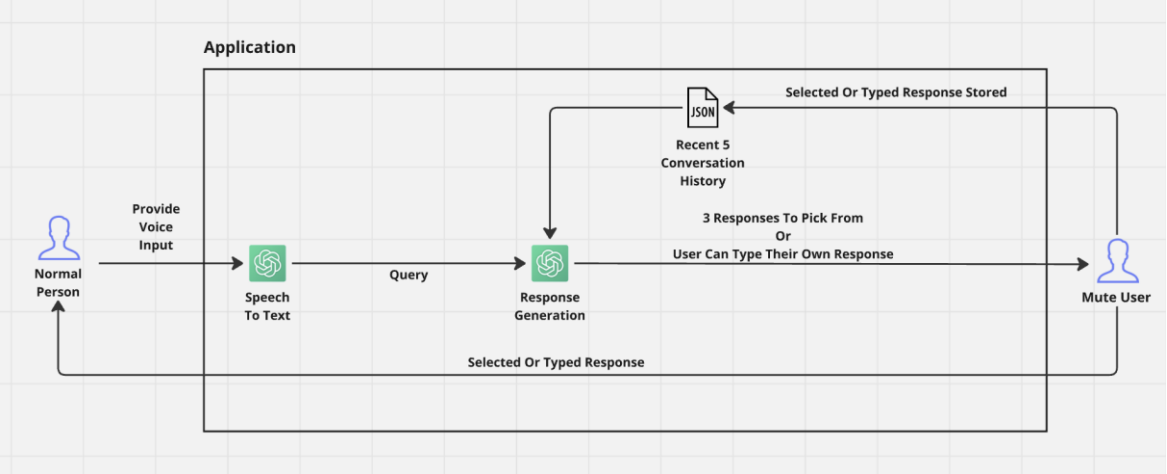


Fig 4A: Version 1 Architecture

- Conversation history provided for Chatgpt to know topic at hand.
- Limited personalization, responses are general matching the current topic at hand. A little unnatural.

4.2 Version 2 Architecture (Current Architecture):

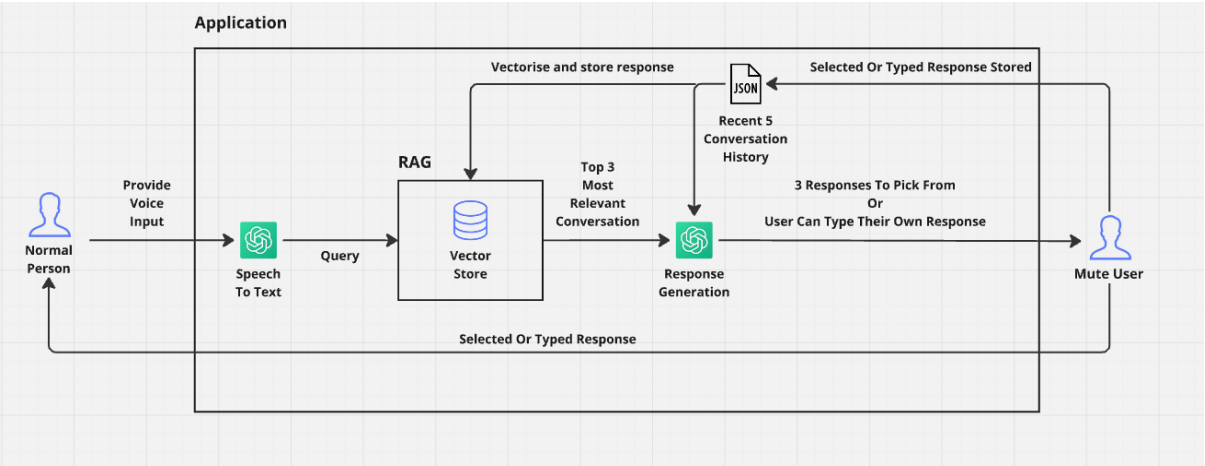


Fig 4B: Version 2 Architecture

- Added RAG to improve generation of personalized responses
- Responses now has information on the mute user's conversations and past experiences. For example, when asking what the mute user is up to, recent activities can be extracted based on RAG.
- Responses are more **natural** and **tailored directly** to the **mute user**. Different mute users with **different past experiences** will experience **different replies** generated for the **same conversation** question. (EG: What are you up to?)

4.3 Version 3 Architecture:

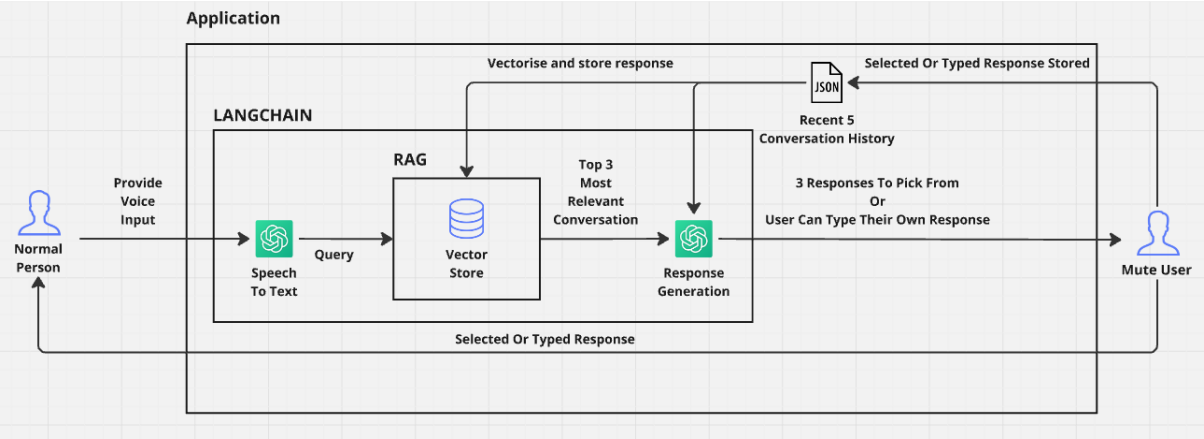


Fig 4C: Version 3 Architecture

- Using of LANGCHAIN allows easy switching of different ai models and RAG vector stores.
- Successfully generates when **no context** is passed to it.
- However, currently unusable due to insufficient token limit (Currently limited to only 1000 tokens) per query. Hence, context from the RAG is unable to be passed into it since it exceeds the limit.

Therefore, I decided to stick with **version's 2 architecture**. That being said, for greater flexibility in the future, **LANGCHAIN could be further explored**.

5. Evaluation & Experiments Completed

In this section, it will contain all experiments conducted leading to the **final architecture**. I will first introduce how to **evaluate responses generated**.

Standard NLP evaluation methods exist but often fall short when evaluating LLMs[12]. Traditional metrics such as BLEU and ROUGE scores are unable to fully capture the nuanced understanding and generation of human language, BLEU for example evaluates the quality of translated text by comparing it with a set of high-quality reference translations. However, its focus is predominantly on the precision of word matches, often overlooking context and semantics. Tone and style are equally important. Hence below are some methods used to evaluate the LLMs.

5.1 Evaluation Methods:

There are numerous new methods that can be used however, due to the nature of the task of generating 3 responses in one string, standard methods may not provide an accurate evaluation. For example, lang-chain has an inbuilt evaluation method with an auto-evaluator with the ability to evaluate a response to a query based on its inbuilt knowledge. However, since our task would generate 3 responses into a single string per query, these responses would have to be split and passed separately when using the auto-evaluator.

Therefore, the below evaluation scores are chosen based on **ease of usage** and **high applicability** to our custom task.

5.1.1 BLEURT Score

This is a learned metric on BERT that has been trained on human judgments. It predicts the quality of text generation more accurately by considering **context** and **semantics**[13]**.**

5.1.2 Deep Eval Score (Specifically G-Eval)

This score can integrate easily into CI/CD pipelines. It provides a diverse evaluation metric allowing for testing of **hallucinations**, **answer relevance and bias.** An AI dashboard could also be used to view all test results.

Deep Eval has a wide package, but I would be focussing on using G-eval to conduct analysis on evaluating the outputs of the model. This score would be more accurate than BLEURT since[14]:

- It uses an llm to evaluate the result where you can specify your **own custom criteria** and evaluation steps the llm should take when comparing the actual output and your ground truth / expected outputs.
- It further provides reasons as to how it evaluates the responses.

5.1.3 RAGAS Score

Lastly, to further evaluate RAG systems, RAGAS score can be used. It is a leading library for evaluating RAG pipelines. It evaluates two different aspects of the RAG system, **the generation component** (How well the LLM is answering the query) and retrieval (How relevant the content retrieved is to the query).

5.2 Experiment 1: Exploring the Addition of A RAG System:

To test the effectiveness of having a RAG system, I explored the effects of it on the responses generated.

To prepare the RAG System I first had to build our own **vector datastore of (Mock data).** Mock data of past conversations were created. These Json conversations are vectorized and stored **locally** as vectors in a vector store. (Figure 5.2A)

Example of a mock conversation: (For more information refer to Appendix 5.2A or mock data under assets in the code.)

Fig 5.2A: Snippet of Conversation in Mock Data

Retrieval process for RAG: Currently, the retrieval process is based on **vector-based similarity comparisons** between the **query** and the **conversations stored**. Depending on the query, the **top 3** most relevant conversations are extracted and passed as context to the model. One conversation consists of a back and forth reply between the mute user and the normal person. (EG: “Response 1” in Figure 5.2A).

5.2.1 Results for No RAG

This is the current **prompt used** for the generation of responses:

You are an assistant whom will facilitate the conversation between a mute and a normal person. The mute person's name is Roydon and the normal person is indicated as other person. You should be generating 3 responses which the mute person could choose from and the responses generated should follow the context of the conversation. The topic should be interpreted from the conversation. If no topic is interpreted, provide default responses that a person would start with such as greetings. The responses should be what a person would say and should not include actions in a third person view. Your persona would be from the perspective of the mute person.

In the case the responses are not chosen, the mute person could type their own response. Do take note of this response and continue the conversation from the response selected or typed out by the mute person.

Ensure the responses generated will allow the conversation to flow smoothly. It must be in English. The 3 generated response would be in the format of 1 single string "Response 1: what you generated Response 2: what you generated Response 3: what you generated" All in one line.

Result: (Figure 5.2B, 5.2C, 5.2D)

It is observed that context and history is remembered thus having the **Json file of the recent history was useful**. History of working location Harbourfront was interpreted by the model (Figure 5.2D), but **responses generated are very general or irrelevant to the mute user.** For example, in (Figure 5.3C), mute user may not be listening to Jchou. Responses generated **may not be useful** for the mute user.

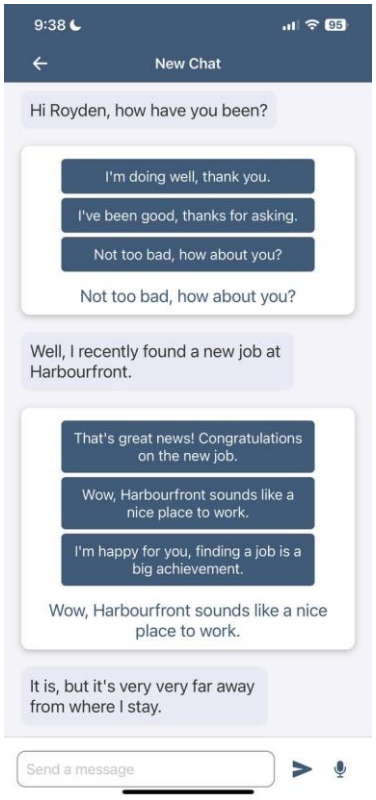


Fig 5.2B: General Responses

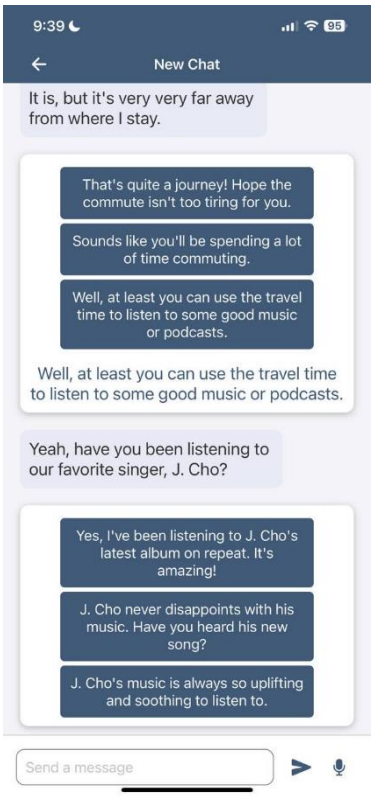


Fig 5.2C: Irrelevant Responses

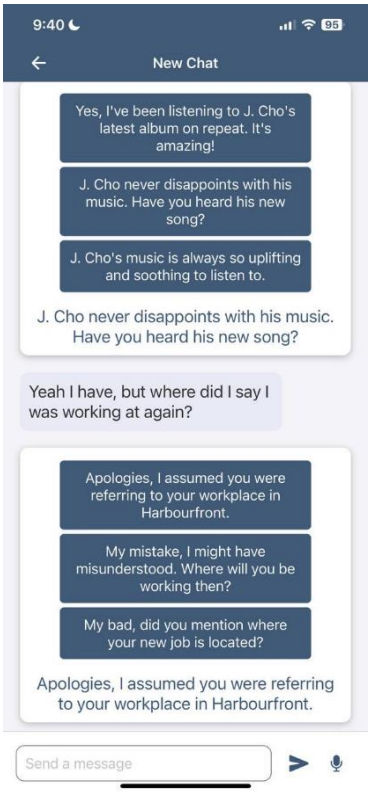


Fig 5.2D: Remembers History

5.2.2 Results With RAG

Next lets observe results for RAG. The RAG system would first extract conversations from the vector database to be **contexts for the GPT Model.**

When the normal person asks “What are you up to?”, context extracted include mostly conversations about Arsenal, having a new pet dog and having a bad trip. The 3 conversations extracted are (Appendix 5.2B).

{ "Roydon": "Hey there! Did you catch the Arsenal game last night? What a thrilling match!",

"John": "Hey Roydon! Yes, I watched it. Arsenal played really well, didn't they?" }

{ "Roydon": "Guess what, I just got a new pet dog!",

"Jacob": "That's awesome! What breed is it?" }

{ "Roydon": "Not really. I was so angry the whole time, I couldn't enjoy anything.",

"Dory": "I'm sorry to hear that, Roydon. Maybe you can plan a better trip next time." }

It can be observed **that Generated Responses** are personal based on **mute user’s past experiences** making the responses **more applicable** for the mute user to choose from. (Figure 5.2E)

Fig 5.2E: Relevant Responses

5.2.3 Using Bleurt Score, G-Eval and RAGAS to compare between having a RAG system and without:

To further prove that having a RAG system improved the responses generated, I would use the above evaluation methods to evaluate the responses generated. For evaluation to be done, preparation must be done in advance. **A dataset** would be created containing the following columns: (Refer to Appendix 5.2B2)

‘question’: Holds a list of queries to be sent into the GPT model for generation. In our case, it would be what the **normal person** says in a conversation (EG: How have you been.?).

‘answer’: Holds a list of generated outputs from the GPT model. In our case, it would be the 3 responses in **one string** for the mute user to choose from.

‘contexts’: Holds a list of contexts extracted from RAG for each query. In our case, it would be conversations from the vector store. Blank strings if there is no RAG.

‘ground_truth’: Holds a list of expected outputs. In our case, it would be the response we want the GPT model to generate.

The following columns: **‘question’** and **‘ground_truth’** are **set by me**. The column **‘answer’** is **generated by GPT** and **‘contexts’ is extracted** from **the vector store based on the vector-based similarity algorithm.** This **dataset** would later be passed into the **evaluation methods above** to get evaluation scores**.**

To ensure **sufficient coverage** the mocked questions (conversations) set by me covers all the **important topics or hobbies** that the mute person has based on the mock data **in the vector datastore such as** having a new pet or being an arsenal fan etc. Therefore, this would test the machines ability in **generating responses** that ties closely to the mute person’s **past experiences** for greater applicability.

Complex questions / conversations are also included such as having 2 questions within one query. (EG: Hey Roydon, how have you been? What are you up to).

Bleurt Score:

The Bleurt scores were not great for both systems with and without RAG (Appendix 5.2C, Appendix 5.2D). For Bleurt, the scores that are closer to 1 would mean a good, generated response since it closely matches that of the expected output. In our case, **almost all of the scores** were **negative.**

Despite the poor scores, having a RAG system still returns better scores (**less negative**) than without.

That being said, the scores could be poor only because of **the difficulty in evaluating our task**. Since our generated response **contains 3 responses within a string**, it is difficult **to evaluate it since getting one response correct should give a fairly high score**. (As a user I would only need to pick one response since that is the one I would reply with). However, bleurt would need all 3 responses to match since it compares the entire string with the expected output.

Deep-Eval Score (G-eval):

To complement Bleurt, we would use G-eval under the deep-eval package. This is because G-val allows us to set our own criterias for evaluation. We can do so by creating our own metric:

We specify the following evaluation steps for G-eval (Appendix 5.2E):

Step 1: "Check whether the main content of the responses generated in 'actual output' are similar to the responses in the 'expected output'",

Step 2: "As long as one of the main content of the responses generated is similar to any of the expected output, the test case is considered correct. For example, if response 1 content is on a pet dog and it matches response 3 content of also a pet dog, give it a high score. The order of the responses is not important.",

Step 3: "Evaluate mainly based on main content but do still give a higher score depending on similarity of responses."

It is important to **recognise that the setting of the criteria** affects the evaluation heavily. Previously, the evaluation steps used were not as accurate (Appendix 5.2F, Appendix 5.2G) and hence the criteria has been updated to the one above. An example was added to the criteria to **better assist in the evaluation.**

Another great thing about G-Eval is that it provides reasons for each evaluation. The reasons provided **gives a good indication** on how **accurate the evaluation steps are**. An example would be below (Figure 5.2F, 5.2G, 5.2H)

question	answer
Woah really how is Arsenal doing right now then?	Response 1: They are currently showing great potential this season, hopefully they can maintain this form. Response 2: Arsenal is doing well at the moment, let's hope they continue to climb up the table. Response 3: Arsenal is picking up momentum and performing better with each game.

Fig 5.2F: Question and GPT Generated Answer

ground_truth	G-Eval Scores
Response 1: Arsenal is doing well, did you catch the match yesterday? Response 2: Arsenal is doing great and Aubameyang is a true asset to the team. Response 3: Arsenal is doing alright since Ben White is a great addition to the team.	0.965023

Fig 5.2G: Ground Truth and G-Eval Score

Reasons
Responses are similar to the expected output in terms of discussing Arsenal's current performance and potential.

Fig 5.2H: Reason for evaluation

Using the above criteria, we can then **pass the dataset** into G-eval for evaluation. It only requires the **expected output (ground_truth)** and **actual output** generated by **the GPT model.**

Results (Refer to Appendix 5.2H for top 10 scores):

Similarly, **a score closer to 1** represents a better score for G-eval. The scores can be seen to be much better compared to **Bleurt scores** since the evaluation can be **customized to better fit the task**. Similarly, having **a RAG system yields better scores than without.** (Figure 5.2I)

```
Average Score for Non-RAG DataFrame: 0.3438411719893059
Average Score for RAG DataFrame: 0.47205539395967344
```

Fig 5.2I: G-Eval Scores For No RAG vs RAG

RAGAS Score:

The above 2 evaluation methods only evaluate the **responses generated by ChatGPT** and not context extracted from the vector datastore.

RAGAS provides many metrics that can be used to evaluate both the generation capabilities of the model and the context extraction capabilities of the RAG system[15].

The below metrics were selected and used:

'answer relevancy': How relevant is the generated answer to the question.

'answer correctness': How similar the generated answer is with the ground_truth

'context precision': Whether all of the ground-truth items present in contexts are ranked highly or not. It represents the signal to noise ratio of retrieved context.

'context recall': How relevant the context retrieved is to the ground truths.

Results: (Refer to Appendix 5.2I for top 10 scores):

It still consistently shows that having a **RAG system yields better scores. (Figure 5.2J)**

For the current retrieval of contexts, recall can be further improved but precision seems to be good suggesting that the ranking of contexts is well done but more improvement can be done to match the contexts to the ground truth.

Answer correctness and relevancy seems similar to the score of G-Eval.

```
Non-RAG Average Answer Relevancy: 0.09371908683966357
Non-RAG Average Answer Correctness: 0.28901515842812675
Non-RAG Average Context Precision: 0.0
Non-RAG Average Context Recall: 0.0
RAG Average Answer Relevancy: 0.4601393128812699
RAG Average Answer Correctness: 0.41325875752084756
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.5
```

Fig 5.2J: RAGAS Scores For No RAG vs RAG

Evaluation decisions:

Since G-eval and RAGAS can **better fit the task in evaluating relevancy** of the generated responses, for the next few experiments it is decided that we will use the following for evaluation:

- **Deep-eval (G-eval) & RAGAS:** To measure the **relevancy** of the **generated responses** to the ground truth.
- **RAGAS:** To measure the **relevancy** of the **context** extracted from RAG to the ground truth.

BLEURT Score has been **removed** for upcoming experiments since it is unable to accurately evaluate the task at hand.

5.3 Experiment 2: Prompt Engineering:

It can be observed that the current scores for correctness are still very low with both G-eval and RAGAS giving scores of less than 0.5. (Generated output does not match well with ground truth).

For further improvements it could be useful to revisit how the prompt has been constructed. Two changes have been made to improve **both latency** and **relatedness** of the **generated responses** to the ground truth.

Firstly, the prompt was updated to include examples. This leverages on the technique known as **few-shot prompting** which uses examples for ChatGPT to better understand **the task at hand.** Only a few examples are needed for this to be effective. (In our case, I only used one example to test if one example would be sufficient so as to **limit the number of tokens used**).

Secondly, the length of the prompt has also been reduced to improve latency in the generation **of the 3 responses**.

New prompt:

You are an assistant whom will faciliate the conversation between a mute and a normal person. The mute persons name is Roydon and the normal person is indicated as other person. You should be generating 3 responses which the mute person could choose from and the responses generated should follow the context of the conversation. The responses should be what a person would say and should not include actions in a third person view. Your persona would be from the perspective of the mute person. Snippets of conversation would be given below in the section of Context. Use the conversations to assist in the generation the 3 responses. Primarily the topic should be inferred from the question asked but if no topic can be inferred, infer the topics from the conversations given in the context. The conversations are seperated by "{{" and "}} Context: {contexts}
For example, if the context above contains "{{"Roydon": "Recently my new pet dog has been so fun!", "Jacob": "That\'s awesome! What breed is it?"}}"

If the user asks "What have you been up to?"

An example of the 3 generated response would be in the format of 1 single string "Response 1: I have been playing with my new pet dog. Response 2: Nothing much, I recently brought my new pet dog to a park. Response 3: Its been tiring lately after getting a new pet dog.

By including an example of what is to be expected, the responses generated are better able to utilize the contexts extracted from RAG.

Results: (Refer to Appendix 5.3A and Appendix 5.3B for G-Eval and RAGAS top 10 scores):

Analyzing the G-Eval scores, it can be observed that the prompt engineered RAG system performs the best. (Figure 5.3A)

```
Average Score for Non-RAG DataFrame: 0.3438411719893059
Average Score for RAG DataFrame: 0.47205539395967344
Average Score for RAG Prompt Engineered DataFrame: 0.554508111381623
```

Fig 5.3A: G-Eval Scores Comparisons

Similarly, RAGAS Scores also improved for the prompt engineered RAG system. Correctness scores have improved tremendously. (Figure 5.3B)

```
Non-RAG Average Answer Relevancy: 0.09371908683966357
Non-RAG Average Answer Correctness: 0.28901515842812675
Non-RAG Average Context Precision: 0.0
Non-RAG Average Context Recall: 0.0
RAG Average Answer Relevancy: 0.4601393128812699
RAG Average Answer Correctness: 0.41325875752084756
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.5
RAG Prompt Engineered Average Answer Relevancy: 0.45274102089224694
RAG Prompt Engineered Average Answer Correctness: 0.6452190922203027
RAG Prompt Engineered Average Context Precision: 0.89999999991
RAG Prompt Engineered verage Context Recall: 0.5166666666666666
```

Fig 5.3B: RAGAS Scores Comparisons

Despite the **score improvements**, it can still be observed that the **context recall** score is still very low. Hence, more could be explored on **improving the RAG system** and how the contexts are extracted. This would further improve the generation of response. For example, for the question: What have you been up to Roydon, the 3 context extracted are: (Figure 5.3C)

```
['{"Roydon": "Hey there! Did you catch the Arsenal game last night? What a thrilling match!", "John": "Hey Roydon! Yes, I watched it. Arsenal played really well, didn\'t they?"}
{"Roydon": "You too, Xavier."}
{"Roydon": "I can\'t believe what happened to me in Thailand.", "Xavier": "What happened?"}']
```

Fig 5.3C: Conversations extracted from vector datastore

The 2nd context does not add much value and instead could skew the generated responses. Hence, more could be explored in improving the extraction of context.

5.4 Experiment 3: Improving RAG Searches:

It'll first be good to understand what affects RAG retrieval. RAG retrieval is mainly affected by **3 areas**, **the chunk size**, **how the information is embedded** and lastly, the **algorithm** used to retrieve this information**. Chunk size** refers to the size of each document and how these documents overlap when stored in the vector datastore. **Larger chunk size** would mean **more information** is contained within each document. In this case, I would not explore changing of chunk size as I would like to **keep it small**. Currently each document is mapped to a single conversation which is small and short. This would further **limit the token count** used when generating responses from ChatGPT hence reducing cost and latency[16].

Next, an area worth exploring would be **how embeddings would affect the retrieval process**. Different embeddings of text **results in different information extracted due to embedding dimension**. With the rise of **open-source embeddings** such as bge by BAAI general embedding model, it would be interesting to explore its comparison to OpenAI's embedding model Ada. According to recent studies, it was shown that **OpenAI's embedding model was dethroned in Massive Text Embedding Benchmark (MTEB) by BAAI** general embedding model[17] .

Lastly, the algorithm used to **retrieve the information** would greatly affect the **information extracted from the RAG datastore**. The most common retrieval of information would be through **semantic search** which matches the similarities between vectors. It is a **search method** that aims to improve the accuracy and relevance of search results by understanding the context and meaning behind a search query.

However, it could be useful to **explore hybrid search** which combines the benefits of both **semantic and keyword-based search**. Another popular method would be **reranking**. These 2 methods have been recently researched on and have been proven to be effective in improving most RAG systems.

5.4.1 Improving RAG Searches Through BGE Embedding And Hybrid Search:

We'll first be exploring Hybrid Search. It is a technique that combines multiple search algorithms to improve the accuracy and relevance of search results by combining dense and sparse vectors in its model[18]. Sparse vectors are used in keyword searches.

Hybrid search **offers** the following (1) Exact Keyword matching, (2) Semantic Similarity Search, (3) Complex-long tail query ability, (4) Multi-modal search, (5) Multi-language search and (6) Personalized search thus having **both the benefits** of **keyword** and **semantic** search. (Appendix 5.4A)

For hybrid search, I would be using the BM25 algorithm for **keyword search**. This algorithm generates a sparse vector. BM25 (Best Match 25) is an information retrieval algorithm used to rank and score the **relevance of documents** to a particular search query. The semantic search algorithm and BM25 algorithm are combined together using an EnsembleRetriever. Both are first given equal weightage of 0.5. Refer to Appendix 5.4B for the code.

To further improve searches, I also tried changing the **text embeddings** from **text-ada-2 embeddings** to **bge-base embeddings**. This required a new vector store where the documents are embedded differently. Refer to Appendix 5.4C for the code.

Results:

The results for G-Eval is shown below in Figure 5.4A.

Few Shot **(Best system from before)**: **Open-Ai Embedding with few shot prompting**, **vector**/semantic **search**.

Ensemble Open Ai: **Open-Ai Embedding with few shot prompting, hybrid search**

Ensemble HF: **bge Embedding from hugging face with few shot prompting, hybrid search**

```
Average Score for Few Shot: 0.5545081113816229
Average Score for Ensemble Open Ai: 0.602516814829001
Average Score for Ensemble HF: 0.6355191874880344
```

Fig 5.4A: G-Eval Scores

Similarly for RAGAS:

```
=========================Few Shot=========================
Non-RAG Average Answer Relevancy: 0.452734835146927
Non-RAG Average Answer Correctness: 0.6440070072030675
Non-RAG Average Context Precision: 0.89999999991
Non-RAG Average Context Recall: 0.475
=========================Ensemble Open AI=========================
RAG Average Answer Relevancy: 0.468133850976659
RAG Average Answer Correctness: 0.4664450979906115
RAG Average Context Precision: 0.69999999993
RAG Average Context Recall: 0.4
=========================Ensemble HF=========================
RAG Average Answer Relevancy: 0.3743571687636189
RAG Average Answer Correctness: 0.5205250062605699
RAG Average Context Precision: 0.69999999993
RAG Average Context Recall: 0.5
```

Fig 5.4B: RAGAS Scores

It can be observed that for an **Ensemble Retriever (hybrid search)** with bge embedding, the **G-eval score improved tremendously**.

Context Recall also improved. However, answer correctness and relevancy fell. This could be due to the **weights of the ensemble retriever**. This would be explored further in the upcoming sections. For now, with the improvement in G-eval and Recall Scores, the Ensemble HF would be our **best model/system.**

5.4.2 Improving RAG Searches Through Re-Ranking:

Another technique which could potentially improve our searches would be to complement our current searches with a **reranking algorithm**. Often, **semantic searches** result in information loss as we are **compressing this information into a single vector**. Because of the information loss, we often see that the **top few vector search documents will miss relevant information**. Using a reranking algorithm, it **reorders the most relevant items at the top**. Hence often, you would **expand your semantic search options** and further use a re-ranker to cut down on the number of options.

For our case, I attempted to use Cohere AI Reranker. Refer to Appendix 5.4D for the code. Cohere AI Reranker is well known for improving RAG performance. Most RAG systems use a **two-stage retrieval** as a means of increasing search quality. In these two-stage systems, a **first-stage model** (an embedding model or retriever) retrieves a set of **candidate documents** from a larger dataset. Then, a **second-stage model** (the reranker) is used to **rerank those documents retrieved by the first-stage model**[19].

Results are as follows:

Ensemble HF: Our best model as before with bge embeddings, hybrid search with few shot prompting

Ensemble HF Rerank: Similar to the one above just adding **the Cohere AI Rerank** after hybrid search.

G-eval scores:

```
Average Score for Ensemble HF: 0.6355191874880344
Average Score for Ensemble HF Rerank: 0.5903976058086038
```

Fig 5.4C: G-Eval Scores For Reranking

RAGAS Scores:

```
=======================Ensemble HF=======================
RAG Average Answer Relevancy: 0.3743571687636189
RAG Average Answer Correctness: 0.5205250062605699
RAG Average Context Precision: 0.69999999993
RAG Average Context Recall: 0.5
=======================Ensemble HF Rerank=======================
RAG Average Answer Relevancy: 0.284599889611891
RAG Average Answer Correctness: 0.5720345380788269
RAG Average Context Precision: 0.69999999993
RAG Average Context Recall: 0.4666666666666666
```

Fig 5.4D: RAGAS Scores For Reranking

It is observed that for the system with the re-ranker, the G-Eval score fell. The **recall** score of RAGAS also fell despite the increase in correctness. Hence, instead of increasing **greater latency** by adding the re-ranker, since there were no **significant improvements,** I decided to continue to use the **hybrid search without re-ranking.**

5.4.3 Improving RAG Searches Through Adjusting The Weights Of The Ensemble Retriever:

To further explore if the **weightage parameter** would affect the results of the **hybrid search**, I conducted two extra experiments of changing the weightages to be **0.6, 0.4** and **0.4, 0.6** for the ensemble retriever for **semantic search** and **keyword search respectively**. Refer to Appendix 5.4E for the code.

Results are as follows:

0.5,0.5 is the original weightage of our **best model** (bge embeddings, hybrid search with few shot prompting without reranking).

G-Eval Scores:

```
Average Score for Ensemble HF 0.5,0.5: 0.6355191874880344
Average Score for Ensemble HF 0.6,0.4: 0.6276375029535538
Average Score for Ensemble HF 0.4,0.6: 0.5456360238792106
```

Fig 5.4E: G-Eval Scores For Weightage Changes

RAGAS Scores:

```
========================Ensemble HF 0.5, 0.5========================
Non-RAG Average Answer Relevancy: 0.3743571687636189
Non-RAG Average Answer Correctness: 0.5205250062605699
Non-RAG Average Context Precision: 0.69999999993
Non-RAG Average Context Recall: 0.5
========================Ensemble HF 0.6, 0.4========================
RAG Average Answer Relevancy: 0.36526090687001056
RAG Average Answer Correctness: 0.5677870677191184
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.6833333333333333
========================Ensemble HF 0.4, 0.6========================
RAG Average Answer Relevancy: 0.27873358142029897
RAG Average Answer Correctness: 0.5779282570940589
RAG Average Context Precision: 0.49999999995
RAG Average Context Recall: 0.26666666666666666
```

Fig 5.4F: RAGAS Scores For Weightage Changes

It can be observed that **0.6,0.4** has the best **context precision and context recall** score. There has also been an increase in answer correctness. Despite a slight drop in G-eval score, overall there has been an increase in other scores. Hence it would also be worth exploring further on 0.7,0.3, 0.8, 0.2 splits to observe changes in scores from a change in weightage.

Further exploration was done and the scores are as follows:

G-Eval Scores:

```
Average Score for Ensemble HF 0.6,0.4: 0.6276375029535538
Average Score for Ensemble HF 0.7,0.3: 0.6556917153416697
Average Score for Ensemble HF 0.8,0.2: 0.610807453001268
```

Fig 5.4G: G-Eval Scores For Next Weightage Changes

RAGAS Scores:

```
========================Ensemble HF 0.6, 0.4========================
Non-RAG Average Answer Relevancy: 0.36526090687001056
Non-RAG Average Answer Correctness: 0.5677870677191184
Non-RAG Average Context Precision: 0.89999999991
Non-RAG Average Context Recall: 0.6833333333333333
========================Ensemble HF 0.7, 0.3========================
RAG Average Answer Relevancy: 0.6425841014958626
RAG Average Answer Correctness: 0.5492757708033762
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.7166666666666666
========================Ensemble HF 0.8, 0.2========================
RAG Average Answer Relevancy: 0.36909307640225575
RAG Average Answer Correctness: 0.5602780987652676
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.6833333333333333
```

Fig 5.4H: RAGAS Scores For Next Weightage Changes

It can be observed that **0.7, 0.3** weightage gave the highest g-eval score of **0.655**.

All 3 had similar **RAGAS correctness score**, however, **0.7, 0.3 weightage** seemed to have the **best context recall** and **answer relevancy scores**. Hence, we would be proceeding with a **0.7, 0.3 weightage.**

5.4.4 Best System So Far:

G-Eval Scores takes priority compared to answer relevancy and answer correctness when measuring the how **personalized responses generated are.** This is because it can more accurately **evaluate our custom task.** With the **0.7, 0.3 weightage system giving** us the best **G-eval**, **recall and precision scores,** the following would be **our best RAG system so far:**

Embeddings: bge embeddings for vector store,

Search: Hybrid Search (Ensemble retriever, semantic search and BM25 keyword search combined)

Weightage: 0.7: Semantic**, 0.3:** Keyword search

Prompt: Few shot prompting

Re-ranking: No re-ranking

5.5 Experiment 4: Further Improving RAG Searches Through Meta Data:

To further improve the **context retrieved** out of RAG, it could be useful to **introduce meta data** that could **enhance content extracted**. In many RAG applications, most of the retrieved information often contains **additional meta data such as source information**, **timestamps** or **topic labels**. Leveraging this metadata could be a powerful technique to improve the **relevance** and **quality** of the **generated responses**[20].

The metadata filter technique involves using the available metadata to selectively filter, weight, or prioritize the retrieved information during the RAG process. For example, the system could give higher priority to information from **specific topic areas** that are **more relevant to the input, recent time periods**, or authoritative sources. By incorporating the metadata filter, the RAG system can **better identify** and **utilize** the **most relevant** and **trustworthy information**, leading to **more informed and reliable responses**.

To incorporate meta filtering, I had to first embed **relevant meta data** to the **content** stored in the **vector datastore**. One of the meta data embedded would be the **topic** of the **conversation**. For example, for the following conversation, the **topic of pets** would be **attached** to it since the **main topic** of the **conversation** is **about pets**:

```
page_content='{"Roydon": "Guess what, I just got a new pet dog!", "Jacob": "That's awesome! What breed is it?"}
```

Fig 5.5A: Conversation Snippet From Vector Datastore

```
metadata={'label': 'Response 1', 'source': 'pet.json', 'topic': 'pets'}
```

Fig 5.5B: Meta Data Attached To Above Conversation

The topic of the conversation would **value add to the search** since we would be able to better obtain useful contexts that would be relevant to the user's question. For example, when the user is having a **conversation about pets**, it would be useful to obtain **contexts/conversations** related to pets. However, without **meta data filtering**, other conversations **could have been included in the search**.

Another example would be when a general question is asked: What have you been up to Roydon. Below is the context retrieved for the ensemble retrieval. It can be observed that the context retrieved were of all travel information.

	question	contexts
0	What have you been up to Roydon?	["{"Roydon": "I tried to, but everywhere I went, I just kept getting ripped off by the locals.", "Dory": "That must have been frustrating. Did you try any of the stre {"Roydon": "Well, everything that could go wrong in Thailand did go wrong. But in Japan, I can\'t wait to explore the beautiful temples and try authentic Japar {"Roydon": "Easy for you to say. You weren\'t the one stuck in a foreign country with nothing going right.", "Dory": "I know, but sometimes these things happe

Fig 5.5C: Conversations Retrieved From Vector Datastore

For these general enquiries it could be useful to have a wider range of contexts retrieved for the mute user to have more diverse options to choose from. With our custom topic filtering, we can make it such that it retrieves conversations that are more diverse:

	question	contexts
0	What have you been up to Roydon?	["{"Roydon": "I tried to, but everywhere I went, I just kept getting ripped off by the locals.", "Dory": "That must have been frustrating. Did you try any of the street food at least?"} {"Roydon": "Hey there! Did you catch the Arsenal game last night? What a thrilling match!", "John": "Hey Roydon! Yes, I watched it. Arsenal played really well, didn\'t they?"} {"Roydon": "No, I had to borrow money from a friend to get back home.", "Xavier": "That must have been a really stressful experience."}"]

Fig 5.5D: Diverse Conversations Retrieved From Vector Datastore

5.5.1 Methodology of topic filtering:

For the topic filtering, **3 new functions** were created. When the normal person speaks, the text would first be fed into these functions.

(1) **GetTopic function:** Purpose is to obtain the main topic of what the normal person said to the mute. For example, if the normal persons asks how are you Roydon? It would be marked as a general topic. Refer to Appendix 5.5A for the code. The topic is obtained by using the following prompt engineering on ChatGPT with the help of few-shot prompting for more accurate results: (Appendix 5.5B)

You would be assisting in identifying topics from a snippet of conversation. I would supply the conversation directly. Interpret the main topic of the conversation and return the main topic.

Do not give multiple topics such as football/soccer. Only give one main topic.

For example if the conversation is: {"Roydon": "Can't wait for the new football season to start, hoping for a great one for Arsenal!", "John": "Hey Roydon! Yeah, it's always exciting to see how your team will perform."}

football

Example 2:

{"Roydon": "I'm planning to go on a trip to Japan next year", "John": "That's awesome! Japan is such a beautiful country."}

travel

Example 3: If no main topic can be determined such as a greeting

{"Roydon": "Hey there! How are you doing?", "John": "Hey Roydon! I'm doing great, how about you?"}

general

With a small number of tokens, it is able to accurately decipher the main topic of a given conversation.

(2) singularize_and_lower function: Next, some preprocessing has to be done in order to ensure the topics can be **matched more accurately**. Refer to Appendix 5.5C for the code. To reduce mismatches due to **upper case**, **plurality** and **present participles**, the NLTK toolkit can be used to generate **singular nouns**.

For example, the words **crocheting**, **crotchets** and **Crotchet** should all be **the same topic** of **crotchet**.

Before trying the **NLTK toolkit**, I tried out the popular **inflect package.** Refer to Appendix 5.5D for the code. The **inflect package** is a popular package in python that could accurately generate singular nouns. However, the package **itself is insufficient**, since it does not process **present participles.** Hence, the word crocheting **would not** be **transformed** to **crotchet.**

This is where the NLTK toolkit specifically **WordNetLemmatizer package** comes into play. Using the lemmatizer, I passed it **the topic**, together with its **nltk pos tag** of a **noun** so that the lemmatized word **would be a singular noun.** The lemmatizer would hence lemmatize using WordNet's built-in morphy function. It would however return the **input word unchanged** if it cannot be found in WordNet.

Hence an additional check was further put into the code for special cases like "crotcheting". The remove_ing function (Appendix 5.5C for code) is created to remove the ing from these words. Lastly, the word would be passed into the is_valid_word function (Appendix 5.5C for code) to ensure that it is a valid word before returning that topic **as a singular noun**.

Refer to Appendix 5.5E for the results of comparison between NLTK toolkit and inflect package.

(3) filter_list function: Lastly, the **filter list function** would handle the **meta filtering**. Refer to Appendix 5.5F for the code. It processes **2 different conditions**:

Condition 1: If the topic is general. If the topic interpreted is of **general type**, it would be useful to **obtain documents of various topic types** so that the mute user has a **wider range** of options to choose from. Hence a check is created **such that no documents/conversations with duplicated topics should be selected unless there are no other unique topic type**s.

Condition 2: If the topic is not general, it will **then check for documents / conversations** with the **same topic** as the **one interpreted** (from what the normal person said) to pick documents **that are most relevant**. For example, if the normal person mentioned something related to the topic of travel, the documents / conversations that are **prioritized and extracted from the vector datastore** will all have the topic of travel.

Lastly, if less than **3 documents have been selected** based on **the topic matching**, it would then **select new documents based on their original ranking** obtained from t**he ensemble retrievers** to be added on to those that have been extracted out **making the total count 3**. To ensure **no duplicated documents** are extracted a **check has been put in place as well**.

5.5.2 Usage of topic filtering:

To summarize, the usage of topic filtering would be as follows:

- The ensemble retriever would get the relevant documents (set us 10 right now).
- The topic of the **conversation/query** is interpreted.
- Out of the 10, they would then be **fed into filter_list function to be cut down via the metadata filtering technique**. Finally leaving with **3 conversations** to be used as context.

Refer to Appendix 5.5G for the code.

5.5.3 Results Of G-Eval Scores And RAGAS for new meta data topic filtering:

To test the effectiveness of meta data topic filtering, we integrated its usage with our **current best system** in the section above.

G-Eval Scores are as follows:

```
Average Score for Ensemble HF 0.7,0.3: 0.6556917153416697
Average Score for Ensemble HF Meta Filtered: 0.6873390817852749
```

Fig 5.5E: G-Eval Scores For Meta Filtering

RAGAS Scores are as follows:

```
========================Ensemble HF 0.7, 0.3========================
RAG Average Answer Relevancy: 0.6425841014958626
RAG Average Answer Correctness: 0.5492757708033762
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.7166666666666666
========================Ensemble Meta Filtered========================
RAG Average Answer Relevancy: 0.6723204037179018
RAG Average Answer Correctness: 0.568161100684651
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.7666666666666666
```

Fig 5.5F: RAGAS Scores For Meta Filtering

Compared to our previous best model (Ensemble retriever of weightage 0.7, 0.3), it can be observed that all scores besides precision have had a slight improvement. Thus, this shows that we should incorporate the meta data topic filtering.

5.6 Experiment 5: Further Improving RAG Searches Through Query Processing:

Lastly, it is also useful to **explore query preprocessing** for **complex queries**. So far based on our testing, most conversations are assumed to **be only simple queries**. For example, when a user poses a query with **multiple questions bundled together**, directly **using the entire query** for embedding comparison **might not give the best result**. This is because a single conversation extracted from the **vector database** may not provide context/answers to all the questions within the user's query. In these instances, it's more effective to split the user query and perform RAG on each individual question[21].

With this in mind, it would be useful to create a function to **preprocess queries** that are **too long**.

The process query function was created and works as follows: Refer to Appendix 5.6A for the code.

- An NLTK toolkit specifically the **PunktSentenceTokenizer package** is used to break down the query/conversation into multiple sentences. The PunktSentenceTokenizer is a pretrained model and could be further retrained for specific cases. Hence it was selected. In this case due to **simplistic conversations in english**, the **pretrained model of the Tokenizer** is sufficient. However, when used for other languages, **this tokenizer** must be trained on a large collection of plaintext in the target language before it can be used. The tokenizer divides a text into a list of sentences by using an unsupervised algorithm to identify abbreviation words, collocations, and words that start sentences.
- After sentences have been obtained, we would have to filter for questions so as to split by questions. Non-questions should be appended to the back of questions since most of the time, these are useful information that provides context to the question. An example is as follows:

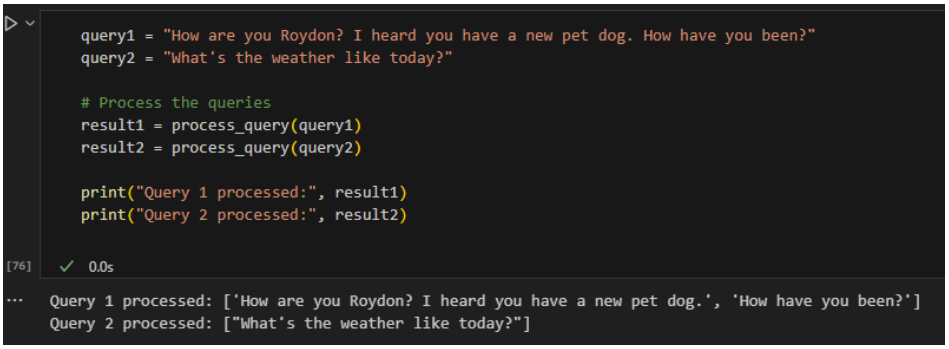


Fig 5.6A: Examples Of Breaking Down Complex Queries

- **Lastly,** the query preprocessing is integrated to the overall code (best system) as follows:

A for loop is employed to loop through all the questions within the **complicated query** to obtain the **top 3 conversations/context** from the **vector datastore** for each question. Refer to Appendix 5.6B for the code.

5.6.1 Results Of G-eval and RAGAS Scores:

G-Eval Scores are as follows:

```
Average Score for Ensemble HF Meta Filtered: 0.6873390817852749
Average Score for Ensemble HF Meta Filtered Query Altered: 0.7141140981676445
```

Fig 5.6B: G-Eval Scores Of Query Processing

RAGAS Scores are as follows:

```
=======================Ensemble Meta Filtered=======================
RAG Average Answer Relevancy: 0.6723204037179018
RAG Average Answer Correctness: 0.568161100684651
RAG Average Context Precision: 0.89999999991
RAG Average Context Recall: 0.7666666666666666
=======================Ensemble Meta Filtered Query Altered=======================
RAG Average Answer Relevancy: 0.6798901851761995
RAG Average Answer Correctness: 0.6569532556501141
RAG Average Context Precision: 0.9
RAG Average Context Recall: 0.7666666666666666
```

Fig 5.6C: RAGAS Scores Of Query Processing

It can be observed that there has been an increase in **all scores** while context recall **remains the same**. Thus, it would be good to integrate the query processing functionality into our current system,

That being said, more tests might need to be done as the **current complex queries** being tested are made up of **multiple questions that relate to the same topic**. (EG : Hi Roydon how is your new pet? Where do you plan to bring him to play?). Despite it being a complex query, both questions are on the topic of a pet. Hence the **contexts retrieved** are still **very similar** despite **splitting up the queries** which hence **could lead to minimal improvements**.

5.7 Experiment Conclusions:

To summarize, based on the **experiments conducted above** we found that the **current best RAG System** should have the below in place. Refer to section 3.2 for the overview on **response generation** using the **final system** incorporating all these **functionalities** below:

Query Processing: Processing of **complicated queries** by splitting multiple queries and obtaining contexts for **each query.**

Embeddings: bge embeddings for vector store.

Search: Hybrid Search (Ensemble retriever, semantic search and BM25 keyword search combined).

Weightage: 0.7: Semantic**, 0.3:** Keyword search.

Meta-filtering: Meta filtering by topics

Prompt: Few shot prompting.

Re-ranking: No re-ranking.

6. Testing

Before user studies can be conducted some testing was done to ensure that the application is ready.

6.1 Backend Testing

To ensure all Apis work as intended, fastApi allows **easy testing** by allowing users to try out the Apis built. Similar to **postman**, I was able to enter **required inputs** to obtain **expected outputs.** Below is an example of testing out the /post-audio-response/ Api.

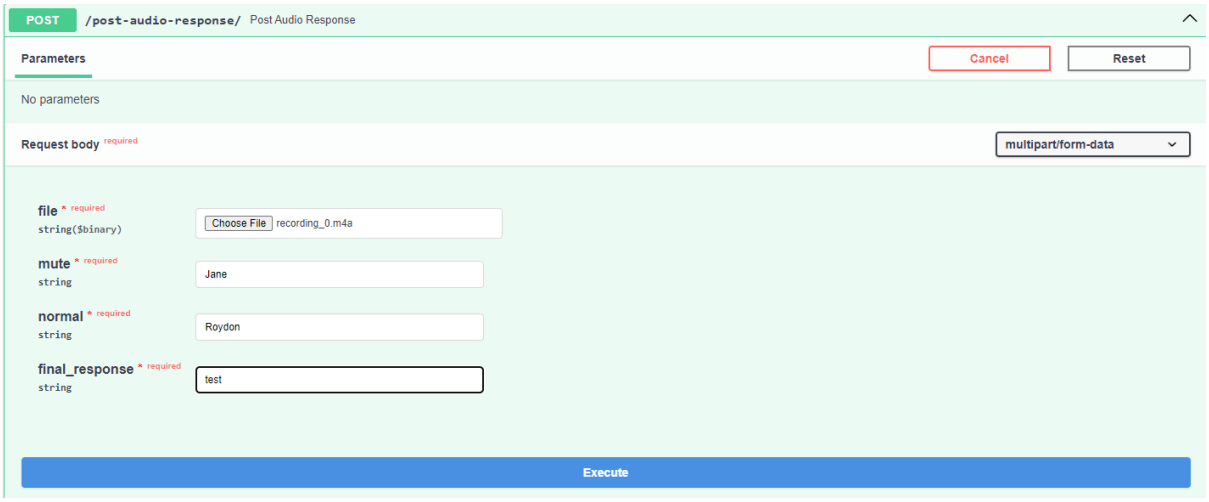


Fig 6.1A: Testing of backend apis

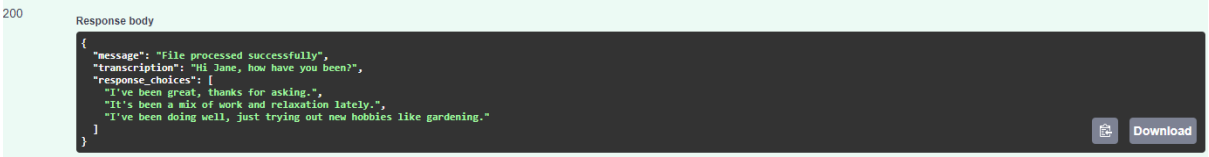


Fig 6.1B: Success Result Of Test

Other apis are tested similarly.

6.2 FrontEnd Testing

Since Expo allows for testing through an actual application, most of the testing were done **directly on my iPhone.** This allows for a good judgment on how the application functions on **real devices.**

7. User Studies

Since there is a **limit** to the **accuracy** of evaluation scores from RAGAS and G-eval, it is important to conduct a few user studies. These studies serve as a purpose to test the application. With the **vagueness** of how **correct** the **responses generated** are, **user studies** would provide **better feedback** on areas of improvements as well as areas that are well done. To do so, some setup **has been done before hand** for user studies. The following setup was done:

1. Building of **mock user profiles:** These **user profiles** contain **demographic information, hobbies** as well as the **generation of some mock conversations**. The mock conversations are **past conversations** that the user **has had with others**. These mock user profiles are **made known** to the **participants** so that they know what to **expect when the responses** are generated. (Participants are to take on the persona of the mock user profiles). A total of **5 different user profiles** were created.
2. Survey: A survey would be given to **participants** after **trying out the application**. This survey will contain the **following questions** to evaluate the applications. (Appendix 7A)

The user studies would be about 1-2 hours long. Users will play two roles:

(1) They will be the **mute user**. Their persona would be according to the **mock user profile** shown to them and they would **pick the responses** accordingly.

Goal: Allow users to observe if responses generated are **personalized to their past** experiences. I would lead the conversation to try to **cover all their hobbies**.

(2) **(Role reversal)** They will be the **normal person**. They would then be leading the conversations, and I would **play the role of the mute** selecting the responses accordingly.

Goal: Allows the user to observe how the **application responds to random topics/questions**. Also allows the users to observe if the application **learns from its mistakes** since users can **ask the same questions** that the **application failed** to **generate personalized responses before** (in role 1).

Instructions for the user study will be given to participants before the start. **(Appendix 7B)**

7.1 Key Findings from User Study:

Upon the completion of user study, it provided **validation** on the good points of the application as well as **areas to work on**. In summary the **good areas** are:

- A clean user interface **that is simple and intuitive**.
- Responses generated were **highly personalized** to hobbies, interests and past experiences.
- Having 3 options was good as it was **not too overwhel**ming and provided enough for the user to choose from.
- Having a way for the user to **still type their responses** was good.
- How the application **grows** and **adapts** to user responses were good.

Refer to Appendix 7C For Survey Results and Appendix 7D for Summarized User Study feedback.

Areas To Work On:

- **Regenerate responses.**

  From one of the user studies (Yu Min)(**Appendix 7.3C**), an incident occurred where the responses generated was **irrelevant** and **unnatural** to the **conversation**. However, since it was stored in the recent history, **subsequent responses** generated **continued to be unnatural and similar to the one before** regardless of **new question**s or **conversations.**

  However, upon clearing the history manually, **the responses generated returned back to being personalized and relevant** to the conversation. Hence there needs to be a way to **clear the history** and **regenerate responses.**

  Having a way to **generate responses** multiple times would also provide more options for the users.

- **Edit existing responses.**

  From the user studies, it can be observed that sometimes the responses generated are good but may not be **exact to what the user** would like to say.

  Having a way to **edit the existing options** would **provide greater convenience** to the mute users instead of needing to re type out fully what they would like to say.

- **Rate responses generated.**

  From the user studies, many users have mentioned that they would like an option to rate the 3 responses.

  A possible way could be to like or dislike the options generated so that the model knows in the future when asked similar questions, which responses should not be generated.

- **User interface for responses itself** seems a **little clumped up** and **design** can be **improved**.

- **Security concerns** due to privacy of past conversations.

  To address this concern, one way would be to ensure that instead of **using the api** of open ai's gpt-3.5 model, since we already **know this model works** with the current prompt engineering and RAG set up, we could store the model locally instead.

  With the model and the vector stores being local, there would be a **lower possibility** of personal content being leaked out.

Personal Observation:

Besides the points above, I would also like to highlight what I found out from the user studies.

- Slower response generation **only at the start:** Usually responses generated are **less than 5s** but when first using it (no history at all) could take up to **10s** only for the **first response generated**.
- **Responses generated are less personalized for general topics.**

For responses generated, it could be observed that generally when it comes to a conversation on a **particular topic such as a particular country** (like Iceland) the person travelled to etc, the **model does very well** in generating personalized responses. The context extracted out are usually related to Iceland and the model does well in obtaining relevant activities/ attractions the user has visited. However, when it is a **more general topic such as what sports the user enjoys,** the model could struggle in terms of personalization giving more **general responses**.

Hence more research could be done on perhaps increasing **contexts for general topics** are expanding the number of options.

- **The growth of the application with the users:**

Currently since all user studies involve some mock data, the application is **not completely starting from scratch**.

Despite saying that, the adaptability in the model can still be observed since during the user studies, new information **was added to the mock data** based on the **user responses**. For example for Cheryl's user study:

When Cheryl was playing as the mute user and I was the normal user:

Where did you bungee jump at? (**Appendix 7.3K**)	1. I tried bungee jumping at a location with a beautiful scenic view to make the experience even more memorable. 2. The bungee jumping spot I chose was known for its great heights and breathtaking surroundings. 3. Bungee jumping was at a popular site known for its safety measures and experienced staff to ensure a thrilling yet secure adventure.	**(Typed out herself)** It was in Singapore at Sentosa.	Audio To Text: 1.1s Response: 3s Total: **4.1s**

During role reversal, where I was now the mute person Cheryl asked the question again but the responses were different: **It learned that the mute user bungee jumped at Sentosa.**

Anyways recently you went Bungee Jumping right? Where did you go? **(Appendix 7.3E)**	1. I went bungee jumping in Singapore at Sentosa. 2. The bungee jumping experience was in Singapore at Sentosa. 3. It was at Sentosa in Singapore where I went bungee jumping. **Important Contexts:** {"Matt": "Where do you bungee jump at?", "Yu Min": "It was in Singapore at Sentosa!"}	I went bungee jumping in Singapore at Sentosa.	Audio To Text: 4.4s Response: 2.8s Total: **7.2s**

More examples could be found as you read through the other user studies and responses generated in the section below. It could be observed that the model considers what the mute user replies and when asked similar question, the responses generated would be similar to how the mute would reply **thus successfully learning their responses and providing convenience in reducing repeated typing.**

However, **more user studies could be done perhaps** to test the **growth of the application (from scratch)** as the user starts conversing more with the help of the application.

It would be interesting to observe how the **general responses** become **more personalized over time**, but a **longer user study** would need to be conducted.

7.2 Participants Of User Study:

There will be a total of 20 participants. Their profiles and feedback are highlighted below.

I tried to pick users from a **wide range of profiles/experiences** to ensure the app has been tested more widely. These users are mainly from:

(1**) NTU's Welfare Service Club specifically those volunteering for the intellectually disabled (RSPID)**: Under the club, there some **intellectually disabled** who are **mute** and hence these volunteers would have experience in **interacting with those who are mute**. One of the mute at **the welfare home is highly functioning** but is able to speak only **very few phrases**, **very slowly,** at **very low volume** only after feeling the vibrations from the **volunteer's throat**.

Hence these volunteers could provide a **good evaluation** on how the **app is** since the **target audience** of the app is similar to those of whom they have interacted with before.

(2) **Software Engineers**: It would be good to get the opinions of fellow software engineers.

(3) **Others:** Other users have a wide range of **occupation** and could provide a **different perspective** on the application. Some include nurses as well as users who may have interacted with those who are mute.

Participants Breakdown is as follows:

Gender	WSC RSPID	Software Engineers	Others
Male: 11 **Female: 9**	***7***	***5***	***8***

Refer to Appendix 7D for summarized responses from the participants.

7.3 3 Examples Of User Study:

I have selected **3 participants'** user study to showcase some of the results during the user studies. These results are worth highlighting to showcase some of the good and bad areas of the application. For the full user studies refer to Appendix 7.3A,B,C below.

Tick represents the hobbies that were covered in the test. (**Note the full user studies are found in the Appendix and this section is only a small snippet of it).

Participant 1 - Sin Yee: Appendix 7.3A

User Profile/Persona:

Hobbies: (1) Crotchet ✅ (2) Gardening ✅ (3) Gym ✅ (4) Works at the hospital as a nurse ✅

Recent History: (1) Crotcheting new projects, (2) Picked up gardening as a hobby, (3) Planning a trip to Australia, (4) Goes to anytime fitness, (5) Trying to cope with workload

Bad Areas:

(1) Long latency at the start of the application as mentioned in the above section.

Normal Person	Responses Generated	Response Selected	Time
Hi Sin Yee, how have you been doing? (**Appendix 7.3A**)	1. Hi Gary, I've been keeping busy with work and my hobbies. 2. Hi Gary, I've been doing well, thank you for asking. 3. Hi Gary, it's been a mix of tiring and fulfilling lately.	Hi Gary, I've been keeping busy with work and my hobbies.	Audio To Text: 2s Response: 7s Total: **10s**

(2) Some irrelevancy to past experiences resulting in the user having to type their response. **But very minimal (~10-20%).** Some examples are shown below:

Oh no! Why has work been so hectic though? **(Appendix 7.3M)**	1. - 2. The increase in patients and emergencies has kept us all on our toes, but being able to provide care makes it worthwhile. 3. -	**(Wrote out herself)** There has been a pandemic lately!	Audio To Text: 1.3s Response: 3s Total: **3.3s**
Well you better take care then. Are you planning for any holiday trips to relax? **(Appendix 7.3N)**	1. Planning a holiday trip to unwind and recharge sounds like a perfect idea after the busy days at work. 2. It's important to take breaks and rejuvenate, a holiday trip to relax and destress would be ideal. 3. Taking care of myself is essential, a holiday trip to relax and unwind is definitely on the cards to recharge for the upcoming challenges at work.	**(Wrote out herself)** Im thinking of Australia.	Audio To Text: 1.6s Response: 3s Total: **4.6s**

Good Areas

(1) Most of the time the application does well **in utilizing the context** to generate more **personal responses.** Response time is **generally 5s.** Some examples are shown below and more can be found in the **Appendix 7.3A**.

Well what hobbies have you been doing recently then? **(Appendix 7.3A)**	1. I've picked up gardening as a new hobby. 2. I've started crocheting recently, it's been really fun! 3. I've been trying out painting as a new hobby, it's quite relaxing. **Important contexts:** {"Jackson": "What have you been up to Sin Yee!", "Sin Yee": "Hey Jackson, guess what? I've picked up gardening as a new hobby!"} {"Xavier": "Hey Sin Yee, I heard you've taken up crocheting as a hobby! That's awesome!", "Sin Yee": "Yes, I'm really excited to start! Do you have any tips for beginners?"}	I've started crocheting recently, it's been really fun!	Audio To Text: 1.1s Response: 3.3s Total: **4.4s**

Whoa nice! What have you been growing in your garden then? (Appendix 7.3C)	1. I've started with some herbs like basil and mint. 2. I'm growing a variety of vegetables like tomatoes, cucumbers, and bell peppers. 3. Currently, my garden has carrots, lettuce, and radishes growing **Important Contexts:** {"Jackson": "That's really cool! What are you planning to grow in your garden?", "Sin Yee": "I've started with some herbs like basil and mint, and I'm thinking of planting some vegetables too."}	I've started with some herbs like basil and mint.	Audio To Text: 1.5s Response: 2.7s Total: **4.2**s

Participant 2 - Cheryl: Appendix 7.3B

User Profile / Persona:

Hobbies: (1) Cooking ✅ (2) Hiking ✅ (3) F1 Race ✅ (4) Travelling ✅ (5) Sports/Marathon ✅

Recent History: (1) Cooking a meal for boyfriend, (2) Planning a hike, (3) Getting F1 Tickets, (4) Planning an Iceland Trip, (5) Taking part in a marathon, (6) Badminton match

Bad Areas:

(1) Similarly, there was a long latency at the very start for the generation of response

Normal Person	Responses Generated	Response Selected	Time
Hi Cheryl, how have you been doing? (Appendix 7.3A)	1. Hi Mary! I've been busy planning a trip to Iceland to see the Northern Lights and explore. How about you? 2. Hi Mary! Lately, I've been focused on training for a marathon. It's been challenging but rewarding. What have you been doing? 3. Hi Mary! I've been enjoying some time outdoors, training for a marathon. How about you, what have you been up to?	Hi Mary! I've been busy planning a trip to Iceland to see the Northern Lights and explore. How about you?	Audio To Text: 2.5s Response: 11s Total: 13.**5s**

(2) There were also some responses that were irrelevant resulting in the participant needing to type out their own responses. But again, these were minimal. Some responses were also **very long winded.**

Yeah, I love going overseas during December of the winter holidays. Besides travelling, you mentioned that you're travelling for a marathon? (**Appendix 7.3E**)	1. Joining a marathon sounds like an exciting challenge! How do you usually prepare for it, and are there any specific goals you have in mind for this one? 2. That's impressive! How do you manage your training schedule with your other commitments, and do you have any tips for staying motivated during tough training sessions? 3. Running a marathon requires a lot of dedication and hard work. How do you stay focused on your training and maintain a healthy lifestyle while preparing for the race?	**(Wrote out herself)** Yeah I am currently training for it.	Audio To Text: 2.1s Response: 3s Total: **5.1s**
Are you going to watch it live at the grandstands?	1. It would be incredible to watch the F1 race live from the grandstands! The atmosphere and energy at the track must be electrifying, providing a one-of-a-kind experience. 2. Watching the F1 race live from the grandstands offers a prime view of the fast-paced action and allows for a more immersive experience. It's definitely a fantastic way to enjoy the race. 3. Watching the F1 race from the grandstands provides a unique vantage point to witness the speed and skill of the drivers up close. The excitement and buzz of the crowd make it an unforgettable experience. Enjoy the race from the grandstands!	**(Wrote it herself)** Yeah I managed to get tickets! What about you?	Audio To Text: 1.7s Response: 3.5s Total: **5.2s**

Good Areas

(1) Similarly, subsequent **response times were fast** and the application does well in using the contexts for **generating responses** that were **personal.**

Well, I've been doing great. Thanks for asking. Which part of Northern Lights or Iceland are you planning to visit? (**Appendix 7.3B**)	1. I'm looking forward to visiting Thingvellir National Park in Iceland for the best view of the Northern Lights. Are you planning any trips yourself? 2. I've heard that Thingvellir National Park is a popular spot to see the Northern Lights, so that's definitely on my list. Have you been to Iceland before? 3. Thingvellir National Park is where I hope to witness the stunning Northern Lights. Have you thought about any travel destinations for your next adventure? **Important Contexts:** {"Cheryl": "Yes, I've been researching the best spots for Northern Lights viewing, and I've heard that Thingvellir National Park is a popular choice due to its clear skies. I can't wait to witness the dancing lights in the sky! Have you thought about what souvenirs you might bring back from Iceland?", "Matt": "Thingvellir National Park sounds like a perfect spot to catch the Northern Lights! As for souvenirs, I've heard that Icelandic wool products like sweaters and blankets are popular choices. Are you planning to bring back any unique souvenirs to remember your trip by?"}	I've heard that Thingvellir National Park is a popular spot to see the Northern Lights, so that's definitely on my list. Have you been to Iceland before?	Audio To Text: 1.3s Response: 3.4s Total: **4.7s**

Nice that sounds great. Where are you planning to hike? (Appendix 7.3J)	1. I'm thinking of exploring MacRitchie Reservoir Park. 2. I'm considering checking out Bukit Timah Nature Reserve. 3. I might go for a hike at Sungei Buloh Wetland Reserve. **Important Contexts:** {"Jackson": "Another great spot is Bukit Timah Nature Reserve for a more challenging hike with steep slopes.", "Cheryl": "Sounds like a good workout! Are there any facilities like restrooms along the trail?"}	I'm considering checking out Bukit Timah Nature Reserve.	Audio To Text: 3.5s Response: 2.5s Total: **6.0s**

Participant 3 – Yu Min: Appendix 7.3C

User Profile / Persona:

Hobbies: (1) Sky diving ✅ (2) Bungee Jumping ✅ (3) Swimming ✅

Recent History: (1) Sky diving trip (2) Bungee Jumping in Singapore, (3) Planning a party with friends, (4) Cooking for boyfriend, (5) Learning a new sport: Floorball

Bad Areas:

(1) During this user study, **the response generation went bad** after a **poor response generation.** This is because after the result was stored in the **history of recent 5 conversations,** the GPT Model became affected by it and **subsequent response** was repeatedly bad. This was only solved by resetting the history (Deleting the content in the Json file).

As observed below, the 2nd conversation had the **same poor response as** the 1st conversation where the responses generated **were completely irrelevant to the question asked.**

(Issue): Had to reset recent 5 conversation storage. Well when is your next bungee jumping trip? (Appendix 7.3H)	1. Next time, would you consider inviting friends to join you in the bungee jumping adventure? 2. Have you researched any new bungee jumping locations or heights for your next trip? 3. Are you planning to do anything different or special for your next bungee jumping experience?'	-	Audio To Text: 1.0s Response: 3.4s Total: **4.4s**

Anyways, how has your swimming lessons been going? **(Appendix 7.3I)**	1. Have you noticed any improvement in your swimming skills since starting the lessons? 2. What has been the most challenging part of your swimming lessons so far? 3. Are there any specific techniques or strokes that you find particularly enjoyable during the lessons?'		Audio To Text: 1.1s Response: 2.5s Total: **3.6s**

Refer to the recovery in Appendix 7.3C

Good Areas

(1) Besides the **poor responses generated above,** all other responses generated were well done with **most being personal**. After the reset of the history, the generation of responses also recovered well and continued to be **personal.**

Normal Person	Responses Generated	Response Selected	Time
Hi Yumin, what's up? **(Appendix 7.3A)**	1. Not much, just been experimenting with new recipes for my boyfriend. 2. Hey Matt! I've been thinking about trying out a new sport, any suggestions 3. Hey Matt! I'm planning a party for our friends, do you know any good bars to host it at? **Important Contexts:** {"Xavier": "Hey Yu Min, have you thought about picking up a new sport lately?", "Yu Min": "Hey Xavier! Yes, I've been thinking about it. Any recommendations?"} {"Matt": "Hey Yu Min, have you been cooking for your boyfriend lately?", "Yu Min": "Hey Matt! Yes, I have been trying out some new recipes for him."} {"Yu Min": "Hey Mary, I'm planning a party for our friends. Do you have any suggestions for a good bar to host it at?", "Mary":	Not much, just been experimenting with new recipes for my boyfriend.	Audio To Text: 1.9s Response: 5.1s Total: **7.0s**

	"Hi Yu Min, that sounds like a fun idea! How about checking out the new rooftop bar downtown? It has a great view and ambiance."}		
(Recovery): Recovered after reset Anyways how has your swimming lessons been going? **(Appendix 7.3J, 7.3K, 7.3L)**	1. My swimming lessons are going well, I've been learning different strokes and techniques. 2. I've been making progress with my swimming lessons, starting with water safety and now moving on to more advanced skills. 3. Swimming lessons have been challenging but rewarding, especially as I work towards swimming laps confidently. **Important Contexts:** {"Natty": "That's a brave decision! Facing your fears head-on is always a good thing. How are your lessons structured?", "Yu Min": "The lessons start with basic water safety and gradually progress to different swimming strokes and techniques."}	My swimming lessons are going well, I've been learning different strokes and techniques.	Audio To Text: 1.7s Response: 2.9s Total: **4.6s**

8. Improvements Made After User Studies

The user studies provided good feedback on the areas that should be worked on including **3 new features** that the application should have **namely** (1) **Refresh / regenerate responses**, (2) **Like/Dislike Responses**, (3) **Edit existing responses**. With the need for these 3 functionalities, 2 new apis were created namely: (1) regenerate-responses, (2) rank-responses.

8.1 Regenerating Of Responses

From the user studies, regeneration of responses would be a great feature for users to refresh all responses that was generated if none of them are to their liking. To do so, I added a feature that wipes the recent 5 history of messages and provide 3 new responses for users to pick from. Below is a demonstration of how it works:

To test the functionality, I first made **mock past history data mocking the recent conversations** stored in the **Json file**. (Appendix 8A) This data would hence influence the responses generated.

After clicking on the refresh button, the regenerate-responses api would be called. The api would first wipe the memory of **recent history (recent 5)** and **proceed to regenerate 3 new responses** using the same way as before. (Appendix 8B)

It can be observed that the responses have been **refreshed and regenerated successfully** thus granting users now with an **option to regenerate and obtain new responses**.

Fig 8A: Previous Responses

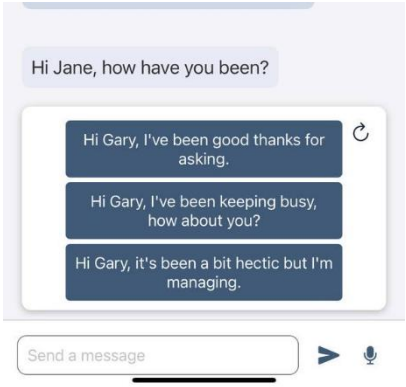


Fig 8B: New Responses After Refresh

8.2 Rate responses generated

From the user studies, it would be good for the users to be able to **provide feedback** towards the generation of certain responses and to **regenerate specific bad responses**. To do so I added 2 buttons beside each response as seen in the figure. Each response is **initially given a score of 0.5** representing it as a **neutral response**. When the user clicks on a **thumbs up**, the **score of the response** would be **changed to 1**. When the user clicks on **a thumbs down**, the **score of** the response would be **updated to 0**. **This response would then be regenerated** based on the **existing scores of the other responses.**

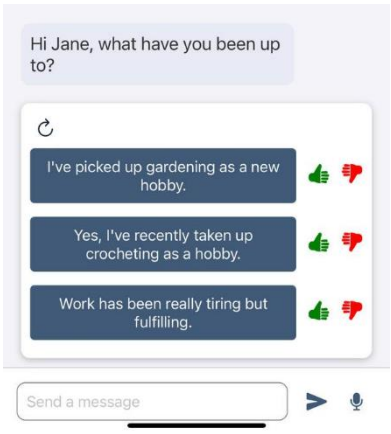


Fig 8C: Like & Dislike

When clicking on **the thumbs down button**, the **rank-responses api** is called. **The query, previous responses** and **scores are fed into the api** including the **response that was thumbs downed**. These will then be **fed into ChatGPT** to **regenerate the response**. The prompt used for the regeneration is different. Using the scores, **ChatGPT should regenerate the response to be similar to the response that was rated good**.

The response regeneration took about **2.98s**.

The current prompt for response generation is:

You are an assistant whom will faciliate the conversation between a mute and a normal person. The mute persons name is {mute} and the normal person is indicated as {normal}. You previously generated the following responses: {responses}. For the query of: {query.

The score of the responses are as follows respectively with a max score of 1 being a good response, 0.5 being neutral and below 0.5 being a bad response: {scores} You have to generate a response that would replace the {badResponse} using the other responses (above 0.5) as a guide. Use responses with higher scores first (those with score of 1).

Do not repeat any of the previously generated responses.

The responses should be what a person would say and should not include actions in a third person view. Your persona would be from the perspective of the mute person.

For example, If the query is: "Hey Jane, what are you up to?" and the responses previously generated were: ["Hi Gary, I've been keeping busy with work and some new hobbies.", "Hey Gary, I've been learning how to cook new recipes and exploring different cuisines. It's been a fun and delicious experience!", "Hey Gary, I've been learning gardening, how about you?"] and the scores are [0.5, 0, 1] respectively, then, the good response is: "Hey Gary, I've been learning gardening, how about you"? and the bad response is "Hey Gary, I've been learning how to cook new recipes and exploring different cuisines. It's been a fun and delicious experience!" The generated response should not be about cooking and should be about gardening (using the good response as a guide to infer the topic). Do not not repeat any previous responses like "Hi Gary, I've been keeping busy with work and some new hobbies." Do not explain reasoning and just give the new response:

I've been learning gardening, it's been so fun!.

Another example: If the query is: "Hey Jane, what are you up to?" and the responses previously generated were: ["Hi Gary, I've been dancing recently.", "Hey Gary, I've been hiking recently!", "Hey Gary, I've been learning gardening, how about you?"] and the scores are [0, 1, 0.5] respectively, then, the good response is: "Hey Gary, I've been hiking recently!"? and the bad response is "Hi Gary, I've been dancing recently." The generated response should not be about dancing and should be about hiking (using the good response as a guide to infer the topic). Do not repeat any previous responses like "Hey Gary, I've been hiking recently!" Do not explain reasoning and just give the new response: :

"I've been hiking recently and its my new hobby".

The prompt used **may need exploring** to **better improv**e the regeneration. But currently below is an example of how the new functionality works: When clicking on the **thumbs down** button for the **3rd response**, its score would **first be changed to 0** before getting **fed into the api**. (Figure 8D)

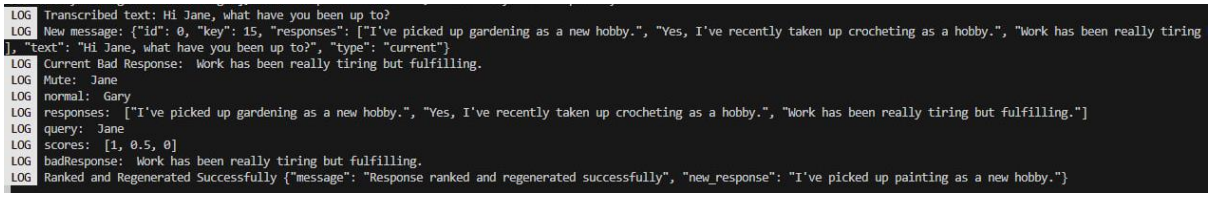


Fig 8D: Information Fed Into API

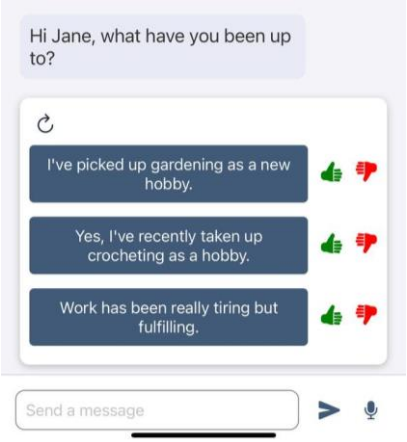


Fig 8E: Response before thumbs down

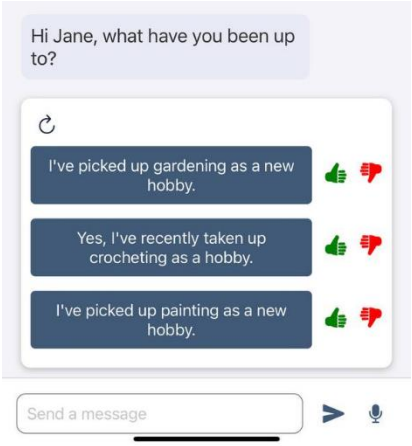


Fig 8F: Response after thumbs down

Note that before clicking the **thumbs down button on the third response**, the user clicked the **thumbs up button on the first response**, changing its **score to a 1** (Figure 8D). It can be observed that the GPT model **picked up on the sharing of hobbies and hence tried to regenerate the 3rd response into a hobby that the user possibly has** (Figure 8E, 8F).

That being said, **the response regenerated** could have **been better if it was about gardening** since the 1st response was **thumbs up**. **Perhaps more could be explored** in terms of **better prompt engineering** the **regeneration of a specific response**.

8.3 Edit existing responses

From the user studies, it would be essential for users to be able to **edit certain responses before picking them**. Often the generated responses may not be **exactly what the user would say.** By allowing them to edit these responses it would **firstly prevent the need to retype messages that are similar** to what was generated and secondly, **provide a more accurate response for ChatGPT to learn from** so as to **better personalize the user experience**. Below is an example of how the functionality works: It shows the **user editing the third option** to include that **he is going to AnyTime Fitness**.

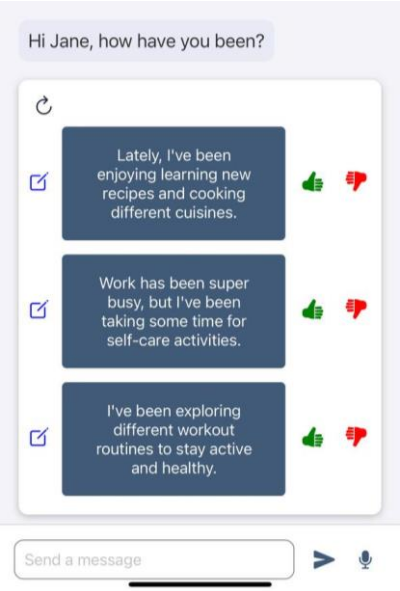


Fig 8G: Before edit

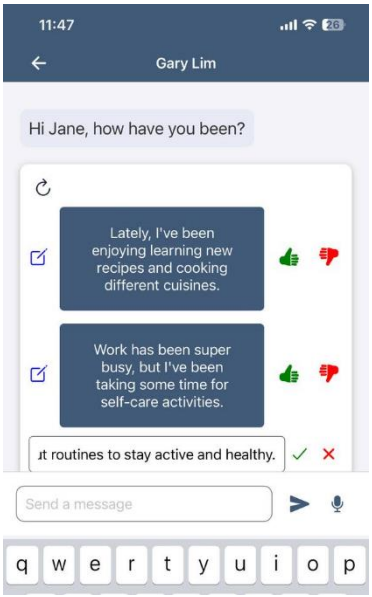


Fig 8H: Tapping On Edit Button

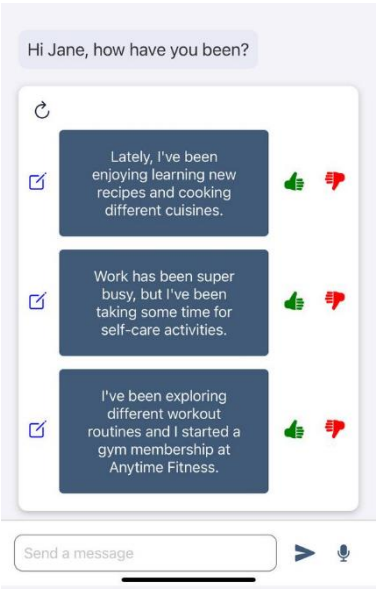


Fig 8I: After edit

9. Conclusion And Next Steps

Throughout the course of this project, the author had the privilege to explore the applicability of generative ai in **assisting the mute communicate** by implementing a **mobile application** that is able to generate **human-like personalized responses.**

The application has done well in the following areas:

(1) It is able to generate **personalized responses** that relates to the users' past experiences
(2) It provides users with a **convenient way** to **regenerate** or **alter generated responses** so as to match the message they are trying to convey to the normal person.
(3) It is able to obtain feedback and grow with the users **so as to eventually mimic the user's** experiences and personality.

This experience has not only allowed the author to apply theoretical knowledge to real-world challenges but also to expand his technical skills and understanding of **mobile development and** highly relevant technologies such as **generative ai.**

The challenges encountered along the way have been invaluable learning opportunities. With generative ai **still being in its infancy stage**, the research conducted here has been challenging since most of these areas are relatively new and **are still being explored on.** This project has hence allowed the author to conduct **useful experiments** that could prove useful in the testing of **new RAG systems.**

Looking forward, the groundwork laid by this project not only sets the stage for further enhancements to **the mobile application for mute users, but also contributes to the broader field of** generative ai in healthcare**.** It is the author's hope that the insights and learnings from this project will inspire and inform future developers and researchers in their endeavors to harness technology for compassionate and personalized care.

As the author reflects on this journey, he is grateful towards his mentor, growth opportunities as a software developer and learning opportunities in **highly relevant fields** like generative ai.

The author eagerly anticipates the continued evolution of the mobile application and **its contributions** to provide mute users with a **convenient, personalized way to reply others.**

Below are **some possible areas to explore** and **improve the application** on.

9.1 Improving Response Generation:

9.1.1 Improving Response Generation Through New meta data Time:

We previously explored using **topic** as a meta data. Another useful meta data to be embedded **would be the time of conversations**. The concept of time tagged to each conversation is important since the more recent ones could be more relevant in majority of situations.

9.1.2 Improving Latency:

Despite **most responses being generated fairly quickly,** the added latencies from speech to text and text to speech results in **about 5-7s** before reading the speech out to the normal person. Perhaps more can be done to speed up **the generation of the texts**.

There are also occasions where the generation of responses take about **~10s.**

There are a few ways to **reduce latency**[23], [24]:

(1) **Reduce** number of tokens **outputted.** Due to **linearity of response time,** 30-100ms of latency is added for **every token generated.** Although the number of **input tokens may affect latency,** it has relatively little effect **compared to output tokens.**

Currently, some of the responses generated **are extremely long-winded** (Appendix 7.3B). Better prompt engineering such as instructing the model to **"be brief"** and also removing newlines or other whitespaces where irrelevant **could help speed up the process.**

(2) **Stream output** and **Use Stop Sequences:**

An interesting solution would be to stream **the output to the user immediately,** hence increasing the **perceived speed of the application.**

To further improve the **actual speed, stop sequences** can be used. **At times when the model is unsure,** it could take longer to respond. Experiments could be done to explore prompting the model to **output unsure** if it is not going to be able to produce a **useful answer.** By setting this string as a **stop sequence,** the **responses generated would be faster.**

9.1.3 Trying a different model:

There was insufficient time to try out different models such as bard or llama which could have been better in both **latency and response generation.** Exploring on hosting the model locally could also be useful in targeting **security concerns.**

9.2 Improving Functionalities:

9.2.1 User Switching Through Cloud Storage:

Currently, the vector store and Json files are **all local.** However, since it is local it is currently more difficult to implement a system for user switching.

It could be useful to create a new storage solution in the cloud containing **the previous conversations** so that these information can be loaded **separately for different users with different accounts.** (Since different users will have different people they are speaking to and each will have their own set of conversations)

9.2.2 Like/Dislike Functionality:

There is still much to explore for the like or dislike functionality:

(1) Prompt Engineering: Better prompt engineering can be **explored and configured** since the current **regeneration** of the bad response **does not fully capture the extent of the scores** of the other responses. Based on section 8.2, it can be observed that despite the first response **having a score of 1** and response 2 **having a score of 0.5** both seems to equally influence it. The **regenerated response** could have **better related** to response 1 if it was **on the topic of gardening**

(2) Memory system: A better memory system could be created so that when regenerating, **it is aware of all previously generated responses** so that it would not repeat itself.

9.3 Increasing Test Coverage:

9.3.1 Creating More Complex Test Cases:

Currently the test cases are still mostly simple ones. It would be useful to test more complex conversations especially longer ones with greater number of questions.

The current complex conversations are also made up of multiple questions of the same topic. It could be useful in **combining multiple questions** with **different main topics** to observe a greater variety of context retrieved out for each question through the query transformation function which breaks down the complicated conversation.

9.3.2 Improving User Studies:

Expand user studies to a **wider demographic** to gain more perspectives.

Having longer user studies where they are able to try the app out over a longer period of time could also be useful **to observe how the app** changes over time **since it should be better accustomed to** the user's habits and experiences over time. However, due to how short the current user studies are, **this could not be fully observed.**

Further, testing the **app from scratch** without any past mock data or conversations **(A Cold Start)** could be useful to observe initial **generation of responses** and how these responses change **the more the user uses it.**

9.4 Proposed Timeline For Next Steps Would Be:

Objective	Estimated Time
Response Generation: Trying out different models and storing model locally	**1 month**
Response Generation: Adding Time Meta Data	**1 month**
Response Generation: Reduce Latency by shortening output	**0.5 month**
Incorporating streaming of output & stop sequences	**0.5 month**
Cloud Storage	**0.5 month**
Like/Dislike Functionalities: Improving regeneration of response	**1 month**
User Study and Testing	**2 months**
Additional Functionalities To Consider	
Auto-complete text when user is typing	-
Lip-reading models	-
Explore Non-Connectivity (No internet)	-

References

[1]: ScienceDirect. (n.d.). *Mutism*. Mutism - an overview | ScienceDirect Topics. https://www.sciencedirect.com/topics/neuroscience/mutism

[2]: SpeechPathology. (2022, May 11). *Mutism: SLP graduate programs and the study of mutism*. https://www.speechpathologygraduateprograms.org/mutism/

[3]: Ability Central, A. (2024, February 7). *What keeps someone from talking? information you should know about muteness*. https://abilitycentral.org/article/what-keeps-someone-talking-information-you-should-know-about-muteness#:~:text=Muteness%20is%20a%20rare%20condition%20that%20can%20be%20either%20temporary,permanent%2C%20but%20many%20are%20misunderstood.

[4]: Masrur Sobhan. (n.d.). (PDF) a communication aid system for deaf and mute using vibrotactile and visual feedback. https://www.researchgate.net/publication/336877461_A_Communication_Aid_System_for_Deaf_and_Mute_using_Vibrotactile_and_Visual_Feedback

[5]: Munir, M. B., Alam, F. R., Ishrak, S., Hussain, S., Shalahuddin, Md., Islam, M. N., Muhaimin Bin MunirMilitary Institute of Science and Technology, D., Fariha Raisa AlamMilitary Institute of Science and Technology, D., Shadman IshrakMilitary Institute of Science and Technology, D., Sonaila HussainMilitary Institute of Science and Technology, D., Md. ShalahuddinMilitary Institute of Science and Technology, D., & Muhammad Nazrul IslamMilitary Institute of Science and Technology, D. (2021, September 22). A machine learning based Sign Language Interpretation System for communication with deaf-mute people: Proceedings of the XXI international conference on human computer interaction. ACM Other conferences. https://dl.acm.org/doi/10.1145/3471391.3471422

[6]: *What are the different types of sign language?*. Sign Solutions. (2024, June 5). https://www.signsolutions.uk.com/what-are-the-different-types-of-sign-language/

[7]: *Proloquo4Text*. AssistiveWare. (n.d.). https://www.assistiveware.com/products/proloquo4text

[8]: Limited, T. B. (2011, January 6). *Predictable*. App Store. https://apps.apple.com/us/app/predictable/id404445007

[9]: *CHATGPT and disability: Benefits, concerns, and future potential*. Rocky Mountain ADA. (n.d.). https://rockymountainada.org/resources/research/chatgpt-and-disability-benefits-concerns-and-future-potential

[10]: Exploring ai chatbot for spontaneous word retrieval in aphasia. (n.d.). https://aphasia.talkbank.org/publications/2023/Purohit23.pdf

[11]: Farid, A. (2024, February 20). *Building apps with Expo & React native: Pros & cons*. Upstack Studio. https://upstackstudio.com/blog/expo-react-native/

[12]: Agastya, A. (2024, January 4). *Decoding LLM performance: A guide to evaluating LLM applications*. Medium. https://amagastya.medium.com/decoding-llm-performance-a-guide-to-evaluating-llm-applications-e8d7939cafce

[13]: Google-Research. (n.d.). Google-Research/Bleurt: Bleurt is a metric for natural language generation based on transfer learning. GitHub. https://github.com/google-research/bleurt

[14]: Cleary, D. (2023, November 30). *Can you use llms as evaluators? an LLM evaluation framework*. Medium. https://medium.com/@dan_43009/can-you-use-llms-as-evaluators-an-llm-evaluation-framework-8681b400b110

[15]: *List of available metrics*. Ragas. (2024, October 3). https://docs.ragas.io/en/latest/concepts/metrics/available_metrics/

[16]: Anello, E. (2024, April 12). *How to improve rag performance: 5 key techniques with examples*. DataCamp. https://www.datacamp.com/tutorial/how-to-improve-rag-performance-5-key-techniques-with-examples

[17]: Lamers, R. (2023, August 9). *OpenAI's embedding model dethroned in MTEB by Baai General Embedding Model*. OpenAI's embedding model dethroned in MTEB by BAAI general embedding model. https://codingwithintelligence.com/p/openais-embedding-model-dethroned

[18]: Splore. (2023, November 3). *Different types of search explained: AI, keyword, Hybrid Search & More*. Medium. https://medium.com/@sploredotcom/different-types-of-search-explained-ai-keyword-hybrid-search-more-d5e24f74ef4d

[19]: Improve rag performance using cohere Rerank | AWS machine learning blog. (n.d.-b). https://aws.amazon.com/blogs/machine-learning/improve-rag-performance-using-cohere-rerank/

[20]: Sharma, H. (n.d.). *Techniques to enhance retrieval augmented generation (RAG)*. Community.aws. https://community.aws/content/2gp2m3BJcl9mSMWT6njCIQNiz0e/techniques-to-enhance-retrieval-augmented-generation-rag?lang=en

[21]: Santhosh, S. (2024, February 4). *How to improve rag(retrieval augmented generation) performance*. Medium. https://medium.com/@sthanikamsanthosh1994/how-to-improve-rag-retrieval-augmented-generation-performance-2a42303117f8

[22]: *Getting started*. ElevenLabs. (n.d.). https://elevenlabs.io/docs/api-reference/getting-started

[23]: Taivo Pungas. (2024, January 23). *Making GPT API responses faster*. https://www.taivo.ai/__making-gpt-api-responses-faster/#:~:text=There%20is%20a%20way%20to,can%20get%20faster%20answers%2C%20guaranteed.

[24]: gfbane23. (2023, November 9). *How can I improve response times from the openai API while generating responses based on our knowledge base?*. OpenAI Developer Forum. https://community.openai.com/t/how-can-i-improve-response-times-from-the-openai-api-while-generating-responses-based-on-our-knowledge-base/237169/5

Appendix

Appendix 3A: Code for text-to-speech

```python
# Text To Speech Response
@app.post("/text-to-speech/")
async def text_to_speech(response: str = Form(...)):
    audio_output = convert_text_to_speech(response)
    if not audio_output:
        raise HTTPException(status_code=500, detail="Failed to convert text to speech")

    audio_file_path = "C:\\Roydon\\Github\\FYP_Application\\MuteCompanion\\backend\\static\\response.mp3"
    with open(audio_file_path, "wb") as f:
        f.write(audio_output)

    return {"message": "Audio saved successfully"}
```

Eleven labs api:

```python
# Convert Text To Speech
def convert_text_to_speech(message):
    data = {
        "text": message,
        "voice_settings":{
            "stability": 0, # Adjusts the amount of emotion in the voice,
            "similarity_boost": 0,
        }
    }

    rachel_bot_id = "21m00Tcm4TlvDq8ikWAM"

    # Construct headers:
    headers = {
        "Content-Type": "application/json",
        "xi-api-key": ELEVEN_LABS_API_KEY,
        "accept": "audio/mpeg"
    }

    # Construct api endpoints.
    endpoint = f"https://api.elevenlabs.io/v1/text-to-speech/{rachel_bot_id}"

    # Send request
    try:
        response = requests.post(endpoint, json=data, headers=headers)
    except Exception as e:
        return

    # Handle Response
    if response.status_code == 200:
        return response.content
    else:
        return
```

Automatically play audio

```
async function playAudioFromResponse() {
  const audioUrl = `http://${config.apiBaseUrl}:8000/static/response.mp3`

  try {
    const { sound } = await Audio.Sound.createAsync(
      { uri: audioUrl },
      { shouldPlay: true }
    );

    await Audio.setAudioModeAsync({
      allowsRecordingIOS: false,
      playsInSilentModeIOS: true, // needed else in silent mode would not play
    });

    await sound.playAsync();
  } catch (error) {
    console.error("Failed to load or play audio:", error);
  }

}
```

Appendix 5.2A:

Conversation between Roydon And Yas on his horrible trip to Thailand


```
{
    "Response 1": {
        "Roydon": "Hey Yas, I'm still fuming about my terrible experience in Thailand, but you know what? I'm super excited about my upcoming trip to Jap
        "Yas": "Oh no, what happened in Thailand? But that's great to hear about Japan! What are you looking forward to the most?"
    },
    "Response 2": {
        "Roydon": "Well, everything that could go wrong in Thailand did go wrong. But in Japan, I can't wait to explore the beautiful temples and try aut
        "Yas": "That sounds like a rough time in Thailand, but Japan sounds amazing! Have you planned out your itinerary yet?"
    },
    "Response 3": {
        "Roydon": "Not yet, but I'm thinking of visiting Tokyo, Kyoto, and Osaka. I want to experience both the bustling city life and the serene country
        "Yas": "That sounds like a perfect balance! I'm sure you'll have a fantastic time exploring all those places."
    },
    "Response 4": {
        "Roydon": "I hope so! I've heard so many great things about Japan, and I'm determined to make the most of this trip.",
        "Yas": "Your positive attitude will surely make this trip unforgettable. I can't wait to hear all about your adventures when you get back!"
    },
    "Response 5": {
        "Roydon": "Thanks, Yas! I'm going into this trip with an open mind and a heart full of optimism. I know Japan will not disappoint.",
        "Yas": "Absolutely, positivity attracts positivity! I'm sure Japan will welcome you with open arms and unforgettable experiences."
    },
    "Response 6": {
        "Roydon": "I couldn't agree more! I'm ready to leave all the negativity from Thailand behind and embrace the beauty and culture of Japan.",
        "Yas": "That's the spirit, Roydon! Japan is waiting for you with open arms, ready to show you the best it has to offer."
    },
    "Response 7": {
        "Roydon": "I can't wait to immerse myself in everything Japan has to offer and create lasting memories that will overshadow my Thailand trip.",
        "Yas": "Your positive outlook will surely make this trip one for the books! Japan is lucky to have you as a visitor."
    },
    "Response 8": {
        "Roydon": "Thank you, Yas! I'm grateful for the opportunity to experience Japan and create new memories that will last a lifetime.",
        "Yas": "Cherish every moment and embrace the journey with open arms. Japan is about to become your new favorite destination!"
    },
```


Appendix 5.2B:

Conversations extracted from RAG to be Contexts.


```
ill allow the conversation to flow smoothly.\n\n        It must be in english. \n\n        Use the following previous conversations to assist in generating the 3 res
ponses:\n\n        {"Roydon": "Hey there! Did you catch the Arsenal game last night? What a thrilling match!", "John": "Hey Roydon! Yes, I watched it. Arsenal played
 really well, didn\'t they?"}{"Roydon": "Guess what, I just got a new pet dog!", "Jacob": "That\'s awesome! What breed is it?"}{"Roydon": "Not really. I was so angry
 the whole time, I couldn\'t enjoy anything.", "Dory": "I\'m sorry to hear that, Roydon. Maybe you can plan a better trip next time."}\n\n        An example of the 3
 generated response would be in the format of 1 single string "Response 1: what you generated Response 2: what you generated Response 3: what you generated" all in o
ne line.\n        '}, {'role': 'user', 'content': 'Roydon says: nullOther person says: Well, what have you been up to, Royden?'}]
Start is: 1
response_choices: Response 1: "I've been working on a new project lately."
Response 2: "I've been spending more time with my new pet dog."
Response 3: "I've been exploring different hobbies in my free time."
['"I\'ve been working on a new project lately."', '"I\'ve been spending more time with my new pet dog."', '"I\'ve been exploring different hobbies in my free time."'
]
```

Appendix 5.2B, Part 2:

Example of **1 row of** the **dataset** fed into evaluation metrics:

Question: Hows your new pet dog? What breed is he?

Answer (Generated): Response 1: He's doing great, thanks for asking! He's a golden retriever. Response 2: My new dog is wonderful, he's a golden retriever. Response 3: The new addition to my family, a golden retriever, is such a joy to have around.

Contexts (Extracted): '{"Roydon": "Guess what, I just got a new pet dog!", "Jacob": "That\'s awesome! What breed is it?"} … and more

GroundTruth: Response 1: He brings so much joy to my life. He is a golden retriever, and he's the cutest thing ever! Response 2: He brings so much joy to my life. He is a golden retriever, and he's the cutest thing ever! Response 3: He brings so much joy to my life. He is a golden retriever, and he's the cutest thing ever!

(** Note: Some groundtruth like the one above is repeated 3 times mostly for the other eval metrics. Unlike G-Eval where you can custom set your own metrics, the other would match based on how similar the answers generated are to the ground truth. **Although users may only have 1 answer in their mind** when replying to a query/conversation**,** since the **generated answer** generates 3 responses, **the ground truth needs to have 3 as well** for matching purposes.

Appendix 5.2C: BLEURT Score Non-RAG (First 8)

```
[-0.5593798160552979, -0.5643997192382812, -0.8871408700942993, -0.7929092049598694,
```

```
-0.8116310834884644, -1.0318307876586914, -0.49550342559814453, -0.9399088621139526,
```

Appendix 5.2D: BLEURT Score RAG (First 8)

```
[-0.7020672559738159, -0.5123528242111206, -0.8473215699195862, -0.09329019486904144,
```

```
-0.6893448829650879, -0.5007555484771729, -0.8085434436798096, -0.7549819350242615,
```

Appendix 5.2E: G-Eval Criteria

```python
correctness_metric = GEval(
    name="Relevance",
    #criteria="Determine whether the actual output matches the expected output as close as possible.",
    # NOTE: you can only provide either criteria or evaluation_steps, and not both
    evaluation_steps=[
        "Check whether the main content of the responses generated in 'actual output' are similar to the responses in the 'expected output'",
        """As long as one of the main content of the responses generated is similar to any of the expected output, the test case is considered correct.
        For example, if response 1 content is on a pet dog and it matches response 3 content of also a pet dog, give it a high score.
        The order of the responses is not important.""",
        "Evaluate mainly based on main content but do still give a higher score depending on similarity of responses."
    ],
    evaluation_params=[LLMTestCaseParams.INPUT, LLMTestCaseParams.ACTUAL_OUTPUT, LLMTestCaseParams.EXPECTED_OUTPUT],
    model="gpt-3.5-turbo",
)
```

Appendix 5.2F: G-Eval Old Criteria

```python
correctness_metric = GEval(
    name="Relevance",
    criteria="Determine whether the actual output matches the expected output as close as possible.",
    # NOTE: you can only provide either criteria or evaluation_steps, and not both
    evaluation_steps=[
        "Check whether the responses generated in 'actual output' are similar to the responses in the 'expected output'",
        "As long as one of the responses generated is similar to the expected output, the test case is considered correct",
        "As long as the main content is similar, it is considered okay"
    ],
    evaluation_params=[LLMTestCaseParams.INPUT, LLMTestCaseParams.ACTUAL_OUTPUT],
    model="gpt-3.5-turbo",
)
```

Appendix 5.2G: G-Eval Example Results From Old and New Criteria

Old Criteria: Scores are high, but inaccurate since the answers and ground truth don't really match.

	Scores	Reasons
0	0.638462	One of the responses is similar to the expected output, mentioning activities and interactions with people.
1	0.730661	One of the responses generated is similar to the expected output.
2	0.719481	Two out of the three responses generated are similar to the expected output.
3	0.878221	One of the responses generated (Take him for a walk in the park) is similar to the expected output.

	question
0	What have you been up to Roydon?
1	Woah really how is Arsenal doing right now then?
2	Nice what breed is your new pet dog?
3	So what you planning to do with your pet dog?

answer
Response 1: Not much, just relaxing at home. Response 2: I've been keeping busy with work and hobbies. Response 3: Just been spending time with family and friends.
Response 1: They are doing well this season Response 2: Arsenal is currently struggling a bit Response 3: Arsenal has been on a winning streak
Response 1: He's a golden retriever Response 2: He's a labrador Retriever Response 3: He's a German Shepherd
Response 1: Take him for a walk in the park Response 2: Teach him some new tricks Response 3: Give him a nice bath

ground_truth
Response 1: I've been watching Arsenal games hoping they will win. Response 2: I've been looking at a trip to Japan. Response 3: I just got a new pet dog. How about you?
Response 1: Arsenal is doing well, did you catch the match yesterday? Response 2: Arsenal is doing great and Aubameyang is a true asset to the team. Response 3: Arsenal is doing alright sinc
Response 1: He is a golden retriever, and he's the cutest thing ever! Response 2: He is a golden retriever, and he's so playful! Response 3: He is a golden retriever, and he's so fluffy!
Response 1: I'm planning to take him on walks and teach him some tricks. Response 2: I'm planning to take him to the park and play fetch with him. Response 3: I'm planning to take him to the

New Criteria: Scores are lower and reasons explains why accurately.

A	B	
	Scores	
0	0.193265	None of the main content in the actual output matches any of the main content in the expected output.
1	0.393131	The main content of the responses does not align with the expected output, but some similarities can be found.
2	0.725914	The main content of the responses in the 'actual output' is similar to the responses in the 'expected output'.
3	0.474375	One of the main content in the actual output (Response 3) matches a content in the expected output (Response 3).

A	B
	question
0	What have you been up to Roydon?
1	Woah really how is Arsenal doing right now then?
2	Nice what breed is your new pet dog?
3	So what you planning to do with your pet dog?

answer
Response 1: I've been keeping busy with work and spending time with family Response 2: Not much, just relaxing at home and reading some books Response 3: I've been working on a new project and trying
Response 1: They are having a decent season so far. Liverpool is currently leading the table though. Response 2: Arsenal is struggling a bit, but they have potential to turn it around. Response 3: Arsenal has h
Response 1: He's a Labrador Retriever Response 2: She's a German Shepherd Response 3: It's a Golden Retriever
Response 1: Take him for a walk in the park Response 2: Teach him some new tricks Response 3: Bring him to the dog park

ground_truth
Response 1: I've been watching Arsenal games hoping they will win. Response 2: I've been looking at a trip to Japan. Response 3: I just got a new pet dog. How about you?
Response 1: Arsenal is doing well, did you catch the match yesterday? Response 2: Arsenal is doing great and Aubameyang is a true asset to the team. Response 3: Arsenal is doing alright since Ben White is a great a
Response 1: He is a golden retriever, and he's the cutest thing ever! Response 2: He is a golden retriever, and he's the cutest thing ever! Response 3: He is a golden retriever, and he's the cutest thing ever!
Response 1: I'm planning to take him on long hikes on the mountain. Response 2: I'm planning to take him to the beach and watch him splash in the waves. Response 3: I'm planning for play dates with other dogs.

Appendix 5.2H: G-Eval Results (Top 10)

Non-rag (Just prompt engineering):

	Scores	Reasons
0	0.193265	None of the main content in the actual output matches any of the main content in the expected output.
1	0.393131	The main content of the responses does not align with the expected output, but some similarities can be found.
2	0.725914	The main content of the responses in the 'actual output' is similar to the responses in the 'expected output'.
3	0.474375	One of the main content in the actual output (Response 3) matches a content in the expected output (Response 3).
4	0.279621	None of the main content in the actual output responses match with the main content in the expected output responses.
5	0.010841	The main content of the responses generated in 'actual output' does not match the main content of the responses in the 'expected output'.
6	0.355077	The main content of the responses generated is not similar to the responses in the expected output. Actual output talks about creating channels for dog training t
7	0.008525	None of the main content in the actual output responses are similar to the main content in the expected output responses.
8	0.331011	The main content of the responses generated in 'actual output' does not match the main content of the responses in the 'expected output'.
9	0.666652	The main content of response 3 is similar to the expected output, mentioning a Golden Retriever as the breed.

With rag implemented:

A	B	C
	Scores	Reasons
0	0.196816	Only one response in the actual output is somewhat similar to one of the expected responses (response 3 mentioning a new project and a new pet dog).
1	0.317145	The main content of the actual output responses does not closely match the main content of the expected output responses.
2	0.702204	The main content of the responses generated in 'actual output' are similar to the responses in the 'expected output'.
3	0.641938	Two of the main content in the actual output match with the expected output.
4	0.733519	The main content of the responses in 'actual output' are similar to the responses in 'expected output'.
5	0.190779	The main content of the responses in the actual output does not match the main content of the responses in the expected output.
6	0.303652	The main content of the actual output responses about having a pet dog matches the expected output responses about having a pet dog, but the specific c
7	0.409588	None of the main content of the responses in 'actual output' are similar to the responses in the 'expected output'.
8	0.465797	One of the main contents of the responses (Response 1) is similar to one of the expected outputs.
9	0.759117	The main content of the responses in the actual output is similar to the main content of the responses in the expected output.

Appendix 5.2I: RAGAS Results (Top 10)

Non-rag (Just prompt engineering):

answer_relevancy	answer_correctness	context_precision	context_recall
0	0.210734963	0	0
0	0.371508139	0	0
0	0.23018674	0	0
0	0.226273349	0	0
0	0.222118755	0	0
0	0.202416468	0	0
0	0.53835768	0	0
0.937190868	0.205021816	0	0
0	0.465978361	0	0
0	0.217555312	0	0

With rag implemented:

answer_relevancy	answer_correctness	context_precision	context_recall
0	0.211641853	0	0.333333333
0	0.218801925	1	0.333333333
0	0.533986303	1	1
0.90197829	0.638223798	1	0.666666667
0.995869797	0.211999336	1	1
0.911640096	0.221126607	1	0.333333333
0.857624438	0.462087405	1	0.5
0	0.432102901	1	0.333333333
0	0.59602954	1	0
0.934280508	0.606587908	1	0.5

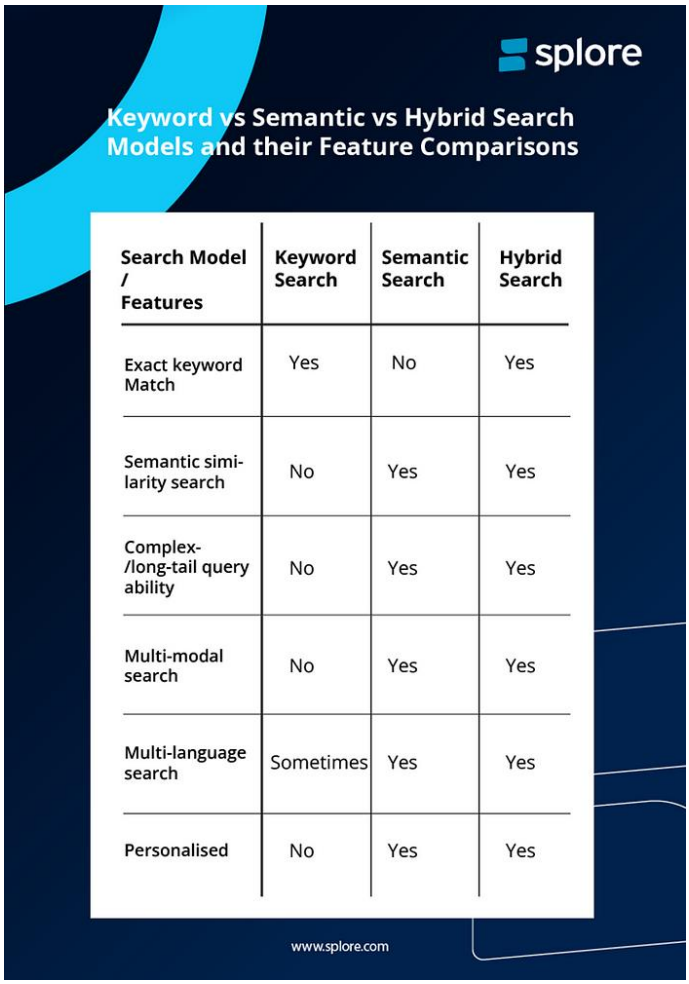
Appendix 5.3A: G-Eval Results (Top 10)

G-Eval Scores	Reasons
0.235683379	None of the responses are similar.
0.489139596	Two main content points match with expected output.
0.749030158	The main content of the responses generated in 'actual output' are similar to the responses in the 'expected output'.
0.62871178	Response 3 in the Actual Output matches the main content of Response 1 in the Expected Output regarding taking the dog on long hikes.
0.747395413	The main content of the responses in the 'actual output' is very similar to the responses in the 'expected output'.
0.622447904	The main content of the responses generated in 'actual output' mostly matches the main content of the responses in the 'expected output'.
0.15493468	The main content of the responses in 'actual output' are not similar to the responses in 'expected output'.
0.625682188	The main content of the actual output responses is similar to the main content of the expected output responses, mentioning the negative experiences in Thailand.
0.480523428	The main content of the responses in the 'actual output' does not align with the main content of the responses ir the 'expected output'.
0.811532588	The main content of the responses in the 'actual output' is similar to the responses in the 'expected output'.

Appendix 5.3B: RAGAS Results (Top 10)

answer_relevancy	answer_correctness	context_precision	context_recall
0	0.205554716	0	0.333333333
0	0.977673102	1	0.333333333
0.941946419	0.66168658	1	1
0.868114466	0.873321757	1	0.666666667
0	0.972729473	1	1
0.911640096	0.21749295	1	0.333333333
0.867220225	0.565412307	1	0.5
0	0.661872483	1	0
0	0.783908103	1	0
0.938489003	0.532539451	1	1

Appendix 5.4A: Benefits Of Hybrid Search



Search Model / Features	Keyword Search	Semantic Search	Hybrid Search
Exact keyword Match	Yes	No	Yes
Semantic simi-larity search	No	Yes	Yes
Complex-/long-tail query ability	No	Yes	Yes
Multi-modal search	No	Yes	Yes
Multi-language search	Sometimes	Yes	Yes
Personalised	No	Yes	Yes

Appendix 5.4B: Code for hybrid search

```
# Initiate retriever
retriever_vectordb = loaded_faiss_vs_hf_v1.as_retriever(search_kwargs={"k": 3})
keyword_retriever = BM25Retriever.from_documents(documents)
keyword_retriever.k =  3
ensemble_retriever = EnsembleRetriever(retrievers=[retriever_vectordb,keyword_retriever],
                                       weights=[0.5, 0.5])
```

Appendix 5.4C: Code for bge embedding

```
embeddings=HuggingFaceInferenceAPIEmbeddings(
    api_key=os.environ['HUGGING_FACE_ACCESS_TOKEN'],
    model_name='BAAI/bge-base-en-v1.5'
)
```

Appendix 5.4D: Code for Cohere Reranking

```
# Cohere reranking
from langchain.retrievers.document_compressors import CohereRerank
from langchain.retrievers import ContextualCompressionRetriever

compressor = CohereRerank(cohere_api_key=os.environ['COHERE_API_KEY'])
compression_retriever = ContextualCompressionRetriever(
    base_compressor=compressor, base_retriever=ensemble_retriever
)
```

Appendix 5.4E: Code for Weights Of Ensemble Retriever

```
ensemble_retriever = EnsembleRetriever(retrievers=[retriever_vectordb,keyword_retriever],
                                       weights=[0.6, 0.4])
```

```
ensemble_retriever = EnsembleRetriever(retrievers=[retriever_vectordb,keyword_retriever],
                                       weights=[0.4, 0.6])
```

Appendix 5.5A: Code for GetTopic Functionality In Topic Filtering

```
def getTopic(meta_content, query):
    # Learning instructions
    instruction = {
        "role": "system",
        "content": meta_content,
    }

    #print("Query is: " + query)

    # Initialize messages
    messages = []

    # Add learn instruction to message array
    messages.append(instruction)

    user_message = {
            "role": "user",
            "content": query
    }

    messages.append(user_message)

    openai.api_type = 'openai'
    openai.api_key = os.environ["OPENAI_API_KEY"]
    openai.organisation= os.environ["OPEN_AI_ORG"]

    raw_response = openai.chat.completions.create(
            model="gpt-3.5-turbo",
            messages = messages,
    )
    topic = raw_response.choices[0].message.content

    return topic
```

Appendix 5.5B: Prompt Engineering For Identification Of Topics

```
# Initiate meta content
meta_content = """
You would be assisting in identifying topics from a snippet of conversation. I would supply the conversation directly.
Interpret the main topic of the conversation and return the main topic.

Do not give multiple topics such as football/soccer. Only give one main topic.

For example if the conversation is:
        {"Roydon": "Can't wait for the new football season to start, hoping for a great one for Arsenal!", "John": "Hey Roydon! Yeah, it's
        always exciting to see how your team will perform."}

        football

        Example 2:
        {"Roydon": "I'm planning to go on a trip to Japan next year", "John": "That's awesome! Japan is such a beautiful country."}

        travel

        Example 3: If no main topic can be determined such as a greeting
        {"Roydon": "Hey there! How are you doing?", "John": "Hey Roydon! I'm doing great, how about you?"}

        general
"""
✓ 0.0s
```

Appendix 5.5C: singularize_and_lower Function

```
# Filtering functions
def singularize_and_lower(topic):
    # Initialize the WordNet Lemmatizer
    lemmatizer = WordNetLemmatizer()

    # Lowercase the topic
    topic = topic.lower()

    # Lemmatize the word with correct POS tag
    pos = get_wordnet_pos(topic)
    topic_lemmataized = lemmatizer.lemmatize(topic, pos)

    # Extra check for words like crotcheting
    topic = remove_ing(topic_lemmataized)
    processed_words = [is_valid_word(word) for word in topic.split()]
    if not all(processed_words):
        return topic_lemmataized

    return topic
```

```
def get_wordnet_pos(word):
    """Map POS tag to first character lemmatize() accepts."""
    tag = nltk.pos_tag([word])[0][1][0].upper()
    tag_dict = {"J": wordnet.ADJ,
                "N": wordnet.NOUN,
                "V": wordnet.VERB,
                "R": wordnet.ADV}

    return tag_dict.get(tag, wordnet.NOUN)
```

```
def remove_ing(topic):
    if topic.endswith('ing'):
        return topic[:-3]
    return topic
```

```
def is_valid_word(word):
    """Check if a word is valid by looking it up in WordNet."""
    return bool(wordnet.synsets(word))
```

Appendix 5.5D: Inflect Package For Preprocessing Topics

```
import inflect

def singularize_and_lower(topic):
    # Create an inflect engine for handling plurals
    engine = inflect.engine()

    # Lowercase the topic
    topic = topic.lower()

    # Singularize the topic (convert plurals to singular), returns false if not noun
    topic = engine.singular_noun(topic) if engine.singular_noun(topic) else topic

    return topic

print(singularize_and_lower("pets"))
```

✓ 0.0s

```
pet
```

Appendix 5.5E: NLTK Package For Preprocessing Topics

```
def singularize_and_lower_lemmatize(topic):
    # Initialize the WordNet Lemmatizer
    lemmatizer = WordNetLemmatizer()

    # Lowercase the topic
    topic = topic.lower()

    # Lemmatize the word with correct POS tag
    pos = get_wordnet_pos(topic)
    topic_lemmataized = lemmatizer.lemmatize(topic, pos)

    # Extra check for words like crotcheting
    topic = remove_ing(topic_lemmataized)
    processed_words = [is_valid_word(word) for word in topic.split()]
    if not all(processed_words):
        return topic_lemmataized

    return topic
```

```
print("lematization results: ", singularize_and_lower_lemmatize("crotcheting"))
print("inflect results: ", singularize_and_lower("crotcheting"))
```

```
lematization results:  crotchet
inflect results:  crotcheting
```

```
print("lematization results: ", singularize_and_lower_lemmatize("flying"))
print("inflect results: ", singularize_and_lower("flying"))
```

```
lematization results:  fly
inflect results:  flying
```

```
print("lematization results: ", singularize_and_lower_lemmatize("crotchets"))
print("inflect results: ", singularize_and_lower("crotchets"))
```

```
lematization results:  crotchet
inflect results:  crotchet
```

Appendix 5.5F: filter_list Function

```python
def filter_list(docs_rel, topic):
    # Filter according to the topic
    filtered_docs = []
    final_docs = []
    general_topic = {}

    if singularize_and_lower(topic) == "general":
        for doc in docs_rel:
            if singularize_and_lower(doc.metadata['topic']) not in general_topic:
                general_topic[singularize_and_lower(doc.metadata['topic'])] = 1
                filtered_docs.append(doc)
            else:
                continue
    else:
        for doc in docs_rel:
            if(singularize_and_lower(doc.metadata['topic']) == singularize_and_lower(topic)):
                filtered_docs.append(doc)

    if len(filtered_docs) > 2:
        final_docs = filtered_docs[:3]
        return final_docs
    else:
        count = 3 - len(filtered_docs)
        final_docs = filtered_docs
        position = 0
        for i in range(count):
            if(position == len(docs_rel)):
                break
            if(docs_rel[position] in filtered_docs):
                i = i-1
                position += 1
                continue
            else:
                final_docs.append(docs_rel[position])# need to change so that it wont be same obtained
                position+=1

        return final_docs
```

Appendix 5.5G: Summary of topic filtering usage

```python
# Get contexts for query
docs_rel=ensemble_retriever.get_relevant_documents(query)

topic_interpreted = getTopic(meta_content, query)

final_docs = filter_list(docs_rel, topic_interpreted) # Still top 3
```

Appendix 5.6A: process_query Function

```python
def process_query(query):
    # Initialize the Punkt tokenizer
    tokenizer = PunktSentenceTokenizer()

    # Tokenize the text into sentences
    sentences = tokenizer.tokenize(query)

    # Check if the query is a simple one (contains only one sentence)
    if len(sentences) == 1:
        return sentences  # Return the single sentence wrapped in a list

    # Combine into segments
    combined = ''
    segments = []

    # Flag to check if last sentence was a question
    last_was_question = False

    # Loop through each sentence and decide whether to start a new segment
    for sentence in sentences:
        # If last sentence was a question and current isn't directly a question,
        # start a new segment
        if last_was_question and sentence.strip().endswith('?'):
            segments.append(combined.strip())
            combined = sentence + ' '
            last_was_question = False
        else:
            combined += sentence + ' '

        # Check if current sentence ends with a question mark
        if sentence.strip().endswith('?'):
            last_was_question = True

    # Append the last segment if there's any remaining text
    if combined:
        segments.append(combined.strip())

    return segments
```

✓ 0.0s

Appendix 5.6B: Integrating the process_query function

```python
# Get contexts for query
contexts = ""
query_split = process_query(query)
for i in query_split:
    # Obtain top 3 filtered docs
    docs_rel=ensemble_retriever.get_relevant_documents(query)
    topic_interpreted = getTopic(meta_content, query)
    final_docs = filter_list(docs_rel, topic_interpreted) # Still top 3
    for context in final_docs:
        contexts += context.page_content
```

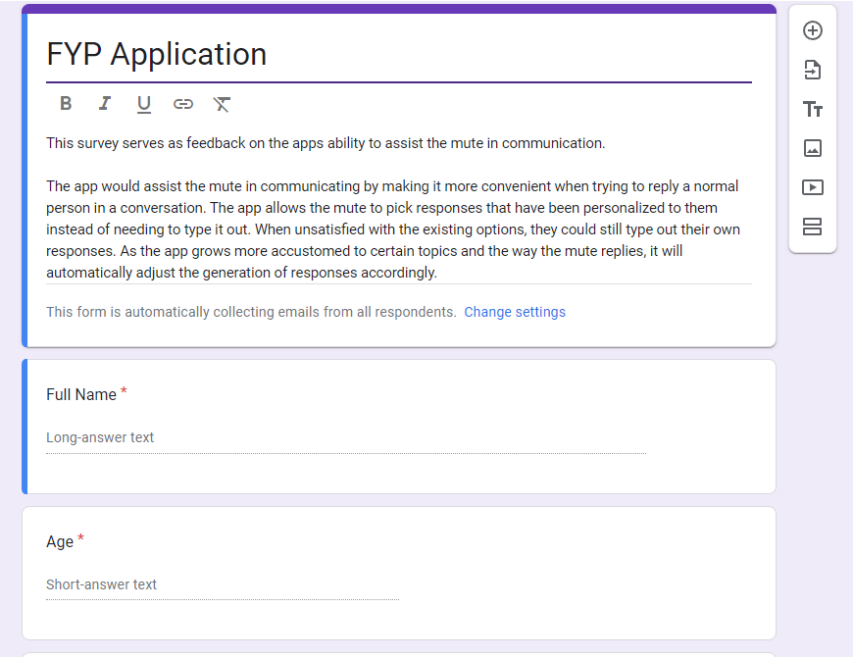
Appendix 5.7A: Final System for context retrieval

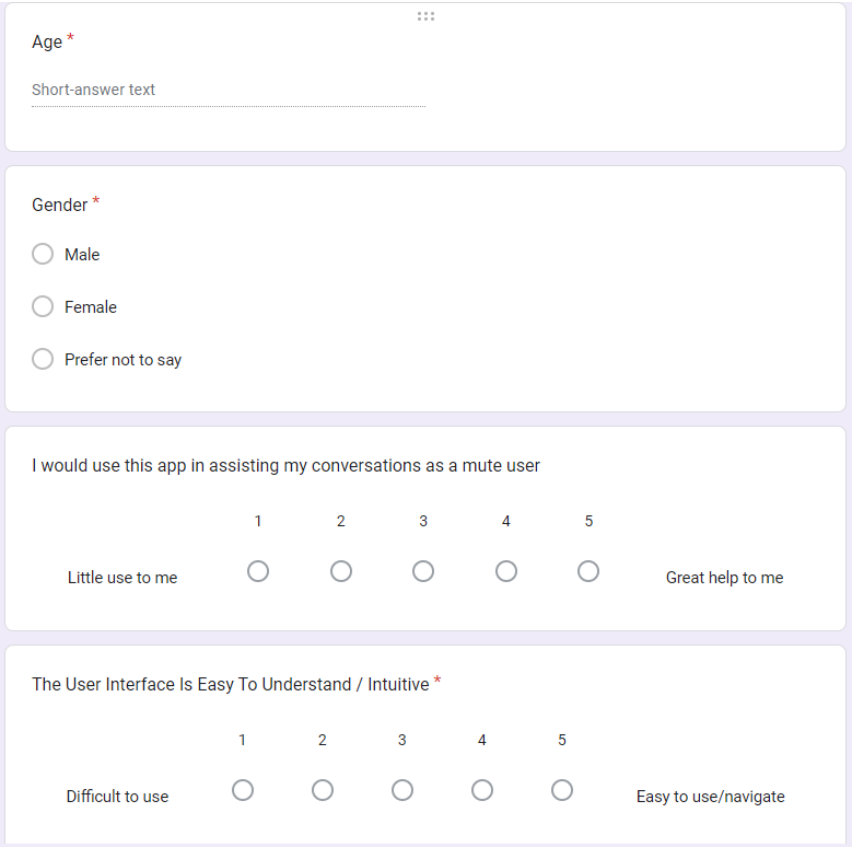
```
# for con in context:
#     contexts += con.page_content

contexts = ""
query_split = process_query(query)
for i in query_split:
    # Obtain top 3 filtered docs
    try:
        docs_rel=ensemble_retriever.invoke(i)
        topic_interpreted = getTopic(meta_content, i)
        final_docs = filter_list(docs_rel, topic_interpreted) # Still top 3
        global final_topic
        final_topic = topic_interpreted
        for context in final_docs:
            contexts += context.page_content
    except Exception as e:
        print("Entered fail safe")
        context = fail_safe_retriever.similarity_search(query, k=3)
        for con in context:
            contexts += con.page_content
```

Appendix 7A: Survey for user study

https://forms.gle/fL6Kvw11Jfmp9wmg6





Back To Content Page

The user interface is visually appealing *

	1	2	3	4	5	
Poor visual appeal	○	○	○	○	○	Visually Appealing

The responses generated was something I would pick when replying (On an average)

	1	2	3	4	5	
Picked little responses (Typed a lot)	○	○	○	○	○	Selected all responses

The responses generated were personalized and contained non-general information. *

	1	2	3	4	5	
Responses were general	○	○	○	○	○	Highly Personalize (Based on past convos)

What did you like about the application

Long-answer text

Areas of improvement on responses generated

Long-answer text

Areas of improvement to application

Long-answer text

Appendix 7B: Instruction for User Study

This user study would take about 1-2 hours. After the study, a feedback form would be given to you for completion.

The user study would require you to first read through these mock data. The mock data are conversations that "you" had before with others. These data are used as your **"past experiences"** and are used to personalize the responses generated.

You would also be shown your hobbies and your past experiences summarized. For example:

User Profile:

Hobbies: (1) Crotchet (2) Gardening (3) Gym (4) Works at the hospital as a nurse

Recent History: (1) Crotcheting new projects, (2) Picked up gardening as a hobby, (3) Planning a trip to Australia, (4) Goes to anytime fitness, (5) Trying to cope with workload

You would then need to play 2 roles:

(1) You would take on the role of the **mute user**

- I would play the role of the normal person who will lead the conversation
- Pick the responses according to your persona (hobbies and past experiences)
- I would try to lead the conversation in a way that would cover all your hobbies so that you can observe how the responses are generated.

(2) A role reversal would be done: You would now play the role of the normal person, and I would be the mute user. I would do the selecting of responses.

- You can lead the conversation in any direction including reusing previously asked questions.
- This is to test how the application responds to random topics.
- This also tests how the application responds to the same questions asked and if it learns from mistakes it made previously.

Appendix 7C: Survey Results

https://docs.google.com/spreadsheets/d/1_54-NbDVEMqZqkuKEMjzq_OHt-XRXj-KGjbb9vL9hX0/edit?usp=sharing

Average Scores: Out Of 5

I would use this app in assisting my conversations as a mute user	The User Interface Is Easy To Understand / Intuitive	The user interface is visually appealing
4.6	4.4	3.6

The responses generated was something I would pick when replying (On an average)	The responses generated were personalized and contained non-general information.
4.4	4.55

Appendix 7D: Summary Of User Study Feedback

	Participant Profiles	**Important Feedback**
1	**Cheryl:** Age: 20 Gender: F - Volunteers under Welfare Service Club for the intellectually disabled	**Good areas:** - The responses were very **personalized**. - **Convenient** for the mute since they **do not** need to type out **repeated conversations**. **Improvement:** - Currently unable to **edit generated responses**. - **Rating** of responses would be good. - The **design** of the generated responses may need to be improved. For long conversations generated the words can seem a little overwhelming.
2	**Yu Min:** Age: 20 Gender: F - Volunteers under Welfare Service Club for the intellectually disabled	**Good areas:** - **Convenient** for the mute since they **do not** need to **repeat** what they said to others before. - The responses generated were **interesting** since they contained my **hobbies and recent activities** done. **Improvement:** - With the **reset issue (Refer below to Yu Min's)**, there needs a button to maybe clear and regenerate responses - Unable to **edit generated responses**. - The **design** can be improved. - Allow the **rating of responses**.
3	**Kistina:** Age: 20 Gender: F	**Good areas:** - **Speeds up communication** with the mute by a considerable amount.

	- Volunteers under Welfare Service Club for the intellectually disabled	- Responses generated are **non-general** and related to **past experiences**. Makes it easier for the mute to pick instead of typing. **Improvement:** - **Unable** to **edit** generated responses.
4	**Clarence:** Age: 21 Gender: M - Volunteers under Welfare Service Club for the intellectually disabled	**Good areas:** - Responses generated are **personal**. **Improvement:** - **Unable** to **edit** generated responses. - **Unable** to **regenerate** new responses. - **Design** of the application could be better **especially for the responses**
5	**Chee Seng:** Age: 23 Gender: M - Volunteers under Welfare Service Club for the intellectually disabled	**Good areas:** - **Convenient** for the mute since they can now **communicate faster** and **may not need to repeat themselves.** - Helps **facilitate** communication with others who may not have interacted with the mute before. - **Personalized responses** generated makes it easier for the mute to pick. - **Application grows with the mute** user the more they use it. **Improvement:** - **Unable** to **generate new responses** again to the same conversation if **unsatisfied**. - **Unable** to **edit generated responses** since mute user may just want to change a part of it. - **Rating of responses** would be good.
6	**Rigel:** Age: 25 Gender: M - Volunteers under Welfare Service Club for the intellectually disabled	**Good areas:** - **Convenient** to **pick personalized responses** that are related to my **hobbies** and **past experiences**. **Improvement:** - There could be a way to **regenerate responses**. - There could be a way to **change / edit responses** instead of needing **to type it out fully**.
7	**Tiffany:** Age: 20 Gender: F - Volunteers under Welfare Service Club for the intellectually disabled	**Good areas:** - Responses generated were **related to my hobbies** and **past experiences**. - Having **3 responses** to pick made it such that it's **not too overwhelmin**g yet having a **good number of choices.** **Improvement:** - There could be a way **to change / edit responses** instead of needing to **type it out fully**.

		- Maybe a way to **rate responses generated**.
8	Lee Hang: Age: 25 Gender: M - Software Engineer at Citibank	**Good areas:** - The responses generated were **related** to what I **experienced before** and **conversations** I had with others. - **Convenient for the mute** to pick since the conversations were relatable. - The **mute does not have to repeat himself** unnecessarily. **Improvement:** - **Unable** to **generate new responses** to the same question. - **Unable** to **edit generated responses**. The mute may want to change some parts of it.
9	Gregory: Age: 25 Gender: M - Software Engineer at JP Morgan.	**Good areas:** - The responses **generated contained my hobbies** and **experiences**. **Improvement:** - **Unable** to **refresh** and **get new responses**. - **Unable** to **edit generated responses**. - Maybe an **auto complete functionality** for users typing.
10	Chiam Chuen: Age: 25 Gender: M - Software Engineer at Citibank	**Good areas:** - The responses generated were **personal** - **3 responses** given was a **good number** **Improvement:** - Perhaps a way to **regenerate** or **edit responses**
11	Kai Sheng: Age: 25 Gender: M - Software Engineer at OKX	**Good areas:** - The responses generated were **relatively fast** and **contained information** about my **hobby** and **past interests** - **Convenient** for the mute user with **enough options** to choose from - **Overall, a great application** for those who are mute **Improvement:** - Sometimes the **generated responses** may not be **fully** what the user would pick. Hence a way **to edit them** could be good
12	Sydney: Age: 23 Gender: F	**Good areas:** - Responses generated felt **personalized** - **Convenient** for mute users **without them needing to repeat** certain topics/conversations

	- Software Engineer at Bank Of America	**Improvement:** - A way to **edit** the **generated responses** - A way to **regenerate** responses - A way for the **normal person to also type** if they would like to
13	Sin Yee: Age: 25 Gender: F - Nurse at NUHS	**Good areas:** - The responses **generated** was shocking as most were **very personalized**. - The **storage functionality** was good to **store past conversations individually** for each person. - **Convenient** for the mute since the they **do not need to repeat** same conversations to others. **Improvement:** - For some cases, the **generated responses** were **overall good** but there might be **parts** that the **mute would like to edit or change**. - **Currently** not able to **regenerate new responses**. - Perhaps the **ability** to **favourite certain responses.** - There should be a way to **rate the responses.**
14	Ryan Teo: Age: 24 Gender: M - Computer Science Student	**Good areas:** - The responses were **generated very quickly** - **Highly personalized** - **Convenient** and **easy** for a mute user to use. **Improvement:** - **Regenerate responses** - **Ability** to **like** or **dislike** responses
15	Qian Wei: Age: 21 Gender: F - Psychology Student	**Good areas:** - The responses were **related** to my **hobbies** and **interests** - **Very convenient** and **multiple options** to choose from **Improvement:** - Some of the responses generated may not be what I would want to **say fully** but **majority** of it is **right**. If there is an **option** to **edit the responses** it would be **good**.
16	Eammon: Age: 23 Gender: M - Mechanical Engineering Student	**Good areas:** - Very **convenient** and **easy** to use - With **high personalization**, mute users are **more likely** to **pick** the **responses** rather than **type themselves** - The app **grows** with you **the more you use** it

		Improvement: - Maybe a way to **generate responses multiple times**. - A way to **edit existing responses**
17	Yi An: Age: 24 Gender: M - Firmware Test Engineer	**Good areas:** - The responses **generated feels personalized** to **past experiences** - The **more** you use the app the **better the responses generated** - The app is very **intuitive** and does **not require much instructions** to use **Improvement:** - There could be **security concerns** regarding personal information
18	Yi Yin: Age: 22 Gender: F - Banking & Finance Student	**Good areas:** - The application **was easy to use** and **provides a lot of convenience** for mute users. - Responses generated were **personal** and made it **easy to pick from** instead of needing to type out. - **Clean user interface** - **Application** grows more **accustomed** to how I would **speak the more I use it**. **Improvement:** - A way to **generate responses multiple times** or a way to **edit the current responses**.
19	Justin: Age: 23 Gender: M - Banking & Finance Student	**Good areas:** - The application was **easy** to use. - **Convenient for the mute especially** since responses generated felt **personalized**. **Less likely** for the mute to **type out.** - Besides generated responses, there was still a way for the mute to **type out their own responses**. **Improvement:** - A way to **rate the responses** generated.
20	Ganlin: Age: 25 Gender: F - Data Analyst	**Good areas:** - The responses generated was related to my hobbies and past experiences - Convenient and easy to use **Improvement:** - There should be a way to edit existing responses since I may like the generated responses but not fully - A way to like or dislike the responses generated

Appendix 7.3A: Sin Yee User Study

Normal Person	Responses Generated	Response Selected	Time
Hi Sin Yee, how have you been doing? (Appendix 7.3A)	1. Hi Gary, I've been keeping busy with work and my hobbies. 2. Hi Gary, I've been doing well, thank you for asking. 3. Hi Gary, it's been a mix of tiring and fulfilling lately.	Hi Gary, I've been keeping busy with work and my hobbies.	Audio To Text: 2s Response: 7s Total: **10s**
Well what hobbies have you been doing recently then? (Appendix 7.3A)	1. I've picked up gardening as a new hobby. 2. I've started crocheting recently, it's been really fun! 3. I've been trying out painting as a new hobby, it's quite relaxing. **Important contexts:** {"Jackson": "What have you been up to Sin Yee!", "Sin Yee": "Hey Jackson, guess what? I've picked up gardening as a new hobby!"} {"Xavier": "Hey Sin Yee, I heard you've taken up crocheting as a hobby! That's awesome!", "Sin Yee": "Yes, I'm really excited to start! Do you have any tips for beginners?"}	I've started crocheting recently, it's been really fun!	Audio To Text: 1.1s Response: 3.3s Total: **4.4s**
Whoa, crotcheting is kind of hard but what have you recently been crotcheting then? (Appendix 7.3B)	1. I made a cozy scarf for winter. 2. I crocheted a cute amigurumi animal as a gift for a friend 3. I recently finished crocheting a colorful blanket for my bed.	**(Wrote out herself)** Well I just made my very first bee but was thinking on starting on a new project do you have any suggestions?	Audio To Text: 1.1s Response: 3.7s Total: **4.8s**
Nice what about the tools necessary do you have all your crotcheting tools?	1. Yes, I've got all my crochet hooks and yarn ready to go. 2. I have a complete set of crochet tools, including stitch markers and a yarn needle. 3. I've prepared all my crochet supplies, like scissors and a tape measure, for my tote bag project. **Important contexts:** {"Xavier": "That's great to hear! To start, you'll need a crochet hook and some yarn. Have you got those already?", "Sin Yee": "Yes, I've got	Yes, I've got all my crochet hooks and yarn ready to go.	Audio To Text: 2s Response: 2.6s Total: **4.6s**

	them ready! What's the first thing I should try to make?"}		
Nice! Besides crotcheting, what other hobbies have you started doing? (**Appendix 7.3C**)	1. I have been learning how to paint with watercolors. 2. I recently started gardening and growing my own vegetables. 3. I've taken up cooking as a new hobby and experimenting with different recipes. **Important Contexts:** {"Jackson": "What have you been up to Sin Yee!", "Sin Yee": "Hey Jackson, guess what? I've picked up gardening as a new hobby!"}	I recently started gardening and growing my own vegetables	Audio To Text: 1.7s Response: 4.1s Total: **5.8s**
Whoa nice! What have you been growing in your garden then? (**Appendix 7.3C**)	1. I've started with some herbs like basil and mint. 2. I'm growing a variety of vegetables like tomatoes, cucumbers, and bell peppers. 3. Currently, my garden has carrots, lettuce, and radishes growing **Important Contexts:** {"Jackson": "That's really cool! What are you planning to grow in your garden?", "Sin Yee": "I've started with some herbs like basil and mint, and I'm thinking of planting some vegetables too."}	I've started with some herbs like basil and mint.	Audio To Text: 1.5s Response: 2.7s Total: **4.2**s
Have you faced any difficulties while gardening so far? (**Appendix 7.3E**)	1. I've had some challenges with pests eating my plants, but I'm finding ways to manage them. 2. Watering the plants properly has been a bit tricky, but I'm getting the hang of it. 3. Keeping the soil nutrient-rich has been a struggle, but I'm learning about composting to improve it. **Important Contexts:** {"Jackson": "That sounds like a great idea! Have you encountered any challenges so far?", "Sin Yee": "Yes, I've been struggling a bit with keeping the pests away from my plants, but I'm learning as I go."}	I've had some challenges with pests eating my plants, but I'm finding ways to manage them.	Audio To Text: 1.1s Response: 3s Total: **4**s

Wow where do you get these suggestions from? **(Appendix 7.3F)**	1. I often find these ideas by reading gardening blogs and watching videos online. 2. Some of these suggestions come from fellow gardeners in online gardening communities I'm a part of. 3. I get tips and advice from books on organic gardening and attending workshops at local gardening centers.	I often find these ideas by reading gardening blogs and watching videos online.	Audio To Text: 2.6s Response: 3s Total: **5.6**s
Do you have any suggestions for a beginner like me? **(Appendix 7.3G)**	1. Starting with easy-to-grow plants like herbs and lettuce can be a good way to gain confidence in gardening. 2. A simple beginner project could be growing cherry tomatoes or peppers in containers on your patio or balcony. 3. Consider investing in quality gardening tools like a trowel, pruners, and gloves to make your gardening experience more enjoyable. **Contexts extracted are all irrelevant:** {"Xavier": "A good beginner project is a simple dishcloth or a scarf. It helps you practice the basic stitches. Have you looked up any tutorials online?", "Sin Yee": "I haven't yet, but I'll definitely check them out. Thanks for the suggestion!"} {"Xavier": "No problem! YouTube has some great tutorials for beginners. Let me know if you need any help understanding the stitches.", "Sin Yee": "Will do! I'm excited to get started and create something with my own hands."} {"Matt": "I'm thinking of starting with some cardio and strength training to build up my fitness level. Do you have any tips for beginners like me?", "Sin Yee": "Definitely! Start slow and gradually increase the intensity of	Starting with easy-to-grow plants like herbs and lettuce can be a good way to gain confidence in gardening.	Audio To Text: 1.3s Response: 2.7s Total: **4s**

	your workouts. It's also important to listen to your body and rest when needed. Have you set a specific fitness goal in mind?"}		
Okay, thank you for the tips. I've also heard you recently started gyming? **(Appendix 7.3H)**	1. Yes, I've been hitting the gym to work on building strength and endurance. 2. I've been incorporating both cardio and strength training into my workout routine at the gym. 3. Gym sessions have been great for destressing and improving my overall fitness level. **Important Contexts:** {"Matt": "I'm thinking of starting with some cardio and strength training to build up my fitness level. Do you have any tips for beginners like me?", "Sin Yee": "Definitely! Start slow and gradually increase the intensity of your workouts. It's also important to listen to your body and rest when needed. Have you set a specific fitness goal in mind?"}	Yes, I've been hitting the gym to work on building strength and endurance.	Audio To Text: 1.3s Response: 3s Total: **4.3s**
Well which gym do you go to? **(Appendix 7.3I)**	1. I go to AnyTime Fitness - they have state-of-the-art equipment and 24/7 access, which fits well with my schedule. 2. I've chosen AnyTime Fitness because they offer personal trainers to help me reach my fitness goals. 3.AnyTime Fitness is where I work out, it's convenient and helps me stay committed to my fitness journey. **Important Contexts:** {"Matt": "I'm planning to visit the gym this weekend to check out their facilities and maybe sign up for a membership. I'm feeling really excited about this new chapter in my life.", "Sin Yee": "That's great to hear! Once you become a member, we can even plan some workout sessions together to keep each other accountable. It's always more fun	**(Wrote out herself)** AnyTime Fitness. Its great there!	Audio To Text: 2.1s Response: 2.9s Total: **5.0s**

	to exercise with a friend. Are you ready to take this step towards a healthier lifestyle?"}		
Nice will you want to go gym together sometimes then? **(Appendix 7.3K)**	1. Exercising together sounds like a great idea, it can help us stay motivated and make workouts more enjoyable. 2. Working out with a friend like you would definitely keep me motivated and accountable towards reaching my fitness goals. 3. I'm up for it! Having a gym buddy like you will make our workouts more fun and challenging. **Important Contexts:** {"Matt": "I'm planning to visit the gym this weekend to check out their facilities and maybe sign up for a membership. I'm feeling really excited about this new chapter in my life.", "Sin Yee": "That's great to hear! Once you become a member, we can even plan some workout sessions together to keep each other accountable. It's always more fun to exercise with a friend. Are you ready to take this step towards a healthier lifestyle?"}	Exercising together sounds like a great idea, it can help us stay motivated and make workouts more enjoyable.	Audio To Text: 3.4s Response: 2.9s Total: **6.3s**
Great! So recently how has work been? **(Appendix 7.3L)**	1. Work has been hectic lately, but I find joy in helping and caring for the patients at the hospital. 2. It's been challenging at work, but the sense of fulfillment from making a positive impact keeps me going. 3. Work has been tiring but rewarding, especially knowing that I'm making a difference in the lives of others at the hospital. **Important Contexts:** {"Natty": "Hey Sin Yee, how's work at the hospital been lately?", "Sin Yee": "Hi Natty, work has been really tiring but fulfilling."}	Work has been hectic lately, but I find joy in helping and caring for the patients at the hospital.	Audio To Text: 1.5s Response: 15s Total: **16.5s**

Oh no! Why has work been so hectic though? **(Appendix 7.3M)**	1. - 2. The increase in patients and emergencies has kept us all on our toes, but being able to provide care makes it worthwhile. 3. -	**(Wrote out herself)** There has been a pandemic lately!	Audio To Text: 1.3s Response: 3s Total: **3.3s**
Well you better take care then. Are you planning for any holiday trips to relax? **(Appendix 7.3N)**	1. Planning a holiday trip to unwind and recharge sounds like a perfect idea after the busy days at work. 2. It's important to take breaks and rejuvenate, a holiday trip to relax and destress would be ideal. 3. Taking care of myself is essential, a holiday trip to relax and unwind is definitely on the cards to recharge for the upcoming challenges at work.	**(Wrote out herself)** Im thinking of Australia.	Audio To Text: 1.6s Response: 3s Total: **4.6s**

Role Reversal:

Normal Person	**Responses Generated**	**Response Selected**	**Time**
Hi Sin Yee, how have you been doing? What are you up to? **(Appendix 7.3A)**	1. I've been busy with work and my hobbies, recently I started gardening. 2. I enjoy helping and caring for patients at the hospital. I've also started gardening as a hobby. 3. Work has been hectic, but gardening has brought me joy.	Work has been hectic, but gardening has brought me joy.	Audio To Text: 2s Response: 5s Total: **7s**
Really? Why has work been so hectic though? **(Appendix 7.3B)**	1. Lately, there's been an increase in patients at the hospital due to the ongoing pandemic. 2. The hospital has been short-staffed, so the workload has been quite overwhelming. 3. Many new protocols and procedures have been implemented at work, making it more demanding and stressful.	Lately, there's been an increase in patients at the hospital due to the ongoing pandemic.	Audio To Text: 2.5s Response: 4.8s Total: **7.3s**
Well, take care then. Don't overwork yourself. Are you planning to travel to any countries? **(Appendix 7.3C)**	1. Thank you for your concern, I'll make sure to take care of myself. I hope to travel to Japan once it's safe to do so. 2. I appreciate your advice, I'll remember not to overwork myself.	**(Wrote out herself)** Im thinking of Australia.	Audio To Text: 2.1s Response: 3.7s Total: **5.8s**

	I'm looking forward to visiting Italy when travel restrictions are lifted. 3. Your concern means a lot, I'll try to find a good balance. I dream of exploring New Zealand when it's possible to travel again.		
Nice! Which part of Australia are you planning to visit? **(Appendix 7.3D)**	1. I was thinking of exploring the beautiful beaches in Gold Coast. Have you been there before? 2. I'm considering visiting the vibrant city of Melbourne. Any recommendations on must-see attractions? 3. I'm intrigued by the stunning landscapes of Tasmania and plan to include it in my travel itinerary. Heard it's a nature lover's paradise. **Important Contexts:** {"Mary": "Hiking in the Blue Mountains sounds like a fantastic adventure! I'll make sure to pack my hiking boots. Any other must-see places in Australia on your list?", "Sin Yee": "We should definitely visit the Great Barrier Reef for some snorkeling and to witness the vibrant marine life."} {"Mary": "Hey Sin Yee, I've been thinking about our trip to Australia. Any ideas on where we should go first?", "Sin Yee": "Hi Mary! I was thinking we should start in Sydney. There's so much to see and do there."} {"Mary": "That's a great idea! It'll be convenient and we can immerse ourselves in the local culture. Should we plan any day trips from Sydney?", "Sin Yee": "We could take a day trip to the Blue Mountains for some hiking and breathtaking views of the landscape."}	I'm intrigued by the stunning landscapes of Tasmania and plan to include it in my travel itinerary. Heard it's a nature lover's paradise.	Audio To Text: 2.1s Response: 2.9s Total: **5.0s**
Have you ever thought about visiting the Great Barrier reef?	1. The Great Barrier Reef sounds amazing! I'll definitely add it to my list of places to visit in Australia. 2. I've always wanted to snorkel in the Great Barrier Reef! It's a dream	I've always wanted to snorkel in the Great Barrier Reef! It's a dream	Audio To Text: 2.0s Response: 3.0s

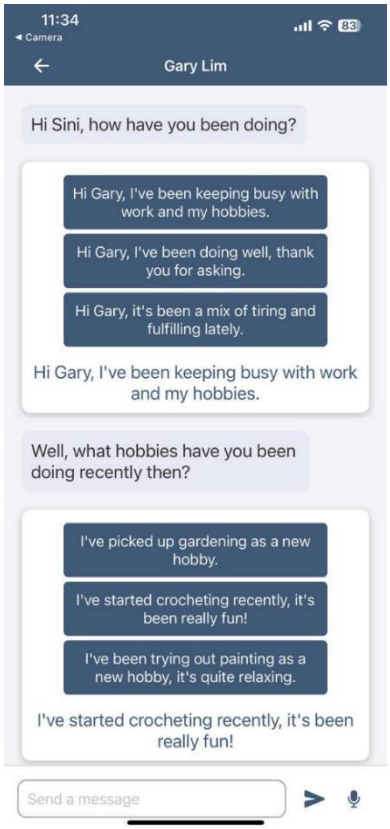
(Appendix 7.3E)	destination for many nature enthusiasts. 3. Exploring the Great Barrier Reef would be a fantastic addition to my travel plans. I'm excited to witness the vibrant marine life there. **Important Contexts:** {"Mary": "Hiking in the Blue Mountains sounds like a fantastic adventure! I'll make sure to pack my hiking boots. Any other must-see places in Australia on your list?", "Sin Yee": "We should definitely visit the Great Barrier Reef for some snorkeling and to witness the vibrant marine life."}	destination for many nature enthusiasts.	Total: **5.0s**
Great, but in Sydney, what else do you plan to do? **(Appendix 7.3F)**	1. I'm excited to visit Bondi Beach for some relaxation and maybe even try surfing to experience the local beach culture. 2. I'll definitely be taking a day trip to the Blue Mountains for some hiking and breathtaking views of the landscape. 3. Exploring the vibrant food scene in Sydney and visiting iconic landmarks like the Sydney Opera House are on the top of my list. **Important Contexts:** {"Mary": "I'm so excited! What other places do you recommend we visit in Sydney?", "Sin Yee": "We should also check out Bondi Beach for some relaxation and maybe even try surfing!"} {"Mary": "That's a great idea! It'll be convenient and we can immerse ourselves in the local culture. Should we plan any day trips from Sydney?", "Sin Yee": "We could take a day trip to the Blue Mountains for some hiking and breathtaking views of the landscape."}	I'll definitely be taking a day trip to the Blue Mountains for some hiking and breathtaking views of the landscape.	Audio To Text: 1.5s Response: 3.0s Total: **4.5s**

Nice that sounds amazing. Besides travelling, you mentioned you were currently gardening? **(Appendix 7.3G)**	1. Yes, I've been learning about different gardening techniques and watching videos to improve my skills and knowledge. 2. I've been expanding my garden and exploring the world of planting vegetables to see where this hobby will lead me. 3. Gardening has brought me much joy and satisfaction. I plan to continue nurturing my garden for a long time to come.	Yes, I've been learning about different gardening techniques and watching videos to improve my skills and knowledge.	Audio To Text: 1.3s Response: 3.2s Total: **4.5s**
Besides videos, how do you learn gardening? **(Appendix 7.3H)**	1. I also enjoy reading articles and books to gain more in-depth knowledge about different plants and gardening methods. 2. Attending gardening workshops and joining online gardening communities have been helpful in broadening my gardening expertise. 3. Experimenting hands-on in my garden and learning from trial and error has been a valuable teacher in my gardening journey. **Important Contexts:** {"Jackson": "That's fantastic! Gardening can be so fulfilling. Have you been doing any research or reading up on gardening tips?", "Sin Yee": "Yes, I've been watching videos and reading articles to learn more about different plants and gardening techniques."}	I also enjoy reading articles and books to gain more in-depth knowledge about different plants and gardening methods.	Audio To Text: 1.8s Response: 3.4s Total: **5.2s**
Besides gardening, I heard you recently started gyming? **(Appendix 7.3I)**	1. Yes, I've been hitting the gym to work on building strength and endurance. 2. That's right, I've recently started going to the gym to focus on my fitness goals. 3. Indeed, I've taken up gyming to improve my overall health and fitness levels	That's right, I've recently started going to the gym to focus on my fitness goals.	Audio To Text: 1.7s Response: 3.1s Total: **4.8s**

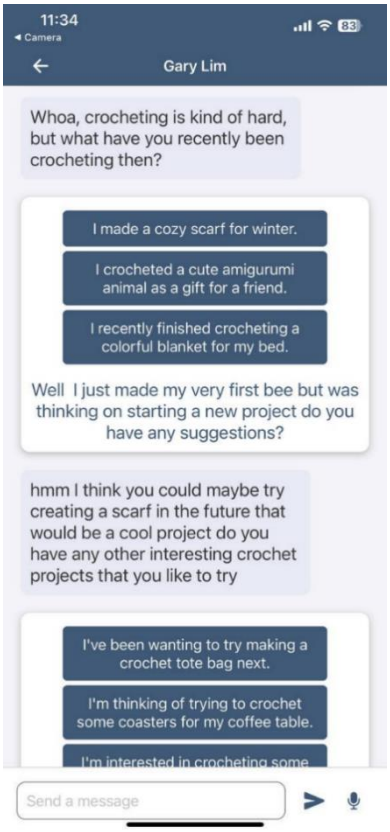
What do you usually do at the gym? **(Appendix 7.3J)**	1. I usually start with some cardio to warm up, then move on to strength training exercises. 2. My gym routine typically involves a mix of cardio exercises like running or cycling, along with weightlifting and resistance training. 3. At the gym, I focus on a variety of workouts such as weightlifting, bodyweight exercises, and some cardio to improve my overall fitness. **Important Contexts:** {"Matt": "I'm thinking of starting with some cardio and strength training to build up my fitness level. Do you have any tips for beginners like me?", "Sin Yee": "Definitely! Start slow and gradually increase the intensity of your workouts. It's also important to listen to your body and rest when needed. Have you set a specific fitness goal in mind?"}	I usually start with some cardio to warm up, then move on to strength training exercises.	Audio To Text: 1.1s Response: 3.1s Total: **4.2s**

Appendix 7.3A: Screen Shots A-D

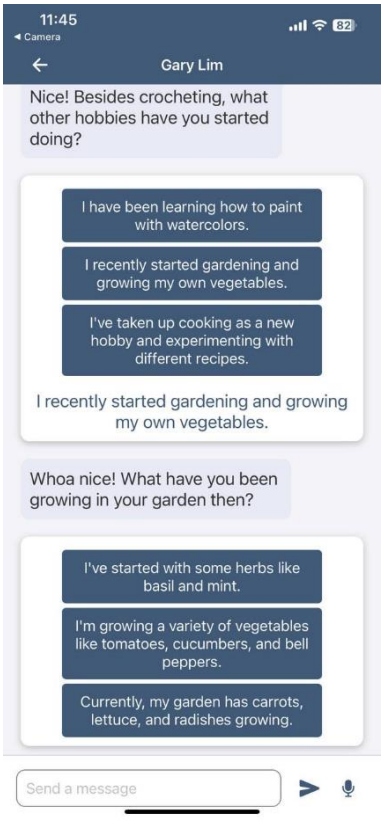
Appendix_A



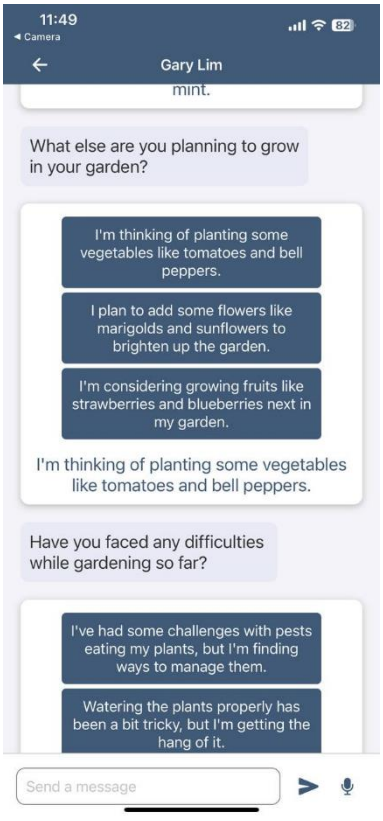
Appendix_B



Appendix_C



Appendix_D

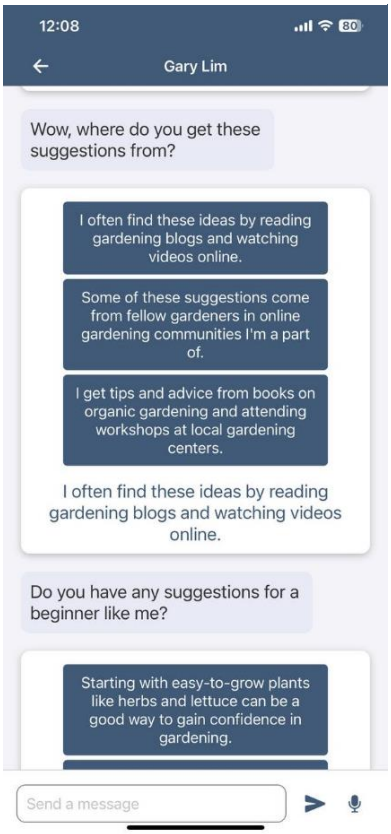


Back To Content Page

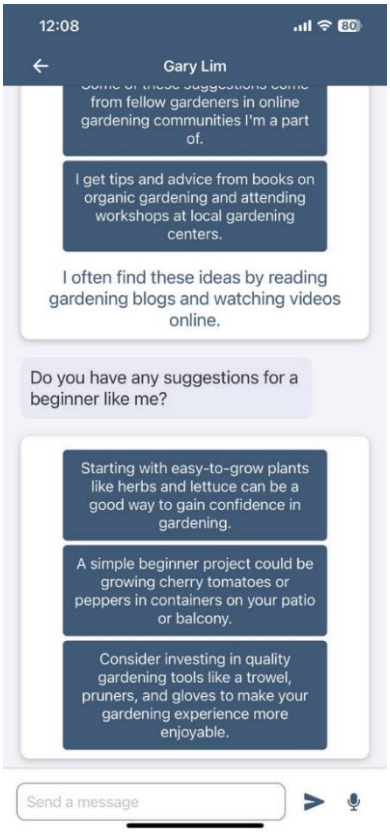
Appendix 7.3A: Screen Shots E-H

Appendix_E

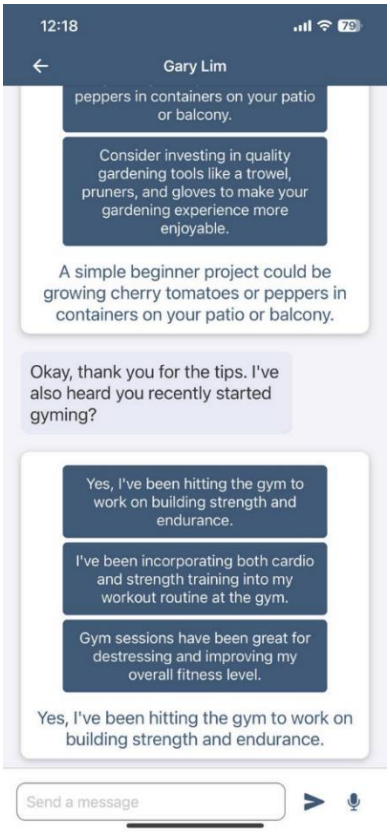
Appendix_F



Appendix_G



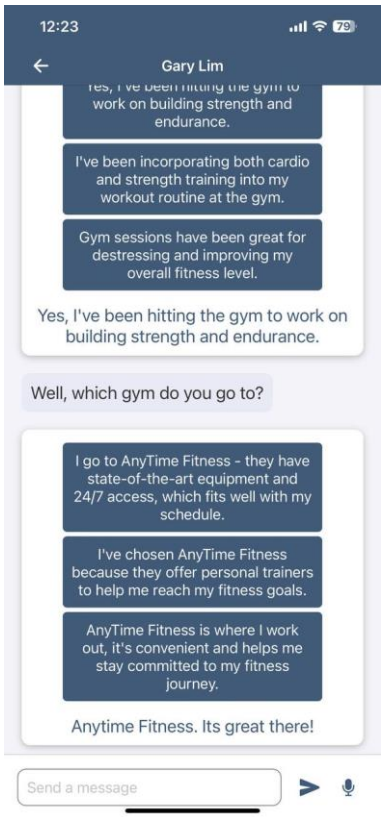
Appendix_H



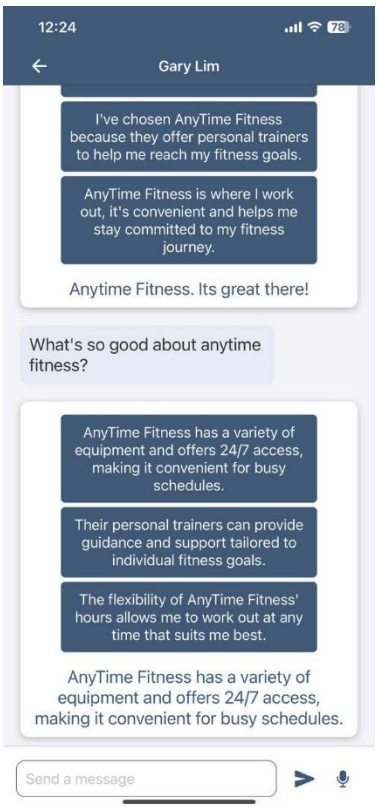
Back To Content Page

Appendix 7.3A: Screen Shots I-L

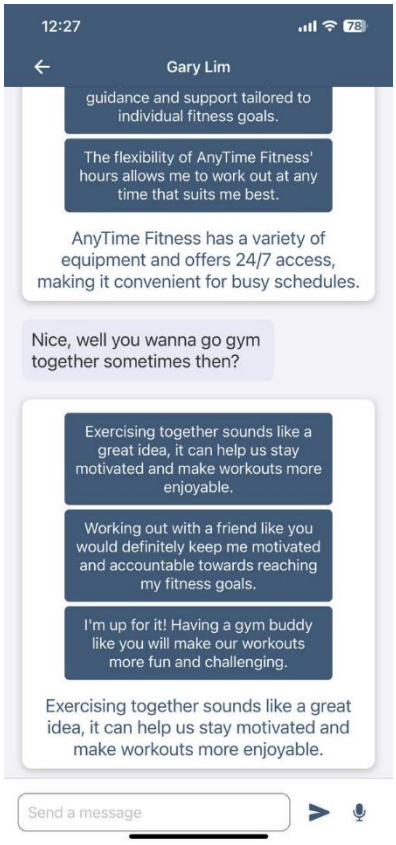
Appendix _I



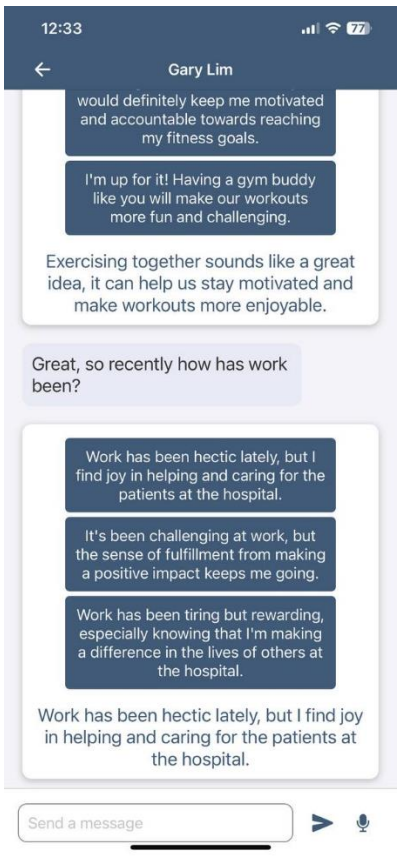
Appendix _J



Appendix _K



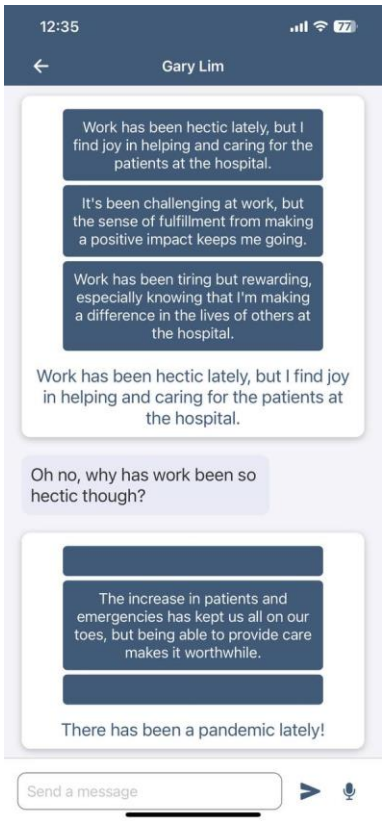
Appendix _L



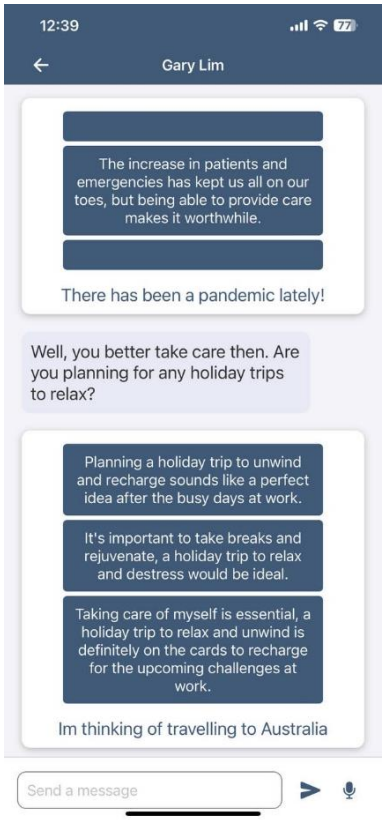
Back To Content Page

Appendix 7.3A: Screen Shots M-N

Appendix _M

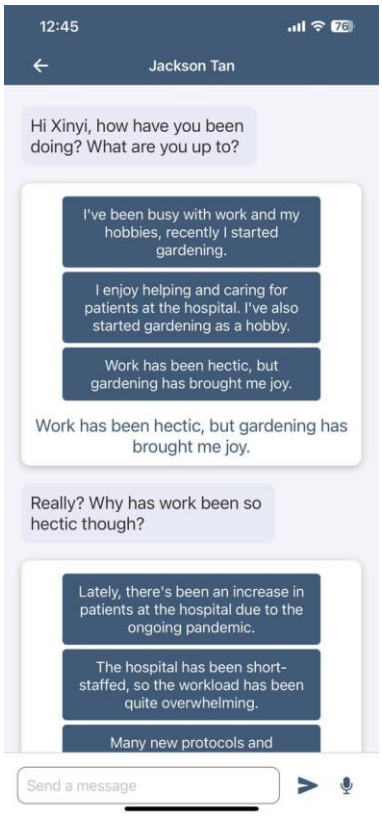


Appendix _N



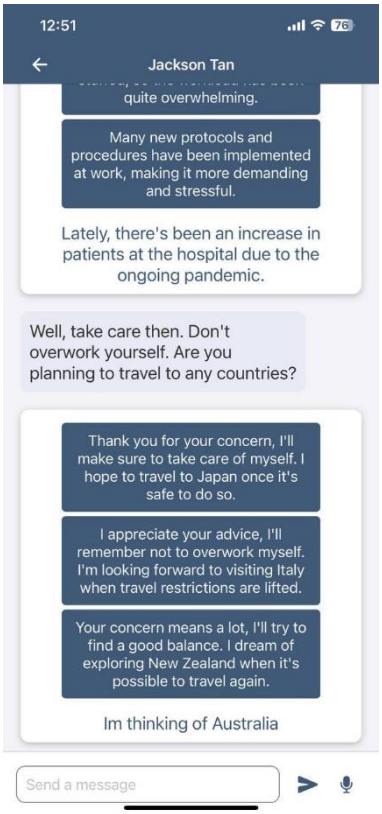
Appendix 7.3A: Screen Shots A-D (Role Reversal)

Appendix _A

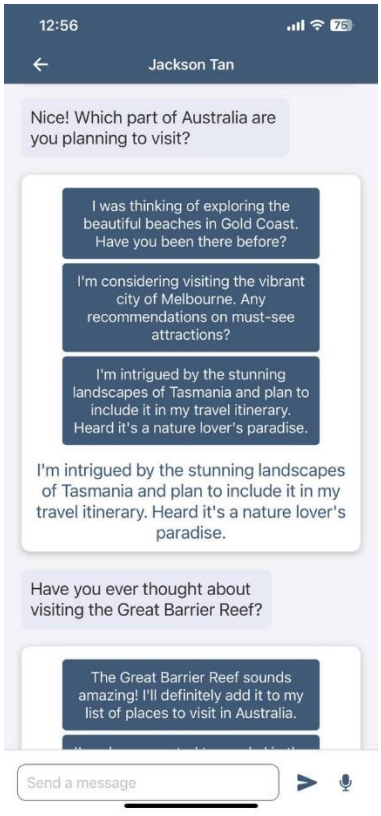


Appendix _B

Appendix _C



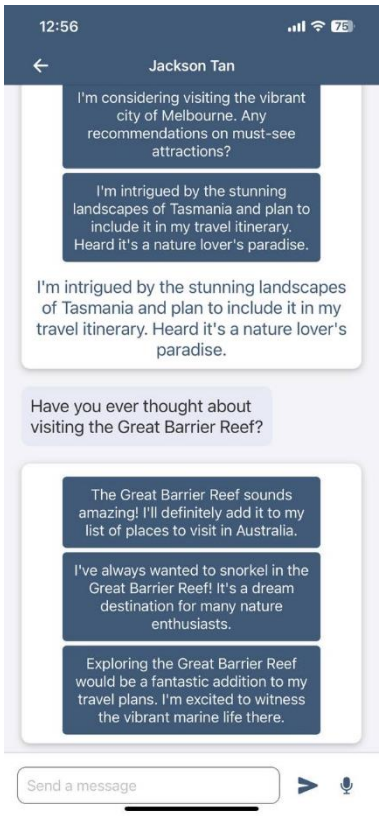
Appendix _D



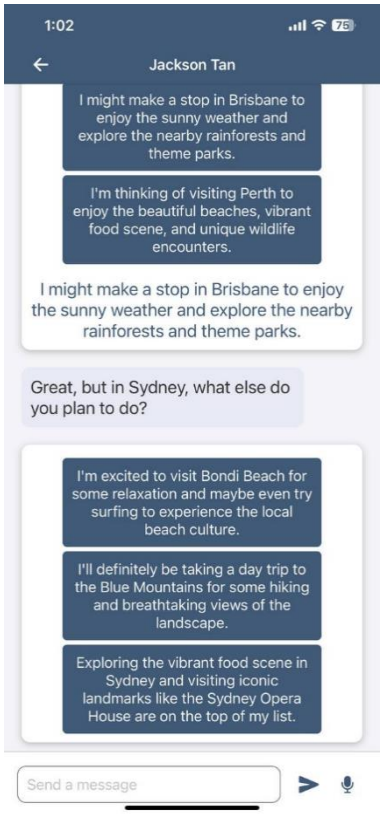
Back To Content Page

Appendix 7.3A: Screen Shots E-H (Role Reversal)

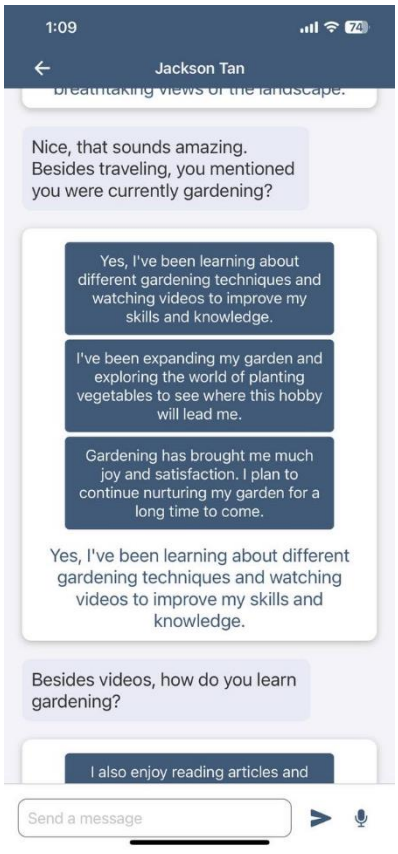
Appendix_E



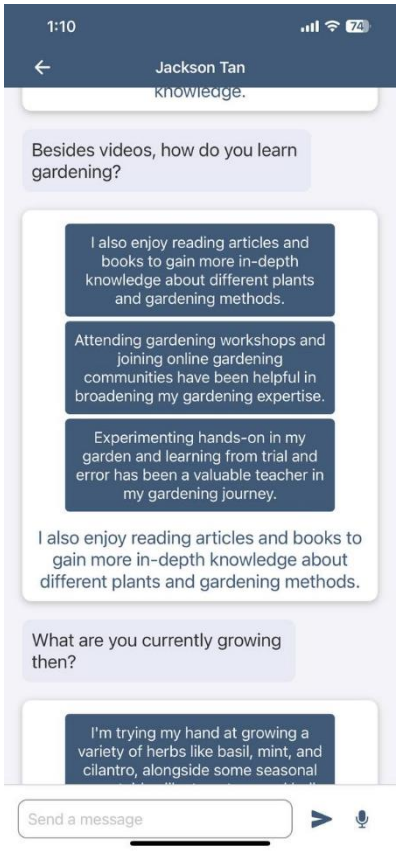
Appendix_F



Appendix_G



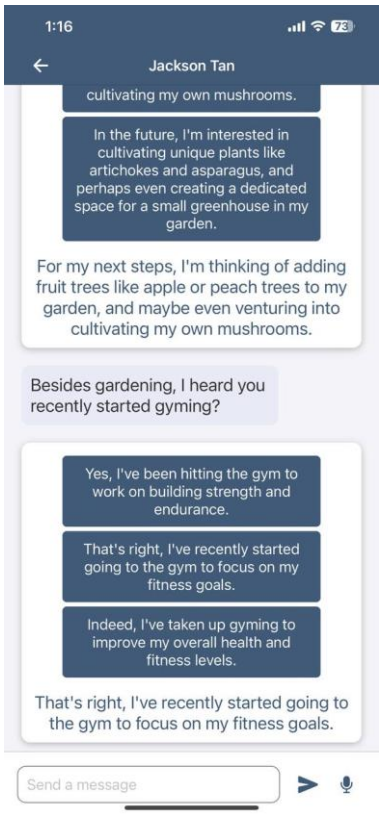
Appendix_H



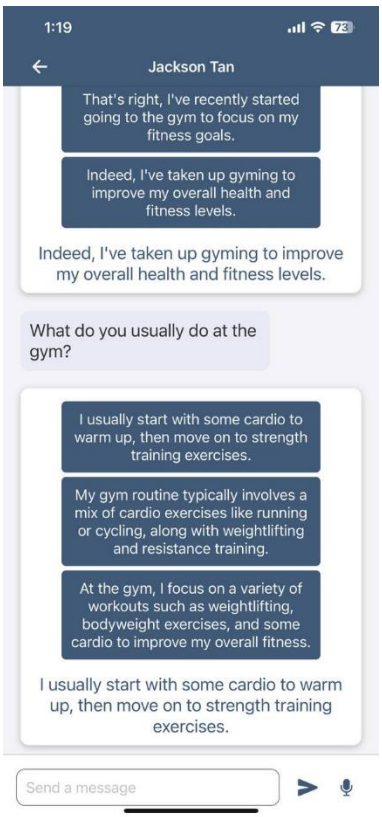
Back To Content Page

Appendix 7.3A: Screen Shots I-J (Role Reversal)

Appendix _I



Appendix _J



Back To Content Page

Appendix 7.3B: Cheryl User Study

Normal Person	Responses Generated	Response Selected	Time
Hi Cheryl, how have you been doing? (Appendix 7.3A)	1. Hi Mary! I've been busy planning a trip to Iceland to see the Northern Lights and explore. How about you? 2. Hi Mary! Lately, I've been focused on training for a marathon. It's been challenging but rewarding. What have you been doing? 3. Hi Mary! I've been enjoying some time outdoors, training for a marathon. How about you, what have you been up to? **Important Contexts:** {"Cheryl": "Hey Matt, I'm thinking of traveling to Iceland to catch the Northern Lights and explore the other attractions there. Have you ever been?", "Matt": "Hey Cheryl, that sounds like an amazing trip! I haven't been to Iceland yet, but I've heard it's absolutely stunning. When are you planning to go?"} {"Natty": "Hey Cheryl, I heard you're planning to do a marathon! That's awesome, how's the training going?", "Cheryl": "Hey Natty! Yeah, I'm excited but also a bit nervous. Training has been tough but I'm pushing through."}	Hi Mary! I've been busy planning a trip to Iceland to see the Northern Lights and explore. How about you?	Audio To Text: 2.5s Response: 11s Total: 13**.5s**
Well, I've been doing great. Thanks for asking. Which part of Northern Lights or Iceland are you planning to visit? (Appendix 7.3B)	1. I'm looking forward to visiting Thingvellir National Park in Iceland for the best view of the Northern Lights. Are you planning any trips yourself? 2. I've heard that Thingvellir National Park is a popular spot to see the Northern Lights, so that's definitely on my list. Have you been to Iceland before? 4. Thingvellir National Park is where I hope to	I've heard that Thingvellir National Park is a popular spot to see the Northern Lights, so that's definitely on my list. Have you been to Iceland before?	Audio To Text: 1.3s Response: 3.4s Total: **4.7s**

	witness the stunning Northern Lights. Have you thought about any travel destinations for your next adventure? **Important Contexts:** {"Cheryl": "Yes, I've been researching the best spots for Northern Lights viewing, and I've heard that Thingvellir National Park is a popular choice due to its clear skies. I can't wait to witness the dancing lights in the sky! Have you thought about what souvenirs you might bring back from Iceland?", "Matt": "Thingvellir National Park sounds like a perfect spot to catch the Northern Lights! As for souvenirs, I've heard that Icelandic wool products like sweaters and blankets are popular choices. Are you planning to bring back any unique souvenirs to remember your trip by?"}		
Wow, that sounds amazing. I haven't been to Iceland before. So what else are you planning to do in Iceland? (**Appendix 7.3C**)	1. Apart from exploring Thingvellir National Park, I'm also excited to hike the glaciers and maybe even try some local cuisine. What about you, do you have any travel plans in mind? 2. In Iceland, besides the Northern Lights, I plan to go glacier hiking and see some breathtaking waterfalls. Have you thought about your next travel destination? 3. Iceland has so much to offer! I'm thinking of visiting the ice caves and relaxing in geothermal pools. Do you enjoy outdoor adventures or prefer a more relaxed trip when traveling? **Important Contexts:**	Iceland has so much to offer! I'm thinking of visiting the ice caves and relaxing in geothermal pools. Do you enjoy outdoor adventures or prefer a more relaxed trip when traveling?	Audio To Text: 2s Response: 3.9s Total: **5.9s**

	{"Cheryl": "I've seen pictures of the ice caves and they look incredible! Glacier hiking sounds like a thrilling adventure, I'll definitely consider adding that to my itinerary. Have you started planning your trip to Iceland yet?", "Matt": "I haven't started planning yet, but your excitement is making me want to start looking into it! Do you have any tips or recommendations for someone planning a trip to Iceland for the first time?"}		
I love outdoor activities, but why do you choose to go to Iceland in this period of time? (**Appendix 7.3D**)	1. I chose to visit Iceland now because it's the best time to witness the Northern Lights and experience the beauty of winter activities. Have you ever considered a winter getaway yourself? 2. I selected this time to visit Iceland for the chance to see the Northern Lights and enjoy winter adventures like glacier hiking. Have you ever traveled during the winter season? 5. I'm going to Iceland now to experience the magic of the Northern Lights and the winter activities the country offers. Have you ever thought about a winter vacation destination?	I chose to visit Iceland now because it's the best time to witness the Northern Lights and experience the beauty of winter activities. Have you ever considered a winter getaway yourself?	Audio To Text: 2.8s Response: 4.2s Total: **7.0s**
Yeah, I love going overseas during December of the winter holidays. Besides travelling, you mentioned that	1. Joining a marathon sounds like an exciting challenge! How do you usually prepare for it, and are there any specific goals you have in mind for this one? 2. That's impressive! How do you manage your training schedule with your	**(Wrote out herself)** Yeah I am currently training for it.	Audio To Text: 2.1s Response: 3s Total: **5.1s**

you're travelling for a marathon? (**Appendix 7.3E**)	other commitments, and do you have any tips for staying motivated during tough training sessions? 3. Running a marathon requires a lot of dedication and hard work. How do you stay focused on your training and maintain a healthy lifestyle while preparing for the race?		
Well, what do you do to train for the marathons? (**Appendix 7.3F**)	1. I incorporate a mix of long runs, speed workouts, and strength training to prepare for the marathon. It's essential to follow a structured training plan and listen to my body to avoid injuries. 2. I make sure to follow a training schedule that includes gradually increasing my mileage, cross-training, and incorporating rest days to allow my body to recover. It's important to also focus on nutrition and hydration during training. 3. To train for the marathon, I follow a well-rounded routine that includes running several times a week, strength training exercises, and proper nutrition to support my body's needs for endurance. **Important Contexts:** {"Natty": "It's good that you're aware of the importance of nutrition. Maybe consider consulting a nutritionist to ensure you're fueling your body properly for the training.", "Cheryl": "That's a great suggestion, Natty. I'll look into that. Have you ever done a marathon before? Any tips for a newbie like me?"} {"Natty": "Yes, I've done a few marathons. One tip I can give you is to pace yourself during the race. It's easy to get caught up in the excitement and start too fast.", "Cheryl": "That's valuable advice, Natty. I'll make sure to start at a comfortable pace and not burn out too quickly. How do you stay motivated during the long training runs?"}	I incorporate a mix of long runs, speed workouts, and strength training to prepare for the marathon. It's essential to follow a structured training plan and listen to my body to avoid injuries.	Audio To Text: 1.8s Response: 3.1s Total: **4.9s**

Sounds great. Anyways, are you watching F1? (**Appendix 7.3H**)	1. Catching the F1 game definitely sounds like an exciting plan! It's always thrilling to experience the live action and atmosphere of such a prestigious racing event. 2. Watching F1 races can be an exhilarating experience, especially with all the adrenaline and speed involved. It must be exciting to witness the skill and precision of the drivers on the track. 3. F1 races are known for their high-speed drama and intense competition. It's amazing to see the technology and talent on display during the races. Enjoy the excitement of watching the F1 action! **Important Contexts:** {"Mary": "Hey Cheryl, I heard you're interested in catching the F1 game.", "Cheryl": "Yeah, I've always wanted to experience it live."}	Catching the F1 game definitely sounds like an exciting plan! It's always thrilling to experience the live action and atmosphere of such a prestigious racing event.	Audio To Text: 1.1s Response: 3s Total: **4.1s**
Are you going to watch it live at the grandstands?	1. It would be incredible to watch the F1 race live from the grandstands! The atmosphere and energy at the track must be electrifying, providing a one-of-a-kind experience. 2. Watching the F1 race live from the grandstands offers a prime view of the fast-paced action and allows for a more immersive experience. It's definitely a fantastic way to enjoy the race. 3. Watching the F1 race from the grandstands provides a unique vantage point to witness the speed and skill of the drivers up close. The excitement and buzz of the crowd make it an unforgettable experience. Enjoy the race from the grandstands!	**(Wrote it herself)** Yeah I managed to get tickets! What about you?	Audio To Text: 1.7s Response: 3.5s Total: **5.2s**
Yup definitely. Anyways, what are you planning to do for your boyfriend's birthday? (**Appendix 7.3I**)	1. Cheryl's idea of cooking a special meal for her boyfriend's birthday is so thoughtful and sweet! It's a great way to show love and appreciation through a homemade meal. 2. Cheryl's plan to cook a special meal for her boyfriend's birthday shows her	**(Wrote it herself)** Im planning to cook for my boyfriend.	Audio To Text: 3.3s Response: 1.5s Total: **4.8s**

	thoughtfulness and effort to make the occasion memorable. It's a lovely gesture that he will surely appreciate. 3. Cheryl putting effort into making a special meal for her boyfriend's birthday is such a heartfelt way to celebrate and create lasting memories. It's a personal touch that adds to the significance of the occasion.		
Nice that sounds great. Where are you planning to hike? (**Appendix 7.3J**)	1. I'm thinking of exploring MacRitchie Reservoir Park. 2. I'm considering checking out Bukit Timah Nature Reserve. 3. I might go for a hike at Sungei Buloh Wetland Reserve. **Important Contexts:** {"Jackson": "Another great spot is Bukit Timah Nature Reserve for a more challenging hike with steep slopes.", "Cheryl": "Sounds like a good workout! Are there any facilities like restrooms along the trail?"}	I'm considering checking out Bukit Timah Nature Reserve.	Audio To Text: 3.5s Response: 2.5s Total: **6.0s**
That sounds fun. What is there to see at the reserve? (**Appendix 7.3J**)	1. There are beautiful lush greenery and a variety of wildlife to spot. 2. You can see native plants, rare birds, and even monkeys along the trails. 3. The reserve offers stunning views from the summit of Bukit Timah Hill.	You can see native plants, rare birds, and even monkeys along the trails.	Audio To Text: 1.8s Response: 2.8s Total: **4.6s**
Great, anyways have you been cooking recently Cheryl? (**Appendix 7.3K**)	1. I've been trying out new recipes in the kitchen. 2. I've been experimenting with different cuisines lately. 3. I'm planning to bake some homemade bread this weekend. **Important Contexts:** {"Cheryl": "Hey Gary, I'm thinking of cooking a special meal for my boyfriend's birthday.", "Gary": "That's a great idea, Cheryl! What are you planning to make for her?"}	I've been experimenting with different cuisines lately.	Audio To Text: 2.4s Response: 2.6s Total: **5.0s**

	{"Cheryl": "I'm thinking of making her favorite dish, homemade pasta with a creamy mushroom sauce.", "Gary": "Sounds delicious! Have you cooked this dish before, or is it a new recipe for you?"}		
Nice, your boyfriend's birthday just passed right? Do you try to cook anything for him? (**Appendix 7.3K**)	1. I made his favorite dish, lasagna, from scratch. 2. I baked a special cake for him and cooked a romantic dinner. 3. I surprised him with a homemade gourmet meal and dessert. **Important Context:** {"Cheryl": "I'm thinking of making her favorite dish, homemade pasta with a creamy mushroom sauce.", "Gary": "Sounds delicious! Have you cooked this dish before, or is it a new recipe for you?"} {"Cheryl": "It's a new recipe for me, but I've been practicing it for the past week to get it just right.", "Gary": "I'm sure she'll appreciate the effort you're putting into making her birthday special. Have you thought about dessert yet?"}	**(Wrote out herself)** Yeah I made him homemade pasta and dessert!	Audio To Text: 1.1s Response: 4.1s Total: **5.2s**
That sounds amazing. Anyways, back to Iceland, are you planning to buy any souvenirs? (**Appendix 7.3L**)	1. I'm thinking of getting some Icelandic wool products as souvenirs. 2. I might bring back some local handmade ceramics from Iceland. 3. I've heard Icelandic cuisine is unique, so maybe I'll buy some food souvenirs. **Important Contexts:** {"Cheryl": "Icelandic wool products sound like cozy souvenirs to bring back! I'm also thinking of getting some local handmade ceramics as a reminder of my trip. Have you looked into the local cuisine in Iceland? I've heard they have some unique dishes.", "Matt": "Local handmade ceramics would make for beautiful souvenirs! I've heard that Icelandic cuisine is quite interesting, with dishes like fermented shark and	I'm thinking of getting some Icelandic wool products as souvenirs.	Audio To Text: 1.1s Response: 3.4s Total: **4.5s**

	lamb soup. Are you adventurous when it comes to trying new foods, or do you prefer sticking to familiar dishes?"}		
That sounds great. Bring some for me as well. Okay. What are you most excited about for Iceland? **(Appendix 7.3M)**	1. I'm excited to see the Northern Lights and explore the ice caves. 2. Glacier hiking sounds like a thrilling adventure that I can't wait to experience. 3. The idea of trying unique Icelandic dishes and local cuisine excites me the most. **Important Contexts:** {"Cheryl": "I've seen pictures of the ice caves and they look incredible! Glacier hiking sounds like a thrilling adventure, I'll definitely consider adding that to my itinerary. Have you started planning your trip to Iceland yet?", "Matt": "I haven't started planning yet, but your excitement is making me want to start looking into it! Do you have any tips or recommendations for someone planning a trip to Iceland for the first time?"}	Glacier hiking sounds like a thrilling adventure that I can't wait to experience.	Audio To Text: 2.1s Response: 3.4s Total: **4.5s**

Role Reversal:

Normal Person	Responses Generated	Response Selected	Time
Hi Cheryl, how have you been? What are you up to? **(Appendix 7.3A)**	1. I've been researching hiking spots in Singapore. 2. I've been experimenting with different cuisines lately. 3. I'm planning to do a marathon, training has been tough but I'm pushing through. **Important Contexts:** {"Natty": "Hey Cheryl, I heard you're planning to do a marathon! That's awesome, how's the training going?", "Cheryl": "Hey Natty! Yeah, I'm excited but also a bit nervous. Training has been tough but I'm pushing through."} {"Jackson": "Hey Cheryl! What have you been up to?", "Cheryl": "Hey Jackson, I've been wanting to go hiking in Singapore. Any recommendations on where to go?"}	I'm planning to do a marathon, training has been tough but I'm pushing through.	Audio To Text: 1.3s Response: 8.9s Total: **10.2s**
That sounds amazing. How far are you planning to run? **(Appendix 7.3A, 7.3B)**	1. I'm aiming to run a full marathon, 42.195 kilometers. 2. I'm planning to start with a half marathon, 21.0975 kilometers. 3. I hope to complete a 10K run for my first marathon.	I hope to complete a 10K run for my first marathon.	Audio To Text: 1.9s Response: 5.2s Total: **7.1s**
Well, all the best then, but besides marathons, are you planning to do any other sports? **(Appendix 7.3B)**	1. I 've been thinking about trying rock climbing or maybe swimming. 2. I'm considering joining a local cycling group for some variety. 3. I might explore yoga or Pilates for flexibility and strength training.	**(Write herself)** I've been playing badminton quite a bit.	Audio To Text: 1.1s Response: 2.4s Total: **3.5s**

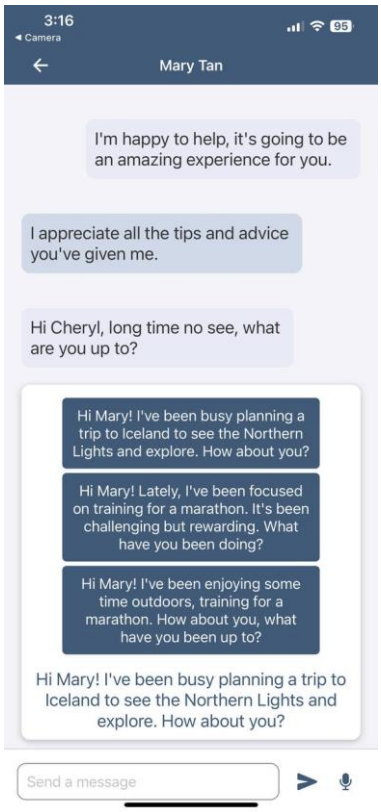
Whoa, that sounds fun. When are you gonna be playing badminton again? **(Appendix 7.3C)**	1. I have a game scheduled for this weekend, want to join? 2. Planning to play on Thursday evenings, feel free to drop by. 3. I usually play on Sundays, you're welcome to come and join the game. **Important Contexts:** {"Xavier": "Hey Cheryl, I heard you're organizing a badminton match. Count me in!", "Cheryl": "That's right, Xavier! I thought it would be a fun way for us to stay active."}	I have a game scheduled for this weekend, want to join?	Audio To Text: 1.1s Response: 3.3s Total: **4.4s**
Honestly, sure, that sounds fun. But how good are you at badminton? I'm not an easy opponent. **(Appendix 7.3D)**	1. I've been playing for a while and enjoy the challenge. Let's see how it goes! 2. I like to think I hold my own on the court, looking forward to a competitive match. 3. I'm always up for a good game, let's test each other's skills on the court.	I've been playing for a while and enjoy the challenge. Let's see how it goes!	Audio To Text: 2.6s Response: 2.9s Total: **5.5s**
Let's go then. Anyways, F1 is coming up. Are you planning to watch it at the grandstands? **(Appendix 7.3E)**	**1**. I've been thinking about it, the live experience must be amazing. 2. I'm considering getting grandstand tickets for a better view of the track. 3. I heard it's an exciting event, leaning towards watching from the grandstands. **Important Contexts:** {"Mary": "With grandstand tickets, you'll have a prime view of the track.", "Cheryl": "I'll make sure to bring my camera to capture the moments."}{"Mary": "You can choose between grandstand tickets or general admission tickets.", "Cheryl": "I think I'd prefer grandstand tickets for a better view."}	I've been thinking about it, the live experience must be amazing.	Audio To Text: 2.5s Response: 5s Total: **7.5s**

	{"Mary": "Hey Cheryl, I heard you're interested in catching the F1 game.", "Cheryl": "Yeah, I've always wanted to experience it live."}		
Right, but it's so hard to get the tickets do you know anyone who is able to get the tickets? **(Appendix 7.3F)**	1. I can check with my friends to see if anyone has insider access to tickets. 2. Let me ask around and see if there's anyone who can help us secure the tickets. 3. I'll reach out to my contacts and see if we can find someone who has access to the tickets. **Important Contexts:** {"Mary": "Have you looked into getting tickets yet?", "Cheryl": "Not yet, I'm not sure where to start."} {"Mary": "I can help you with that. There are different ticket packages available.", "Cheryl": "That's great, I want to make sure I get a good view of the race."} {"Mary": "You can choose between grandstand tickets or general admission tickets.", "Cheryl": "I think I'd prefer grandstand tickets for a better view."}	I can check with my friends to see if anyone has insider access to tickets.	Audio To Text: 1.6s Response: 5.3s Total: **6.9s**
Indecisive huh? **(Appendix 7.3G)**	1. Haha, you could say I appreciate the diversity and skills of all the drivers. 2. I guess you can call it being a fan of the sport as a whole! 3. I just find it hard to pick just one when there are so many talented drivers out there.	I just find it hard to pick just one when there are so many talented drivers out there.	Audio To Text: 2.6s Response: 2.7s Total: **5.3s**

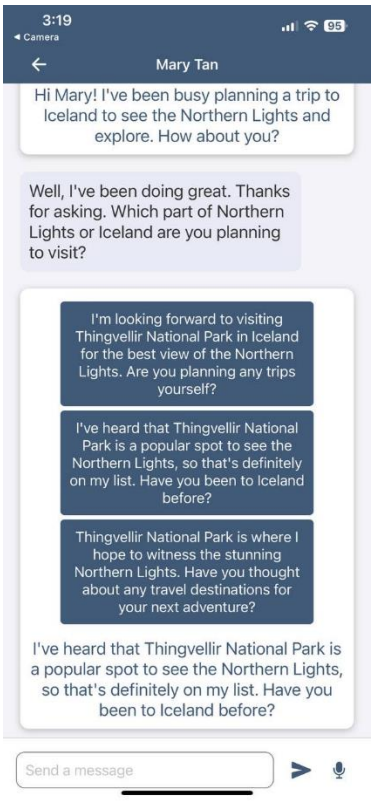
Anyways, why do you like F1 so much? **(Appendix 7.3H)**	1. The adrenaline rush and technical aspects of the sport always keep me intrigued. 2. The thrill of high-speed races and the strategic element of F1 really captivate me. 3. I find the combination of speed, skill, and innovation in F1 to be truly fascinating.	**(Write herself)** Well its always been my dream to be an f1 driver.	Audio To Text: 2.4s Response: 3s Total: **5.4s**
Thanks. What do you think about Iceland? Are you planning to go there? **(Appendix 7.3J)**	1. - 2. Hey Matt, that sounds like an exciting adventure! Iceland is known for its stunning landscapes and unique experiences. When are you thinking of going? 3. Iceland is on my travel bucket list too! The Northern Lights and natural wonders there seem truly magical. When are you planning to visit?'	Iceland is on my travel bucket list too! The Northern Lights and natural wonders there seem truly magical. When are you planning to visit?'	Audio To Text: 1.0s Response: 4.6s Total: **5.6s**

Appendix 7.3B: Screen Shots A-D

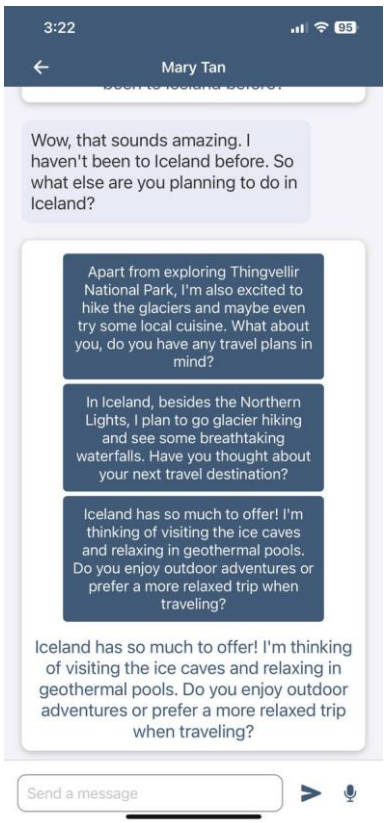
Appendix_A



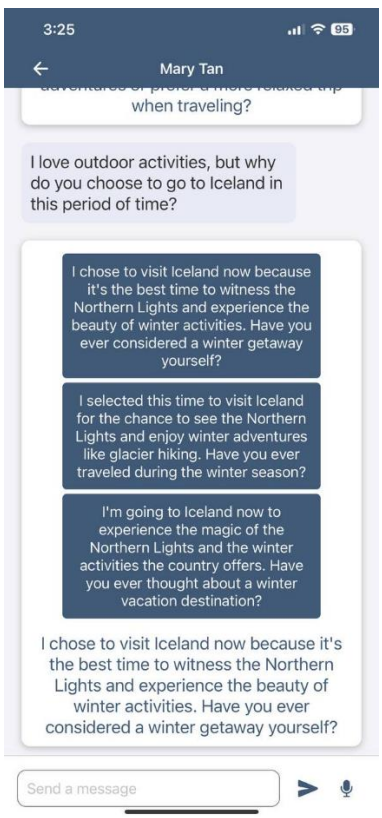
Appendix_B



Appendix_C



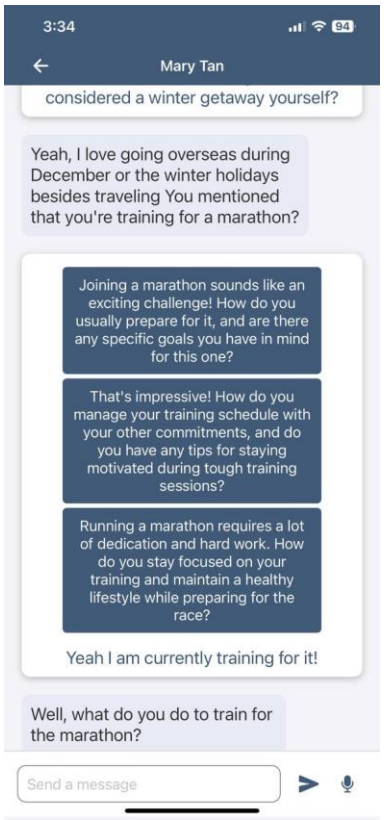
Appendix_D



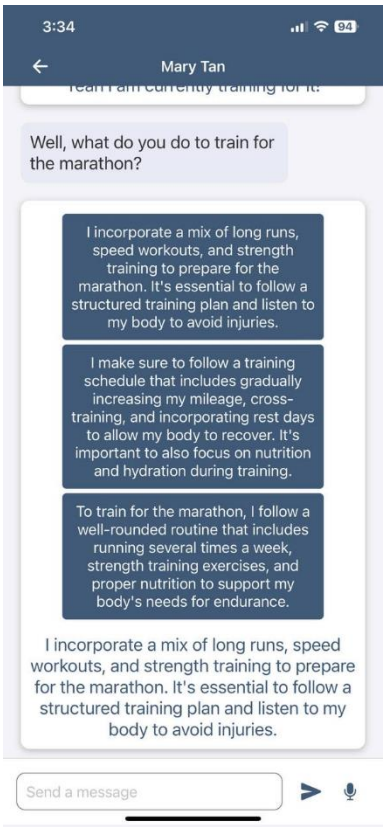
Back To Content Page

Appendix 7.3B: Screen Shots E-H

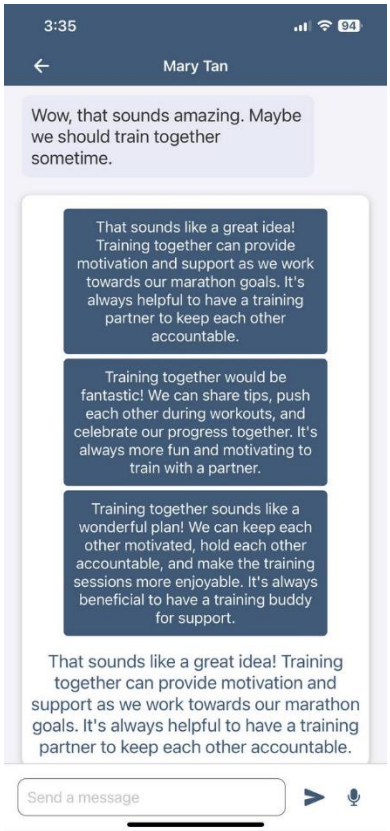
Appendix_E



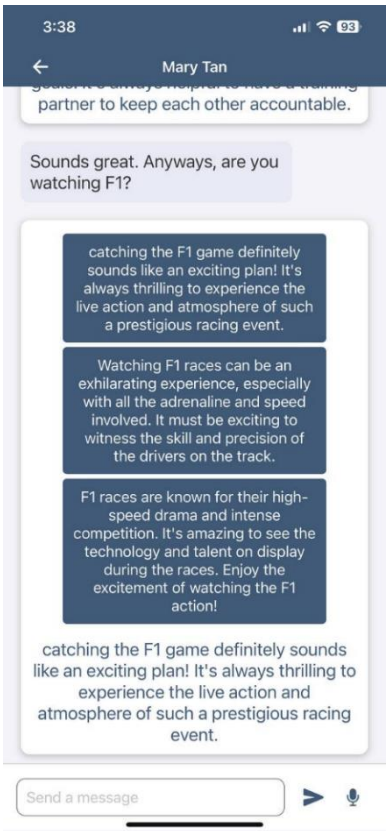
Appendix_F



Appendix_G

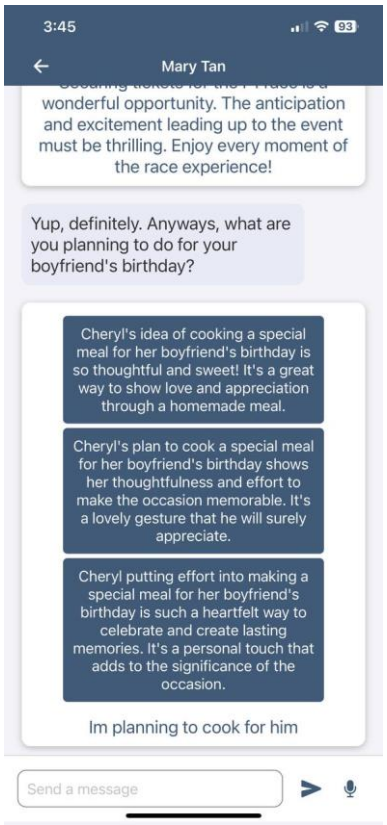


Appendix_H

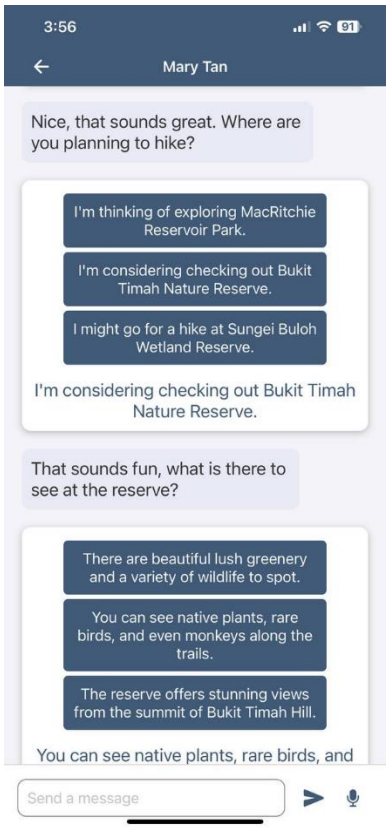


Appendix 7.3B: Screen Shots I-L

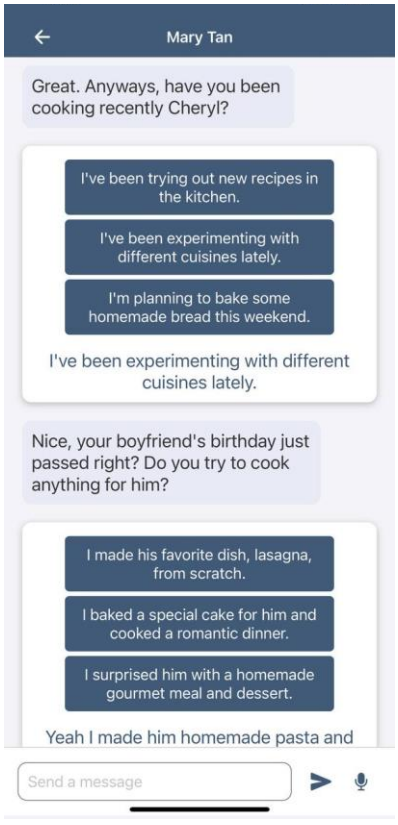
Appendix _I



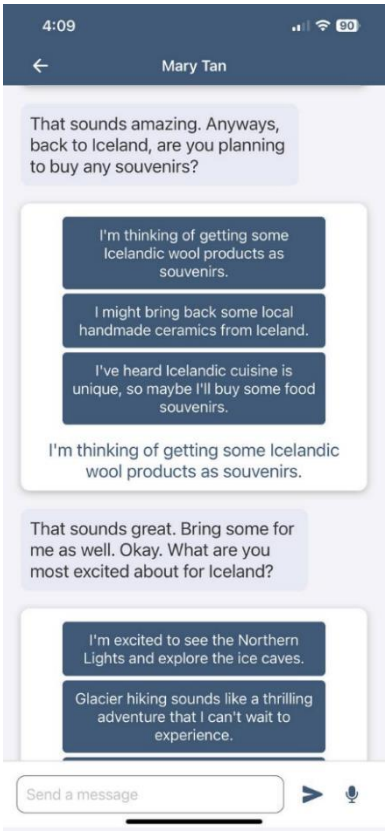
Appendix _J



Appendix _K



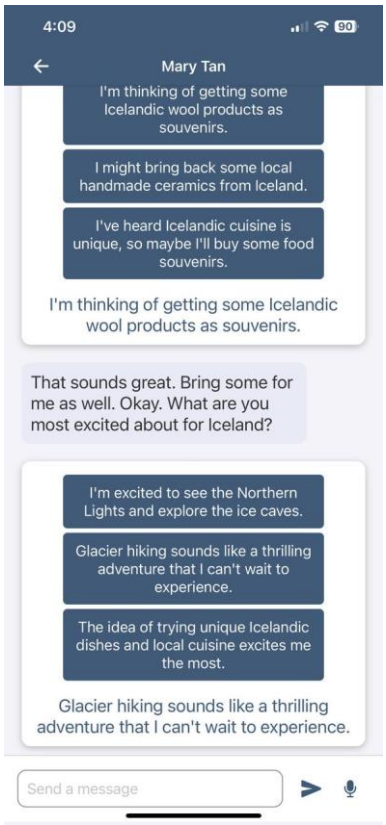
Appendix _L



Back To Content Page

Appendix 7.3B: Screen Shots M

Appendix_M

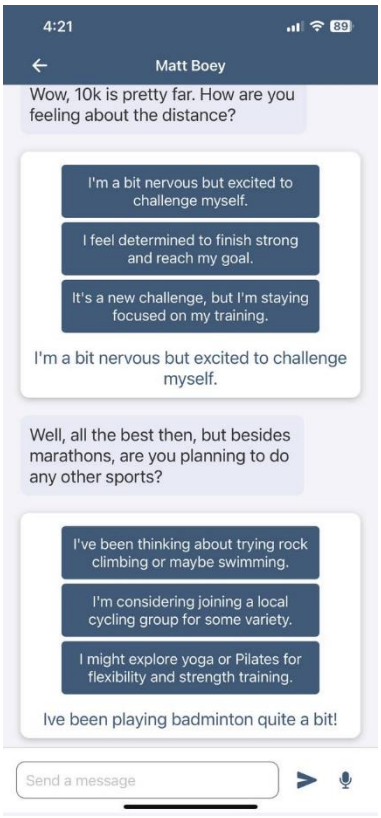


Back To Content Page

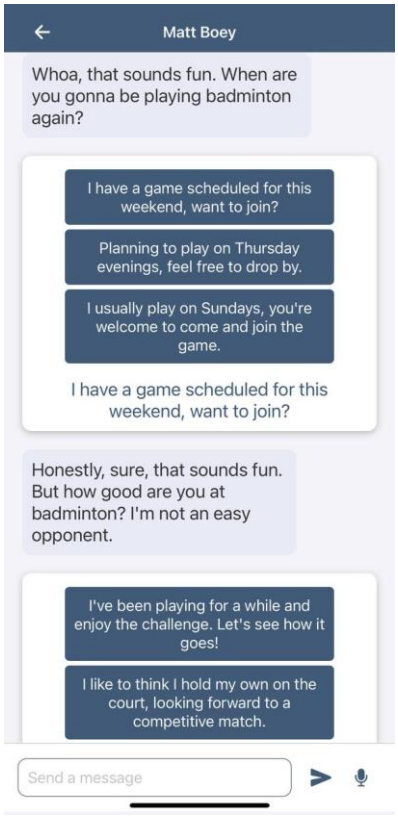
Appendix 7.3B: Screen Shots A-D (Role Reversal)

Appendix_A

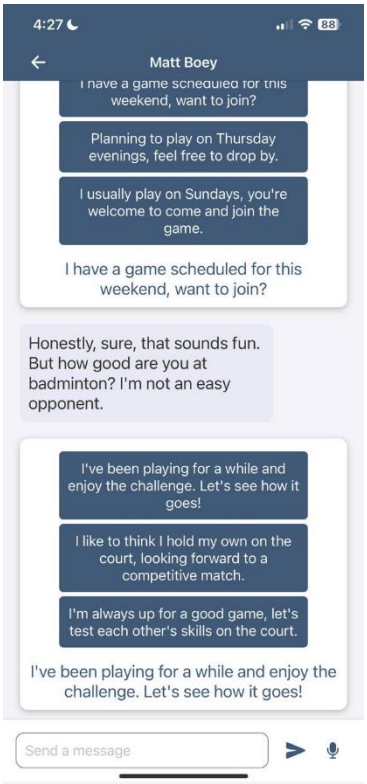
Appendix_B



Appendix_C



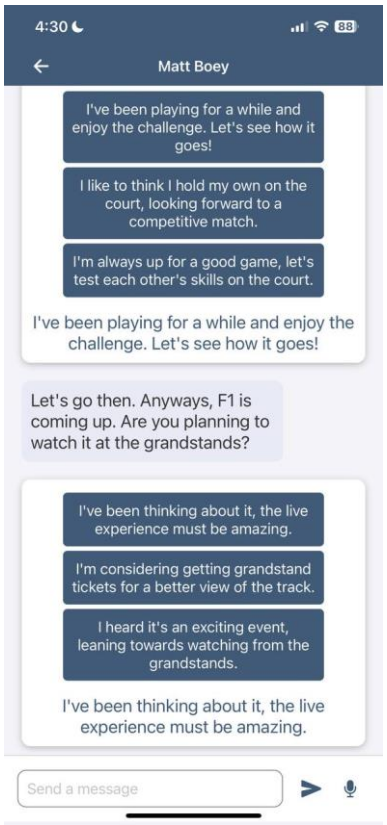
Appendix_D



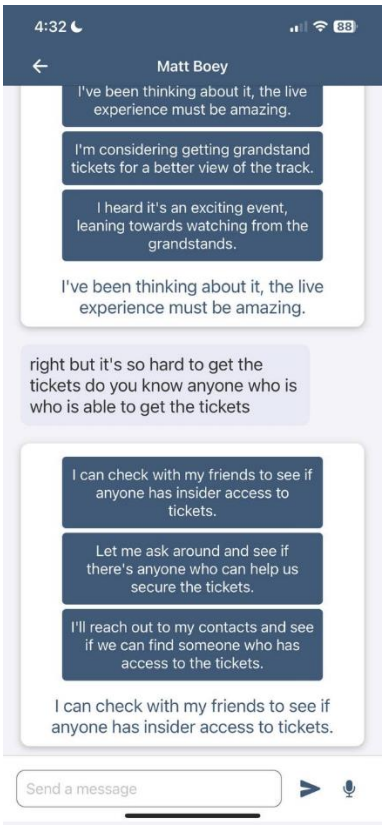
Back To Content Page

Appendix 7.3B: Screen Shots E-H (Role Reversal)

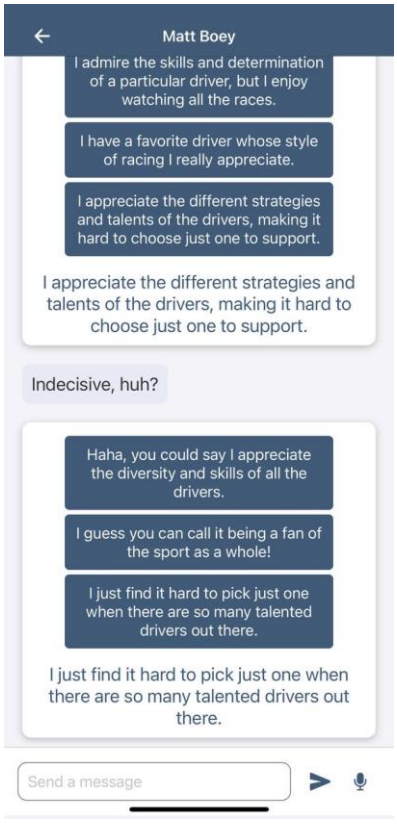
Appendix_E



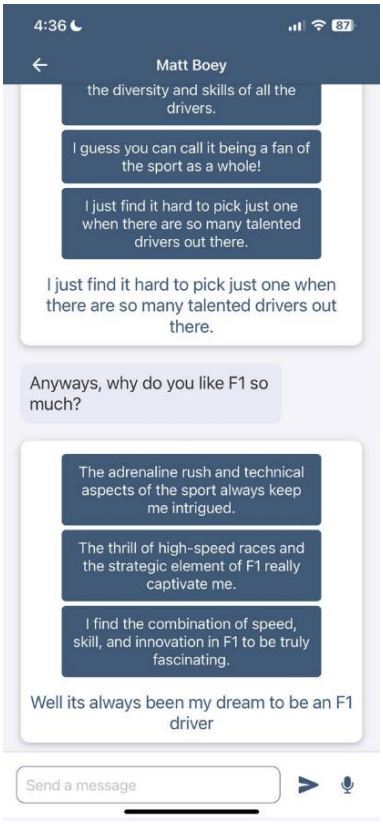
Appendix_F



Appendix_G



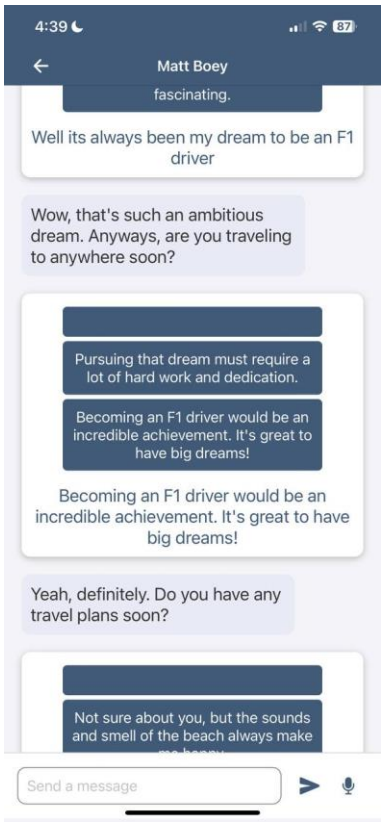
Appendix_H



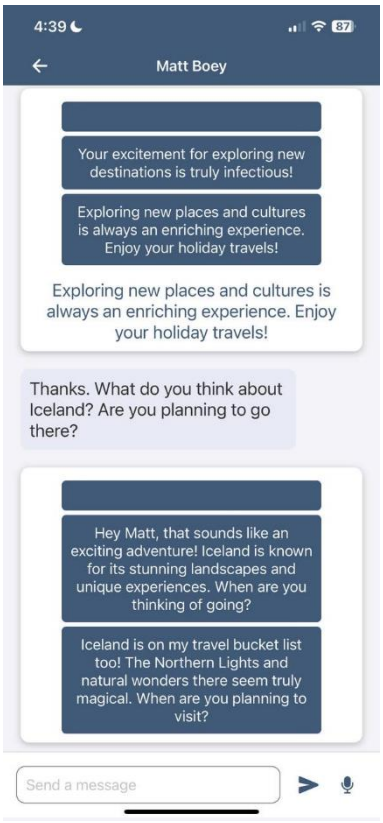
Back To Content Page

Appendix 7.3B: Screen Shots I-J (Role Reversal)

Appendix _I



Appendix _J



Appendix 7.3C: Yu Min User Study

Normal Person	Responses Generated	Response Selected	Time
Hi Yumin, what's up? (**Appendix 7.3A**)	1. Not much, just been experimenting with new recipes for my boyfriend. 2. Hey Matt! I've been thinking about trying out a new sport, any suggestions 3. Hey Matt! I'm planning a party for our friends, do you know any good bars to host it at? **Important Contexts:** {"Xavier": "Hey Yu Min, have you thought about picking up a new sport lately?", "Yu Min": "Hey Xavier! Yes, I've been thinking about it. Any recommendations?"} {"Matt": "Hey Yu Min, have you been cooking for your boyfriend lately?", "Yu Min": "Hey Matt! Yes, I have been trying out some new recipes for him."} {"Yu Min": "Hey Mary, I'm planning a party for our friends. Do you have any suggestions for a good bar to host it at?", "Mary": "Hi Yu Min, that sounds like a fun idea! How about checking out the new rooftop bar downtown? It has a great view and ambiance."}	Not much, just been experimenting with new recipes for my boyfriend.	Audio To Text: 1.9s Response: 5.1s Total: **7.0s**
That sounds fun. What have you tried experimenting so far? (**Appendix 7.3A**)	1. I've been making a lot of Asian fusion dishes lately, he really enjoys those. 2. Lately, I've been baking different types of bread from scratch, it's been a fun challenge. 3. I've been trying out new dessert recipes, his sweet tooth is never satisfied! **Important Contexts:** {"Matt": "That sounds amazing! I'm sure your boyfriend appreciates	Lately, I've been baking different types of bread from scratch, it's been a fun challenge.	Audio To Text: 2.5s Response: 2.8s Total: **5.3s**

	all the effort you put into cooking and baking for him. Do you have any other hobbies you enjoy together?", "Yu Min": "We both enjoy hiking and exploring new places together. It's a great way for us to stay active and bond."}		
Well what's his favourite dish? (**Appendix 7.3B**)	1. He really enjoys my homemade lasagna. 2. He loves when I make spicy garlic shrimp pasta for him. 3. His favorite dish is the classic roast chicken with all the trimmings. **Important Contexts:** {"Matt": "I bet he loved it! Does he have a favorite dish that you make for him?", "Yu Min": "He really enjoys my spicy garlic shrimp pasta."}	He loves when I make spicy garlic shrimp pasta for him.	Audio To Text: 4.1s Response: 3.4s Total: **7.5s**
That's great. Maybe you should make it for me sometime. But anyways, any other exciting activities? (**Appendix 7.3C**)	1. I've been thinking about taking a cooking class to improve my skills. 2. I've been considering planning a surprise trip for his birthday, maybe to a cozy cabin getaway. 3. I'm thinking about joining a local hiking group to explore more outdoor trails. **Important Contexts:** {"Matt": "That's wonderful to hear. Cooking for someone you care about is such a meaningful gesture. Do you have any other culinary skills you'd like to master?", "Yu Min": "I'd love to improve my baking skills and learn how to make intricate desserts in the future."}	**(Wrote herself)** Im planning to go skydiving!	Audio To Text: 2s Response: 7.4s Total: **9.4s**
Wait that sounds so fun. Where are you planning to skydive at? (**Appendix 7.3C**)	1. I'm considering skydiving in Switzerland, the views there are breathtaking.	I'm considering skydiving in Switzerland, the views there are breathtaking.	Audio To Text: 1.2s Response: 5s

	2. I'm looking into skydiving in New Zealand, the landscapes there are stunning. 3. I'm thinking of skydiving in Hawaii, the experience would be unforgettable. **Important Contexts:** {"Gary": "I bet it will be! Have you decided where you're going to go for the skydiving?", "Yu Min": "I'm thinking of going to Switzerland. The views there are supposed to be amazing."}		Total: **6.2s**
That sounds amazing. Anyways, besides skydiving, what else are you gonna do in Switzerland? (**Appendix 7.3D, 7.3E**)	1. I'm thinking of exploring the beautiful lakes and trying some local Swiss cuisine. 2. I plan to visit historical sites, like castles and museums, to learn more about the country. 3. Besides skydiving, I'm looking forward to hiking in the Swiss Alps and taking in the stunning views. **Important Contexts:** {"Gary": "Hiking in the Swiss Alps sounds like a dream come true. Have you started planning out your itinerary yet?", "Yu Min": "Not yet, but I'm looking forward to researching and putting together a detailed plan for the trip."}	Besides skydiving, I'm looking forward to hiking in the Swiss Alps and taking in the stunning views.	Audio To Text: 2.1s Response: 2.8s Total: **4.9s**
Are you gonna visit any other countries in Europe? (**Appendix 7.3F**)	1. Yes, I'm planning to extend my trip and do a Europe tour to explore more countries. 2. I have my eye on visiting Italy next, the history and culture there are fascinating. 3. I'm considering adding France to my itinerary to see iconic landmarks like the Eiffel Tower. **Important Contexts:** {"Gary": "Italy is definitely a must-visit! Do you have any other	I'm considering adding France to my itinerary to see iconic landmarks like the Eiffel Tower.	Audio To Text: 1.7s Response: 6s Total: **7.7s**

	activities planned for your Europe tour?", "Yu Min": "I'm hoping to do some hiking in the Swiss Alps and maybe even visit the Eiffel Tower in Paris."}		
That sounds so fun. Anyways what sports have you been doing recently? (**Appendix 7.3G**)	1. Yes, I'm planning to extend my trip and do a Europe tour to explore more countries. 2. I have my eye on visiting Italy next, the history and culture there are fascinating. 3. I'm considering adding France to my itinerary to see iconic landmarks like the Eiffel Tower.	I just started taking swimming lessons	Audio To Text: 1.8s Response: 2.6s Total: **4.4s**
Well how has the swimming been then? (**Appendix 7.3H**)	1. Swimming has been thrilling and I'm making progress with each lesson. 2. Learning how to swim has been both refreshing and a great workout. 3. Swimming lessons have been enjoyable and I'm excited to improve my skills in the water. **Important Contexts:** {"Natty": "Hey Yu Min, have you been learning how to swim lately?", "Yu Min": "Hey Natty, yes, I actually just started taking swimming lessons."}	Swimming lessons have been enjoyable and I'm excited to improve my skills in the water.	Audio To Text: 1s Response: 4.5s Total: **5.5s**
Well what made you want to learn swimming? (**Appendix 7.3I**)	1. I've always enjoyed being in the water and wanted to become a stronger swimmer. 2. I decided to learn swimming to overcome my fear of deep water and feel more confident. 3. The idea of being able to swim freely in any body of water motivated me to start swimming lessons. **Important Contexts:**	I decided to learn swimming to overcome my fear of deep water and feel more confident.	Audio To Text: 1.8s Response: 2.7s Total: **4.5s**

	{"Natty": "I'm sure you'll get the hang of it soon. What made you decide to start learning how to swim?", "Yu Min": "I've always been a bit afraid of the water, so I wanted to overcome that fear and learn a valuable life skill."} {"Natty": "That's fantastic progress! Do you have any specific goals you want to achieve by the end of your swimming lessons?", "Yu Min": "I hope to be able to swim a few laps confidently without feeling anxious about being in the water."}		
Nice, all the best. Oh yeah I wanted to ask how was your bungee jumping experience? (**Appendix 7.3J**)	1. Bungee jumping was such a thrilling adventure! It pushed me out of my comfort zone. 2. The bungee jumping experience was both terrifying and exhilarating. It definitely tested my limits. 3. Trying bungee jumping was a mix of fear and excitement. I felt a rush like never before when I took the leap. **Important Contexts:** {"Jackson": "Hey Yu Min, I heard you went bungee jumping in Singapore! How was it?", "Yu Min": "Hey Jackson! Yes, it was such an exhilarating experience. I was so nervous at first."}	The bungee jumping experience was both terrifying and exhilarating. It definitely tested my limits.	Audio To Text: 1.9s Response: 3.2s Total: **5.1s**
Where did you bungee jump at? (**Appendix 7.3K**)	1. I tried bungee jumping at a location with a beautiful scenic view to make the experience even more memorable. 2. The bungee jumping spot I chose was known for its great heights and breathtaking surroundings. 3. Bungee jumping was at a popular site known for its safety measures and experienced staff to	**(Typed out herself)** It was in Singapore at Sentosa.	Audio To Text: 1.1s Response: 3s Total: **4.1s**

	ensure a thrilling yet secure adventure.		
Would you do it again? (**Appendix 7.3L**)	1. I'm already thinking about where my next bungee jumping adventure will be. 2. Absolutely! I would definitely take the plunge again for that rush of adrenaline. 3. The experience was unforgettable, and I would love to bungee jump again in the future. **Important Contexts:** {"Jackson": "That's awesome to hear! Would you do it again in the future?", "Yu Min": "Absolutely! I'm already thinking about where my next bungee jumping adventure will be."}	I'm already thinking about where my next bungee jumping adventure will be.	Audio To Text: 1.8s Response: 2.7s Total: **4.5s**
That sounds amazing. What else would you do at South Africa? (**Appendix 7.3M**)	1. Apart from bungee jumping, I would definitely go on a safari to see the incredible wildlife up close. 2. I'm looking forward to experiencing the vibrant culture and trying out some traditional South African cuisine. 3. Exploring the beautiful beaches, going on a hiking adventure, and maybe even trying some local adventure sports!	Apart from bungee jumping, I would definitely go on a safari to see the incredible wildlife up close.	Audio To Text: 2.5s Response: 3.1s Total: **5.6s**
That sounds exciting, do you love wildlife? (**Appendix 7.3N**)	1. Absolutely! Being in close proximity to nature and wildlife gives me a sense of peace and wonder. 2. Yes, I have always been fascinated by animals and their behaviors in their natural habitats. 3. The beauty of wildlife and the opportunity to observe them in	The beauty of wildlife and the opportunity to observe them in their own environment is truly a humbling experience for me.	Audio To Text: 1.9s Response: 2.8s Total: **4.7s**

	their own environment is truly a humbling experience for me.		

Role Reversal:

Normal Person	Responses Generated	Response Selected	Time
Hi Cheryl, how's the party planning going? **(Appendix 7.3A)**	1. Party planning is going well, thanks for asking! 2. I've been working on some details for the party, it's coming together. 3. The party planning is in full swing, just finalizing some things.	I've been working on some details for the party, it's coming together.	Audio To Text: 1.5s Response: 7s Total: **8.5s**
Well what have you planned out so far? **(Appendix 7.3A)**	1. I have planned the decorations and the food menu so far. 2. The party will have a theme and some fun activities for everyone. 3. The guest list is finalized and invitations will be sent out soon.	The party will have a theme and some fun activities for everyone.	Audio To Text: 2.8s Response: 1.9s Total: **4.7s**
Well where are you planning to have the party at? **(Appendix 7.3B)**	1. The party will be at a rented venue with a spacious area for dancing. 2. I've booked a cozy outdoor space for the party with a beautiful view. 3. The party will take place at a friend's house with a lovely backyard setup. **Important Contexts:** {"Yu Min": "I'll make sure to designate a dancing area then. Should I plan any surprises or special activities for the party?", "Mary": "A surprise performance or a themed costume contest could add an element of excitement to the party. It'll keep everyone entertained."}	**(Typed out herself)** It will be held at a bar but we're still thinking of a better location.	Audio To Text: 2.5s Response: 3.1s Total: **5.6s**

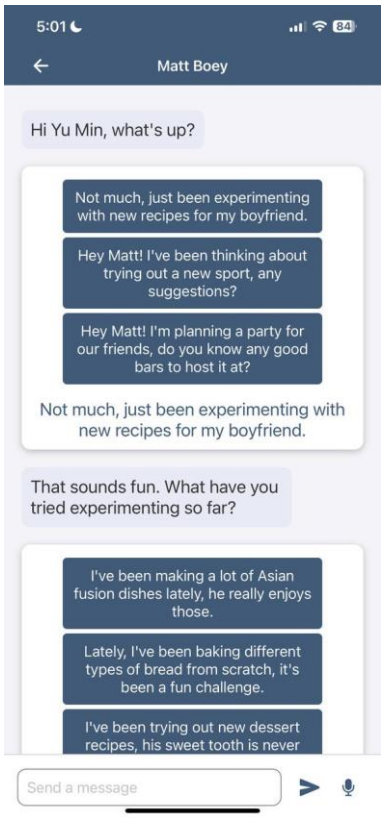
Hmm, I think the location depends on what kind of activities you're planning. What are you planning to do at the party? **(Appendix 7.3C)**	1. We are considering having a themed costume contest and maybe a surprise performance. 2. I thought of having a dance area and maybe some party games for entertainment. 3. We're planning on having music, drinks, and possibly a photo booth for the party. **Important Contexts:** {"Yu Min": "I'll make sure to designate a dancing area then. Should I plan any surprises or special activities for the party?", "Mary": "A surprise performance or a themed costume contest could add an element of excitement to the party. It'll keep everyone entertained."} {"Yu Min": "I like the idea of a themed costume contest! It'll be a fun way to get everyone involved. Should I set a specific start and end time for the party?", "Mary": "Setting a start time will give your guests an idea of when to arrive, but the end time can be more flexible depending on how the party is going."}	We are considering having a themed costume contest and maybe a surprise performance.	Audio To Text: 2.5s Response: 2.7s Total: **5.2s**
That sounds so fun! Any photo booths planned? **(Appendix 7.3D)**	1. Yes, we will have a photo booth with props for everyone to enjoy. 2. There will be a photo booth area with different backgrounds for group photos. 3. The photo booth will have fun props and accessories for memorable pictures. **Important Contexts:** {"Yu Min": "That's a thoughtful idea, Mary. I'll prepare some party favors for our friends. Should I have a designated area for taking group photos at the party?", "Mary": "Having a designated	There will be a photo booth area with different backgrounds for group photos.	Audio To Text: 3.4s Response: 2.9s Total: **6.3s**

	photo area with props could be a fun way to capture group photos and memories. It'll be a popular spot for everyone."}		
Anyways recently you went Bungee Jumping right? Where did you go? **(Appendix 7.3E)**	1. I went bungee jumping in Singapore at Sentosa. 2. The bungee jumping experience was in Singapore at Sentosa. 3. It was at Sentosa in Singapore where I went bungee jumping. **Important Contexts:** {"Matt": "Where do you bungee jump at?", "Yu Min": "It was in Singapore at Sentosa!"}	I went bungee jumping in Singapore at Sentosa.	Audio To Text: 4.4s Response: 2.8s Total: **7.2s**
Well how was it compared to sky diving? **(Appendix 7.3F)**	1. Bungee jumping was more intense and sudden compared to the gradual thrill of skydiving. 2. Bungee jumping was a mix of fear and excitement, whereas skydiving felt like a surreal experience. 3. Bungee jumping was a quick adrenaline rush, while skydiving provided a longer lasting thrill.	Bungee jumping was a quick adrenaline rush, while skydiving provided a longer lasting thrill.	Audio To Text: 1.1s Response: 2.7s Total: **3.8s**
Which would you actually prefer if you had to choose? **(Appendix 7.3G)**	1. I think I would prefer skydiving for the breathtaking views and the sense of freedom. 2. Personally, I would choose skydiving for the overall experience and the feeling of flying. 3. I lean towards skydiving because of the peaceful moments during the descent and the panoramic scenery.	I would actually pick Bungee Jumping because it's a much shorter thrill. My heart can't take it.	Audio To Text: 1.0s Response: 2.6s Total: **3.6s**

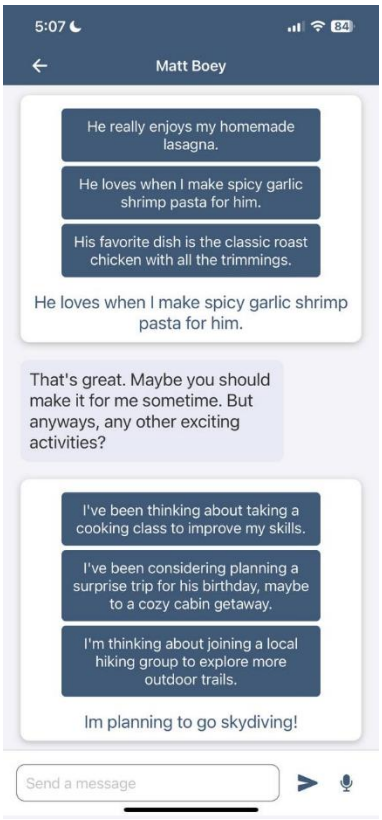
(Issue): Had to reset recent 5 conversation storage. Well when is your next bungee jumping trip? **(Appendix 7.3H)** Anyways, how has your swimming lessons been going? **(Appendix 7.3I)**	1. Next time, would you consider inviting friends to join you in the bungee jumping adventure? 2. Have you researched any new bungee jumping locations or heights for your next trip? 3. Are you planning to do anything different or special for your next bungee jumping experience?' 1. Have you noticed any improvement in your swimming skills since starting the lessons? 2. What has been the most challenging part of your swimming lessons so far? 3. Are there any specific techniques or strokes that you find particularly enjoyable during the lessons?'	-	Audio To Text: 1.0s Response: 3.4s Total: **4.4s** Audio To Text: 1.1s Response: 2.5s Total: **3.6s**
(Recovery): Recovered after reset Anyways how has your swimming lessons been going? **(Appendix 7.3J, 7.3K, 7.3L)**	1. My swimming lessons are going well, I've been learning different strokes and techniques. 2. I've been making progress with my swimming lessons, starting with water safety and now moving on to more advanced skills. 3. Swimming lessons have been challenging but rewarding, especially as I work towards swimming laps confidently. **Important Contexts:** {"Natty": "That's a brave decision! Facing your fears head-on is always a good thing. How are your lessons structured?", "Yu Min": "The lessons start with basic water safety and gradually progress to different swimming strokes and techniques."}	My swimming lessons are going well, I've been learning different strokes and techniques.	Audio To Text: 1.7s Response: 2.9s Total: 4**.6s**

Appendix 7.3C: Screen Shots A-D

Appendix_A

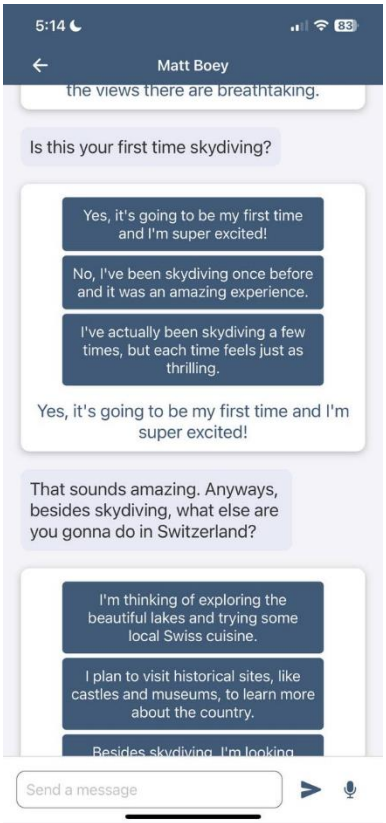


Appendix_B



Appendix_C

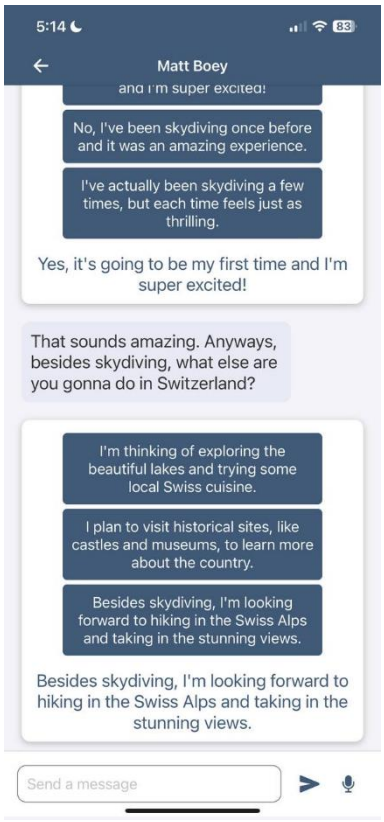
Appendix_D



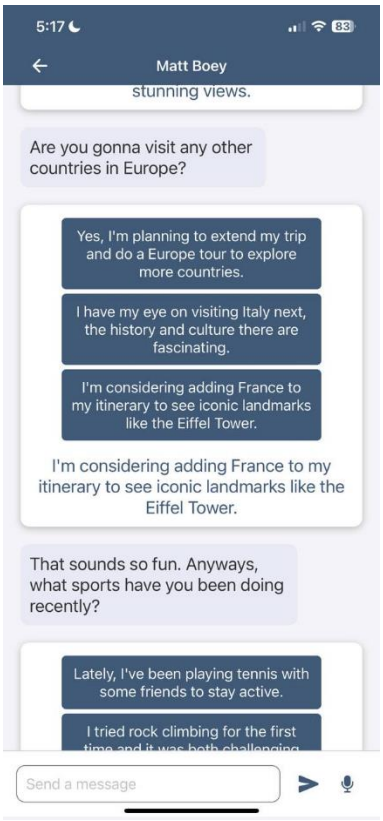
Back To Content Page

Appendix 7.3C: Screen Shots E-H

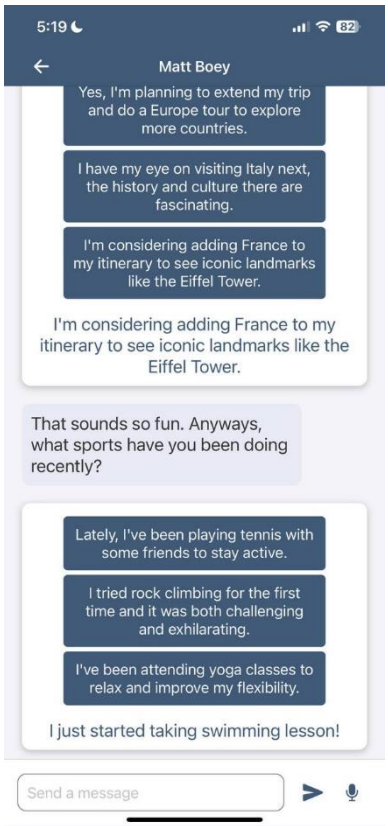
Appendix_E



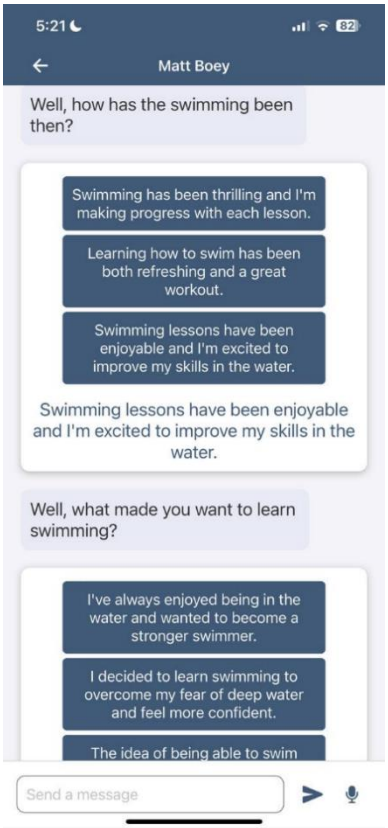
Appendix_F



Appendix_G



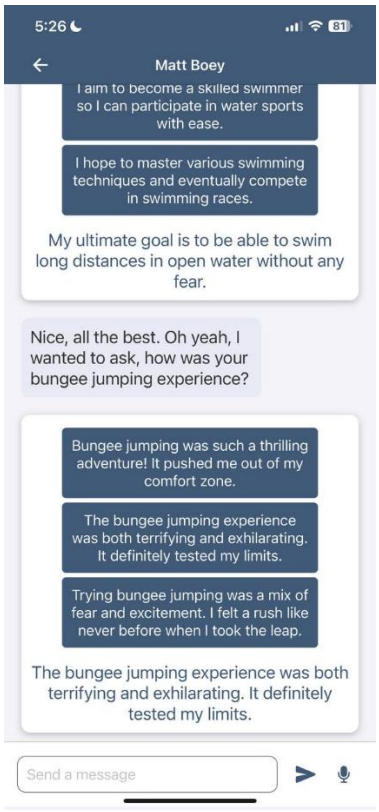
Appendix_H



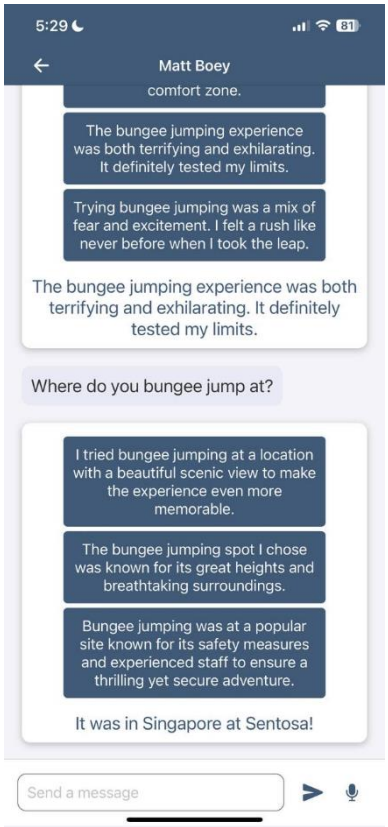
Appendix 7.3C: Screen Shots I-L

Appendix _I

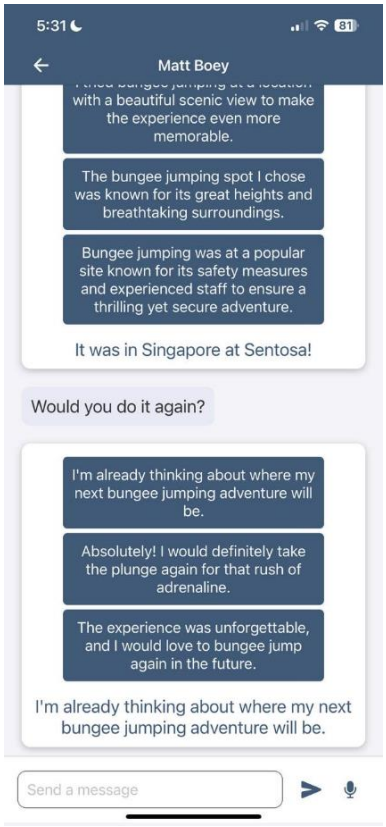
Appendix _J



Appendix _K



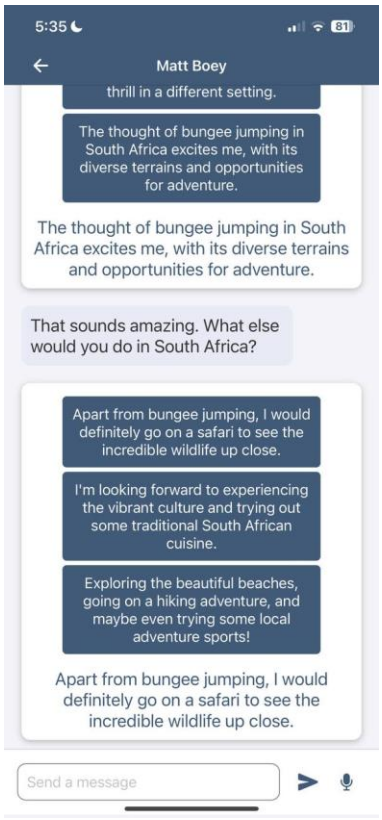
Appendix _L



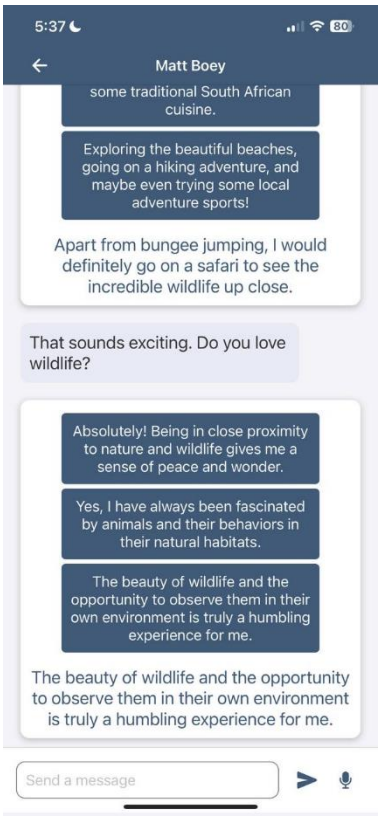
Back To Content Page

Appendix 7.3C: Screen Shots M-N

Appendix_M

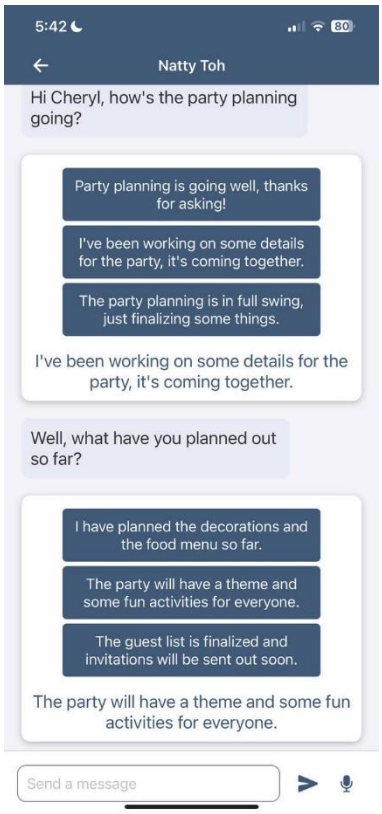


Appendix_N

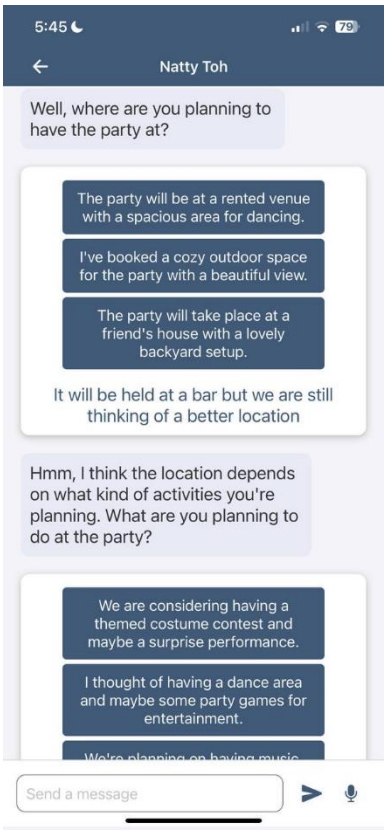


Appendix 7.3C: Screen Shots A-D (Role Reversal)

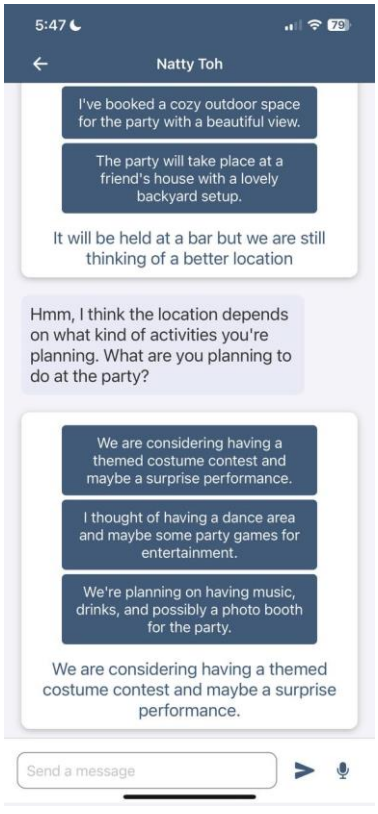
Appendix_A



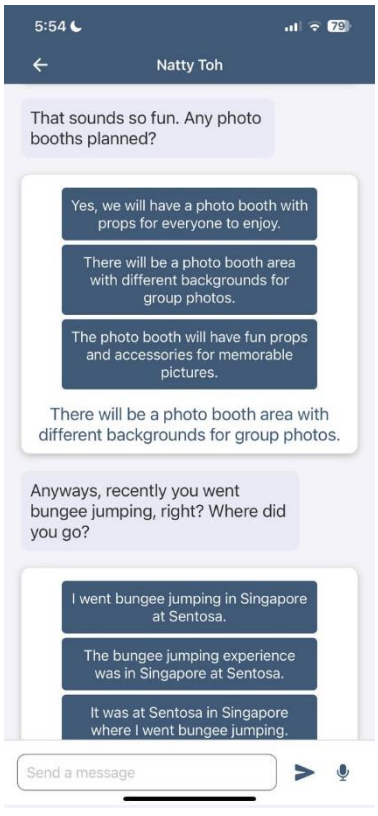
Appendix_B



Appendix_C



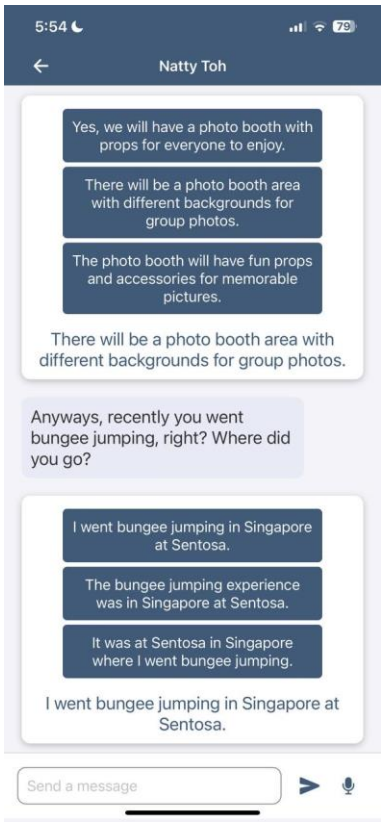
Appendix_D



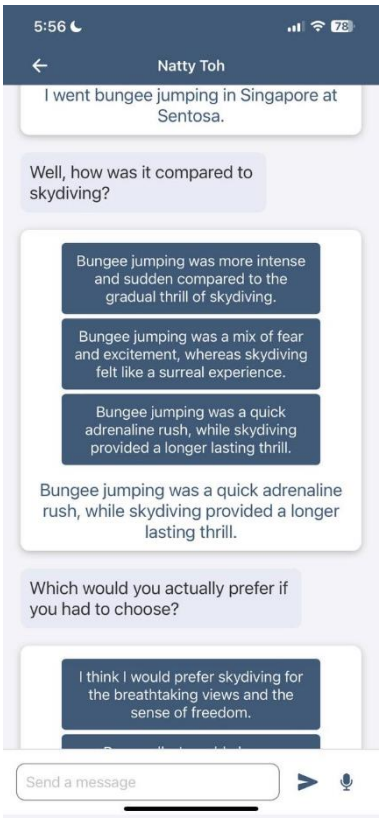
Back To Content Page

Appendix 7.3C: Screen Shots E-H (Role Reversal)

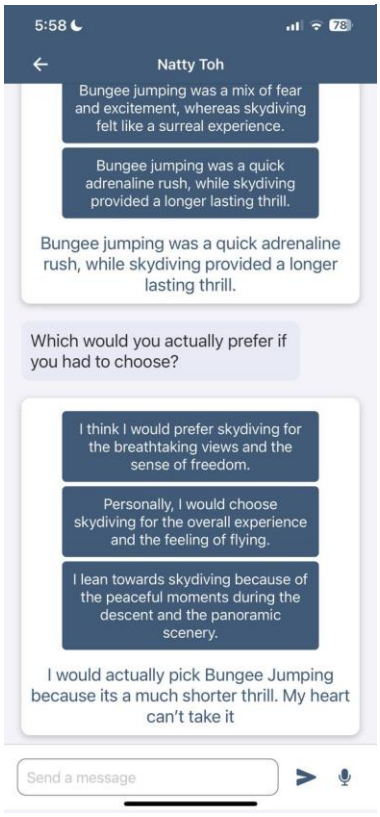
Appendix_E



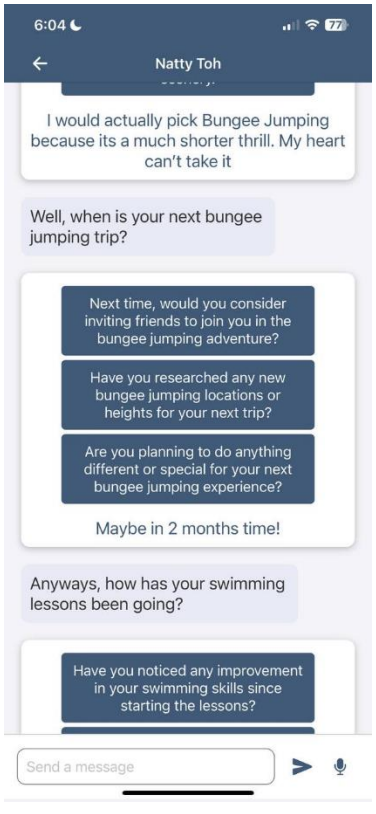
Appendix_F



Appendix_G



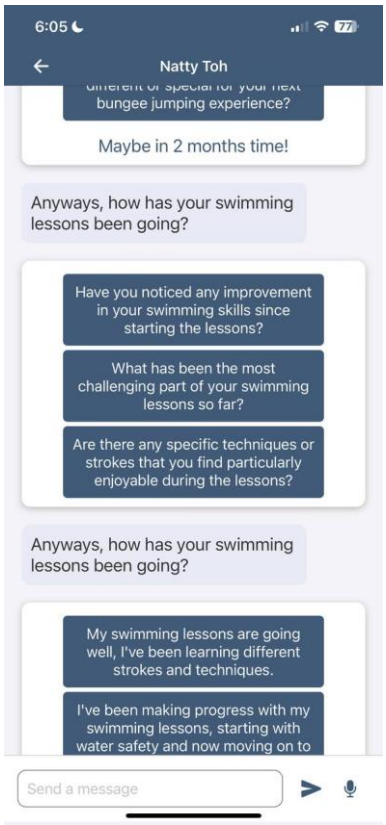
Appendix_H



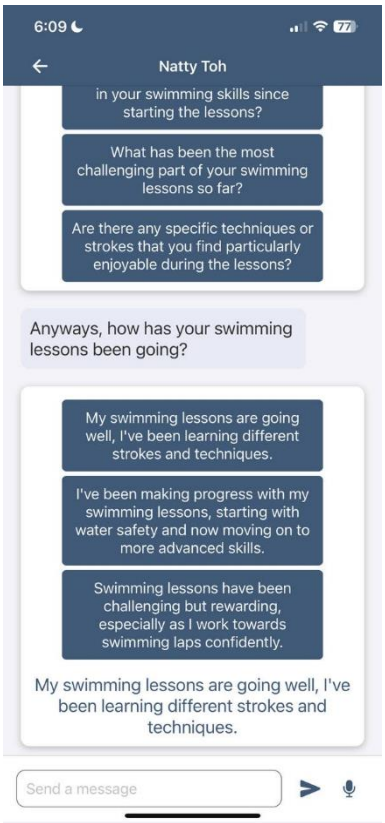
Back To Content Page

Appendix 7.3C: Screen Shots I-L (Role Reversal)

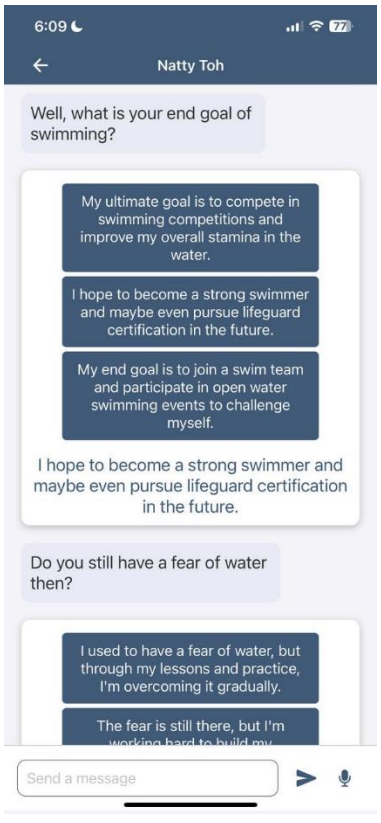
Appendix _I



Appendix _J



Appendix _K



Appendix _L

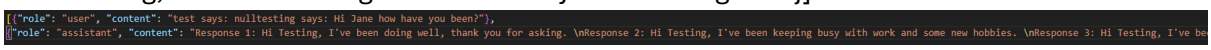
Back To Content Page

Appendix 8A: Mock Past History Data

Old data:

[{"role": "user", "content": "test says: nulltesting says: Hi Jane how have you been?"},

{"role": "assistant", "content": "Response 1: Hi Testing, I've been doing well, thank you for asking. \nResponse 2: Hi Testing, I've been keeping busy with work and some new hobbies. \nResponse 3: Hi Testing, I've been feeling a bit tired lately but overall good."}]



Appendix 8B: New history data after reset

Responses generated stored in the history below:

[{"role": "user", "content": "Jane says: nullGary says: Hi Jane, how have you been?"}, {"role": "assistant", "content": "Response 1: Hi Gary, I've been good thanks for asking.\nResponse 2: Hi Gary, I've been keeping busy, how about you?\nResponse 3: Hi Gary, it's been a bit hectic but I'm managing."}]