Detailed Proposal

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 (Leader)**

15th March, 2024

Keywords: LLM, sandwich architecture, blockchain, chatbot, mental support

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.

LITERATURE DISCUSSION

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.

RESEARCH AIM

To address these issues, we propose a mental health-supporting mobile app chatbot with a sandwich architecture that utilizes large language models (LLM). This architecture respects the Privacy (Data) Protection Ordinance (PDPO) using artificial intelligence technology and employs sophisticated authentication methods to foster chatbot app security. The design features an LLM acting as a moderator between users, providing them with a human-like conversational experience with the chatbot without directly connecting one another while assuring the absence of content with triggering languages. From the perspectives of users, they are talking with a bot with a persona built from multiple users, which mimics the experience of conducting genuine conversations with an empathetic individual. To promote accessibility among individuals who speak different languages, the chatbot integrates translation into the LLM layer to overcome language barriers and implements blockchain technology to provide user anonymity and data privacy (Song, I. et al, 2024).

PROPOSED METHODOLOGY

Anti-Harassment Sandwich Architecture

Harassment is common in the mental support apps available in the current market, yet a regulation system that blocks abusive messages without knowing the context of the conversation may hinder positive and constructive communications that are therapeutic to mental health support-seeking users. On the other hand, such software can indeed generate indiscernible human-like messages between humans given a context with the current LLM technology. However, its major drawback is its reliance on relevant user inputs in the generation of messages, where the LLM has limited ability to spontaneously generate meaningful discussions. Therefore, the sandwich architecture we are proposing places an LLM between users as a moderator (Figure 1). The design filters inappropriate messages and responds spontaneously with a decent message based on learned stories and responses from other users. From the users' perspective, individuals are sharing their real-life stories with each other through the layer of a chatbot, preserving the confidentiality of their social interactions.

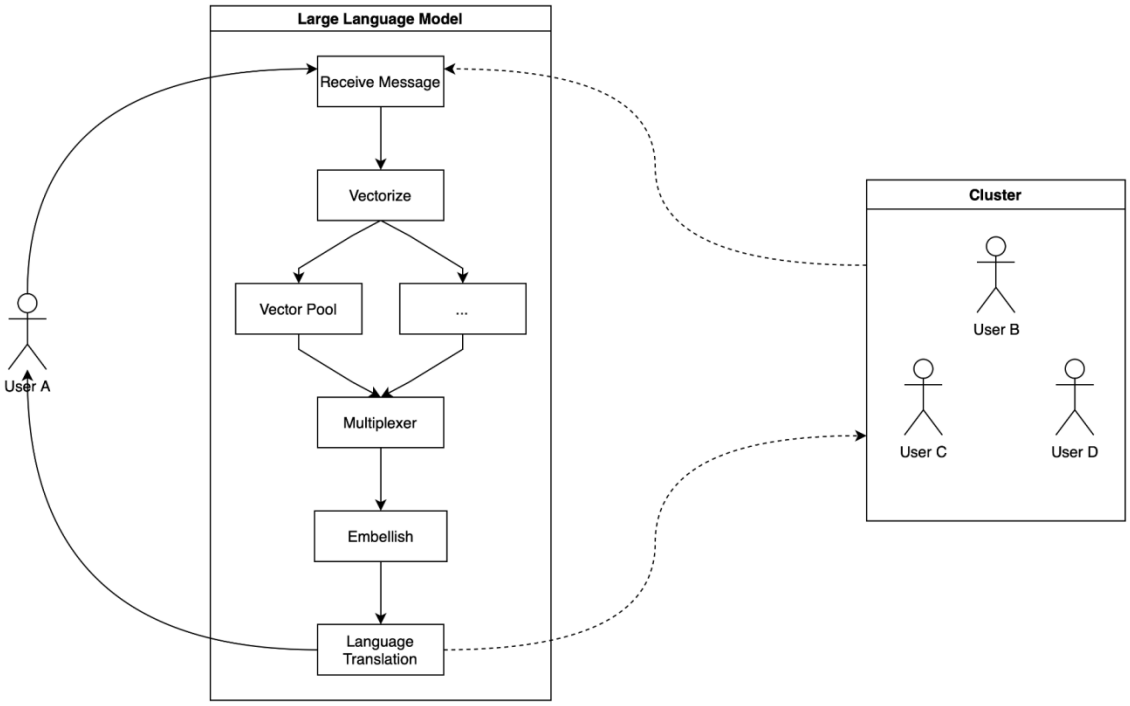

Figure 1: Workflow of the Large Language Model among Users

1) Users provide inputs to the large language model (LLM) to build the story backbone of the chatbot.
2) Upon receiving the message of User A, the LLM paraphrases the received inputs, while preserving the core elements of the stories; and proposes appropriate responses with additional story backgrounds contributed by other existing or historical users, namely User Ns within the same cluster.
3) The integrated embellished messages that are determined to be harassment-free and appropriate by the LLM are forwarded to User A, and User B, respectively.

The following outlines the technical aspect with regard to the proposed sandwich architecture used in the mental health support mobile app.

Clustering

The sandwich architecture is built upon real-time user interactions, aiming to offer conversations that go beyond the ordinary. However, it's common for users to abruptly leave a conversation without notice, creating a recurring challenge where peers must restart conversations repeatedly until they find a suitable match for engaging in meaningful discussions and receiving valuable assistance. In order to tackle this issue, users are grouped into clusters, and the chatbot maintains a vector pool. This pool serves the purpose of providing ongoing inspiration and significantly mitigating the negative impact caused by users who quickly join and exit conversations without fully participating.

When a new user joins the application, a greeting routine is initiated to assess their current mood level. This mood level serves as one of the indicators for the matching system, which identifies the most suitable cluster with available slots. Each cluster is associated with an overall topic vector, allowing the system to select the cluster with the highest semantic similarity to seamlessly integrate the new user. Additionally, the topic vector of each cluster dynamically evolves as conversations progress and bring together users with high similarity.

To prioritize user well-being, a well-being score is calculated by considering the user's past experiences and the current state of the selected cluster. This score plays a crucial role in determining the optimal cluster placement for the user, aiming for a slightly higher level of happiness to facilitate their mental improvements. By carefully considering the user's background and the dynamics of each cluster, the system strives to create an environment that promotes positive mental states and fosters personal growth.

Prompt Engineering

We propose an implicit prompt template for the chatbot that integrates vectors from users within the same cluster into a prompt instruction to be sent to the LLM. The template consists of four modules: background, cluster story base, local history base, and embellishment.

The background module, tailored to each individual user, initializes the chatbot's role background based on the user's mental needs and the topic of the cluster, establishing the fundamental tone of the chatbot. The cluster story base module is populated with the most relevant corpus captured from other users' chat history, providing a reference for human-like responses during instant chatting scenarios. The local history base module stores the historical chat messages between the user and the chatbot, enabling the chatbot to provide consistent reactions when discussing topics specific to the user. The embellishment module further processes the response by translating it into different languages and filtering inappropriate content, thus expanding the cluster corpus base to accommodate users with different language backgrounds.

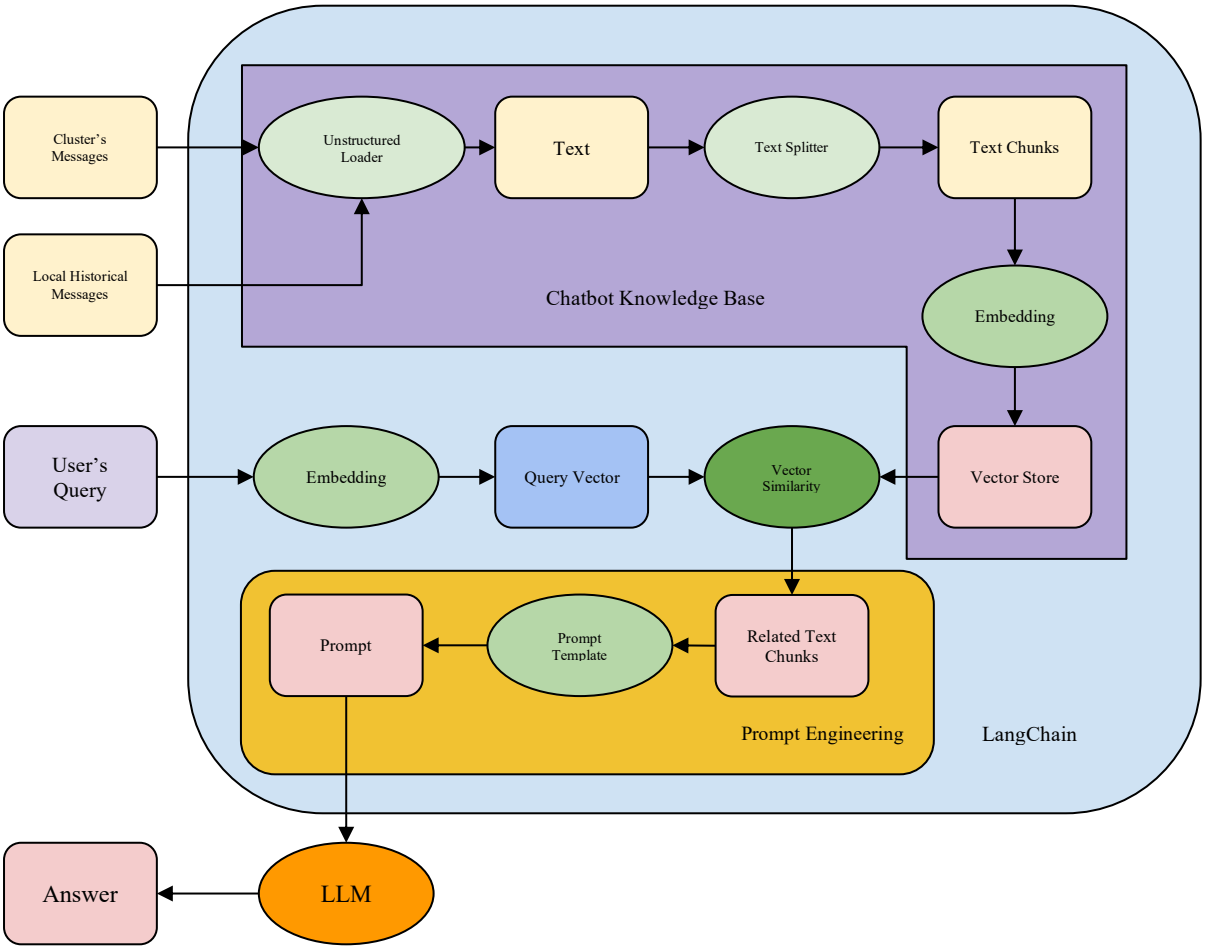

Figure 2

Multiplexer

Upon joining a cluster, a comprehensive summary of the user's individual history base is merged with the existing cluster history base, providing 2 main sources of corpora for the chatbot. Within the cluster, users contribute their own stories, expanding the collective knowledge of the cluster. To create a reply, the chatbot leverages messages or segments of messages with the highest semantic similarity from the cluster store base as a semantic enhancer, guiding the chatbot's final input to the prompt template and determining the response given to each user. The

multiplexer and the two sources of chatbot knowledge base work together to ensure a consistent response from the LLM, regardless of cluster transitions.

SCOPE

Introduction

The purpose of this project scope statement is to define the boundaries and objectives of developing the **"Mental Support Chatbot"** application. The chatbot leverages a **Large Language Model (LLM)** and employs a unique **User-LLM-User Sandwich Architecture**. The application aims to provide mental health support through human-like conversational interactions.

Project Objectives

The primary purpose of this project is to create an intelligent chatbot that assists users in managing their mental well-being. By combining user input, LLM-generated responses, and insights from other users, the chatbot will offer personalized support and encouragement.

Audience

The primary objective of this application is to offer mental support to its users. However, it's important to acknowledge that this application does not meet the standards of a professional healthcare application due to time limitations and the vast scope of the psychological research field. Consequently, the scope of this application focuses on providing an outlet for users to express their frustrations and alleviate stress in their daily lives. The target audience for this application primarily consists of mobile app users, specifically young adults and individuals who are experiencing moderate mental symptoms, such as depression, and anxiety, seeking a safe and secure environment to express their emotions.

Project Boundaries and Features

Functionality

While the long-term vision for this application is to develop a comprehensive healthcare platform encompassing features such as mood tracking, recovery plans, and collaborations with government entities to enhance support, the initial development phase has a more fundamental functional scope. The primary objective is to facilitate fluent and spontaneous conversations among users. Consequently, the user interface will be designed to have a maximum of two layers and will be available in English, Mandarin, and Cantonese for this project.

Integration

The core functionality of this application relies on a sufficient number of concurrent users to ensure optimal performance. To expand the user base, the application will be compatible with both Android and iOS platforms, leveraging the advantages offered by multi-platform frameworks. Acknowledging the limitations of computational resources, the application will utilize a commercially available LLM API rather than developing a custom LLM from scratch. This approach ensures the provision of the LLM component's services in a more efficient and resource-friendly manner.

1. User Registration and Login

- Users can create accounts, providing necessary details for personalized interactions.
- Secure login mechanisms ensure user privacy and data protection.

2. Chatbot Interaction

- Users can engage in conversations with the LLM chatbot.
- The chatbot responds with empathy, encouragement, and relevant mental health information.

3. Local History Base

- The app stores chat history locally for each user.
- Historical interactions contribute to personalized responses.

4. User Behavior Summarization

- Analyzes user behavior patterns (e.g., frequency of chat, emotional tone) to understand individual needs.
- Summarized insights feed into the Background module.

5. Background Module

- Aggregates user characteristics and behavior summaries.
- Helps the chatbot tailor responses based on context.

6. Grouping Users by Topics

- Identifies common topics users discuss with the chatbot.
- Groups users with similar interests or concerns.

7. Online Story Base

- Stores anonymous chat histories related to specific topics.
- Enables the chatbot to draw from collective experiences.

8. Chatbot Knowledge Base

- For each user, combine information from local history and online story bases.
- Generates user-specific vectors representing mental health context.

9. Vector Comparison

- Compares user query vectors with chatbot knowledge base vectors.
- Determines relevant responses based on similarity.

10. Prompt Engineering

- Constructs structured answers using predefined prompt templates.
- Ensures coherent and context-aware responses.

11. Security Measures

- Prevents prompt template injection.
- Validates user input to maintain response quality.

Exclusions

1. Integration with external systems or APIs not specified in the scope.
2. Natural language processing beyond the scope of generating answers using prompt engineering.
3. Directly chat between users.

Constraints

4. **Time Constraints**:
    - The project must be completed within the specified timeline.
    - Development, testing, and deployment phases are subject to time limitations.
5. **Resource Constraints**:
    - Availability of development resources (e.g., developers, data scientists).
    - Hardware and infrastructure limitations for deployment.
6. **Scope Stability**:
    - Changes to project scope may impact deliverables and timelines.
    - Scope changes will be managed through formal change requests.
7. **Other**:
    - Compliance with relevant legal and ethical guidelines for mental health support apps.
    - Compatibility with the selected programming languages, frameworks, and tools.

References

The following references were consulted during the preparation of this scope statement:

1. **Project Proposal**: The initial project proposal outlines the vision, goals, and expected outcomes.
2. **Industry Best Practices**: Research on chatbot development, natural language processing, and mental health support.
3. **User Feedback**: Insights from potential users regarding their expectations and needs.

SCHEDULE, MILESTONES, AND ESTIMATED LEARNING HOURS

Week	Stage	Deliverables	Learning Items	Hours (Total learning hours ~300)
7 8	Sprint 1	- **Complete detailed proposal**: agree on project objectives, the scope of the project, outline project schedule, set milestones and estimated learning hours - Analyze current cybersecurity requirements for Chatbot development locally and regulations on utilizing user data - Collect and analyze data sources on mental health related topics- POV from professionals and from patients	Project initiation, research and literature review of LLM, cybersecurity implementation on mobile apps, market research on current demand for mental health supporting technology, qualitative research on existing chatbot apps	18
9 10	Sprint 2	- Train language model using the collected data and conversation flow - Implement user input processing and sentiment analysis - Develop an initial database of supportive responses and coping strategies specific to common mental health associated symptoms/ signs - Implement initial cybersecurity measures (i.e., data encryption algorithms, secure data storage) - Design conversation flow and create a dialogue framework	TypeScript Basics Node.js Basics Express Framework Front-end development basics React Framework	20-30 15-20 10-15 20-30 30-40
MILESTONE 1 Present trained language model & initial chatbot prototype (PROTOTYPE) *The prototype should include the basic functionality of a chatbot app, such as user registration, authentication, and the ability to send and receive messages.				
11 12	Sprint 3	- Improve existing prototype based on feedback given by Yiu - Enhance the chatbot's ability to handle a wider range of topics and concerns - Test chatbot's core functionalities, especially for catching trigger words and security testing - Create a penetration testing report	Input validation Applying security best practices Authentication and authorization implementation Data encryption Deployment and hosting	5-10 10-15 15-20 10-15 10-15

13 14	Sprint 4	- Refine and optimize the chatbot's responses based on feedback from Yiu - Implement additional cybersecurity measures (i.e., access controls, authentication mechanisms) - UI for chatbot - Comprehensive cybersecurity review report - Start working on the Project's website		
MILESTONE 2 Present refined chatbot with improved responses (comparable), implemented security measures, and initial cybersecurity implementation i.e., data encryption				
15 16	Sprint 5	- Refine and optimize the chatbot's responses based on feedback from Yiu - Fix major bugs and handle security-related vulnerabilities - Implement and enhance cybersecurity measures, and further perform vulnerability assessment to verify the implemented measures - Create final report draft 1	Accessibility and Language Support Chatbot Development with LLM (on-going from the beginning) UI design enhancements	10-15 30 15-20
17 18	Sprint 6	- Implement bonus features that are good to have - Test for chatbot's scalability and security - Finalize and format documentation and user guides regarding data privacy and security practices - Perform comprehensive cybersecurity audit report - Create final report draft 2		
MILESTONE 3 Present finalized chatbot with bonus UI features, improved performance, comprehensive cybersecurity implementation, implemented language support features, integrated LLM				
19 20	Sprint 7	- Final testing and bug fixing - Finalize final report draft 3 - Finalize project website	Testing and minor bug fixing (major bugs are troubleshooted on a rolling basis) Final User Feedback Implementation	30 10-15
21	Sprint 8	- Finalize all deliverables		

		- Prepare for presentation	Final Testing and Deployment	10-15
MILESTONE 4 Present the deployed chatbot and verify cybersecurity implementation is in place				

References

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

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

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

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

Laestadius, L., Bishop, A., Gonzalez, M., Illenčík, D., & Campos-Castillo, C. (2022, December 22*). Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika*. https://doi.org/10.1177/14614448221142007

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/

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