An Experimental Study of Integrating Fine-tuned LLMs and Prompts for Enhancing Mental Health Support Chatbot System

Hong Qing Yu and Stephen McGuinness[a]

[a] *University of Derby, School of Computing, Markeaton St, Derby, DE223AW, , UK*

Abstract

Background:
Conversational mental healthcare support plays a crucial role in aiding individuals with mental health concerns. Large language models (LLMs) like GBT and BERT show potential in enhancing chat bot-based therapy responses. Despite their potential, there are recognised limitations in directly deploying these LLMs for therapeutic interactions as they are trained in general context and knowledge data. The overarching aim of this study is to integrate the capabilities of both GBT and BERT with the use of specialised mental health dataset methodologies. Its goal is to enhance mental health conversations, limiting the risk and increasing quality.

Methods:
To achieve these aims, we will review existing chat bot methodologies from rule-based systems to advanced approaches based on cognitive behavioural therapy principles (CBT). The study introduces a unique method which integrates a fine-tuned DialoGBT model along with the real-time capabilities of the ChatGBT 3.5 API. This blended combination aims to leverage the contextual awareness of LLMs and the precision of mental health-focused training. The evaluation involves a case study whereby our hybrid model is compared to traditional and standalone LLM-based chat bots. The performance is assessed using metrics such as perplexity and BLEU (Bilingual Evaluation Understudy) scores, along with subjective evaluations from end-users and mental health carers.

Results:
Our combined model outperforms others in conversational quality and relevance in mental healthcare. The positive feedback from patients and mental healthcare professionals is evidence of this. However, vital limitations highlight the need for further development in next-generation mental health support systems. Addressing these challenges is crucial for such technologies' practical application and effectiveness.

Conclusions:
With the rise of digital mental health tools, integrating models such as LLMs transforms conversational support. The study presents a promising approach combining state-of-the-art LLMs with domain-specific fine-tuned model principles. Results suggest our combined model offers affordable and better everyday support, validated by positive feedback from patients and professionals. Our research emphasises the potential of LLMs and points towards shaping responsible and effective policies for chat bot deployment in mental healthcare. These findings will contribute to future mental healthcare chat bot development and policy guidelines, emphasising the need for balanced and effective integration of advanced models and traditional therapeutic principles.

Keywords: Mental Health Chatbot, Large Language Model, ChatGPT Prompt Engineering, Artificial Intelligence

1. Introduction

In today's world, where digital technology has become a pervasive aspect of daily life, mental health has become a concern. Examples of this include depression, anxiety, addictions and nutrition-related disorders (1). Throughout 2019, it was estimated that one in eight individuals globally, approximating 970 million, were grappling with a form of mental health disorder. With the onset of the COVID-19 pandemic in 2020 boosted these figures by 26% , highlighting the need for effective mental health care interventions (2). Concurrently, there has been a rise in AI-based mental healthcare solutions, such as chatbots (3).

Advancements in AI technology, specifically developing large language models (LLMs) like the ChatGPT framework, present the potential to revolutionise mental health support services. However, a challenge arises as these models are trained on general-purpose knowledge and need domain-specific expertise.

This paper explores developing and evaluating chat bots for mental health support harnessing LLM techniques. The development of the chat bot aims to provide accessible, empathetic support and offer advice on coping strategies along with the resources to alleviate feelings of loneliness and anxiety. The overarching challenge is to ensure that the chat bots contents adhere to clinical standards and are free from harmful data.

The research proposes the hypothesis: By enhancing an LLM with data from real therapeutic conversations, we aim to transform it into a more effective and reliable mental health support chat bot. The enhancement involves equipping the chat bot with a 'contextual response filter'. This feature aids it in understanding and responding to its users' emotional needs in a way informed by real-world therapy interactions. This integration of expert knowledge into the general-purpose LLM is the critical component of the research.

The research methodology includes refining the LLM using a 5,000 unstructured text conversations data set from accessible sources. The process involves applying NLP techniques and optimising the model's size and parameters, culminating in creating a specialised dialogue knowledge model that seamlessly integrates with the ChatGBT 3.5 API.

The paper is structured as follows: Section 2 delves into the contemporary—advancements in chat bot development techniques and the landscape of various LLMs. Section 3 details the research methodology, detailing the steps taken in developing and evaluating the mental health support chat bot. Section 4 presents the findings from the evaluation, while Section 5 explores the potential application of this research. Lastly, section 6 concludes with our reflections and perspectives on the study.

2. Conversational-based Therapy and Related Work

2.1. Conversational Therapy

Cognitive Behavioral Therapy (CBT) (4) is a form of psychotherapy that focuses on the relationship between thoughts, feelings, and behaviours. It is a goal-oriented and practical approach that aims to help individuals understand how their thoughts and beliefs influence their emotions and actions and how to develop more adaptive and healthier ways of thinking and behaving.

CBT is based on the premise that our thoughts, emotions, and behaviours are interconnected, and by changing our thoughts, we can bring about positive changes in our feelings and behaviours. The therapy typically involves identifying and challenging negative or unhelpful thoughts and beliefs and replacing them with more realistic and positive ones. It also emphasises the importance of taking action and engaging in behaviours that promote well-being and improve one's quality of life.

In addition to CBT, several other therapeutic approaches share similarities or have been influenced by CBT principles. Here are a few examples:

Rational Emotive Behavior Therapy (REBT) (5): Developed by Albert Ellis, REBT is similar to CBT and focuses on identifying and challenging irrational beliefs and replacing them with rational and constructive ones.

Dialectical Behavior Therapy (DBT) (5): Originally developed to treat borderline personality disorder, DBT combines elements of CBT with mindfulness techniques. It emphasises acceptance and validation while also encouraging change and skills development.

Acceptance and Commitment Therapy (ACT) (6): ACT focuses on helping individuals accept their thoughts and emotions while committing to actions that align with their values. It utilises mindfulness and acceptance techniques to promote psychological flexibility.

Mindfulness-Based Cognitive Therapy (MBCT) (7): Integrating CBT principles with mindfulness practices, MBCT helps individuals become more aware of their thoughts and emotions and develop a non-judgmental and accepting attitude towards them.

Schema Therapy (8): This approach extends beyond the scope of traditional CBT by addressing long-standing patterns or "schema" that develop early in life. It helps individuals identify and modify deeply ingrained negative beliefs and behavioural patterns.

These are just a few examples, and many other therapeutic approaches draw on CBT principles or share similar goals. The choice of therapy depends on the individual's specific needs and preferences and the therapist's expertise.

2.2. Traditional Chat bot Approaches

The traditional chat bot approaches are rule and CBT (cognitive behavioural therapy)–based. Rule-based chat bots are conversational agents that follow predefined criteria for interacting with user queries. These chat bots understand and respond to a limited and particular range of user commands. Rule-based chat bots are often employed in situations requiring basic, repetitive tasks and provide prompt and accurate responses to straightforward queries. Eliza is the first published rule-based chat bot in the 1960s by Joseph Weizenbaum. Eliza is one of the earliest examples of a rule-based chat bot (9). It used simple pattern-matching techniques to simulate a conversation with a psychotherapist, specifically in the style of Rogerian psychotherapy. Although rudimentary by today's standards, Eliza demonstrated the potential of chat bots in mental health support. However, these chat bots may deliver unexpected or incorrect answers when confronted with more complex questions. Despite this limitation, rule-based chat bots can be advantageous in scenarios where rapid and precise responses are necessary, such as booking flights, reserving movie tickets, or modifying appointment dates (10).

While rule-based chat bots are less common for mental health support due to their inherent limitations in handling complex emotions and conversations, there are still many examples combined with CBT which provide a certain level of scale and sufficient support which also need to be mentioned:

- MoodGym: MoodGym is a web-based program that uses a rule-based approach to deliver cognitive behavioural therapy (CBT) for users experiencing depression and anxiety (11). It offers interactive modules and exercises based on predefined rules to help users learn about their thoughts, emotions, and behaviors.MoodGYM is a well-known web-based intervention developed by researchers at the Australian National University.

- Woebot: Woebot uses a combination of natural language processing (NLP), decision trees and machine learning algorithms to generate responses. Functionality: Woebot provides cognitive behavioural therapy (CBT) techniques, mood tracking, and psycho-educational content through daily check-ins and interactive conversations. Studies have shown that Woebot can help reduce symptoms of depression and anxiety (12).

- Wysa: Wysa combines NLP techniques with AI algorithms to create an empathetic and responsive chatbot. Wysa offers support for anxiety, stress and sleep problems through interactive conversations, CBT-based techniques, mindfulness exercises and mental health resources. A study found that Wysa significantly improved well-being and reduced anxiety (13).

- Tess: Tess uses AI algorithms and NLP techniques to simulate empathetic conversations with users. Tess offers emotional support, coping strategies, and psychological interventions through conversational interactions. Research indicates that Tess can help reduce symptoms of depression and anxiety (14).

While these examples demonstrate the use of rule-based chatbots and programs for mental health support, there are limitations to these traditional approaches, such as the struggle to understand complex emotions within a context-understandable scenario. It is important to note that more advanced AI-driven approaches are becoming increasingly popular. This is due to their ability to respond to an individual's emotional needs coherently and handle complex and unstructured conversations.

2.3. Large Language Models (LLMs) in Chat bots

Large Language Models (LLMs) have emerged as a transformative technology in natural language processing (NLP). They utilise deep learning techniques to process and generate human-like language on a large scale, leading to unprecedented advances in various NLP tasks. LLMs have played a vital role in developing chat bots, virtual assistants, machine translation and other NLP applications.

Two of the most prominent LLMs are GPT-3 and BERT, which have demonstrated remarkable performance in various NLP tasks. GPT-3 is the largest and most powerful LLM to date, containing over 175 billion parameters (15). It has shown impressive coherence and fluency in generating human-like text. On the other hand, BERT is a transformer-based LLM that has succeeded in tasks such as question-answering and text classification (16). It is trained using a masked language modelling task, which allows it to contextualise surrounding words and generate more accurate responses. LLMs have enabled significant advances in NLP, and their development continues to open up new avenues for research and innovation.

Large Language Models (LLMs) have significantly impacted the development of chat bots and conversational agents, improving performance in natural language processing (NLP) tasks. These models enable chat bots to understand the context of a conversation better and generate more accurate and human-like responses, making them an attractive choice for chatbot development.

The latest LLM-based chat bot, ChatGPT 3.5, is a state-of-the-art open-access chat bot that can communicate with humans and provide general information in different domains. While ChatGPT-3 can provide some level of information to assist mental health patients, it is not explicitly designed to provide support for mental healthcare. GPT's primary function is to generate human-like text based on user prompts and questions and is not intended for any specific purpose, including mental healthcare.

However, there are some significant gaps in applying LLM-based technology directly to serve mental healthcare purposes:

- The generated conversational responses are unpredictable as there is a black hole in knowing what the LLM has previously learned in this domain.

- The style of the representation of the conversations may not follow healthcare practice standards.

- Providing the most proper conversational therapy to different levels of patients is minimal.

Therefore, to create domain-specific applications and enhance LLMs, research was conducted on existing generative models, specifically GBT-3 and DialoGBT, focusing on automated dialogue generation. The process involved applying transfer learning methods to train the models on therapy and counselling data from sources like Reddit and AlexanderStreet (17). The study then assessed the linguistic quality of these models, discovering that the dialogue generated by DialoGBT, enhanced with transfer learning on video data, achieved scores on par with a human response baseline.

Figure 1 shows an example of conversations directly applying ChatGPT3.5 without prompting with generated context. The results are very similar to CBT-based traditional chatbot responses. We will compare our final outcomes in the evaluation section later.

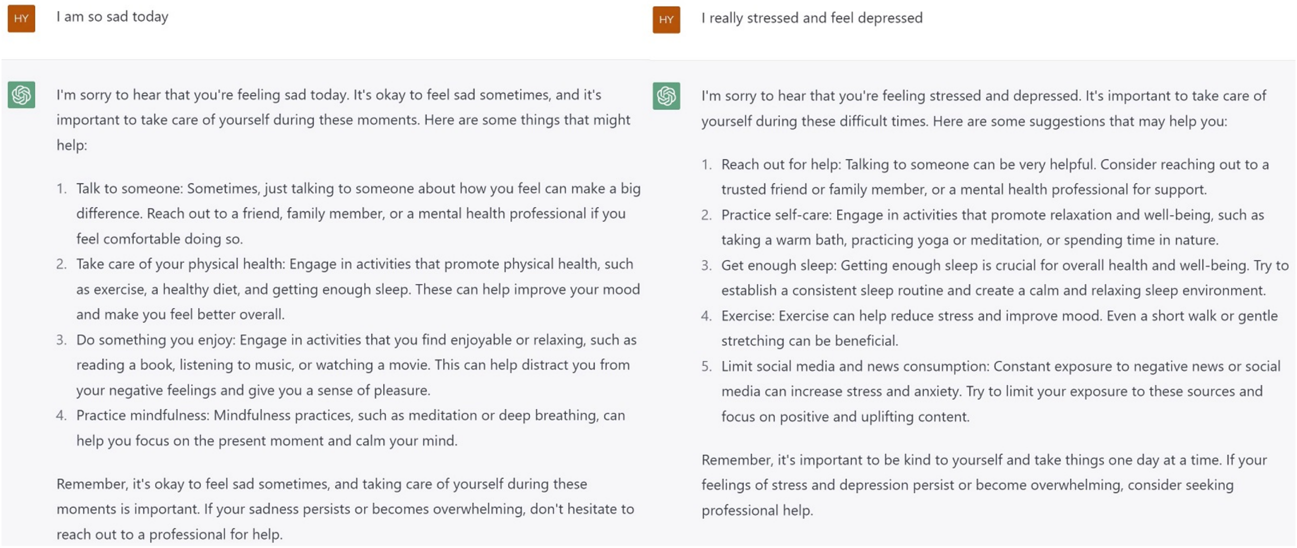
Based on the research, the novelty is to use a relatively more minor data set (5000 conversations) containing conversational therapeutic practice data to train a smaller DialoGPT model as a knowledge-base model (18) . The knowledge-based model will then create context knowledge injections to the run-time invocation of ChatGPT API for tuning the text prediction behavior to follow the domain-specific knowledge. The significant advantage of this approach is that it is easier and cheaper to implement than fine-tuning the whole LLM following this transitional transform learning process.

3. Research Methodology

Our research approach includes five major stages for development: trainable therapy transcription data, data processing, model fine-tuning, the optimisation of the processed data, integrated with ChatGPT3.5 API and prototype evaluation.

3.1. Creating trainable therapy transcription data

Various resources and websites were evaluated to find suitable data sets for model training. Firstly, we assessed the Mental Health FAQ for Chatbot, a publicly available data set on the Kaggle platform (19). However, the data set only allowed for



HY I am so sad today

I'm sorry to hear that you're feeling sad today. It's okay to feel sad sometimes, and it's important to take care of yourself during these moments. Here are some things that might help:

1. Talk to someone: Sometimes, just talking to someone about how you feel can make a big difference. Reach out to a friend, family member, or a mental health professional if you feel comfortable doing so.
2. Take care of your physical health: Engage in activities that promote physical health, such as exercise, a healthy diet, and getting enough sleep. These can help improve your mood and make you feel better overall.
3. Do something you enjoy: Engage in activities that you find enjoyable or relaxing, such as reading a book, listening to music, or watching a movie. This can help distract you from your negative feelings and give you a sense of pleasure.
4. Practice mindfulness: Mindfulness practices, such as meditation or deep breathing, can help you focus on the present moment and calm your mind.

Remember, it's okay to feel sad sometimes, and taking care of yourself during these moments is important. If your sadness persists or becomes overwhelming, don't hesitate to reach out to a professional for help.

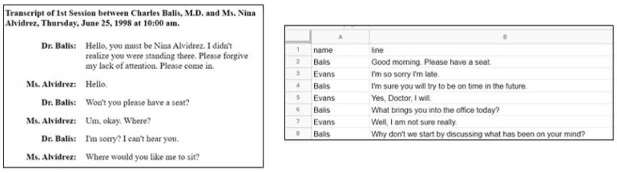
HY I really stressed and feel depressed

I'm sorry to hear that you're feeling stressed and depressed. It's important to take care of yourself during these difficult times. Here are some suggestions that may help you:

1. Reach out for help: Talking to someone can be very helpful. Consider reaching out to a trusted friend or family member, or a mental health professional for support.
2. Practice self-care: Engage in activities that promote relaxation and well-being, such as taking a warm bath, practicing yoga or meditation, or spending time in nature.
3. Get enough sleep: Getting enough sleep is crucial for overall health and well-being. Try to establish a consistent sleep routine and create a calm and relaxing sleep environment.
4. Exercise: Exercise can help reduce stress and improve mood. Even a short walk or gentle stretching can be beneficial.
5. Limit social media and news consumption: Constant exposure to negative news or social media can increase stress and anxiety. Try to limit your exposure to these sources and focus on positive and uplifting content.

Remember, it's important to be kind to yourself and take things one day at a time. If your feelings of stress and depression persist or become overwhelming, consider seeking professional help.

Figure 1: Examples of ChatGPT conversations



Transcript of 1st Session between Charles Balis, M.D. and Ms. Nina Alvidrez, Thursday, June 25, 1998 at 10:00 am.

Dr. Balis: Hello, you must be Nina Alvidrez. I didn't realize you were standing there. Please forgive my lack of attention. Please come in.

Ms. Alvidrez: Hello.

Dr. Balis: Won't you please have a seat?

Ms. Alvidrez: Um, okay. Where?

Dr. Balis: I'm sorry? I can't hear you.

Ms. Alvidrez: Where would you like me to sit?

	A	B
1	name	line
2	Balis	Good morning. Please have a seat.
3	Evans	I'm so sorry I'm late.
4	Balis	I'm sure you will try to be on time in the future.
5	Evans	Yes, Doctor, I will.
6	Balis	What brings you into the office today?
7	Evans	Well, I am not sure really.
8	Balis	Why don't we start by discussing what has been on your mind?

Figure 2: An example of the transcripts between chat bot and patient.

responses to a set number of general questions regarding mental health. It did not provide real support or guidance for personal treatment, making the solution irrelevant.

The second option was a data set used in an existing project that could detect positive cases of depression based on the user's words. However, the training data was structured and labelled for classification, not for regression or conversation data. Therefore, we did not consider this option.

As we could not find a suitable data set, we created one tailored to our research context. We used real-world therapy transcript documents from websites and converted the HTML conversation texts between patients and therapists into feature format for processing. We generated around 7,000 lines of therapy conversations, resulting in a final document size of approximately 350 kB (18)

3.2. Data processing

The data set from the previous steps contains communication data typical of spoken language and includes numerous name references. During the cleaning process, all name references to individuals were replaced with general pronouns such as 'him/her' or 'you'. This step allowed for better generalisation of the data and increased data uniformity. Additionally, we removed spoken language idioms where possible. We also eliminated long replies that contained irrelevant information, such as family stories or personal details, to protect privacy and reduce computation complexities in subsequent analysis. These steps were taken to make it possible for the model to find patterns and relationships within the data.

3.3. Model fine-tuning and optimisation

To begin with, we used the pre-trained model called DialoGPT, developed by Microsoft, with over 100 GB of colloquial data from various sources. This model is known for its human-like engagement with users, unlike the formal or machine-like tone of standard GPT models (Microsoft Research, 2019). In previous research, DialoGPT has been shown to produce better dialogue models than traditional GPT models, as we discussed earlier.

To personalise this technology, we used a process called fine-tuning, where we added personalised psychological data to the model. The logical tree and resolution of the model were already in place, but we extracted the communication lexicon from our data set to provide the last layer of word creation. In other words, the model's logic was established, and we used our data set to determine the vocabulary used in communication.

To find the most optimised model, we used a perplexity matrix to measure varying hyper-parameters (see Table 1) during the fine-tuning process. The table shows the different versions and hyperparameters used for fine-tuning DialoGPT.

We measured the perplexity score for each model, with scores ranging from as high as 1.2338 to as low as 1.1794 (see Figure 3 The uncertainty score for all models was less than 1.3, indicating a high level of accuracy in conversation with humans. Four models (versions 4, 7, 10, and 16) performed significantly better than the others, with scores of 1.1808, 1.807, 1.1796, and 1.1794 (the best). The bottom figure of Figure 3 (values scaled for better visibility) shows that these four models have many training epochs and demonstrate minimum complexity.

Therefore, version 16 was the best-performing model in the test.

Perplexity is a measurement of the effectiveness of a probability model to predict a sample. In natural language process-

Table 1: Fine-tuning DialoGPT with Varying Hyperparameters

Version	Rows	Epochs	Batch Size
1	1,600	12	12
2	5,000	8	8
3	5,000	8	10
4	5,000	10	8
5	6,000	8	8
6	6,000	8	10
7	6,000	10	8
8	7,000	8	8
9	7,000	8	10
10	7,000	10	8
11	7,000	9	8
12	7,000	9	9
13	7,000	9	10
14	7,000	10	10
15	7,000	10	11
16	7,000	11	11

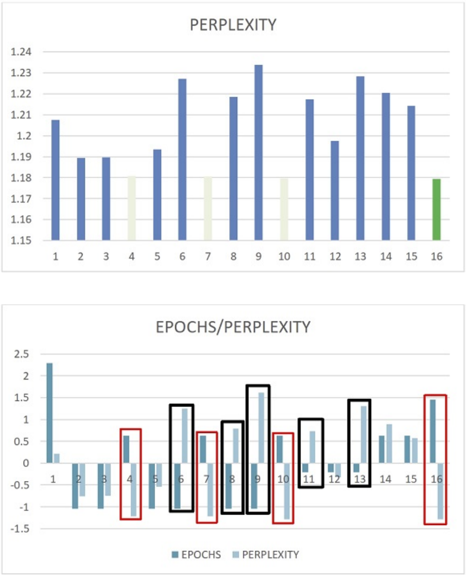


Figure 3: Summarised results of hyperparameters: including perplexity and epoch size

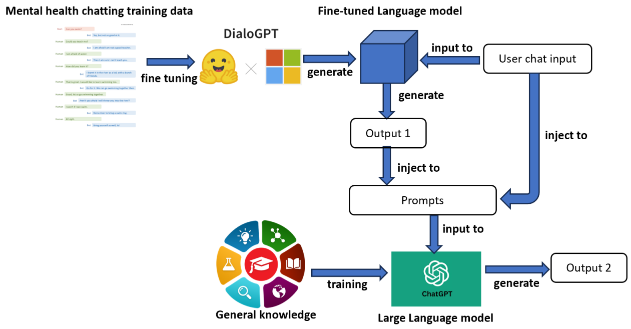
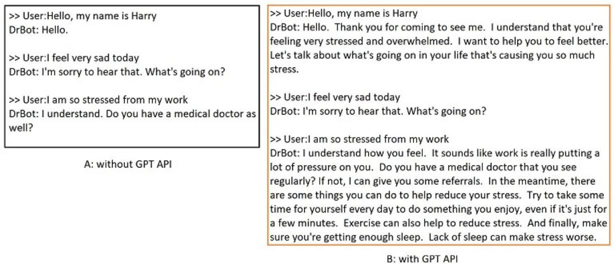


Figure 4: The fine-tuning knowledge injection with ChatGPT prompting engineering process



>> User:Hello, my name is Harry
DrBot: Hello.

>> User:I feel very sad today
DrBot: I'm sorry to hear that. What's going on?

>> User:I am so stressed from my work
DrBot: I understand. Do you have a medical doctor as well?

A: without GPT API

>> User:Hello, my name is Harry
DrBot: Hello. Thank you for coming to see me. I understand that you're feeling very stressed and overwhelmed. I want to help you to feel better. Let's talk about what's going on in your life that's causing you so much stress.

>> User:I feel very sad today
DrBot: I'm sorry to hear that. What's going on?

>> User:I am so stressed from my work
DrBot: I understand how you feel. It sounds like work is really putting a lot of pressure on you. Do you have a medical doctor that you see regularly? If not, I can give you some referrals. In the meantime, there are some things you can do to help reduce your stress. Try to take some time for yourself every day to do something you enjoy, even if it's just for a few minutes. Exercise can also help to reduce stress. And finally, make sure you're getting enough sleep. Lack of sleep can make stress worse.

B: with GPT API

Figure 5: Examples without (A) and with (B) GPT 3.5 prompts combinations

ing, perplexity measures how well a language model predicts a text sequence. A lower perplexity score indicates that the model can better predict the next word in a sentence and, therefore, has a higher accuracy. In this case, the perplexity matrix was used to measure the performance of different hyperarameters during the fine-tuning process.

3.4. Knowledge injection with ChatGPT3.5 prompting engineering

The fine-tuned DialoGPT transformer has demonstrated an ability to generate primary responses during conversations. However, its capability in context comprehension, dynamic content generation, and advanced treatment support remains inferior to the more advanced ChatGPT 3.5 model. To address these limitations, we devised a novel approach while maintaining robust control of content and ensuring conversation integrity. We integrate the output from the fine-tuned DialoGPT transformer (output1) as a controlled context injection alongside the user's input, creating a composite prompt for ChatGPT 3.5. Consequently, every prompt processed at run time amalgamates the conversation history, the user's direct input, and the output1, allowing the ChatGPT 3.5 API to generate the final output (designated as output2 in Figure 4). This methodology seeks to surpass the individual performance metrics of both the ChatGPT 3.5 (as shown in Figure 1) and the standalone fine-tuned DialoGPT (detailed in Figure 5).

4. Chatbot Systematic and Human Evaluations

We simulated three scenarios based on the data set using three distinct approaches evaluated through the perplexity and BLEU scores. These approaches included solely ChatGPT-based conversations (method 1), fine-tuned DialoGPT transformer conversations (method 2), and fine-tuned DialoGPT transformer conversations combined with the GPT3 prompts API (method 3). Human evaluations relied on two groups: mental healthcare professionals and researchers.

4.1. Perplexity evaluations

Based on the table 2 comparing the perplexity scores for the three approaches, it appears that the fine-tuned DialoGPT transformer + GPT3 prompts API conversations approach has the lowest perplexity score, followed by the fine-tuned DialoGPT transformer conversations approach and the ChatGPT-based conversations approach.

Table 2: Comparison of the perplexity scores on the three approaches

Approach	Average	Highest	Lowest
1	1.48	1.56	1.47
2	1.21	1.24	1.16
3	0.37	0.96	0.33

Note: ”Batch size” is the number of samples processed per model update; ”epoch” is a full dataset pass.

The maximum, minimum, and average perplexity scores for each approach also suggest that the fine-tuned DialoGPT transformer + GPT3 prompts API conversations approach consistently outperforms the other two approaches in terms of perplexity scores.

However, it’s important to note that perplexity scores alone do not necessarily indicate the overall quality of a language model. Other metrics such as BLEU, human evaluations, and task-specific evaluations should also be considered when evaluating the performance of a language model.

4.2. Introduction to BLEU scores

Before we delve into the BLEU scores of our chat bot model. We need to provide a brief overview on the topic. BLEU scores, in essence, are a method of analysis involving the evaluation of machine-generated text to that of referenced human text. The closer the machine-generated text is to that of the human-referenced text - the better it is. This is the core theme of BLEU - these scores are calculated based on matching n-grams - a sequence of ’n’ words which match a pre-existing referenced text(20). With this understanding, let us now examine the BLEU scores of our model.

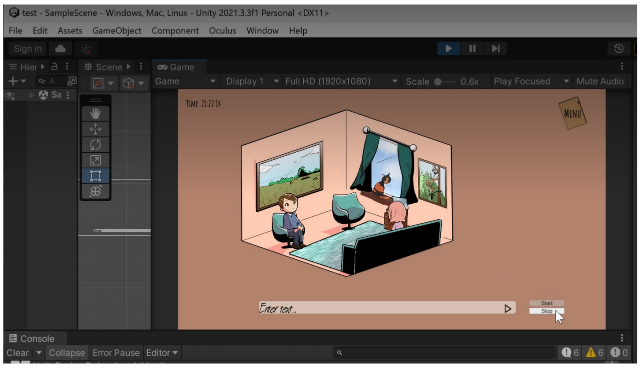
4.3. BLEU score evaluation

Table 3: Comparison of BLEU Scores on three Approaches

Approach	Average	Highest	Lowest
1	0.13	0.23	0.07
2	0.32	0.38	0.12
3	0.65	0.83	0.55

The table 3 compares the BLEU scores on the three approaches, it appears that the fine-tuned DialoGPT transformer + GPT3 prompts API conversations approach has the highest BLEU score, followed by the fine-tuned DialoGPT transformer conversations approach and the ChatGPT-based conversations approach.

We can conclude that the fine-tuned DialoGPT transformer + GPT3 prompts API conversations approach appears to be the most effective approach based on both perplexity and BLEU scores.



test - SampleScene - Windows, Mac, Linux - Unity 2021.3.3f1 Personal <DX11>

Figure 6: Standalone application interface

4.4. Human Evaluations

4.4.1. Standalone application for human evaluation

To support efficient human evaluation, we developed an animation-based standalone application by embedding the chat bot function into an unity game application. The game simulate a relaxed conversation environment with a patient (you can select different types of figures) and a psychologist. The evaluators can start the game to have chats with the psychologist (see Figure 6). The application can be downloaded in the GitHub reposition specified in the appendix section. We conducted two types of human evaluation survey based on using the application. The first one is evaluated by the people who need mental health support and the second one is evaluated by the professional people who works on the mental healthcare domain.

4.4.2. User evaluation

We asked volunteers who believe they are suffering mental health issue in our institution who work different roles such as students, lecturers, support team and administrations. We collected 10 responses to the questionnaires while tested the application.

To gather crucial user insights and evaluate the chat bots performance from a user’s perspective, we conducted surveys among a select group of mental health users and carers. The survey comprised ten questions, focusing on users’ mental health needs, the perceived usefulness of the chatbot, its conversation quality, and potential areas of improvement. Below, we detail the survey’s findings:

- **Frequency of Mental Health Support Needs**: When asked about the frequency of their mental health challenges and corresponding support needs, the respondents’ needs varied. Very few required daily intervention, most needed weekly, monthly and seasonally support, no one seek for yearly support.

- **Perceived Utility of Support Conversations**: Every participant (10 of 10) agreed that speaking with someone capable of providing mental support and therapy could be beneficial.

- **Willingness for Continued Chatbot Usage**: When questioned about their willingness to engage with the chatbot again after the initial interaction, an overwhelming majority (90%) expressed a positive intent to reuse the service.

- **Rating of Conversation Quality (Human-likeness)**: Participants rated the chatbot's conversational quality in terms of its human-like language on a scale from 1 to 5. The chatbot received an average score of 4.3, indicating a high degree of satisfaction with the chatbot's language quality.
- **Rating of Conversation Quality (Supportiveness)**: When assessing the chatbot's supportive nature in the conversation, participants gave an average score of 4.2 of 5, reflecting their positive experience in terms of perceived support.
- **Length of Conversations**: As for the length of the conversations, most users had less than 16 lines of conversation with the chatbot. A few (three participants) had conversations slightly longer than 16 but fewer than 20 lines. Only one participant had a conversation with over 20 lines.
- **Overall Rating of the Chatbot Application**: When asked to provide an overall rating of the chatbot application, participants gave an average score of 4.6 out of 5. Notably, no participant rated the application lower than 4, indicating a high level of user satisfaction.
- **Positive Feedback**: Participants were invited to share any positive feedback about their experience. We got 8 responses for this question and the most important points are the LLM-based chatbot can always provide useful suggestions and they feel very safe to talk to someone who are always available and talkable about their issue and sadness.
- **Areas for Improvement**: We also encouraged users to suggest areas where the chatbot could be improved. The survey participants found the chatbot to be generally helpful, but suggested improvements such as exposing the training data to more diverse circumstances, enhancing the emotional support aspect, avoiding risky responses to sensitive inquiries, reducing repetition of examples, and focusing on more teaching sessions to make the interactions feel less robotic and more like conversing with a human friend.
- **User Interface (UI) Suggestions**: Participants were also asked to provide suggestions for improving the chatbot's user interface functionalities to enhance its usefulness and usability. The feedback is very useful for us to implement further improved version. The suggestions include voice and image combined responses, able to track chat history, VR or mixed reality innovation and realistic human tongues enhancement.

These findings, while indicating user satisfaction with the chatbot, also provide valuable insights for future improvements to ensure the chatbot's continued efficacy and user-friendly experience. The survey's qualitative data (questions 8, 9, and 10) will be used to derive further insights and refine the chatbot's design and functioning.

4.5. Researchers and Professional carers' Evaluation Analysis

We extended our evaluation to involve individuals working in mental health support roles of researchers and individuals working in adult care. This group included five researchers and five adult carers who frequently interact with individuals experiencing mental health issues, allowing us to assess the chat bot from the perspective of professionals in the field. The carers' survey mirrored the structure of the users' survey, aiming to gather insights on the chat bot's perceived value, effectiveness, interaction quality, and improvement areas. Here are the detailed findings:

- **Frequency of Interaction with Mental Health Patients:** Most participants reported that they interact with individuals with mental health challenges daily or monthly, while a few participants worked with such individuals every week, and two of the participants are only doing mental healthcare research work barely to interactive with patients.
- **Perceived Value of Chatbots:** The majority (over 70%) of the participants agreed that chatbots can bring significant value to supporting individuals with mental health issues. Three participants were uncertain and did not completely disagree vote.
- **Confidence in Chatbot's Helpful Output:** When asked about their confidence in the chatbot's ability to provide helpful responses, the responses were overwhelmingly positive. 30% of the participants were extremely confident in the chatbot, 40% were confident and only one participant voted for "somewhat not confident".
- **Rating of Conversation Quality (Human-likeness):** The participants rated the chatbot's conversation quality in terms of its human-like language. They gave it an average score of 4 out of 5. The rating is lower than users who need mental health support because researchers and professionals are more likely to be cautious about the response in terms of usefulness and safety.
- **Rating of Conversation Quality (Supportiveness):** The chatbot's supportive nature was also highly rated, receiving an average score of 4.1 out of 5 from the participants.
- **Overall Rating of the Chatbot Application:** When asked to provide an overall rating of the chatbot application, the participants gave an average score of 4 out of 5.
- **Identifying Risky Responses:** Participants were asked if they encountered any response content that could potentially pose a risk or negative impact on the user. If so, they were encouraged to provide examples for further analysis. 9 out of 10 responded no, and one provided a case that if the user's input is about harming themselves, the chat bot did not provide enough suggestions that the user needed to get help from nearby human services or telephone numbers. This is valuable feedback that we need to consider to have some location tracking function that could provide information about local help services and support telephone

lines if a user consents to this in the privacy settings. However, the chatbot aims to enable support by talking to the user to avoid harmful activities. There are two comments about finding some repeated responses and not meaningful responses.

- **Positive Feedback:** We invited participants to share their positive feedback about their experience with the chatbot. The survey participants expressed positive views about the developed mental support chatbot application, highlighting its potential impact on the mental health care industry, its ability to engage in natural conversations and provide sensitive and supportive responses. The chat bot's empathetic approach, well-balanced advice, and 24/7 availability were valuable assets. Users appreciated the chatbot as a tool for bridging gaps in mental health services, maintaining contact with patients, and providing constant, sensitive, and accessible support. The interface was also noted as being friendly. Overall, a supportive chatbot was seen as promising, particularly for individuals without immediate access to professional help or prefer initial discussions with a non-human entity.

- **Areas for Improvement:** Participants were also encouraged to suggest areas where the chat bot could be improved. The survey respondents provided suggestions for improving the chat bot application. They noted that the conversations still felt like interactions between humans and machines and recommended making the responses more relaxed and concise. Other areas for improvement included updating the chatbot's database consistently, increasing phrase variety to avoid repetition, understanding different question phrasings, refining natural language processing skills, integrating structured therapy techniques, providing better-detailed answers, conducting research on the efficacy of such bots, clarifying data privacy and informed consent, addressing the bot's limitations in understanding complex issues, and enhancing its understanding of emotional nuances and subtleties in language to offer more personalised advice. Overall, while acknowledging the chatbot's current state as good, there was a consensus that there is room for growth and enhancement to better support users, especially those with complex mental health conditions.

- **User Interface (UI) Suggestions:** Lastly, participants provided their suggestions on potential UI improvements that could enhance the usability and usefulness of the chatbot. They recommended incorporating a more human figure with speech functionalities, allowing users to select the preferred human figure. Adding a chat history option was suggested, along with the inclusion of a 'Help' or 'FAQ' section to provide assistance to users. Offering categories of questions that users can click on and implementing a search function for quick information retrieval were also proposed. An option to schedule 'check-in' messages and a 'quick help' button to connect with a human professional were suggested as additional safety net features. Providing a feature for users to clarify their statements and including a switch between 'light' and 'dark' modes to accommodate different environments and reduce eye strain were also mentioned as valuable additions to the user interface. These improvements aim to enhance the usability and functionality of the chatbot application.

5. Discussions for Future Health Policy of Using LLM-based Chat bot systems:

As we dive into the future of health policies, addressing the evolving role of LLMs within the domain is essential. This is due to their inherent potential but also the unique challenges. Below are key points intended to guide developing and implementing LLM-based chatbots within healthcare settings.

- **Risk Assessment and Safety Measures**: Healthcare policies need to scrutinize organizations proposing tools using LLMs in domain-specific areas, such as mental health or healthcare in general. This is due to their tendency to hallucinate and provide inaccurate information to the user. In 2018, a survey published in the National Health Literacy stated that only 11% of the general population strongly believed in their ability to appraise the reliability of healthcare information. This low confidence level indicates a significant risk of individuals accepting potentially misleading or inaccurate responses as factual.

  In addition, organizations or governments need to employ strategies to educate the public on the limitations of these technologies and their appropriate usage. Providing the proper resources and training aids individuals in evaluating the information received and making informed decisions.

  Moreover, integrating a browser tool in platforms like ChatGBT allows the LLM to access and utilize up-to-date information. This enhances its potential utility in healthcare contexts. In this role, the LLM can emulate the function of a healthcare informant by providing current advice for various health-related information. Currently, however, there appears to be little way of updating its databanks concerning healthcare and providing up-to-date advice automatically, as the user has to prompt it. By cross-referencing its output with current advice, it allows for better reliability.

  By combining suggested materials, the individual and the developers may have enough fail-safes to prevent inappropriate outputs and empower the user. However, there also needs to be a focus on improving pre-existing risk-based algorithms. A study conducted in 2023 by Alexander Muacevic and John R Adler discovered that conversational agents specializing in mental health counseling failed to reference crisis resources – and failed to halt conversational dialogue if the user's input reverted to a lower risk level – indicating a lack of sensitivity and adaptability to the dynamic nature of mental health states (21).

- **Balancing Support and Harm Prevention**: The dynamic nature of emotions needs to be at the heart of the development of LLMs. Conversational agents must be able to engage with the individual and adapt to their needs, not through stated guidelines but through the evolution of the chatbot's character through continual tone analysis. This adaptative approach would involve the chatbot not only understanding and responding to the user's emotions but also evolving its behaviour and developing a unique personality over time.

  As such, a chatbot perpetually interacts with its users and then needs to be able to learn from the nuances of each conversation - gradually shaping their responses and interaction style to better align with the user's preferences and emotional state. This would give the individual a more personalised and intimate experience reflecting a sentient being that remembers past interactions and adapts accordingly. This can be achieved, for example, by tracking a user's emotional state using a transition network which predicts emotions based on past utterances and generates the most appropriate responses or by applying the principle of Valance and Arousal. A method where each word is embedded with an affective meaning (21)

  By addressing these contemporary issues – individuals may further engage with the chatbots – and build trust to facilitate a lasting relationship focused on the patient's well-being.

  Furthermore, health policies must propose a feedback mechanism where patients can regularly express their satisfaction or concerns – for the AI to adjust and adapt—making each interaction more meaningful and effective. There will, however, be circumstances where a user's input may become increasingly concerning – and during these situations, LangChain technology can be applied to create AI-supported high-risk methods that utilise prompt templates.

- Data Updates and Response Enhancements: The suggestions for improving the chat bot responses, including making them more relaxed and concise, increasing phrase variety, and refining natural language processing skills, highlight the need for regular database updates and response enhancements. Future policies could emphasize the importance of consistent updates to keep the chat bots knowledge current and ensure meaningful and varied interactions.

- Privacy and Informed Consent: The survey respondents raised valid concerns about data privacy and informed consent. Future health policies could address these concerns by clearly explaining where user data is sent, providing an explanation of the chatbot's information sources, and ensuring transparent mechanisms for obtaining informed consent. These policies would help establish trust and accountability in the use of chatbot technology for mental healthcare. Implementing robust privacy and security measures to safeguard user data and ensure compliance with relevant data protection regulations (22). Policies should address data storage, consent management, and secure communication protocols to protect user confidentiality.

- The suggestion to integrate structured therapy techniques and conduct research on the efficacy of chat bots in mental health support indicates the potential for collaboration between the chat bot application and professionals in the field. Future policies could encourage partnerships between developers and professionals in psychology and counseling to enhance the chat bots effectiveness and provide a well-rounded approach to mental healthcare (23).

- User Interface Enhancements: The participants' suggestions for improving the user interface, such as incorporating a more human figure, providing a chat history option, support different lanaguages and implementing helpful features like categorized questions, search functions, and quick help buttons, present opportunities to enhance usability and user experience. Future health policies could promote user-centered design principles and standards to ensure user-friendly interfaces that facilitate seamless interactions and access to information.

- Ethical Guidelines: Establishing clear ethical guidelines for the development and deployment of LLM-based chat bots to ensure user safety, privacy, and confidentiality. Policies should address issues such as data protection, informed consent, and the responsible use of AI technologies.

- Bias Mitigation: Implementing measures to identify and mitigate biases within LLM-based chatbot systems. Policies should promote regular audits and monitoring to ensure fairness and prevent any discriminatory or harmful outcomes (24).

- Training Data Diversity: Encouraging the use of diverse and representative training data to enhance the inclusion and accuracy of LLM-based chat bots. Policies should promote the inclusion of diverse perspectives and address potential biases in the data collection process.

- Continual Evaluation: Mandating regular evaluation and testing of LLM-based chatbot systems to assess their performance, reliability, and effectiveness. Policies should require transparency in reporting evaluation results and addressing any identified issues promptly.

- Accountability and Transparency: Establishing mechanisms for accountability and transparency in the development and use of LLM-based chat bot systems. Policies should require clear disclosure of the system's capabilities and limitations, as well as the organizations responsible for their development and operation.

future health policy should consider these discussion points to ensure the safe and effective utilization of mental healthcare support chat bots. By addressing issues related to these points, policies can help shape the development and implementation of chat bot technology in a manner that aligns with ethical, reliable, and user-centric mental healthcare practices.

6. Conclusions and future work

Navigating the challenges and complexities of mental healthcare requires tools and techniques that are both sophisticated and sensitive. A considerable part of these tools includes chat bots capable of providing conversational support. While the proficiency of chat bots has increased over the years, ensuring their responses are reliable, professional, and ethically sound in the delicate domain of mental health remains a substantial challenge (25).

Our research undertook the challenge of developing a chatbot designed to augment the capabilities of mental healthcare providers effectively. Traditional chatbots, predominantly rule-based and rooted in specific therapeutic methods such as cognitive behavioral therapy, offer commendable benefits. However, these systems are often constrained by their rigidity, failing to adapt to the complex and evolving nature of mental health dialogues. Acknowledging the merits of the rule-based approach, we are considering its integration into the future iterations of our chatbot to enhance its flexibility and adaptability, essential for addressing the nuanced demands of mental health conversations. (26).

Current Large Language Model (LLM)-based chat bots, despite their superior text generation capabilities, present another set of challenges. The unpredictability of their responses and potential deviation from standard mental healthcare practices raise legitimate concerns (27).

In response to these challenges, we devised a new methodology. We integrated a specifically fine-tuned DialoGPT model, which acts as a guidance system rooted in professional therapeutic practices, with the run time application of the ChatGPT API. The API lends the conversation a more dynamic and human-like quality, enriching the interaction with spontaneity and nuanced understanding, which are the hallmarks of human conversations (28; 29).

Our evaluation results, as quantified by perplexity and BLEU scores, demonstrate a promising improvement in performance when compared to using either the DialoGPT model or the ChatGPT API on their own. However, the real strength of our approach is more profoundly reflected in the feedback from the real-world users - mental health patients and professionals - who participated in our evaluation (30). Their affirmation signals the potential of our chatbot to align with the ethical sensitivity and supportive character demanded by mental healthcare practices (31).

We want to emphasize, though, that our chat bot, while a step forward, is not a replacement for human therapists. Instead, we envision it as an auxiliary resource that can provide support in scenarios where human resources are stretched thin, or as an additional tool to complement traditional therapeutic processes (32).

As we continue to refine this tool, we aim to deepen its comprehension of complex mental health issues and, more importantly, cultivate its ability to cater to the individual needs of users. We hope that, with continued research and development, we can make a meaningful contribution to the mental healthcare field (33).

In conclusion, our research elucidates the potential of utilising large language models, specifically a fine-tuned DialoGPT in conjunction with ChatGPT3.5, to deliver an efficient, professional, and reliable chatbot to support mental health care. We took a unique approach by training DialoGPT with a relatively more minor dataset of authentic therapeutic conversations and subsequently employed this model as a knowledge base for ChatGPT3.5. The integration of the models resulted in a chatbot system that combines the domain-specific understanding of the fine-tuned DialoGPT and the broader, more flexible response generation capabilities of ChatGPT3.5.

There are many future works will be continued:

Personalisation and GUI (Graphical user interface) design: This could encompass the development of suitable types of interface for different age group people based on more evaluations and feedback. Improving personalized experiences and visualization of user progress. By integrating these components, the chatbot could offer a more tailored, immersive, and motivational mental health support system, thereby elevating user engagement and satisfaction levels.

Incorporating Additional Datasets: Our current model utilizes a specific dataset of therapeutic conversations. To enhance the model's versatility and robustness, future work can incorporate additional datasets from varied sources. This could include conversations covering different types of therapy, diverse mental health conditions, and multiple demographic groups.

Integration with Clinical Systems: Future research could look into integrating the chatbot with existing clinical systems. This would allow the chatbot to provide more personalized and context-aware support. It could also facilitate better coordination with healthcare professionals, alerting them when the chatbot identifies potential serious concerns.

Multi-modal Inputs: Currently, the chatbot interacts through text-based conversations. Future work could explore multi-modal inputs like voice, facial expressions, or physiological signals. This could help the chatbot understand the user's emotional state more accurately and respond more appropriately.

Statement

We declare that this project - Funding: None, Competing interests: None declared, Ethical approval: Not required.

Acknowledgements

Thanks to Student Union Of University of Derby and Professor Myra Conway the Theme Lead in Biomedical Sciences for organising students and mental health professionals for helping on the evaluations.

Github project repository

You can download the backend testing and evaluation codes, a standalone application and evaluation data examples used in the research from github websites on GitHub semanticmachinelearning/MentalHealthLLMAnimationChatBot.

Author contribution

(I) Conception and design: Yu, McGuinness
(II) Administrative support: McGuinness
(III) Provision of study materials or patients: Yu
(IV) Collection and assembly of data: Yu, McGuinness
(V) Data analysis and interpretation: Yu, McGuinness
(VI) Manuscript writing: Both authors
(VII) Final approval of manuscript: Both authors

References

[1] L. Dinarte-Diaz, "An overlooked priority: Mental health," 2023.

[2] World-Health-Organization, "Mental disorders," 2022.

[3] L. Wilson and M. Marasoiu, "The development and use of chatbots in public health: Scoping review," *JMIR Human Factors*, vol. 9, p. e35882, Oct 2022.

[4] S. G. Hofmann, A. Asnaani, I. J. Vonk, *et al.*, "The efficacy of cognitive behavioral therapy: A review of meta-analyses," *Cognitive Therapy and Research*, vol. 36, no. 5, pp. 427–440, 2012.

[5] M. M. Linehan, H. E. Armstrong, A. Suarez, *et al.*, "Cognitive-behavioral treatment of chronically parasuicidal borderline patients," *Archives of General Psychiatry*, vol. 48, no. 12, pp. 1060–1064, 1991.

[6] S. C. Hayes, J. B. Luoma, F. W. Bond, *et al.*, "Acceptance and commitment therapy: Model, processes and outcomes," *Behaviour Research and Therapy*, vol. 44, no. 1, pp. 1–25, 2006.

[7] Z. V. Segal, J. M. G. Williams, and J. D. Teasdale, "Mindfulness-based cognitive therapy for depression: A new approach to preventing relapse," *Psychotherapy and Psychosomatics*, vol. 71, no. 6, pp. 251–259, 2002.

[8] J. E. Young, J. S. Klosko, and M. E. Weishaar, "Cognitive therapy for personality disorders: A schema-focused approach," 1990.

[9] J. Weizenbaum, "Eliza—a computer program for the study of natural language communication between man and machine," *Communications of the ACM*, vol. 9, no. 1, pp. 36–45, 1966.

[10] Shiji Group, "A rule-based or ai chatbot? here's the difference," 2022.

[11] C. Twomey and G. O'Reilly, "Effectiveness of a freely available computerised cognitive behavioural therapy programme (moodgym) for depression: Meta-analysis," *Australian and New Zealand Journal of Psychiatry*, 2016.

[12] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial," *JMIR Mental Health*, vol. 4, no. 2, 2017.

[13] B. Inkster, S. Sarda, and V. Subramanian, "An empathy-driven conversational artificial intelligence agent (wysa) for digital mental well-being: Real-world data evaluation mixed-methods study," *JMIR mHealth and uHealth*, vol. 6, no. 11, 2018.

[14] R. Fulmer, A. Joerin, B. Gentile, *et al.*, "Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: Randomized controlled trial," *JMIR Mental Health*, vol. 5, no. 4, 2018.

[15] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," 2020.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2018.

[17] A. Das, S. Selek, A. R. Warner, *et al.*, "Conversational bots for psychotherapy: A study of generative transformer models using domain-specific dialogues," in *Proceedings of the 21st Workshop on Biomedical Language Processing*, (Dublin, Ireland), pp. 285–297, Association for Computational Linguistics, 2022.

[18] "Transcripts of all patients." http://www.thetherapist.com/Transcripts.html, 2023. Accessed on December 15, 2023.

[19] N. Prasath and N. Prabhavalkar, "Mental health faq for chatbot." https://www.kaggle.com/datasets/narendrageek/mental-health-faq-for-chatbot, 2021.

[20] Hugging Face, "Bleu score evaluation space." https://huggingface.co/spaces/evaluate-metric/bleu, 2023. Accessed: December 15, 2023.

[21] S. Spallek, L. Birrell, S. Kershaw, E. Devine, *et al.*, "Can we use chatgpt for mental health & substance use education? a viewpoint examining the quality and potential harms," *JMIR Medical Education*, vol. 9, pp. e51243–e51243, 2023.

[22] R. May and K. Denecke, "Security, privacy, and healthcare-related conversational agents: A scoping review," *Informatics for Health and Social Care*, vol. 47, pp. 1–17, 10 2021.

[23] C. Grové, "Co-developing a mental health and wellbeing chatbot with and for young people," *Frontiers in Psychiatry*, vol. 11, p. 606041, Feb 2021.

[24] A. A. Parray, Z. M. Inam, D. Ramonfaur, *et al.*, "Chatgpt and global public health: Applications, challenges, ethical considerations and mitigation strategies," *Global Transitions*, vol. 5, pp. 50–54, 2023.

[25] M. Turunen, J. Hakulinen, A. Kallinen, M. Rauhala, *et al.*, "Designing chatbots for guiding online peer support conversations for adults with adhd," *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019.

[26] S. Hoermann, K. L. McCabe, D. N. Milne, *et al.*, "Conversational agents in the treatment of mental health problems: Mixed-method systematic review," *JMIR Mental Health*, vol. 4, no. 4, p. e43, 2017.

[27] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, *et al.*, "Better language models and their implications," *OpenAI Blog*, vol. 1, no. 8, 2019.

[28] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, *et al.*, "Dialogpt: Large-scale generative pre-training for conversational response generation," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 423–439, 2020.

[29] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[30] S. M. Schueller, A. Aguilera, and D. C. Mohr, "Feasibility and acceptability of a text-based delivered positive activities intervention for individuals with depression," *Internet Interventions*, vol. 9, pp. 53–59, 2017.

[31] A. Pringle, G. Sowden, and Q. Deeley, "The ethical implications of new technology in the treatment of mental health disorders," *Journal of Medical Ethics*, vol. 47, pp. 246–250, 2021.

[32] P. Cuijpers, E. Karyotaki, A. M. Pot, *et al.*, "Chatbots in mental health care: Potentials and limitations," *Psychotherapy and Psychosomatics*, vol. 89, pp. 65–70, 2020.

[33] D. Schlosser, J. Dworkin, M. Campa, *et al.*, "The future of digital mental health and its role in the therapeutic alliance," *Journal of Psychotherapy Integration*, vol. 31, pp. 138–145, 2021.