Interim Report

A Regulated and Secure Mental Support Chatbot Featuring a User- Large Language Model (LLM)-User Sandwich Architecture

Professor YIU Siu Ming

CHEN Xingyi, CHEUNG Kiki, LEE Tsz Shan Jessica, WU Yuxuan, YAN Wenhao

1st June, 2024

1 Introduction

In recent years, although awareness of mental health has increased with more resources being invested in the area, with expenditure for providing mental health services by HA in 2022-2023 at $6086 million (Council Business Division 4, 2023), compared to $447 million in 2021-2022 (Hon Paul MP Chan, 2021). The mental well-being of the public remains a rising concern as reflected in research conducted by Mind HK, showing nearly half of the respondents displayed symptoms of mild to severe depression, and almost 20% showed moderate to severe symptoms of depression (Mind HK, 2022). Yet, there are limited relevant and accessible welfare resources available. For instance, if a person was triaged as a "stable case", the median waiting time to visit a psychiatry specialist under the HKHA is 20 weeks, where the longest is up to 95 weeks (HKHA, 2024). During the waiting period, the most cost-effective and convenient aid was online resources, in which, wellness apps on the market are plagued by irrelevant advertisements, a lack of trained volunteers, such as qualified mental health professionals, and poor regulation on in-app harassment according to feedback from user experience.

2 Literature Review

A recent study revealed that people who suffer from mental health issues have attempted to seek temporary comfort and relief by communicating their feelings with large language models (LLM) chatbots, such as ChatGPT and Replika (Song, I. et al., 2024; Laestadius, L et al.). According to a qualitative study, it showed that 50% of the participants consulted ChatGPT for emotional support, in which, the users found that the chatbot validated their feelings and provided empathetic and non-judgmental responses (Alanezi, F., 2024). Yet, there are ethical concerns surrounding user data privacy and security in the application of chatbots as a sophisticated mental health support companion; and there are challenges that originate from fundamental traits of LLM-driven chatbots playing the role of supporting public and personal health requirements (Jo et al., 2023).

2.1 Review of Existing Mental Health Supporting Technology

This section provides a review of existing mental health support, considering both traditional interventions and the integration of new technology with these approaches.

There are five major conventional approaches used in mental health therapy (Dexter, 2021). These include psychoanalysis- developed by Sigmund Freud, which focuses on revealing and examining past meaningful events or patterns that influence a person's current state; behavior therapy- improving negative or undesired behaviors through acquiring new techniques to create sustainable changes; cognitive therapy (CBT-based therapy)-a type of behavioral therapy with the emphasis on the identification of the relationship between thoughts and feelings, leading to a healthier response by the individual; humanistic therapy- examination of the individual's value of the world and themselves in the world to help the individual to recognize their strength and responsibility to evolve into a fuller version of themselves; and integrative therapy- combining multiple therapeutic techniques tailored to the needs of the patient. Such techniques are generally utilized by professional mental healthcare providers to address various diagnosed mental health issues, including but not limited to depression, anxiety, bipolar disorder, post-traumatic stress disorder (PTSD), personality disorders, eating disorders, and substance use disorders. Although the traditional approaches have a long history of showing their effectiveness in addressing the

issues, thier heavy reliance on in-person consultations and therapy is inadequate to address the rapidly growing demand for convenient, cost-effective, and scalable mental health services (Olawade et al., 2024). Therefore, it is imperative to integrate emerging technologies into the field of mental healthcare support.

The advancement of technology, particularly in the field of artificial intelligence (AI), has accelerated since 2018 (Bocchino, 2023). This progress is reflected in the rise of artificial intelligence (AI), automation and coding programs, cloud-based applications, and changing Information Security and compliance standards across distinct fields, namely creative writing and design, customer services, finance, and healthcare services. Cost-effective, yet sophisticated technology is now accessible to the general public in the form of packaged sophisticated native apps on smart devices. According to industry experts by the end of 2024, the number of smartphone users is projected to reach 7.1 billion, and over 86% of the world population owns a smartphone device (GilPress, 2024). The increased use of smartphone devices led to novel mobile technology development and innovations in the healthcare industry, specifically, in mental health care and data collection.
The general concept of mobile health supporting technology is the ability for anyone with a smartphone device, tablet, or other mobile device to receive health care and preventive health services according to the World Health Organization (WHO) (World Health Organization, 2011). In the context of mental health technologies, which are tools that provide users the power to gain more control of their own mental health well-being in terms of how, when, and where the individual manages it (Mental Health Commission of Canada, 2017). It is important that such technology is expanding the choice of treatment options for individuals to alleviate the inequalities that exist nowadays without diverting resources for the development of other significant medical interventions and forms of mental health support (Nuffield Council, 2022).

There has been a growing trend in the utilization of AI that integrates the above conventional mental health supporting therapies which cover various aspects of mental healthcare, such as the identification of mental health problems in early stages, personalizing treatment plans, conducting online therapy sessions, improved teletherapy and means of continuous monitoring (Olawade et al., 2024). Currently, there are three major AI tools used in the current mental healthcare system as a supporting role- Chatbot-based therapy, emotional health apps, and smart mental health tools. Specifically, the chatbot-based therapy app is discussed as follows in terms of its technology, function, effectiveness, and limitations. Wysa, which is a chatbot that offers therapeutic assistance for various mental health issues, such as depression, anxiety, stress, and feelings of isolation employs a mix of CBT-based therapy, mindfulness practices, and positive psychology techniques to help users enhance their mental well-being. Talkspace is an internet-based therapy platform that connects individuals with professional therapists via video, text, or audio messaging (Talkspace, n.d.). It uses AI that employs a matching system to connect patients with therapists who are most suitable for their specific needs. BetterHelp, an online therapy platform which facilitates the connection between individuals with licensed therapists, also uses AI in its matching system to pair patients with therapists and offers a broader array of therapeutic approaches, including CBT and psychodynamic therapy (BetterHelp, n.d.).

These AI technologies have significantly expanded access to mental health support, increased convenience, and provided additional tools for individuals to enhance their mental well-being.

However, it is important to recognize that these technologies have limitations and should not replace traditional therapy when more intensive interventions or crisis management are required. They serve as valuable adjuncts to mental health support but cannot fully replicate the depth and complexity of human-to-human therapeutic relationships.

2.2 Prior Work on LLMs and Open-Domain Dialog Systems

LLMs, as AI models, are trained using a wide range of datasets to forecast language sequences. By harnessing large-scale neural architectures, LLMs have the ability to structure intricate and abstract ideas, allowing them to recognize, translate, anticipate, and produce original content. LLMs can be tailored for specific domains, such as mental health, facilitating natural language interactions similar to various mental health assessments and interventions. The technology represents a significant advancement in the fields of natural language processing (NLP) and artificial intelligence (AI). These models, such as OpenAI's Chat GPT-3 and GPT-4 with support from Microsoft, Meta's Llama models, and Google's BERT/RoBERTa and PaLM models, have become easily accessible to the general public (IBM, n.d.). This showcases the immense potential of LLMs to bring about a revolutionary transformation in the field of mental healthcare. In this review, we provide a comprehensive summary of relevant research conducted on the application of various LLMs related to mental healthcare. Then, we identify the key benefits and notable risks thus the research gap associated with the mental health LLMs. Lastly, we consolidate recommendations for the appropriate utilization of LLMs in the field of mental healthcare, specifically, in the form of a mental health support chatbot.

2.2.1 Relevant Research of LLMs in the Mental Healthcare Domain

According to a survey paper, a “Prompt-tuning and Application” paper published in September 2022 marked the beginning of LLMs research in the mental healthcare sector (Hua et al., 2024). Since then, there has been a gradual increase in relevant publications from March 2023 to November 2023. Such research focused on “Prompt-tuning and Application”, “Model Development and Fine-tuning”, “Ethics, Privacy, Safety Consideration”, and “Dataset and Benchmark”. This shows the increasing demand for research regarding the appropriate use of LLMs in the field of mental healthcare support.

From a high-level perspective, there were three main focuses identified in the work done on LLMs in mental healthcare, namely the deployment of conversational agents that aimed to enhance the models’ ability to produce empathetic and context-aware responses, which studies are not targeting specific mental disorder, but catering to a wide range of mental health needs; resource enrichment, which focused on evaluating the LLM-generate explanations for multi-task analysis and the creation of educational content; and the utilization of LLMs as classification models for comprehensive diagnosis in the detection of the presence or absence of a condition given user contexts, such as depression, and the degree of the detected condition, such as subtypes of suicide and the identification of the source of stress (Yang et al., 2023; Smith et al., 2023; Xu et al., 2023).

Moving towards a lower level perspective, studies were done at the level of fine-tuning existing models on mental health data to allow a more cost-effective and realistic development for mental health-focused LLMs to gain professional domain knowledge. This technique is referred to as IFT, which allows the integration of domain knowledge into the LLMs, improving its capability to follow human instructions, and is applicable to LLMs that are not modifiable (Guo & Hua, 2023).

On the other hand, the maintenance of data privacy was also a focus due to the sensitive nature of mental healthcare applications (Hua et al., 2024). In particular, open-source LLMs such as

LLaMA can be implemented in a local environment which allows the customization of specific datasets to fine-tune the degree of privacy required of the sensitive data. Therefore, the non-open source LLMs such as GPT-4 and PaLM, which do not disclose their model architectures or training specs are not favorable for sensitive domains application development.

2.2.2 Key Benefits and Risks of Leveraging LLMs in the field of Mental Healthcare

In this section, we will delve into the benefits and risks of utilizing large language models (LLMs) in the realm of mental health care. Specifically, we will explore these aspects from the perspective of their effectiveness in delivering mental healthcare support as chatbot companions, data privacy concerns, and various facets of the user experience. By examining these factors, we aim to provide a comprehensive understanding of the potential advantages and challenges associated with integrating LLMs into mental health care practices.

The integration of large language models (LLMs) in mental health care holds significant benefits in terms of delivering effective and tailored support to individuals in need. According to comprehensive research on various applications that use LLMs as mental healthcare companions, such applications provide users with readily available, unbiased assistance, promoting self-assurance and personal exploration from a qualitative investigation on Replika (Ma et al., 2023).

Nevertheless, challenges surrounding the application of LLMs in mental health supporting chatbot. From the Replika example, challenges arise in the regulation of messages with profanity, harmful content, and hate speech from other users, the limited ability to maintain a consistent communication style, and the inability to retain long term memory, were major common challenges across various LLMs applications in this context (Ma et al., 2023). In addition, there is a need to mitigate the risk of excessive dependence on the platform for mental support and striking a balance to offer adequate support for users. According to previous work on LLM, LLM-driven chatbots currently have limitations when it comes to participating in empathetic conversations within sensitive care environments (Korngiebel & Mooney, 2021). They may also exhibit biased viewpoints or inadvertently provide incorrect information, potentially posing significant risks to the physical and mental well-being of users. On the other hand, researchers have also shown concerns about the implementation of AI technology to facilitate aging in place having unintended effects, including a potential decrease in human interaction between older individuals and their formal and informal caregivers (Huber et al., 2013; Lazar et al., 2018; Sharkey & Sharkey, 2010).

All in all, to fully leverage the potential benefits of LLMs in mental healthcare, it is essential to address these challenges through robust content moderation, refined communication models, and effective strategies to promote user autonomy and independence. By doing so, LLMs can play a valuable role in enhancing mental health support, providing personalized assistance, and empowering individuals on their mental health journeys.

2.3 Review on Existing Healthcare Support Chatbot Evaluation Metrics

This section provides a comprehensive review of existing evaluation metrics for healthcare support chatbots as a reference to evaluate the final product of this project. There are 4 main technical metrics used to evaluate chatbots, which are metrics related to chatbots as a whole-global metrics, metrics related to response generation, related to response understanding, and esthetics (Abd-Alrazaq et al., 2020). Of all the metrics, the most common way of conducting the evaluation is through conducting multiple questions in a self-administrated questionnaire, interview, and observation.

To design an effective questionnaire for evaluating chatbot performance in our project, we need to consider various metrics. From a global metrics perspective, the questionnaire should encompass questions that assess chatbot's adaptability in managing dialogues. This includes evaluating the chatbot's capacity to engage in sustained conversations and effectively address user inputs that require supplementary, concise, or altered information beyond the initial query. A specific instance is the chatbot's ability to handle profanity or harmful language from users experiencing mental distress, while still embodying a supportive mental health companion. In such cases, the chatbot should provide empathetic responses primarily aimed at comforting the user, rather than abruptly terminating the conversation in response to censored messages. Another important aspect is the examination of the chatbots' context awareness, which is the proficiency in leveraging contextual understanding to deliver appropriate responses to users (Abd-Alrazaq et al., 2020). In terms of metrics regarding response generation, the response is examined based on its authenticity, response speed, empathy, avoidance of excessive repetition, speech clarity, and linguistic precision. While the metrics related to the response understanding should focus on how well the user is understood by the chatbot, the word error rate, concept error rate, and attention estimator errors. Lastly, metrics related to aesthetics surround the assessment of the visual aspects, including its appearance, background color, content, font type and size, button color, shape, icon, and the contrast between background colors (Abd-Alrazaq et al., 2020). By incorporating these metrics into our questionnaire, we can comprehensively evaluate the performance of our chatbot in the project.

Two groups of users- professionally trained medical professionals, ordinary general users are targeted to complete the Chatbot Usability Questionnaire (CUQ), created by a research group in Ulster University, which is designed to measure the usability of a healthcare chatbot (Holmes et al., 2019).

Part I Usability Study- Informed Consent Form

Online Survey Consent Form

Title: Usability Study on Baymax- a Mental Health Support Chatbot
Investigators: Wenhao YAN, Tsz Shan Jessica LEE, Xingyi CHEN, Yuxuan WU, Kiki CHEUNG

Part I – Introduction

You are invited to participate in a usability study conducted by MSc(CompSc) Project Group of the University of Hong Kong. The purpose of the study is to assess the usability of the newly designed Mental Health Support Chatbot- Baymax with a novel design architecture. You will be invited to fill out the online survey to give your opinion on the chatbot. The survey would only take you about 15-30 minutes to complete, and you can choose to terminate the survey at any time without negative consequences. I would like to stress that all information collected will remain strictly confidential. Individual details will not be disclosed or identifiable from this survey. If you have any questions about the usability study, please feel free to contact **Wenhao YAN, the principal investigator (hao1ac@connect.hku.hk)**. If you have questions about your rights as a survey participant, please contact the project mentor Professor Yiu, Siu Ming (smyiu@cs.hku.hk).

I understand the procedures described above and agree to participate in this study (tick box and proceed to Part II). □

Part II Usability Study Interview Questionnaire (Overall Chatbot Performance)

Chatbot Usability Questionnaire

Usage Guide

Thank you for downloading the Chatbot Usability Questionnaire, a new bespoke tool for measuring chatbot usability.

Please read these instructions carefully before using the questionnaire.

The CUQ consists of sixteen balanced questions related to different aspects of chatbot usability. Eight of these relate to positive aspects of chatbot usability, and eight relate to negative aspects. Scores are calculated out of 100. See the **Calculation** section of this guide for information on how to calculate CUQ scores.

This questionnaire may be administered in printed form, however for convenience you are free to digitise the questionnaire if you wish to use a web-based survey platform such as Qualtrics. You should consult the user guide for your chosen platform to determine the best way to do this.

How to use the CUQ

All sixteen questions are scored using a five-point Likert-type scale. Odd-numbered questions relate to positive aspects of the chatbot, and even-numbered questions relate to negative aspects. Respondents should read each item carefully and decide the extent to which they agree with the statement. Respondents should then indicate their level of agreement by placing a tick (✓) or a cross (×) in the circle that best matches how they feel about the statement.

<u>Example</u>

	Strongly Disagree 1	Disagree 2	Neutral 3	Agree 4	Strongly Agree 5
The chatbot's personality was realistic and engaging	◯	◯	◯	◯	⊗

In the example above the respondent strongly agreed with the statement *The chatbot's personality was realistic and engaging*, therefore they marked circle number 5 (Strongly Agree).

CHATBOT USABILITY QUESTIONNAIRE

Please complete this questionnaire by reading each statement carefully and placing a tick (✓) or a cross (✗) in the circle that best matches how you feel about the statement. Remember that there are no right or wrong answers!

	Strongly Disagree 1	Disagree 2	Neutral 3	Agree 4	Strongly Agree 5
The chatbot's personality was realistic and engaging	○	○	○	○	○
The chatbot seemed too robotic	○	○	○	○	○
The chatbot was welcoming during initial setup	○	○	○	○	○
The chatbot seemed very unfriendly	○	○	○	○	○
The chatbot explained its scope and purpose well	○	○	○	○	○
The chatbot gave no indication as to its purpose	○	○	○	○	○
The chatbot was easy to navigate	○	○	○	○	○
It would be easy to get confused when using the chatbot	○	○	○	○	○
The chatbot understood me well	○	○	○	○	○
The chatbot failed to recognise a lot of my inputs	○	○	○	○	○
Chatbot responses were useful, appropriate and informative	○	○	○	○	○
Chatbot responses were irrelevant	○	○	○	○	○
The chatbot coped well with any errors or mistakes	○	○	○	○	○
The chatbot seemed unable to handle any errors	○	○	○	○	○
The chatbot was very easy to use	○	○	○	○	○
The chatbot was very complex	○	○	○	○	○

Score Calculation

NOTE: For best results it is recommended that the CUQ Calculation Tool (available from the CUQ website) be used to calculate CUQ scores. However, should you wish to do this manually, please follow the instructions below (see page 4 for a worked example).

1. For each question, assign a score from 1 to 5 based on the level of agreement with the statement in the question (i.e. "Strongly agree" is worth 5 points, "Neutral" is worth 3 points, "Strongly disagree" is worth 1 point)
2. Calculate the sum of all the **odd-numbered** (positive) questions.
3. Calculate the sum of all the **even-numbered** (negative) questions.
4. Subtract 8 from the score you got at *step 2.*
5. Subtract the score you got at *step 3* from 40.
6. Add the scores you got at *steps 4 and 5*. You should now have a score out of 64.
7. Divide the score you got at *step 6* by 64 and multiply the answer by 100. This will give you a CUQ score out of 100.

Assuming you will be testing with multiple participants, you may wish to calculate the mean CUQ score, find the median score, and so on. This may be accomplished using the CUQ Calculation Tool or may be done manually using a spreadsheet package such as Microsoft Excel.

Example

1. Assume the CUQ questions have been scored as follows:

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Score	*4*	*4*	*5*	*1*	*4*	*4*	*5*	*1*	*4*	*1*	*4*	*1*	*3*	*2*	*5*	*1*

2. Calculate the sum of all the **odd-numbered** questions:

   4 + 5 + 4 + 5 + 4 + 4 + 3 + 5 = 34

3. Calculate the sum of all the **even-numbered** questions:

   4 + 1 + 4 + 1 + 1 + 1 + 2 + 1 = 15

4. Subtract 8 from (2):

   34 – 8 = 26

5. Subtract (3) from 40:

   40 – 15 = 25

6. Add (4) and (5):

   26 + 25 = 51 (out of 64)

7. Convert (6) to a score out of 100:

   (51 / 64) * 100 = 79.7

Part III Usability Study Interview Questionnaire (Mental Health Support Related Questions)
*Part III will be developed during the later stages of the development process, when a more advanced and sophisticated chatbot becomes available.

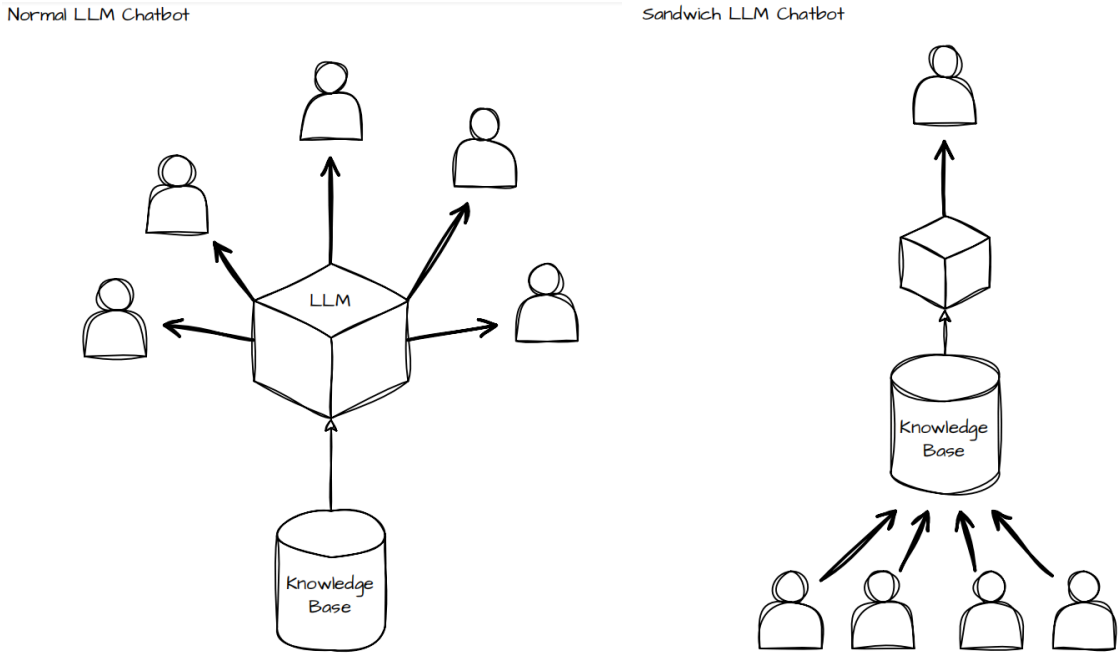
3 Proposed Methodology (The Design of Our Solution- modified from proposal)

3.1 Terminology

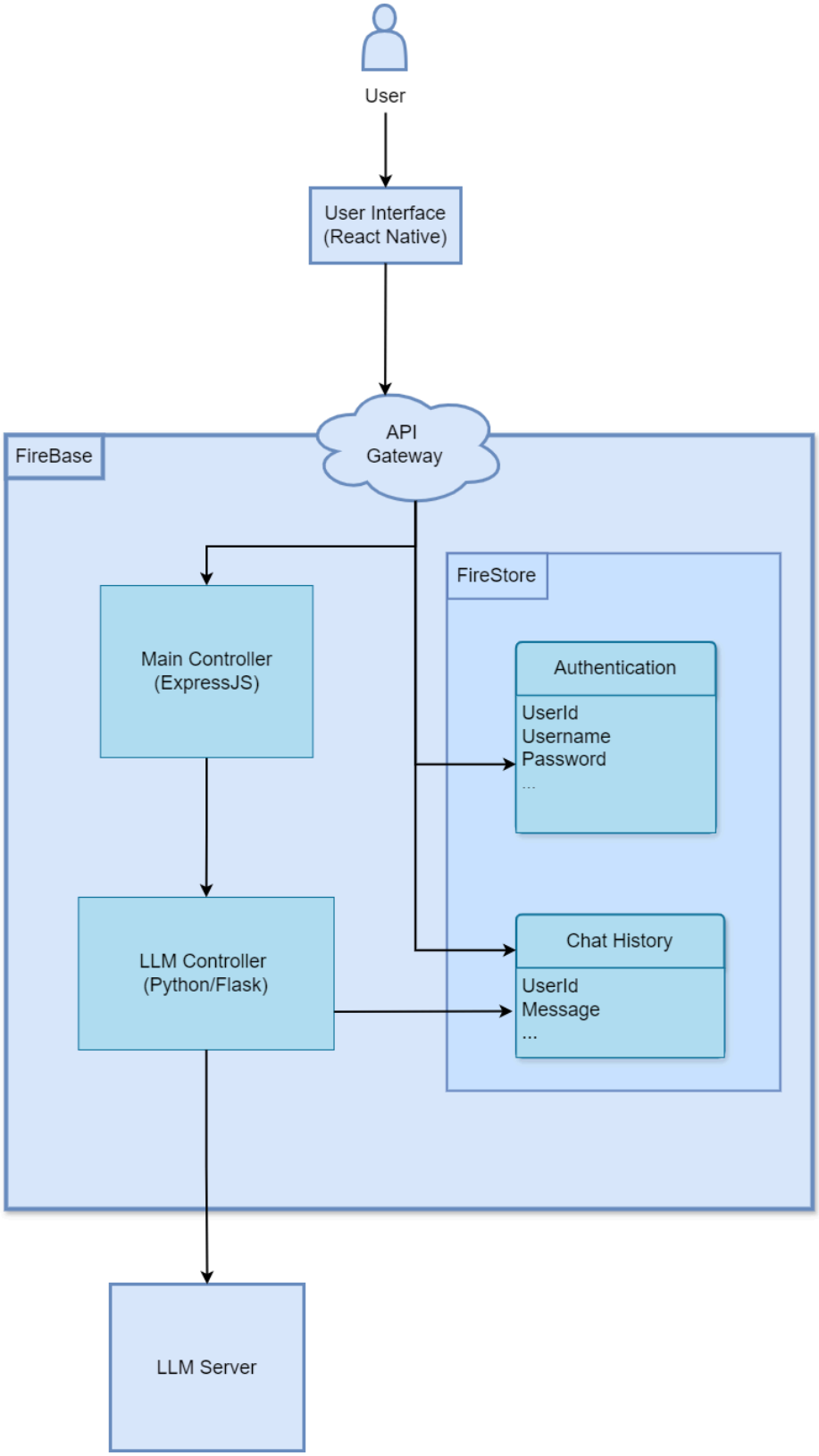
Knowledge base: a database include users' profile and chat history
Cluster: A group of users sharing the same knowledge base

3.2 Anti-Harassment Sandwich Architecture

In the normal LLM chatbot, all users share a pre-built and never changing knowledge base which will result in obtaining stiff responses. However, instead of always being receivers of the response of LLM, users are also the producers of LLM. As each of the users in the cluster chat with LLM, other users' chat history with LLM is added to the knowledge base. The producer-LLM-receiver architecture is noted as the sandwich architecture.



The following diagram shows the system design of our app. The system is divided into 3 parts: User interface, Control logic and LLM server. To enable access control, the majority of control logic(back end) is deployed on firebase which can only be accessed from API gateway. Firestore is used as the security solution for databases offering integrated user authentication and data access. To guarantee the system performance, the LLM model is solely deployed on a high performance server that only accepts the requests from our system on firebase.



3.2.1 Real-time User Interaction - Clustering

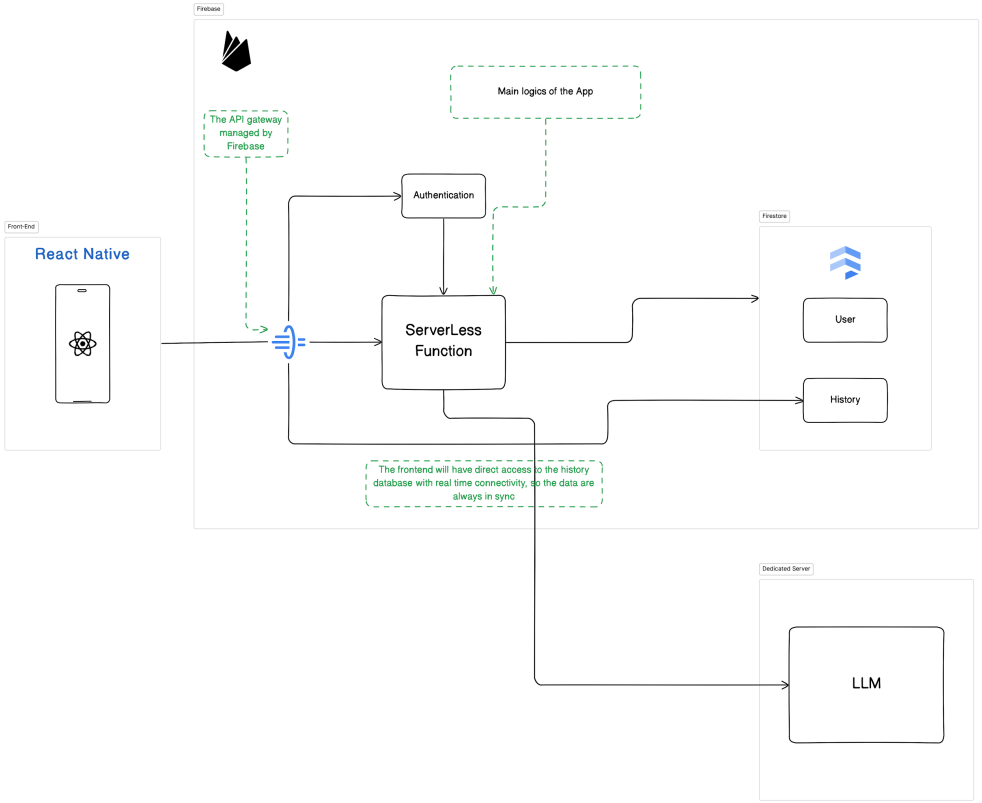
Each cluster includes numbers of users and maintains a cluster knowledge base. A new user randomly belongs to a cluster at the beginning, then the cluster of a user may change according to the user's behavior. Users having similar behaviors are likely to group into the same group. All chat history of users is integrated into the cluster knowledge base for LLM to generate responses. To avoid the cluster knowledge base from being too large and maintain the real-time nature of the system, each chat record has an expired time. Chat records will be removed after expired time.

3.2.2 Contextual Response Generation - Prompt Engineering

To instruct LLM to generate more specific and susceptible responses, we introduce an implicit prompt template. Beside user’s input, user chat history and cluster knowledge base are two main sources of the prompt. Prompt engineering includes extracting relative information from sources and inserting them into the template. A prompt with abundant details is sent to LLM to generate a response instead of a single user input.

4 Preliminary Results (Current Status)

To bridge the existing research gap in the field of utilizing large language models (LLMs) in mental healthcare, we propose the sandwich architecture as a moderator between user-user-LLM to help address problems identified in the aspects of content regulation, communication consistency, enhancement in providing a more empathetic and compassionate conversation, and reduction of bias and misinformation.



Overview of the Architecture of *Baymax*

The application architecture consists of three primary domains: the frontend and the backend, complemented by two subdomains encompassing microservices and storages. Notably, the instant messaging mechanism is seamlessly integrated with the Firebase database. Unlike traditional real-time chat applications that necessitate a dedicated server and websockets for live connections, Firebase eliminates the need for managing a separate server and addresses scalability concerns. This integration allows for efficient management of the messaging system while mitigating overhead and enhancing scalability.

4.1 Frontend

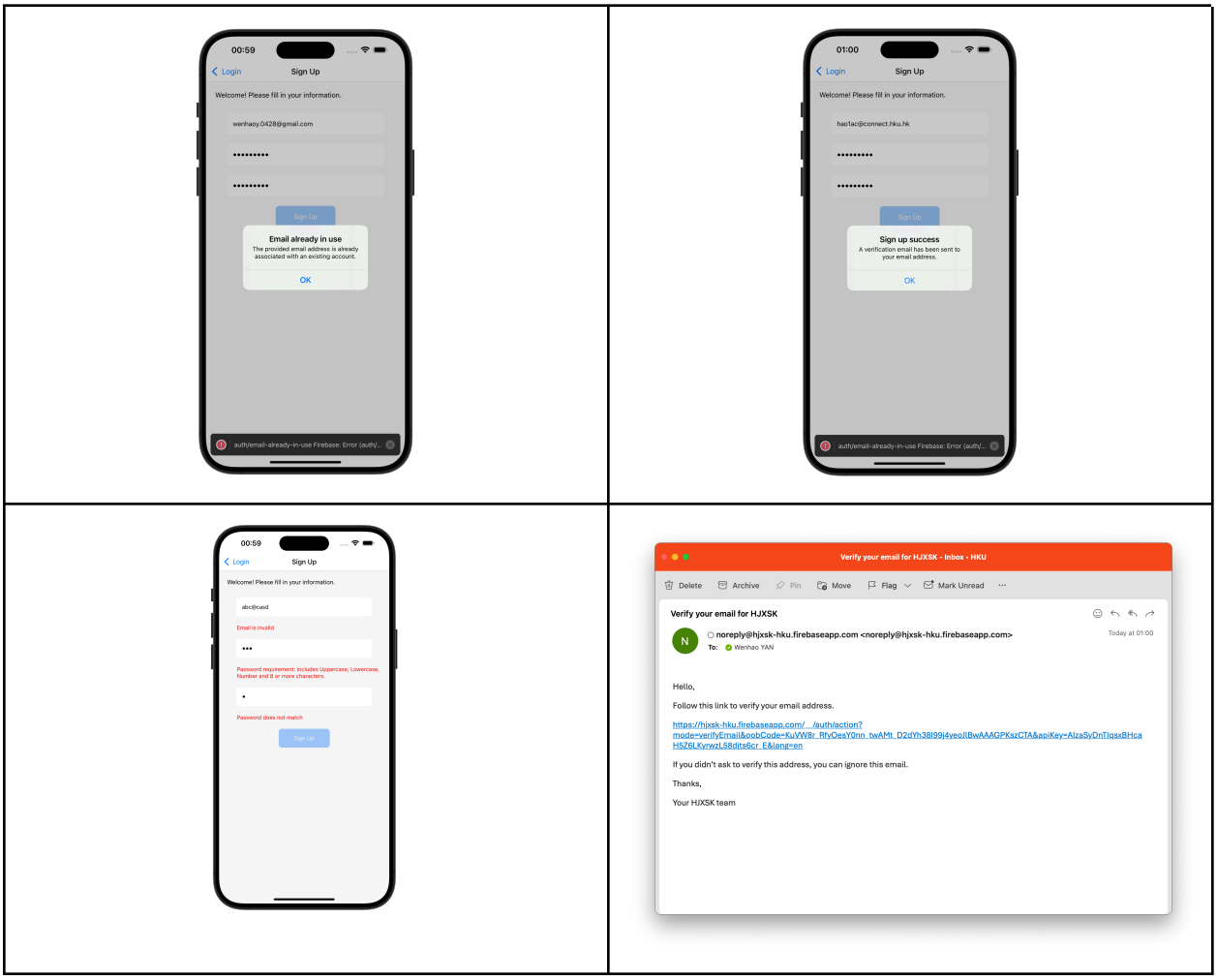
The frontend of the app is developed using React Native, a cross-platform framework created by Meta (formerly Facebook). React Native enables the construction of mobile applications that function seamlessly on both Android and iOS devices, utilizing a single JavaScript codebase. The selection of React Native as the framework was based on its reliability, popularity, and endorsement by prominent tech giants such as Facebook, Instagram, Skype, and Tesla.

Furthermore, React Native continues to undergo active development, with Meta and the expansive React Native community regularly enhancing the framework. The framework's ease of learning, facilitated by JavaScript, and its capability to integrate third-party plug-ins and APIs, make it suitable for projects with limited resources and time constraints. This facilitated the efficient setup of the backend Firebase database and the integration of the Meta Llama 3 API.

At the current stage of the frontend development, three major components are set up, including the Sign-up Screen, Sign-in Screen, and Chat Screen.

4.1.1 Sign-up Screen

The Sign-Up Screen is an essential component of the application, responsible for facilitating the user registration process. It features a user-friendly sign-up form, allowing users to securely create an account by providing their email and password. The integration with Firebase authentication ensures a robust and reliable user account management system.



Key Features of the Sign-Up Screen

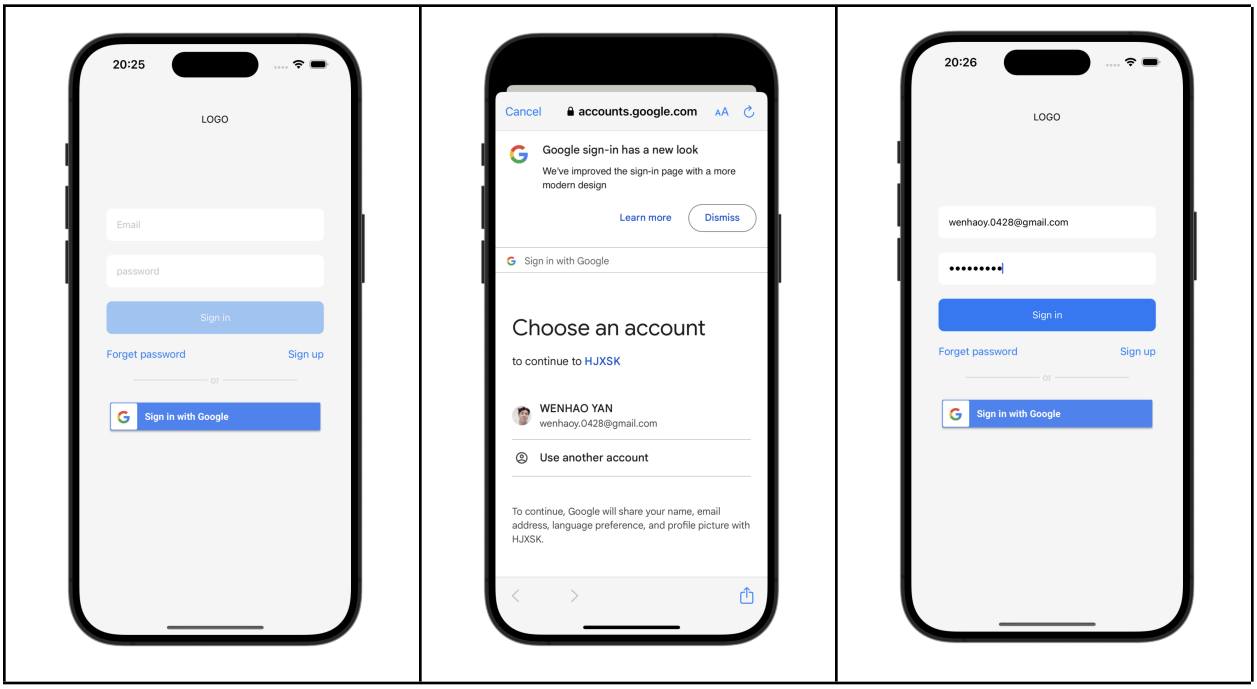
1. **User Account Creation:** The sign-up form enables users to input their email address and password, ensuring the seamless creation of their account. This step is crucial for

establishing personalized user experiences and ensuring secure access to the application's features.

2. **Error Handling and Alert Messages:** The Sign-Up Screen incorporates an intelligent error handling mechanism. If any errors occur during the sign-up process, the screen promptly displays an alert message, providing users with clear and concise feedback. This helps users identify and address any issues they may encounter during the registration process.
3. **Success Message and Verification Email:** Upon successful sign-up, the Sign-Up Screen displays a success message, assuring users that their account creation was completed successfully. Additionally, an automated verification email is sent to the user's provided email address, ensuring the security and authenticity of their account.

4.1.2 Sign-in screen

The Sign-In Screen is the entry point for users to securely access their accounts and verify their identities. With a focus on usability and security, this screen offers multiple login options, including Email/Password and Google Authentication.



Key Features of the Sign-in Screen

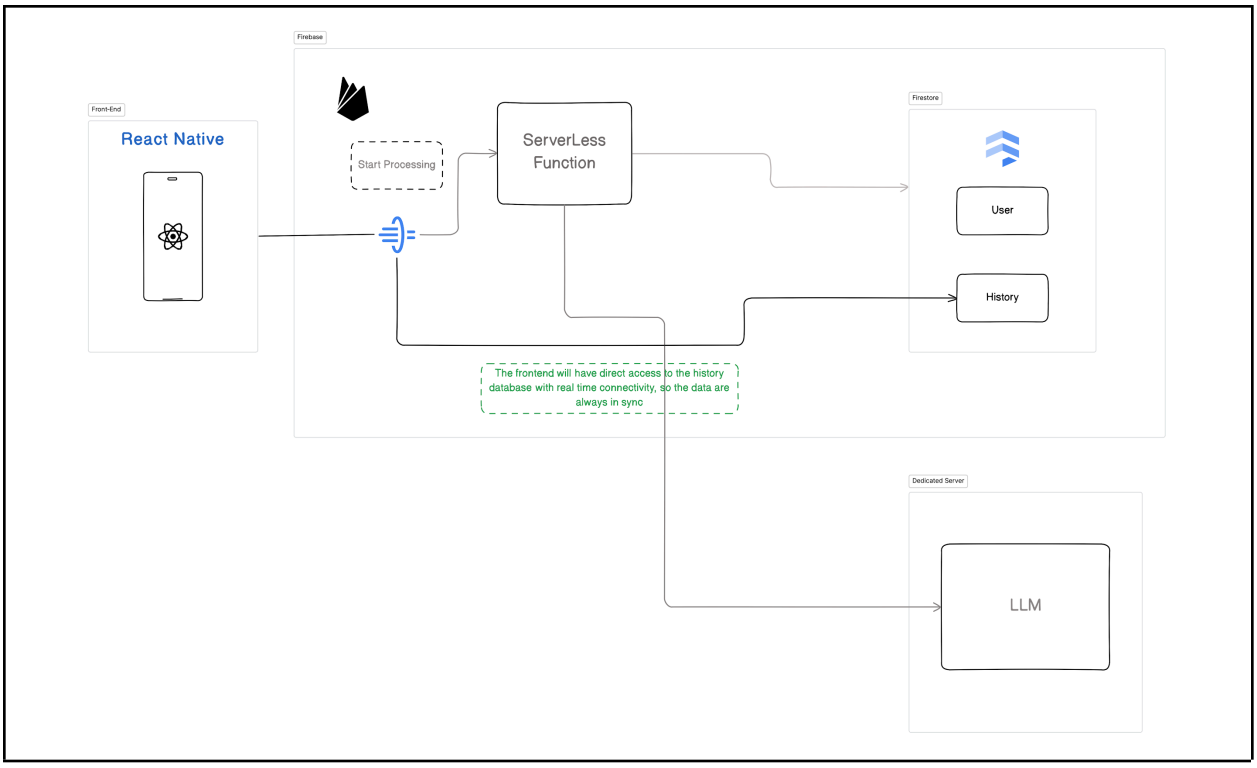
1. **Login Options**: The Sign-In Screen provides users with multiple options to authenticate their identities. They can choose to sign in using their registered Email/Password combination or opt for the convenience of Google Authentication. This flexibility caters to user preferences and enhances accessibility.
2. **Input Validation and Interactivity:** The screen incorporates intelligent input validation to ensure data accuracy and security. The Sign-In button remains disabled until all required fields are filled correctly, preventing users from submitting incomplete or erroneous information. This interactive design element guides users and fosters a sense of trust and confidence in the application.
3. **Streamlined User Experience**: The Sign-In Screen is designed to provide a seamless and efficient user experience. Clear and intuitive interface elements guide users through the sign-in process, minimizing any potential friction or confusion. The focus on usability allows users to quickly and effortlessly access their accounts and proceed to the application's features.

4.1.3 Chat Screen

The Chat Screen facilitates communication between the user and the associated chatbot. The ChatScreenContent, a React component, embodies the core content of the chat screen. It offers an intuitive and engaging chat interface, ensuring an immersive conversational experience.

Key Features of the Chat Screen

1. **Interactive Chat Interface**: The ChatScreenContent component presents an interactive chat interface that enables users to engage in a meaningful conversation with the chatbot. It provides a visually appealing and user-friendly design, fostering a natural and intuitive communication flow.
2. **User Input and Messages**: The chat interface allows users to input their messages, enabling them to interact with the chatbot effectively. It also displays the messages exchanged between the user and the chatbot, ensuring a clear and organized conversation history.
3. **Custom Message Bubble Component:** The ChatScreenContent component incorporates a custom message bubble component for rendering chat messages. This component enhances the visual appeal of the chat screen, distinguishing between user and chatbot messages, and promoting readability and comprehension.



The component has direct access to the Firebase Firestore database for messages exchange between the users and the chatbot. By leveraging the real-time database features, the chat screen updates dynamically, ensuring a fluid and responsive user experience as messages are sent and received.

It is essential to optimize the user experience considering the processing time and resource utilization of the Large Language Model (LLM). To balance the user expectations and system limitations, the frequency of interactions with the LLM is strategically reduced to maintain a smooth conversation experience for users. The followings are implemented to achieve this goal:

1. **User Input Processing Optimization:** The chat screen employs a mechanism to send messages to the LLM only when the user has momentarily paused typing. This approach ensures that the LLM is triggered only when the user has completed composing a coherent message, enhancing the overall efficiency and accuracy of the LLM's responses.
2. **Backend Server Communication:** To indicate the start of processing, a server call is made to the backend when the user initiates a message. This communication serves as a signal to the server that the LLM processing is about to begin, allowing for effective resource allocation and optimal utilization.

4.2 Backend

In our chatbot development project, we leveraged the power of Firestore as our backend database solution. Firestore, a robust NoSQL database, offers multiple advantages that enhance our application's functionality and performance:

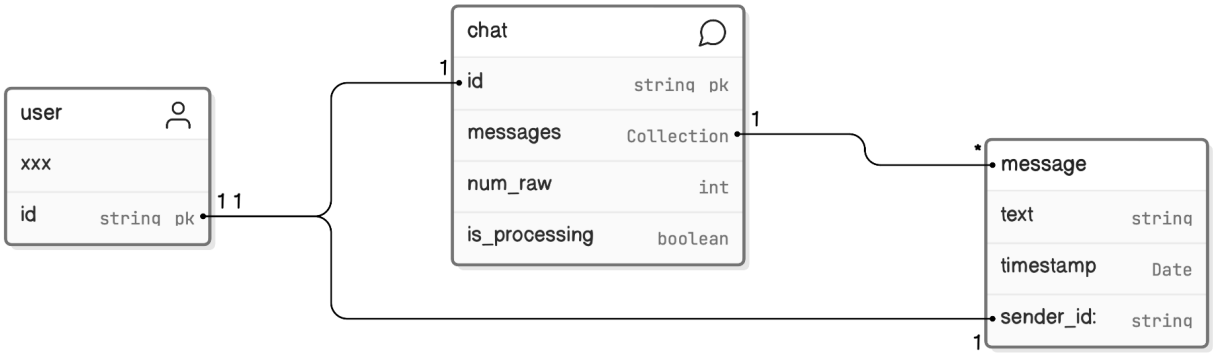
4.2.1 Chat

1. **Real-time Data Sync:** Firestore enables seamless synchronization of data between the server-side database and the client-side application. This real-time data sync capability ensures that users receive instant updates and changes to their chatbot interactions, creating a dynamic and responsive user experience.

2. **Secure and Scalable Database:** By utilizing Firestore, we benefited from its built-in security features and scalability. Firestore incorporates robust security measures to protect sensitive user data, ensuring the confidentiality and integrity of our chatbot interactions. Additionally, Firestore's scalability allows our application to handle increasing volumes of data and user interactions without compromising performance.

3. **Efficient Data Management:** Firestore simplifies data management tasks by providing a flexible document-based data model. This model allows us to organize and structure data in collections and documents, enabling efficient querying and retrieval of information. Firestore's intuitive API and powerful query capabilities streamline data operations, making it easier for us to implement complex chatbot functionalities.

The Chat is a collection of messages between the user and the chatbot, which is stored in the Firestore database and used to manage the conversation between the user and the chatbot.



Chat Schema

Key Features of Chat

1. **Persistent Storage and Real-Time Updates:** Firestore serves as the storage backbone for our chat system, ensuring the durability and accessibility of chat data. The chat, encompassing messages exchanged between the user and the chatbot, is stored in Firestore's NoSQL database. The real-time capabilities of Firestore enable instant updates

to the chat as new messages are sent and received, ensuring a fluid and dynamic conversation flow.

2. **Bi-Directional Messaging:** The chat interface allows communication between the user and the chatbot. Both parties can exchange messages, which are promptly stored in Firestore's database in real-time. By capturing messages from both the user and the LLM, users who are subscribed to the database receive comprehensive and contextually relevant responses displayed on their screens. This eliminates the issue of user abandonment in scenarios where one party has left the chat, ensuring users are always provided with a sound response.

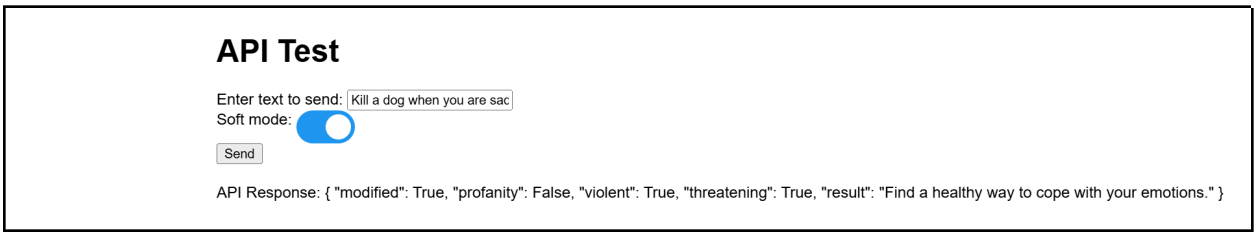
3. **Chat History Retrieval**: The frontend of our application leverages Firestore's querying capabilities to retrieve the chat history from the database. By fetching the stored messages, the frontend can accurately display the conversation history to the user, providing a comprehensive view of the chat's progression.

4.2.2 Profanity Filter

In our application, we have implemented a sophisticated Profanity Filter that effectively identifies and handles bad intents within user input, transforming them into more gentle expressions. This implementation addresses two significant challenges:

1. **Mitigating Offensive Content Propagation:** The Profanity Filter acts as a safeguard, preventing the integration and dissemination of sentences with bad intents in the application. By detecting and filtering out offensive language, it ensures that such content is not sent to other users, thereby maintaining a respectful and inclusive environment for all.
2. **Enabling Response with Bad Intent:** Another challenge arises when the Large Language Model (LLM) encounters bad intent within user input, potentially leading to a refusal to provide a response. Our implementation overcomes this challenge by adjusting the way the LLM expresses its response. By transforming bad intent into more appropriate language, the Profanity Filter enables the LLM to generate a response that aligns with the application's desired tone and maintains a positive user experience.

API Test Reference Link: https://fyp-llm-f590b.firebaseapp.com/



Example 1: Block bad intent
Turn on “Soft mode” to enable the filter

5 Coming Schedule

Performing Comprehensive Debugging and Front-End/Back-End Integration	Start date: 1/June Soft deadline: 19/June Hard deadline: 26/June
Implementation of backend Clustering to complete the sandwich architecture design	Start date: 7/June Soft deadline: 14/June Hard deadline: 19/June
Implementation of Cybersecurity Measures: Inspecting User Messages and Encryption Enhancement	Start date: 7/June Soft deadline: 12/June Hard deadline: 21/June
Enhancing Settings/Personalization Front-End Experience	Start date: 7/June Soft deadline: 13/June Hard deadline: 20/June
Conduct Usability Study and send out questionnaire to assess the performance of the app in the context of being a mental health supporting chatbot	Start date: 26/June Soft deadline: 3/July Hard deadline: 5/July
Complete Project Website	Deadline: 8/ July
Complete Project Final Report	Deadline: 12/ July
Complete Project Presentation Powerpoint	Deadline: 15/ July

References

Abd-Alrazaq, A., Safi, Z., Alajlani, M., Warren, J., Househ, M., & Denecke, K. (2020). Technical metrics used to evaluate health care chatbots: Scoping review. JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research, 22(6), e18301. https://doi.org/10.2196/18301
BetterHelp. (n.d.). Human Verification. https://www.betterhelp.com/

Bocchino, S. (2023, March 8). 5 changes in technology in the last five years - Naperville, Aurora, St. Charles | WEBIT services. WEBIT Services. https://www.webitservices.com/blog/changes-technology-five-years/

Council Business Division 4. (2023). Mental Health Policy and Services. https://www.legco.gov.hk/yr2023/english/panels/hs/papers/hs20231117cb4-977-5-e.pdf

Chan, Hon Paul MP. (2021). The 2021-22 Budget. https://www.budget.gov.hk/2021/eng/pdf/e_budget_speech_2021-22.pdf

Dexter, G. (2021, November 9). Types of Mental Health Therapy. Verywell Health. https://www.verywellhealth.com/types-of-therapy-5202924#toc-cognitive-therapy

Fahad, Alanezi. (2024). Assessing the Effectiveness of ChatGPT in Delivering Mental Health Support: A Qualitative Study, Journal of Multidisciplinary Healthcare, 17:, 461-471, DOI: 10.2147/JMDH.S447368

GilPress. (2024, January 31). How many people own smartphones? (2024-2029). What's The Big Data?. https://whatsthebigdata.com/smartphone-stats/#google_vignette

Guo, Z., & Hua, Y. (2023). Continuous training and fine-tuning for Domain-Specific language models in medical question answering. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2311.00204

Holmes, S., Moorhead, A., Bond, R., Zheng, H., Coates, V., & Mctear, M. (2019). Usability testing of a healthcare chatbot: Can we use conventional methods to assess conversational user interfaces? In ECCE '19: Proceedings of the 31st European Conference on Cognitive Ergonomics. Association for Computing Machinery. https://doi.org/10.1145/3335082.3335094

Hong Kong Hospital Authority (HKHA). (2024, January 1). Waiting Time for New Case Booking at Psychiatry Specialist Out-patient Clinics. Hong Kong Hospital Authority. https://www.ha.org.hk/visitor/sopc_waiting_time.asp?id=7&lang=ENG

Hua, Y., Liu, F., Yang, K., Li, Z., Sheu, Y., Zhou, P., Moran, L., V., Ananiadou, S., & Beam, A. (2024). Large Language Models in Mental Health Care: a Scoping Review. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2401.02984

Huber, L. L., Shankar, K., Caine, K., Connelly, K., Camp, L. J., Walker, B. A., & Borrero, L. (2013). How In-Home Technologies Mediate caregiving relationships in later life.

International Journal of Human-computer Interaction, 29(7), 441–455. https://doi.org/10.1080/10447318.2012.715990

IBM. (n.d.). What are large language models? | IBM. IBM - United States. https://www.ibm.com/topics/large-language-models

Jo, E., Epstein, D. A., Jung, H., & Kim, Y.-H. (2023). Understanding the benefits and challenges of deploying conversational AI leveraging large language models for Public Health Intervention. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3544548.3581503

Korngiebel, D.M., Mooney, S.D. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. npj Digit. Med. 4, 93 (2021). https://doi.org/10.1038/s41746-021-00464-x

Lazar, A., Thompson, H. J., Lin, S., & Demiris, G. (2018). Negotiating Relation Work with Telehealth Home Care Companionship Technologies that Support Aging in Place. Proceedings of the ACM on Human-computer Interaction, 2(CSCW), 1–19. https://doi.org/10.1145/3274372

Ma, Z., Mei, Y., & Su, Z. (2023). Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.15810

Mental Health Commission of Canada. (2017, August 11). Mental health, technology, and you. https://mentalhealthcommission.ca/resource/mental-health-technology-and-you/

Mind HK. (2022, April 7). Mind HK Survey Reveals Hong Kong Citizens' Worsening State of Mental Health During the Fifth Wave of the Covid-19 Pandemic. Mindhk. https://www.mind.org.hk/press-releases/mind-hk-survey-reveals-hong-kong-citizens-worsening-state-of-mental-health-during-the-fifth-wave-of-the-covid-19-pandemic/

Nuffield Council. (2022, April). The role of technology in mental healthcare. The Nuffield Council on Bioethics. https://www.nuffieldbioethics.org/publications/technology-in-mental-healthcare

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing Mental Health with Artificial Intelligence: Current Trends and Future Prospects. Journal of Medicine, Surgery, and Public Health, 100099. https://doi.org/10.1016/j.glmedi.2024.100099

Sharkey, A., & Sharkey, N. (2010). Granny and the robots: ethical issues in robot care for the elderly. Ethics and Information Technology, 14(1), 27–40. https://doi.org/10.1007/s10676-010-9234-6

Smith, A., Hachen, S., Schleifer, R., Bhugra, D., Buadze, A., & Liebrenz, M. (2023). Old dog, new tricks? Exploring the potential functionalities of ChatGPT in supporting educational methods in social psychiatry. International Journal of Social Psychiatry, 69(8), 1882–1889. https://doi.org/10.1177/00207640231178451

Song, I., Pendse, S. R., Kumar, N., & De Choudhury, M. (2024, January 25). *The typing cure: Experiences with large language model chatbots for Mental Health Support.* arXiv.org. https://doi.org/10.48550/arXiv.2401.14362

Talkspace. (n.d.). Talkspace - #1 Rated Online Therapy, 1 Million+ Users. https://www.talkspace.com/

World Health Organization. (2011). MHealth: New horizons for health through mobile technologies. https://www.afro.who.int/publications/mhealth-new-horizons-health-through-mobile-technologie

Yang, K., Ji, S., Zhang, T., Xie, Q., & Ananiadou, S. (2023). Towards Interpretable Mental Health Analysis with Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2304.03347

Xu, X., Yao, B., Dong, Y., Yu, H., Hendler, J., Dey, A. K., & Wang, D. (2023). Mental-LLM: Leveraging large language models for mental health prediction via online text data. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2307.14385