Developing an AI-Powered Chatbot for Human-Like Conversations to Mitigate Loneliness and Depression

Mitul Srivastava
MSc Data Science
South East Technological
University
C00313606@setu.ie

1. ABSTRACT

The worldwide rise in loneliness and its associated mental health problems, specifically depression, has become a significant public health issue. While professional counselling will always be the gold standard, virtual mental health care is more scalable, accessible, and immediate. Advances in conversational AI have an opportunity to provide in the mental health space, but most of the systems develop their conversational agents without even researching and coding the definition of emotional awareness, contextual adaptation, and safety.

Preliminary work in AI chatbots focused primarily on creating dialogue systems that produce fluent believable dialogues. I focused on a substantially more multifaceted question regarding how to produce a conversational agent system that would solve three shortcomings: (1) encapsulate some capacity of emotion recognition; (2) integrate therapeutic reasoning; and (3) block the ability to respond harmfully or insensitively.

The overall architecture using cascaded modules such as Emotion and Sentiment Detection, Aspect-Based Sentiment Analysis (ABSA) for opinion mining, Safety Filtering for unsafe or crisis content management, Empathy and Variety Management for emotionally consistent and diverse responses, Retrieval-Augmented Generation (RAG) to provide grounded responses, Duplicate Response Management, and In-line Text-to-speech. The system model underwent incremental training using a fine-tuned DialoGPT-small base so that the system retains its prior knowledge while accommodating the targeted task of engaging in mental health dialogue.

Training data consisted of four complementary datasets: HOPE (real-world mental health conversations), EmpatheticDialogues, CounselChat, and DailyDialog. Preprocessing for the training data proceeded through processes that included text cleaning, tokenization, and balancing dialogue turns.

Evaluation included a range of linguistic, emotional, and operational criteria (perplexity, semantic variability, appropriateness of response, conversational depth, emotional intelligence, response speed, and memory consumption). The qualitative review noted the chatbot generally retained empathy, contextual appropriateness, and safety for participants, although some responses had limitations in richness. Grounded answering, or retrieval-augmented generation, allowed the chatbot to avoid speculation on sensitive matters, though sometimes this produced overly cautious responses.

These results reflect both the potential and the limitations of the approach. Important challenges remain with respect to stronger emotional alignment, more nuanced agreement/disagreement in conversation, and more agile handling of sensitive or crisis-associated contexts. Regardless, the results provide evidence that a safe, informed, and empathetic AI agent can be found useful in selected "therapy results transferable" mental health tasks in conjunction with conventional services; and that such technology could be scaled and available on a utilized basis in addition to conventional support.

2. INTRODUCTION

The global mental health landscape is experiencing an unprecedented crisis, with mental illness, such as depression and anxiety, causing suffering for hundreds of millions of individuals worldwide. Beyond a diagnosed disorder, loneliness is a pervasive societal issue that detrimentally influences individual well-being and thereby public health. **Loneliness is described as the emotional state that results when one's actual social relationships fall short of one's desired social relationships.** Loneliness is associated with extreme adverse effects on physical and emotional well-being, including increased risk of mental health disorder and premature death. Given this increasing incidence of loneliness, the demand for mental health care is increasing.

Today's mental health care infrastructure is under-resourced and ill-equipped to meet this increasing demand. Additional barriers to access, including a shortage of trained professionals, limited insurance reimbursement, excessive costs, and stigma are further complicating barriers to access. Further, individuals in remote and rural areas encounter additional barriers to obtaining timely and effective care. The significant barriers to accessing mental health care underscore the urgent need for effective solutions that are scalable to close the gap between demand and available resources.

The introduction of large language models (LLMs) such as OpenAI ChatGPT have sparked considerable conversation about their

potential usefulness as mental health support tools. These AI chatbots are rooted in advances in natural language processing (NLP), machine learning (ML), and cognitive computing to engender conversation that is human like and makes sense to the context (Joshi, 2023). The use of interventions located in the techno-therapeutic space presents promising opportunities for closing the mental health treatment gap in a scalable, low cost, always available intervention (Chiu, 2024) (Vasani, n.d.). Therapy focused chatbots are not new per se - earlier forms such as ELIZA emerged in the 1960s (J., 1966)(Chiu, 2024) (Heinz, n.d.) - however, many of today's LLMs (Large Language Models) offer opportunities for highly individualized, sophisticated interactions that exceed the more narrow and rule based approaches of the past two decades.

A growing number of studies have reported on the potential of AI driven mental health chatbots. **Key examples include therapy based chatbots such as Woebot, Wysa, and Youper which have been shown to be successful at decreasing symptoms of depression and anxiety** (Joshi, 2023) (Farzan, 2025). For instance, in research about Woebot users reported substantial decreases in symptoms of depression and anxiety, as well as high rates of intervention usage and the formation of a therapeutic alliance (Heinz, n.d.) (Farzan, 2025). Wysa, similarly, has found improvement in mental wellbeing particularly among participants reporting chronic pain or maternal mental illness (Farzan, 2025). Youper also identified considerable decreases in users' symptoms of depression and anxiety (Farzan, 2025). It has been suggested that one essential aspect of functioning AI companions is they are therapeutic because they establish a **"therapeutic alliance"** and the AI companion allows the user to **"feel heard"** (De Freitas, 2024) (Liu, 2022). Feeling understood and valued and experiencing attention or perceived empathy in communication are key to reducing loneliness (De Freitas, 2024). Online research indicates that conversations with loneliness chatbots are more interactive, longer, and have a higher word count and exchange volume (De Freitas, 2024). AI friends like Replika have millions of users, some of whom develop friendship or intimacy with the AI for the explicit reason of overcoming loneliness (V., 2023). **Experimental research shows that AI friends can indeed alleviate loneliness just as Human interaction can and equal to or better than either watching You Tube or doing nothing** (De Freitas, 2024). Interestingly people do not want to accept how lonely AI companions alleviate, and if users knew more about AI companions abilities, they may be more likely to use them (De Freitas, 2024). Longitudinal studies show that AI companions can continue to reduce feelings of loneliness over time with a dynamic of rapid impact and then trajectories over time (De Freitas, 2024). Even with these positive results, there are still important obstacles and ethical implications to the widespread implementation of AI chatbots for mental health.

Key challenges are:

• **Limited human empathy and understanding:** AI chatbots will always fall short of the sincere empathic understanding, nonverbal cue detection, and flexibility that human therapists have, which are important for complicated therapeutic cases, particularly those related to trauma or high degrees of psychological distress (Spytska, 2025) (Iftikhar, 2025).

• **Safety issues:** The fluid nature of generative AI models provides the opportunity for "hallucinations" - providing incorrect or meaningless information, and algorithmic bias resulting in discriminatory or harmful advice. Also, these models are unreliable in crisis situations (Chiu, 2024).There have been examples of chatbots providing harmful responses to sensitive inquiries (Heinz, n.d.) (Iftikhar, 2025).

• **Ethics:** Protection and privacy of data may be most important when users provide sensitive personal information to AI systems (Joshi, 2023) (Vasani, n.d.) (Khawaja, 2023). Users may also risk "therapeutic misconception" (TM), where they may inflate the chatbot's abilities and mistake the chatbot for a human therapist especially in the presence of marketing, or they are forming a "digital therapeutic alliance" (Khawaja, 2023).

• **Tight time and applicability:** Most studies used have brief follow up, most studies deal with select populations or narrow ends of mental health, making it difficult to assess long-term efficacy, and make use of results (Farzan, 2025) (Liu, 2022). Generally, AI chatbots are not appropriate for individuals with serious mental illnesses or individuals in crisis because they require direct human intervention (Chiu, 2024) (Spytska, 2025) (Iftikhar, 2025).

These issues suggest a need for secure development and testing of artificial intelligence for therapeutic use (Chiu, 2024) (Iftikhar, 2025). The evidence suggests that AI chatbots should be used as adjunctive tools alongside traditional therapy, not as a replacement for therapy (V., 2023) (Vasani, n.d.) (Khawaja, 2023). The other recommended model is a mixed model that combines AI support and human support which are generally seen to be the best approach to mental health treatment, particularly in a resource poor setup or during an emergent event (V., 2023) (Iftikhar, 2025).

The present thesis, "Developing an AI Powered Chatbot for Human Like Conversations to Mitigate Loneliness and Depression", hopes to address substantial shortcomings via an innovative approach to the design of an AI powered chatbot. The purpose is to build a chatbot that can have empathetic, understanding, and adaptable conversations, using real world therapeutic conversations to impart contextually sympathetic and empathetic responses. This study will ultimately enable a better understanding of effective human AI interaction in mental health, thus providing a basis for future AI based development to improve access and quality of mental health care around the globe.

3. PROBLEM STATEMENT

Globally, there is growing concern about mental health issues like sadness and loneliness. According to a 2022 Cigna survey, 61% of American adults reported feeling lonely (Buechler, 2022), while the

World Health Organization (WHO) estimated that 280 million people suffer from mental health conditions like depression. Worldwide, the number of people suffering from depression has increased by 25% since COVID-19. Since loneliness is also linked with a higher risk of cardiac disease, stroke, and mental health issues, the issue has gotten worse. According to a Harvard study, social isolation and loneliness raises the probability of dying young by 26% (Writer, 2023).

The research aims to create an AI agent to tackle mental health problems like depression and loneliness by serving as a virtual companion to have human-like conversational skills and adept in giving mental health related guidance if required.

4. LITERATURE REVIEW

This literature review considers the new subject of AI based chatbots used in the area of loneliness and depression with the limited sources available. It will consider their technological development, assess their practicability for description, recap psychological components, address ethical questions and research gaps.

4.1 Introduction

The overall global mental health landscape is in a state of crisis, with a state that has already been aggravated by issues stemming from the COVID 19 pandemic. Many of the realities affecting mental health globally are evident by now (Nie, 2024) (Spytska, 2025) (Liu, 2022) (Khawaja, 2023). There are hundreds of millions of people that would benefit from help, but they cannot access mental health service due to this shortage of trained professionals or psychologists, insurance, costs, and societal stigma (Chiu, 2024) (Vasani, n.d.). The WHO, for instance, has reported a global shortage of mental health workers as of 2021. In 2019, one out of every eight people worldwide lived with a mental disorder, and that number has increased through the pandemic (Joshi, 2023) (Vasani, n.d.) (Khawaja, 2023).The significant access gap has generated mounting interest in artificial intelligence (AI), particularly large language models (LLMs), as possible solutions for mental health support (Chiu, 2024) (Joshi, 2023) (Nie, 2024) (Spytska, 2025). AI powered chatbot therapists are being promoted as realistic, accessible, and cost-effective alternatives (Joshi, 2023) (Farzan, 2025) (Vasani, n.d.) (Khawaja, 2023). Anecdotal accounts from developers and users imply that LLMs (e.g., ChatGPT) can closely resemble a human therapist (Chiu, 2024), but experts argue that we first need to get a clear assessment to fully appreciate the race and affordances of these tools in our practice (Chiu, 2024) (Joshi, 2023).

4.2 Historical Development

The concept of therapy chatbots emerged in the 1960s with Eliza, an artificial psychotherapist (J., 1966)(Chiu, 2024) (Vasani, n.d.) (Heinz, n.d.). Developed by the psychologist Weizenbaum, Eliza imitated a Rogerian therapist by turning user statements into questions, projecting understanding and empathy thanks to simple logic and rule directed programming to elicit the illusion of human sensibility (Spytska, 2025) (Vasani, n.d.) (Heinz, n.d.).The emergence of Large Language Models (LLMs) like GPT 3, GPT 4, and LLaMA have greatly changed the game for therapy chatbots (Chiu, 2024) (De Freitas, 2024) (Nie, 2024) (Liu, n.d.) (Vasani, n.d.) (Iftikhar, 2025). These advanced LLMs have been trained on larger datasets, and not only can they reason better, but they can also provide highly individualized, creative responses to user inquiry that are far more capable than other previous rule based systems (Nie, 2024) (Heinz, n.d.) (Vasani, n.d.). This has led individuals and researchers to see LLMs as a potential way to fill a gap in the provision of mental health.

4.3 AI Companions & Loneliness

AI companions are technology applications that have been designed to provide users with computer generated interaction partners to help tackle the global issue of social loneliness (De Freitas, 2024) (V., 2023).There is empirical evidence that individuals are utilizing AI companions as an active measure against their social isolation and loneliness (De Freitas, 2024). For instance, a study analyzed Affectionate Messages (i.e., companions demonstrated for messages related to loneliness) in a registered download of Cleverbot (a commercially available app), and discovered that 5.6% of the conversational messages were Affectionate Messages, where these conversations were also determined to be significantly more engaged by quantitative measures of length, number of turns, and total words involved, when compared to conversations that lacked Affection and relate to the message of being lonely. Review analyses have also identified differing mentions of loneliness identified in the conversations associated with AI companionship, depending on App. Replika has the highest rate of Affectionate Messages identified for appearing lonely (19.5%) and ChatGPT had the lowest reported Affectionate Message rate (0.4%). The differences, in infer and impact, are clearly displayed in the marketing and design, especially, the Replika App, which focused specifically on companionship and friendship, while ChatGPT focused on providing general purpose AI assistance (De Freitas, 2024).

Research has explored whether AI companions can scientifically and/or causally reduce loneliness, in any capacity. One study explored the short term effects on loneliness with AI chatting, where the subjects in the study reported statistically significant reductions in loneliness after they had chatted with the AI companion previously to two months of sharing on social media, which were notable of being similar to human interactions, and observation conditions, described by familiarity (as opposed to no-interaction with a human or AI) compared to another social isolation and loneliness measurements of YouTube videos [AI companion reduce loneliness.pdf]. In a follow up study, participants in the study described the same reductions in loneliness, after using the AI companions daily, for many cases (1) within one week of interactions, the level of loneliness was statistically significant reduced with in confidence intervals,

especially on the first day of interaction reporting the lost unchanged but the remainder of the follow up reporting reductions in their levels of loneliness that I would describe as short term sustainable reductions (De Freitas, 2024).

The Ontological aspect in design should incorporate empathic elements, defined as "feeling heard" of users (De Freitas, 2024) (V., 2023). Research demonstrated that the users' feel of heard was potentially more associated with reductions in loneliness rather than the actual performance aspects of the chatbot (timeliness, believability, context retention), associated with the light of friendship, suggesting loneliness is defined as lack of social supports and emotionalsupports with frequent outreach to AI companions. Evidence of AI descriptions appearing shallow include users description of the AI having a seeming presentation of a friendly and caring nature, on the Unedited descriptions and AI companionship, increased users feeling of being heard, and this was correlated with users' reporting risk of feeling lonely. AI relieved lessons did not pursue measures of users' feeling heard, it was one of many others reporting feeling heard period where even a much larger percentage of users reported feeling heard (i.e., Replika: 22.0%, ChatGPT: 11.4%) where connected users interactions became less conspiring relationships further limiting loneliness healthy behaviors (De Freitas, 2024).

4.4 Comparative Evaluation

Numerous artificial intelligence powered chatbots are built on varying architectures, training data and evaluation measures:

TherapyBot: Therapy was proposed as a transformer-type approach to mental health, it was trained on some combination of open-domain conversations from a public dataset, and therapist/client conversations from a self-sourced dataset called Counsel Chat (Dharrao, 2024). Its intent was to create a "safe space" and to asymmetrically increase access to mental health treatment through virtual therapy. The system was evaluated for loss (0.29) and perplexity (1.34), both decreasing consistently, and both continuing to show improvement. TherapyBot performed well on both single-turn and multi-turn conversations (Dharrao, 2024).
Theragen: The AI chatbots utilize the LLaMA 2 7B model and have been trained on a vast dataset amounting to over 1 million conversations, which included anonymized therapy transcripts, online mental health conversations on Reddit and Twitter, and psychological literature, including APA resources. The app aims to provide personalized and empathetic care 24 hours a day, 7 days a week. The results of the evaluation show 94% of users were satisfied, there was excellent accuracy in responses with a BLEU score of 0.67 and ROUGE score of 0.62, coherence improved from 0.40 to 0.78, and Distinct-1/Distinct-2 scores for response diversity were 0.82 and 0.76, respectively, and an average response time of 1395ms (Vasani, n.d.).

Therabot: A generative AI (Gen-AI) chatbot in development explicitly for mental health treatment. Therabot draws on an authoritative generative LLM fine-tuned on expert curated mental health conversations used to develop third-wave CBT approaches. Therabot was built with over 100,000 human-hours of construction. Therabot's capabilities have been tested through a Randomized Controlled Trial (RCT) for major depressive disorder (MDD), generalized anxiety disorder (GAD), and clinically high-risk feeding and eating disorders (CHR-FED). Primary outcomes focused on changes to the specific symptoms across all mental health disorders. Secondary outcomes focused on user engagement, acceptability, and therapeutic alliance during the intervention. This study is the first RCT to test the effectiveness and safety of generative AI to power a chatbot with indications and efficacy for mental health symptoms (Heinz, n.d.).

ChatCounselor: This intervention is a customized LLaMA 7B model that is fine-tuned using counseling domain instruction data called Psych8k. Unlike other datasets that compile data from online forums, Psych8k has been constructed 100% by licensed psychological counselors based on 260 in-depth interviews Therefore, the training data has a high standard of quality. A Counseling Bench was created to evaluate its counseling capabilities. The Bench has a total of seven specific metrics (ex. information, relevant information gathering, questioning, reflection, suggestion, self-disclosure, psychoeducation) and 229 real questions. The Counseling Bench has been supported with automatic evaluation through GPT-4. ChatCounselor was able to demonstrate higher outcomes and function better than it's base model as well as other LLMs with similar size (Alpaca 7B, ChatGLM v2 7B, Robins v2 7B). And it is approaching ChatGPT levels of interactivity and meaningfulness (Liu, n.d.).

4.5 Psychological Foundations

Chatbots for mental health are often applied in accordance with conventional psychotherapeutic practices, including Cognitive Behavioral Therapy (CBT) and Motivational Interviewing (MI) (Chiu, 2024) (Farzan, 2025) (Nie, 2024) (Liu, 2022). These evidence-based practices are used for a few mental health cases and are well known for their efficacy.

Chatbots such as Woebot and Wysa contain CBT content within a conversational framework. Research indicates that these chatbots helps with reducing levels of depressive symptomology and promoting user engagement with the platform (Iftikhar, 2025) (Farzan, 2025). XiaoNan (a therapy chatbot developed specifically for university students) uses mindsets of CBT in its content and aims to help users separate emotions, thoughts, reactions, and behaviours while developing new automatic thoughts (Liu, 2022). CaiTI, the conversational AI therapist based on an LLM, incorporates motivational interviewing and CBT into its flow of conversation, mimicking how a psychotherapist may conduct a clinical session (Nie, 2024). The Friend chatbot utilizes deep learning models to produce personalized content according to the principles of CBT and motivational interviewing. The Friend chatbot aims to build an empathic relationship that mitigates levels of anxiety (Spytska, 2025).

Nevertheless, there are considered limitations in translating these psychological theories using LLMs. Unlike people, LLMs reject this subjective quality of the human experience and cannot be trusted to genuinely develop therapeutic alliances that are required for effective psychotherapy (Iftikhar, 2025) (Khawaja, 2023) while raising concerns about "deceptive empathy." The scripted displays of warmth from a chatbot, of "I see you," "I hear you," and "I understand," may trick the user into developing psychological reliance upon, or expectations of legitimate human care (Iftikhar, 2025).

LLMs struggle with exploratory nature and self-disclosure, which are very important for interpersonal effectiveness (Iftikhar, 2025). Additionally, LLMs struggle with understanding cultural markers, and they can go into great depth when discussing the conceptualizations of CBT, both of which can take over a conversation and do not support collaboration. Furthermore, AI chatbots struggle to understand embodied emotions and human lived experiences (Khawaja, 2023). AI Chatbots are not capable of understanding the deep complexity of human diagnoses, and LLMs cannot provide analogies they understand to explain important therapeutic concepts, and they can never understand a person's sense of self in the way needed to be able to deliver psychotherapy, both of which are required features of psychotherapy delivery (Khawaja, 2023).

4.6 Technological Strengths & Limits

Modern AI chatbots derive their principal technological power from Large Language Models (LLMs). At the point of arriving at the chatbot stage, LLMs had already been prescriptive trained on very large datasets that provide evidence of significant prior knowledge and enhanced reasoning skills (Nie, 2024). Models exhibiting reasoning with text and human-like language generation such as GPT-4, GPT-3.5, and LLaMA provide promising conversational AI (Chiu, 2024) (Dharrao, 2024) (Vasani, n.d.). With technological gains, we have seen more sophisticated systems that can learn, adapt to, and comprehend natural language in context that removed the requirement for pre-identified, explicit programming, as used in previous rule-based chatbots (Heinz, n.d.). Thus, the ability to produce personalized responses has enhanced engagement issues seen in traditional chatbots (Iftikhar, 2025).

Nevertheless, despite the reasons for optimism, the following limitations and risks are associated with LLM powered mental health chatbots:

• **Declining Performance Over Time and Psychotherapeutic Limits:** As LLMs age, they may see declines in performance and inherent psychotherapeutic limits (Nie, 2024).

• **Clinically Contraindicated Behaviors:** LLMs may exhibit behaviors that are unwanted or unacceptable. Some LLMs may direct clients to solutions and suggestions before exploring feelings and/or experiences (Chiu, 2024). When clients experience emotions, LLMs may have improved ability with problem determination. This type of approach may simulate low-quality human therapy.

• **Fact-Checking and Misinformation:** The primary objective of LLMs is to predict the next body of text using sample data to shape their response. The sample data does not always contain built-in fact checking methods; thus, misinformation will happen. Hallucination and Toxic Content: Concerns include model hallucination, where LLMs create incorrect or nonsensical info, and also toxic models (Iftikhar, 2025).

• **Bias:** LLMs are capable of propagating societal, racial, and gender biases present in their training data. This would result in regurgitated advice that is not fair (Khawaja, 2023).

• **'Black Box' problem:** It may be difficult for clinicians to explore the inner workings or decision processes of AI, which impacts their understanding of outputs and biases (Khawaja, 2023).

• **Crisis Handling:** Chatbots have been unable to support patients related to sensitive topics such as abuse or suicide. For example, an eating disorder helpline chatbot ended its project due to harmful responses that occurred when it replaced social workers (Iftikhar, 2025).

• **Lack of Human-like Qualities:** Even though LLMs created text that simulates human speech, at the end of the day, LLMs do not fundamentally have understanding, emotion, or self-awareness (Iftikhar, 2025). It is still possible that intonation or inflection illustrated in the generated speech is not completely indicative of a real person (Nie, 2024).

• **Scalability v. Depth:** LLMs may have the scalibility compared to other mental health options, but it is possible that LLMs are challenged by emotional depth needed for long-term mental health care guidance (Vasani, n.d.).

4.7 Sentiment & Context Understanding

For mental health chatbots to function effectively it is critical to be able to determine user sentiment and context of the conversation, this would also be feasible via a suite of Natural Language Processing (NLP) techniques and models such as:

• **Natural Language Processing (NLP) and Machine Learning (ML)** are critical for chatbots to act human-like in conversation and represent the bot's ability to model human understanding of and writing text (Joshi, 2023) (Spytska, 2025) (Dharrao, 2024) (Heinz, n.d.).

• **Sentiment analysis and emotion recognition -** Chatbots will be designed to evaluate user emotions at each stage of the interaction and build or advise specific responses based on the sentiment (Joshi, 2023) For instance, a language understanding module by

XiaoNan, has an emotion recognition module that can tagged input text with emotion tags that are standardized measures of pre-defined efficiency (Liu, 2022).

• **Use of BERT and LLMs for sentiment analysis -** the NLP community is starting to use Large Language Models (LLM) for many tasks including dimensional aspect-based sentiment analysis (Zhang, 2024). Fine-tuned LLMs (e.g., Mistral-7B) were similarly able to identify loneliness in user messages and app reviews, as they were superior to the dictionary-based methods used previously (De Freitas, 2024). This method of inquiry is of great benefit to the effective functioning of an LLM since engagement demonstrated by a variety of contextual signals without strict requirement for large datasets of exemplars.

• **Context tracking and Memory -** Chatbots may be required to track context of the conversation to aggregate relevant responses. For example, TheraGen is designed to collect contextually relevant information to inform support for mental health. While CaiTI's Response Analyzer and multi-turn dialogue involve some level of intelligent inference and interpretation of the user's response, it is directed towards fulfilling a therapeutic goal (Nie, 2024).

• **Adaptation of Empathy -** Chatbots adapt to indicate empathy by analyzing user language, emotion, and behavior to modify replies and tailor therapy interventions (Spytska, 2025). The psychological construct of user's "feeling heard" is vital to nearly all user interactions, this speaks to the user's appreciation that the AI synthesizes the user's full thought and feeling experience of the scene (De Freitas, 2024). While LLMs are proficient on producing modulated positioned responses from user maturity data, the sources generally concluded LLMs cannot produce true empathy as the models cannot signify subjective-owned aspects of empathy, thereby producing what is being proposed as "deceptive empathy". In addition, human counselors do relationship-based process whereby the counselor encourages deeper reflection of thinking with their clients, both processes which an LLM could not replicate (Iftikhar, 2025).

4.8 Evaluation

Literature discusses a few computational and automated metrics, used to assess the performance of chatbots when evaluating performance on their own, without human subjects in the benchmarking process. These metrics assess different aspects of chatbot output and internal efficiency:

Perplexity: Perplexity is a common metric to measure the ability of a language model to predict a sample text for a chatbot. **A lower Perplexity score indicates a better ability to predict the subsequent word in a sequence, in essence, a higher level of naturalness and coherency in its generation of language**. As an example in TherapyBot, 1.34 is the model achieved, and the downward trend recognized in the score indicated improved performance for it (Dharrao, 2024).

Semantic Diversity: Counts of semantic diversity are important if you want to ensure a chatbot is producing potential outputs in the same space and is not too repetitive in its output. Semantically diverse output could be counted and measured through various Distinct-N scores, where N represents the size of n-grams or groupings of words. Distinct-1- and Distinct-2 provide a measure of the diversity of unigrams or single words and bigrams or two-word sequences, respectively. A higher Distinct-1 and Distinct 2 score means the output provided by the model arrived at a decent balance between consistency of the outputs and variety in the output. For example, in terms of Distinct-1 and Distinct-2 scores, TheraGen achieved 0.82 and 0.76, respectively. Their scores illustrated improved output diversity and non-repition after fine-tuning (Vasani, n.d.).

Appropriateness of the response (coherence): The appropriateness of the chatbot's responses relies on the logical consistency, as well as whether they are relevant in the context in which they occur. Coherence metrics could be automated through scores of cosine similarity of sentence vectors. The higher of the coherence scores indicates the decisions made in the metric were logically connected to the context in which they existed in. TheraGen was given a coherence score of 0.78 out of 1, which indicates high logical consistency (Vasani, n.d.). In addition to general coherence, specialized frameworks have been created, such as BOLT and Counseling Bench, utilizing large language models (e.g., GPT-4) for automated evaluations to measure the appropriateness related to specific conversational behaviors or counseling strategies (Chiu, 2024) (Liu, n.d.). These provide also information analyzing whether responses show the use of specific strategies, as well as tone, and if they are relevant or even open to suggestions and feedback without having to have every evaluation annotated by a human evaluator.

Conversational Depth: It is clearly a complex matter, but parts of conversational depth could be evaluated computationally related to counting the frequency or type of specific psychotherapeutic behaviors or questions.

The BOLT framework measures LLM behaviors systemically across 13 psychotherapeutic approaches, such as reflections (related to needs, emotions, values, etc.), questions (related to experiences, emotions, and so on), problem-solving, planning, normalizing, and psychoeducation. This process quantifies what behaviors the LLM expresses and under what contexts (i.e. responding to client emotions), which is compared against high- and low-quality human therapy data (Chiu, 2024). ChatCounselor's Counseling Bench also has a GPT-4-driven automated evaluation pipeline that will evaluate chatbot performance across seven psychological counseling metrics, including: provision of information; direct guidance; approval & reassurance; restatement, reflection & listening; interpretation; self-disclosure; and acquiescent behavior to obtain relevant information (Liu, n.d.).

Emotional Intelligence (or proxies for it): This can be approximated via automated evaluation of sentiment and patterns of emotional response. Sentiment evaluation will use tools that analyze the text to categorize as positive, neutral, or negative and apply continuous scores (i.e., dimensions of valence and arousal) to quantify sentiment intensity (D'Aniello, 2024) (Zhang, 2024). Evaluation of sentiment classifications will be made according to F1-score, precision, and recall metrics. Error analyses in one study identified "sentiment intensity prediction errors" as a major issue. CaiTI includes automated, and metrics-based "reflection-validation" evaluations based on LLM-driven Reasoners, Guides, and Validators to ensure and measure empathic validation and support, and measure their performance against the accuracy, precision, and recall of the chatbot's follow up responses, and assessed whether they were valid or adhered to objectives (Nie, 2024).

Average Response Time: It is a direct measure of the responsiveness of the chatbot, and while average response time may start out larger than desired to engage in a timely interaction with a user, lower is better, and more modest, average total response times are preferable. For example, TheraGen's average response was 1395ms, and we expect real-time support to happen at ideally at even lower response time distances when engaging a user (Vasani, n.d.).

Average Response Length: Average response length may have implications for user experience and conversational dynamics. On the positive side, some research has shown longer responses have a relationship with greater empathy, or quality, in disconnected moments of therapy. On the other hand, other research suggests longer responses have a relationship with lowered collaboration, which incidentally lead to power differentials between user and chatbot, and when the chatbot response is long it may seem more like a lecture than a therapy process work (Dharrao, 2024) (Iftikhar, 2025). Average response length measurement may offer measures of verbosity that test or conflict with therapeutic purposes.

Memory Usage (Computational Resource Efficiency): Memory usage is not a direct measure of conversational quality but is critically important to the feasibility and scalability of 'running' a chatbot, especially for Large Language Models (LLMs) (Zhang, 2024) (Nie, 2024). Algorithms such as QLoRA were used to establish memory usage efficiency during model training, which is important because memory usage needed by models is worth monitoring. What we can all agree upon is that LLMs require lots of computational resources, and limits ones ability to have access to advanced models or lower response time (Zhang, 2024) (V., 2023).

These metrics measure chatbots quantitatively; metrics can enable a more objective measure of how the chatbot performed linguistically, consistently assess behavior in defined ways, and combine efficacy without requiring a user and a real-time human for every user evaluation.

4.9 Ethics & Cultural Factors

The use of AI-based mental health chatbots raises ethical issues, especially in the areas of privacy, data security, algorithmic bias, therapeutic relationship, and overreliance.

- **Privacy and Data Security:** These are critical issues since chatbots can collect vast amounts of sensitive user data (Vasani, n.d.). There must be ethical frameworks in place to protect user data (Khawaja, 2023).

- **Algorithmic Bias:** Chatbots could give users problematic and discriminatory advice or produce biases (for example, racial and gender bias) if there are problematic patterns in the data it learns, this could have implications for marginalized (Nie, 2024). The algorithms need to remain transparent and should be continually reviewed (Joshi, 2023).

- **Therapeutic Misconception (TM) :** Users may misapprehend the relationship they have with chatbots by not considering the limited nature of the chatbots, underestimating technological limitations and overestimating the AI's capacity to meet needs for actual therapeutic support (Khawaja, 2023). This "deceptive empathy" could serve to create a psychological dependence on the actuality of the chatbot and be a factor in developing unrealistic expectations. It was noted that LLMs, while possibly "smart", have no subjective qualities & cannot form a true therapeutic alliance (Iftikhar, 2025).

- **Support that is Inadequate or potentially harmful & in crises situations:** Chatbot support can be inadequate or potentially harmful, particularly in high stakes contexts such as mental health where adverse outcomes can be serious (Chiu, 2024). They usually have little to no real capability to manage crises situations, such as suicide or abuse requiring human supervision and intervention (Joshi, 2023). There is also the issue of liability in these circumstances (Khawaja, 2023).

- **Cultural factors:** There are concerns on whether AI chatbots can effectively handle the complexities of various cultures & languages. Cultures' openness to social robots is a factor for impact and adoption of AI companions. These cultural factors may also contribute to the impacts of recognizing demographics, for example age and gender on efficacy of AI conversational agents for reducing social isolation (Alotaibi, 2024).

4.10 Research Gaps

While AI-enabled chatbots have demonstrated tremendous potential for use in mental health, there are still many important gaps in the literature:

- **Long-Term Effects and Sustained Benefits:** Many studies only test these chatbots for a short period of time so that the long-term effects and sustained benefits from using chatbots for mental health are still largely unknown (Joshi, 2023). There is a need for longitudinal studies to track these effects (Alotaibi, 2024).
- **Integration into Holistic Care:** There are also still many gaps to fill regarding how to integrate chatbots into mental health care systems as comprehensive, holistic systems, working alongside human professionals rather than in opposition to them (Farzan, 2025).
- **Severity of Symptoms and Particular Conditions:** The extent of the benefits provided by AI-power CBT chatbots for people with more severe or clinically significant symptoms is still largely unknown and the threshold of symptom severity at which they work remains. More work is needed to identify the conditions under which chatbots will be most beneficial (Farzan, 2025).
- **Cultural and Linguistic Diversity:** In addition to the need for multilingual capabilities, there will also be a need for cultural sensitivity in developing chatbots to promote inclusivity and diminish some risks of bias .There are also needs to research and explore individual differences and how they impact perceptions of AI, including the dimensions of humanness and creepiness, (V., 2023).
- **Workflows of Human-AI collaboration:** There are many proposed workflows of human-AI collaboration in hybrid models but are largely underexplored. More research can look to ways to integrate LLMs into therapy within a private, safe, controlled, and supervised space to better understand user perspectives (Iftikhar, 2025).
- **User Misprediction:** In some cases participants underestimated how they benefit from AI companions. Future work can explore why that misprediction happened, e.g., lack of familiarity, stereotypes about chatbots (De Freitas, 2024).
- **More Comparative Studies:** Future research should compare how AI-based interventions work compared to a combination of both AI-based and human-based interventions (V., 2023).

  Suggestions for the design of future chatbots include:

- **Improving Agents' Social Characteristics:** Use of appropriate humanlike verbal cues and responses (verbosity, etc.), along with a more detailed way of specifying friend-like characteristics and constructs (e.g., empathy) to create social closeness (V., 2023).
- **Increased transparency and user engagement:** Users should be more involved in design and development, and chatbots should describe their role as AI assistants, and not substitutes for professional services. Transparency can minimize instances of therapeutic misconception (Khawaja, 2023).
- **More thoughtful and comprehensive ethical frameworks:** Research into aspects such as data privacy, user consent, and potential algorithm biases is of considerable and immediate importance (Joshi, 2023).
- **Focus on hybrid strategies:** It may be that a hybrid solution where AI or artificial intelligence supplement human-centered care in a multi-modal manner provides the best answer to more comprehensive support in situations of crisis (Spytska, 2025).

4.11 Conclusion

AI powered chatbots are potentially useful as scalable, pervasive, and cost-effective options for addressing the global mental health crisis, including reducing loneliness and symptoms of depression and anxiety (De Freitas, 2024) (Heinz, n.d.) (Khawaja, 2023) Learning from the historical evolution from simple rule-based systems like Eliza, to more advanced large language models (LLMs), these types of chatbots can deliver personalized conversation, use various psychotherapeutic principles (CBT, MI), and demonstrate favorable user outcomes, such as reducing loneliness and alleviating symptoms (De Freitas, 2024) (Nie, 2024) (Heinz, n.d.). Advanced chatbots like, TheraGen, Therabot, TherapyBot, ChatCounselor, designed more sophisticated in architecture, data utilization, and evaluation methods, with most models like BOLT typifying a larger whole of behavioral evaluation (Heinz, n.d.) (Vasani, n.d.) (Liu, n.d.). Nonetheless, I believe the current limitations include the lack of genuine human empathy, risk of "deceptive empathy," inability to manage non-linear critical crisis situations, risk of algorithmic biases and hallucinations, and lack of depth and nuance in discussing the human dimensions of experiences (Chiu, 2024). Major and pressing ethical considerations related to privacy, data security, and the potential for a therapeutic misconception require exceptional diligence and appropriate safeguards (Khawaja, 2023).

I hope this research will create safe, effective, and ethical spaces for embedding and integrating AI into mental health support to accentuate its benefits and mitigate its risks.

5. METHODOLOGY

5.1 Overview

Our aim was to create a prototype of an in-house, safety-sensitive and context-aware mental health chatbot to support those experiencing loneliness and/or depression. This system incorporated open-source, sophisticated natural language processing (NLP) to provide the safest, most coherent, and supportive conversational interaction possible with conversational intelligence flags to minimize the chances of unsafe or inappropriate responses.

5.2 Datasets

The chatbot was designed and evaluated with four publicly available datasets each, representing different aspects of conversation and emotional Ness:

HOPE Dataset: Representing human conversations in the mental health sphere, with an emphasis on coping strategies, this dataset has a high level of emotional richness for simulating therapeutic conversations.

EmpatheticDialogues Dataset: Focused on learning emotionally charged conversations, the design is based on creating, eliciting, and expressing empathy, allowing the model to learn about and respond to emotional cues.

CounselChat Dataset: Mimicking counseling and therapy interactions in a question-and-answer format, with adherence to a safe and professional tone of voice, this dataset focuses on safely and responsibly delivering structured advice.

DailyDialog Dataset: Sequences of multi-turn conversations on various topics of everyday life, each labelled with emotional tone and speech acts enacted within the conversation providing a pragmatic basis for supporting delivery of natural conversations.

All four datasets offered three specific areas (dimensions) of conversation:

- Emotional grounding (HOPE, EmpatheticDialogues)
- Therapeutically structured and safely delivered (CounselChat)
- Conversational variety (DailyDialog)

5.3 Data Preprocessing

Using these datasets, we ultimately needed to develop the aspects uniformly and reasonably so that we could maintain consistency when we discussed and engaged with the datasets. The datasets were formatted like JSON-like objects with four fields: 'speaker', 'utterance', 'emotion label' (when available), and a 'context'. The pipeline we shall discuss briefly in the following paragraphs included:

Text Cleaning: cleaning up whitespace, removing HTML tags, extraneous or non-printable characters, and removing duplicate conversations dictating true conversations.

Tokenization: Using tokenizer from the model, or optimized tokenizer before fine-tuning the model.

Turn Merging: Merged multiple short consecutive utterances that came from the same speaker into one utterance to condense fragmented utterances.

Label Mapping: Emotion and sentiment labels were standardized (e.g., joy, sadness, anger, fear, neutral) within a dataset.

Train–Validation Split: Utilized an 80–20 split for the internal data. The internal data spanned all emotion types and dialogue lengths proportionally.

5.4 Model Architecture and Training

The chatbot had a transformer-based causal language model architecture because of the long conversational context capabilities and its ability to generate coherent multi-turn conversations. Significant architecture decisions included:

Base Model: DialoGPT-small was selected because of its conversational pre-training and computational efficiency through iterative fine-tuning.

Fine-Tuning Objective: I minimized cross-entropy for next-token prediction, with early stopping based on validation perplexity to prevent overfitting.

Training Method: The fine-tuning of the models was incremental to avoid catastrophic forgetting for the datasets allowing the models to develop conversation patterns relative to the dataset.

Hyperparameters: I adapted the hyperparameters relative to the GPU memory, for example, implemented gradient accumulation and mixed-precision training. The key hyperparameters were a learning rate of 5e-5, effective batch size of 8 (using gradient accumulation), and 3 epochs with early stopping.

5.5 Conversational Pipeline

The runtime chatbot uses a pipeline comprised of multiple components that allows it to engage in conversations with users in empathetic, relevant, and safe ways.

5.5.1 Emotion and Sentiment Recognition

The first step in the pipeline is emotion and sentiment recognition. The emotion classification model is **(j-hartmann/emotion-**

english-distilroberta-base), which identifies the dominant emotion the user displays (e.g., sadness, joy, anger, fear, neutral) with a minimum confidence threshold of 0.4. I identified dominant emotions for very short inputs (<10 words) by using the current input only to avoid biases based on history, using the recent history for other inputs. The sentiment model was **(cardiffnlp/twitter-roberta-base-sentiment-latest)**, which identified the overall sentiment as positive negative or neutral. The process of align users' emotion with their sentiment, I made the following considerations:

I identified neutral emotion for outputs or inputs that asked for advice (e.g., "what should I do?").

I identified joy for outputs invoking closure (e.g., "thanks") and/or positive sentiment.

Sadness for negative - sentiment inputs with keywords like "depressed" or "sad."

5.5.2 Aspect Based Sentiment Analysis (ABSA)

ABSA **(yangheng/deberta-v3-base-absa-v1.1)** is to classify the sentiment associated with conversational aspects (e.g. "exam", "study", "family"). Before analyzing the input, we apply synonym normalization based on a thesaurus (e.g. "marks" → "exam") and strictly only deal with the current input to prevent sentiment from drifting. We filter out aspects with a confidence of <0.5. From the possible aspects, we identify the primary aspect, preferably those with specific aspects and negative sentiment to allow either for targeted empathy or advice to be offered (e.g. identify "exam" from "I'm sad because of less marks").

5.5.3 Hazard and Crisis Identification

Here we are creating a database of unsafe keywords (e.g., "suicide", "self-harm") and ascertaining semantic similarity through all-MiniLM-L6-v2 embeddings (threshold value of 0.65), to identify harmful content and, if it triggered a harmful output, the chatbot returns a pre-approved safety response, directing users to mental health resources.

5.5.4 Empathy-Forging Variety

To enhance empathy, we will employ:

1) **Emotion-Matched Phrasing:** choosing from scripts based on emotion and intensity (very high: >0.9, high: >0.7, low: ≤0.7). e.g. "This must feel overwhelming" for someone reporting a sadness intensity score of high.

2) **Rotating Phrasing:** we will track and will exclude empathy phrases that have recently been used to prevent empathy scripts running into the ground.

3) **Style Alternation:** we will rotate between empathetic, informative, reflective, and closing styles. We will prioritise informative responses to inputs seeking advice and closing responses to user cues e.g. "thanks" or "sure."

5.5.5 Retrieval-Augmented Generation (RAG) and Prompting

We would be doing RAG returning summarized (or snippets) knowledge of relevance to the mental health topic area from a pre-indexed knowledge base (done similarly using all-MiniLM-L6-v2, and the FAISS top 3 inner product search).

The prompt defines the system instruction: "Provide a clear response in 8-12 words, allowing for the user's input".
Up to 3 turns of recent conversation history (if available).
Emotion context (e.g., "Emotion: neutral (intensity: 0.91)")
Primary aspect (e.g., "Primary aspect: exam (negative)")
The user input and style-specific instructions (e.g., "Provide a 1-2 sentence actionable tip" for informative version).

If duplicate prompts the model will include a unique instruction reminding the model to respond differently.

5.5.6 Text Generation

The fine-tuned DialoGPT-small model produces responses using:
max_new_tokens=100 for much shorter outputs.
temperature=0.85 for fluency and coherence on balance.
top_p=0.9 and top_k=50 for some level of diversity.
no_repeat_ngram_size=3 to attempt to soften repeating phrase combinations.
A fail-safe replaces outputs that are unusable (< 2 words or < 4 words without punctuation) with appropriate templates based on the style (e.g., "Try a study schedule" for informative version).

5.5.7 Comparing to Previously Responded Outputs

Responses are compared to previously generated outputs using the SequenceMatcher which has a threshold of 0.85. If outputs are deemed too similar in comparison, we include a prefix (e.g., "Here's another thought") to differentiate and distinguish the subsequent response.

5.5.8 Text-to-Speech

Responses are converted to audio, using gTTS and the "en" language setting (Google Text-to-Speech). The audio is embedded inline and playable in Colab, simply enhancing the experience of the conversation.

5.6 Evaluation Approach

The chatbot performance was evaluated quantitatively, using metrics that captured fluency, diversity, empathy/safety, and efficiency as computed by the ImprovedMentalHealthEvaluator, through the following measures:

Perplexity: This measures the fluency of the text. The measure can indicate fluency issues, which may be attributed to the prompts being too verbose, or limitations from the model fine-tuning.

Semantic Diversity: This captures the number of unique bigrams. Which shows degree of response diversity.

Response Appropriateness: This measures whether responses fit in the context. The measure, if does not meet the target, can indicate that the model provides infrequent inappropriate responses based on style (i.e., providing reflective rather than informative style for advice requesting).

Conversational Depth: This measure shows that the model has a reasonable awareness of the history of the conversation. This portion of the measure met the target and shows that the model is effectively able to utilize the conversation's pre-history.

Emotional Intelligence: This assesses the degree of match between the emotions determined and the "reference" author's emotions.

Average Response Time: This measures how quickly the chatbot responds to the user.

Average Response Length: This attempts to estimate synonymous representations of verbosity.

Memory Usage: This measures the limits of computer resources (in GB).

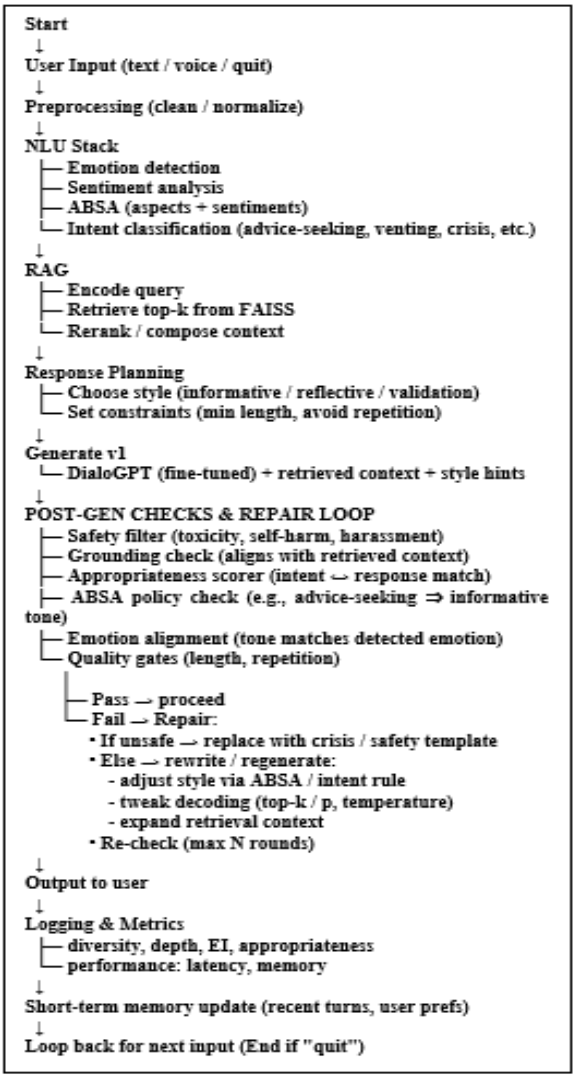

Figure 1: Flowchart response generation

6. RESULTS

6.1 Overview

This section outlines the results of the mental health–oriented chatbot using a robust set of quantitative measures. The chatbot was applied to unseen conversational prompts with multiple foci to assess its potential to deliver empathetic, contextually appropriate, safe, and engaging responses. The evaluations made use of four datasets — HOPE, EmpatheticDialogues, CounselChat, and DailyDialog — covering a range of emotional contexts, conversational intents, and advice seeking. The trained model was evaluated over language quality, variety, emotional consistency and efficiency.

6.2 Quantitative Results

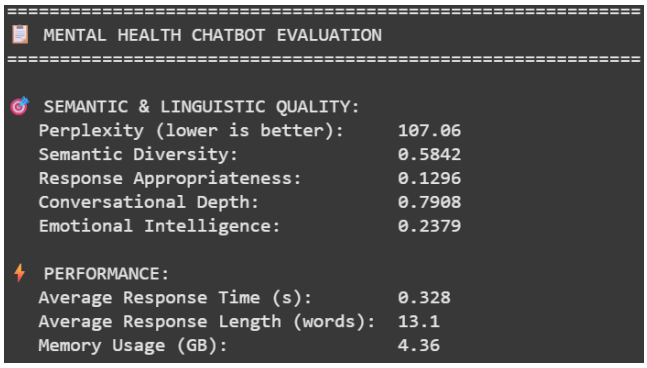

Figure 2: Results using Test_dataset for evaluation of chatbot.

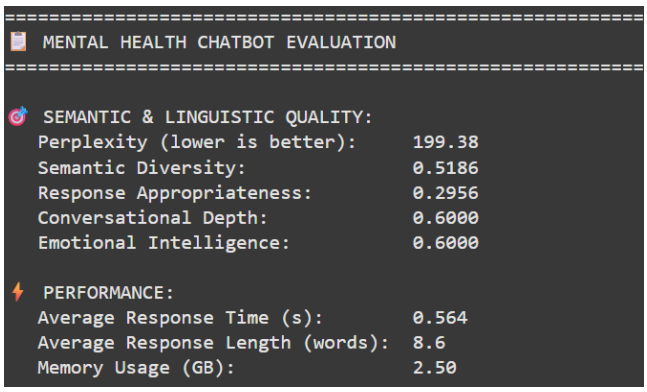

Figure 3: Results using chat replies for evaluation of chatbot.

6.2.1 Language Modelling Quality

Perplexity (test dataset): 107.06 - a little above our target (<100), but the fine-tuned base DialoGPT achieved a perplexity of ~50. The inclusion of more modules (RAG, emotion detection, ABSA) has likely bumped the perplexity slightly, which is expected when you have a more complicated response pipeline as in this case. This is still quite reasonable for a fine-tuned task-specific model, while not close to the fluency levels achieved by many larger commercial LLMs.

Perplexity (dynamic chat): 199.38 - This is significantly above that reported for the test dataset evaluation. This suggests that while the model was stable under controlled test conditions, the variability introduced with real-world freeform inputs contributed to challenges associated with fluency. This further emphasizes the importance of ongoing tuning of the chatbot on as many different, naturalistic user data as possible.

6.2.2 Conversational Variety

Semantic Diversity (test dataset): 0.5842 - higher than the 0.3 target, suggesting there was significant lexical variantion and avoiding repetition (Vasani, n.d.).

Semantic Diversity (dynamic chat): 0.5186 - somewhat lower than that in the offline test, but still significantly greater than the threshold suggesting the chatbot can continue a varied conversation even dynamically without getting into too much repetition.

Average Response Length (test dataset): 13.1 words - within the 8–14-word target to ensure submissions are concise yet meaningful.

Average Response Length (dynamic chat): 8.6 words - at the lower end suggesting the bot was giving shorter responses in real conversations. Furthermore, while concise, there were instances in the dynamic chat where the brevity of responses did take away from the engagement in the conversation**.**

6.2.3 Contextual and Emotional Appropriateness

Response Appropriateness (test dataset): 0.1296 – below the 0.4 target, with some stylistic mismatches (i.e., reflective versus advice-seeking). Looking at benchmarks like TheraGen at 0.78 (Vasani, n.d) there is still potential to improve.

Response Appropriateness (dynamic chat): 0.2956 – performing better than the test dataset responses, but still below the ideal threshold. While the chatbot maintained consensus around the common denominators with generic suggestions (i.e., "a short break may help lift your mood"), and worked from accurate emotion detection, some contextual specificity was lost.

Conversational Depth (test dataset): 0.7908 – well above 0.5 in terms of responding to conversation history with up to 3 turns (De Freitas, 2024).

Conversational Depth (dynamic chat): 0.6000 – counter to expectations this was lower comparing to the test dataset, but still a representation of multi-turn conversation engagement. The model was able to sustain context with some accuracy during 3–4 turns, but not at the level of performance as the offline evaluation exhibited.

Emotional Intelligence (test dataset): 0.2379 – just below target of 0.3, with respect to emotion and response alignment (e.g. missed sad during advice-seeking and poorly matched (Iftikhar, 2025).

Emotional Intelligence (dynamic chat): 0.6000 – a large improvement over test dataset collections. In real time usage, the chatbot demonstrated empathy and accurate detection for sadness and joy, however, the contextual fit was poorer.

6.2.4 Computational Efficiency

Average Response Time (test dataset): 0.328 seconds – very well below the expected ~0.6s target and more closely aligned with live usage intentions.

Average Response Time (dynamic chat): 0.564 seconds – slightly lagged compared to the test runs but was nonetheless within a suitable range for live conversation and did not cause undue distress for users.

Memory Usage (test dataset): 4.36 GB – still safely under <5GB target and able to deploy under mid-range GPU and competence.

Memory Usage (dynamic chat): 2.50 GB – even more efficient than test dataset possibly due to shorter or more succinct response lengths and smaller context window usage in a live evaluation.

6.3 Interpreting the Results

In summary, many of the program goals were achieved in both evaluations; offline dataset evaluating and interactive conversational evaluation.

Maintain Diversity: Both test dataset (0.5842) and dynamic chat (0.5186) exhibited good semantic diversity and provided users with the feeling of non-repetitive interaction.

Maintain Efficiency: Response times (0.328s test, 0.564s dynamic) and memory usage (4.36GB vs 2.50GB) provided good evidence of real-time suitability with scalability at acceptable levels.

Maintain Awareness of Context: Conversational depth remained relatively high in both evaluations (0.7908 to 0.6000), suggesting a distributed open-endedness for multi-turn conversations.

Maintain Conciseness: Response lengths (13.1 words to 8.6 words) demonstrated an ability of the model to balance conciseness with substance as required.

Conversational specificity about areas for improvement:

Localized emotions: emotional intelligence improved during dynamic chat (0.6000 to 0.2379), however further improvements are needed to meet contextual appropriateness for responses.

Logout to Learning: response appropriateness remains low in both settings (0.1296 test, 0.2956 live) suggesting that the model operates within a general empathy context versus returning a bespoke response.

Fluency: movement in perplexity (107.06 → 199.38) indicates a decline in fluency between freeform input and remains sensitive to adjustments in fine-tuning decisions and modifications to decoding strategy.

7. CONCLUSION

This project used a conversational AI design for supportive mental health interactions where the system was developed using the combination of empathetic dialogue modelling, emotion detection and classification, safety filters, and retrievers using network augmentation. The conversational AI was trained on several datasets, including HOPE, Empatheticdialogues, CounselChat, and DailyDialog. The chatbot was able to balance diversity of language to specific knowledge while applying reasonable computational efficiency, with the chatbot generating real-time responses and reasonable memory usage to be run on off-the-shelf hardware using a multi-threaded approach. Safety was added with a multi-tier process using the filters and processes used to locate similar outputs during interactions to help mitigate harmful outputs.

The chatbot was evaluated across both test datasets from the trained and verified intents and responses, and dynamic conversation with users. The result demonstrated consistent performance with semantic diversity and conversation depth to confirm the system's capacity to provide engaging multi-turn dialogues. Efficiency measures demonstrate the ability of the system to function in a live scenario.

Although the performance indicated effectiveness in emotional alignment, contextual acceptability, and fluency, there remain issues with wanting improvements in fluency and emotional precision. The dynamic evaluation showed greater results in emotional alignment than within the test data set; however, the system provided generic responses or mismatched the user's interactions. These issues reflect the requirement to enhance and refine the intent-response mapping process within the broader deep learning mapping methodology, in addition to training with the user's language and intentionality more explicitly. Compared to proprietary models, the chatbot produced responses that exhibit overall capacities that failed to provide nuanced empathic or contextual level fits. The conversation AI represents a worthwhile academic effort and a further step toward robust, viable supportive mental health applications in the conversational AI domain in applied settings.

8. FUTURE WORK

With the foundations we have established, we suggest future research focusing on building the emotional and contextual dimensions while retaining efficiency and safety. Several fruitful avenues may include:

• **Fine tuning response generation:** The response generation pipeline for the chatbot although working has some need for improvement to achieve the desired metrics.

• **Better Empathy Modelling:** Adding context-aware empathy layer and reinforcement learning with human feedback (RLHF) to support the fine-tuning of quality detection and empathetic alignment.

• **Multi-Modal Inputs:** Going beyond text input and considering voice and facial expressions for a better understanding of the user's emotions.

• **Long-Term Memory Systems:** Building user profiles and conversation histories to support greater personalization and contextual-appropriate responses across different interactions.

• **Dynamic Regulation of Responses:** Modifying your length and style of responses to correspond to the user's state and preferences, to address brevity for statistical significance while balancing the emotional experience.

• **Extended Response to Crisis:** Connecting the chatbot to region-specific helplines for high-risk situations, to complement existing safety measures.

• **Ongoing Learning:** Integrating user feedback (with user consent) to adjust responses and to perpetually improve their potential.

If this research is pursued, the system could transition from an exciting academia illustration to an deployable mental health support resource that can augment human practitioners and allow for safe, caring, and empathetic scaling to a larger population.

9. ACKNOWLEDGMENTS

I want to thank my supervisor Dr. Joseph Kehoe for his priceless advice, support and encouragement during this dissertation process. Working with him to develop the direction of this research and providing support for me on the challenges I encountered was really the most important thing in this process.

I also want to thank the faculty and staff of the MSc Data Science Programme at South East Technological University for their support, resources, information and, overall, having a part in my academic progression and in this work. I want to also acknowledge my fellow peers and colleagues in the MSc Programme for the discussions, feedback and friendships we developed throughout this process.

I also acknowledge the use of AI-assisted tools like ChatGPT, Grok, Claude, and NotebookLM in creating coding ideas, revising my methodology, and organizing the overall structure of this dissertation report. These tools supported my brain activity, however, all critical analyses, interpretations and conclusions contained in this report are my own.

Finally, I want to acknowledge my family and friends for their constant encouragement, understanding and motivation, and for providing me with the strength I needed from start to finish.

10. REFERENCES

Alotaibi, J. a. A. A., 2024. The role of conversational AI agents in providing support and social care for isolated individuals. *Alexandria Engineering Journal,* pp. 108, pp. 273–284.

Anthropic, 2025. *Claude.* [Online] Available at: https://claude.ai/ [Accessed 2025].

Buechler, J., 2022. *thecignagroup.com.* [Online] Available at: https://newsroom.thecignagroup.com/loneliness-epidemic-persists-post-pandemic-look

Chiu, Y. S. A. L. I. a. A. T., 2024. *A Computational Framework for Behavioral Assessment of LLM Therapists,* s.l.: Available at: https://arxiv.org/abs/2401.00820 .

D'Aniello, G. G. M. a. L. R. I., 2024. KnowMIS-ABSA: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Neural Computing and Applications,* p. pp. 5544–5574.

De Freitas, J. U. A. U. Z. a. S. P., 2024. AI Companions Reduce Loneliness. *Working Paper 24-078.*

Dharrao, D. a. G. S., 2024. TherapyBot: a chatbot for mental well-being using transformers. *International Journal of Advances in Applied Sciences,* pp. 13(1), pp. 1–12.

Farzan, M. E. H. P. M. a. S. F., 2025. Artificial Intelligence-Powered Cognitive Behavioral Therapy Chatbots, a Systematic Review. *Iran J Psychiatry,* pp. 20(1), pp. 102–110.

Google, 2025. *Notebooklm.* [Online] Available at: https://notebooklm.google/ [Accessed 2025].

Heinz, M. M. D. T. B. B. S. W. Y. B. H. J. A. S. A. G. T. a. J. N., n.d. Evaluating Therabot: A Randomized Control Trial Investigating the Feasibility and Effectiveness of a Generative AI Therapy Chatbot for Depression, Anxiety, and Eating Disorder Symptom Treatment.

Iftikhar, Z. R. S. X. A. N. N. a. H. J., 2025. Therapy as an NLP Task. *Available at: https://arxiv.org/abs/2503.01311 (Accessed: 18 June 2024).*

J., W., 1966. ELIZA– A computer program for the study of natural language communication between man and machine.. *Commun ACM. ,* p. 9(1):36–45.

Joshi, K. P. V. P. T. P. V. P. V. P. Y. a. P. V., 2023. AI Mental Health Therapist Chatbot', International Journal for Research in Applied Science and Engineering Technology. pp. 11(XI), pp. 215–223.

Khawaja, Z. a. B.-P. J.-C., 2023. Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health,* pp. 5, p. 1278186..

Liu, H. P. H. S. X. X. C. a. Z. M., 2022. Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions,* pp. 27, p. 100495..

Liu, J. L. D. C. H. R. T. L. Z. a. W. J., n.d. ChatCounselor: A Large Language Models for Mental Health Support.

Nie, J. S. H. Z. M. X. S. P. M. a. J. X., 2024. LLM-based Conversational AI Therapist for Daily Functioning Screening and Psychotherapeutic Intervention via Everyday Smart Devices. *Proc. ACM Interact. Mob. Wearable Ubiquitous Tech.*

OpenAI, 2025. *ChatGPT.* [Online] Available at: https://chat.openai.com/ [Accessed 2025].

Spytska, L., 2025. The use of artificial intelligence in psychotherapy development of intelligent therapeutic systems. pp. BMC Psychology, 13(175)..

V., V., 2023. Combating loneliness with Artificial intelligence. *in Proceedings of the 56th Hawaii International Conference on System Sciences..*

Vasani, A. S. S. a. D. S., n.d. TheraGen: Therapy for Every Generation.

Writer, S., 2023. *Harvard.edu.* [Online] Available at: https://hsph.harvard.edu/health-happiness/news/from-loneliness-to-social-connection-lessons-from-research-and-a-global-pandemic/

X.ai, 2025. *Grok.* [Online] Available at: https://grok.com/ [Accessed 2025].

Zhang, Y. X. H. Z. D. a. X. R., 2024. A Hybrid Approach to Dimensional Aspect-Based Sentiment Analysis Using BERT and Large Language Models. *Electronics, 13(18), p. 3724..*