INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

ROBERT GORDON UNIVERSITY ABERDEEN

Voice-Assisted AI Mental Health Companion

Literature Review Document by

Rikzy Jezuli

Supervised by

Ms. Dileeka Alwis

Submitted in partial fulfilment of the requirements for the BSc(Hons) in Artificial Intelligence and Data Science degree at the Robert Gordon University.

November 2024

Table Of Contents

2.1 Introduction

The integration of Artificial Intelligence (AI) in mental health care has emerged as a significant field of study, driven by the need for scalable and cost-effective solutions. With the growing prevalence of mental health issues, traditional mental health services face limitations in accessibility and resource allocation (Denecke et al., 2020; Xu et al., 2024). The purpose of this review is to explore the application of AI, particularly Large Language Models (LLMs) and chatbots, in providing mental health support. LLMs such as GPT-3 and specialized AI-based chatbots are being increasingly adopted to offer personalized therapeutic interventions, leveraging advancements in natural language processing (Na, 2023; Crasto et al., 2021).

Mental health chatbots, grounded in Cognitive Behavioural Therapy (CBT), have shown promise in enhancing emotional regulation and reducing symptoms of anxiety and depression (Omarov et al., 2023; Pandey et al., 2022). However, despite these advancements, challenges remain in terms of ethical concerns, data diversity, and model accuracy (Bill & Eriksson, 2023). This literature review addresses key themes, including therapeutic approaches, advancements in LLMs, and the challenges and ethical considerations involved. By analyzing current trends and research, this review aims to contribute to the ongoing development of effective AI-driven mental health interventions.

2.2 Development and Evolution of AI in Mental Health Support

2.2.1 Early Approaches in AI-Based Mental Health Tools

Early developments in AI-driven mental health tools primarily focused on rule-based systems and foundational chatbots. Notable examples include pioneering chatbots like ELIZA and PARRY, which aimed to replicate a human-like interaction through predefined scripts and rules (Pandey et al., 2022). Although these chatbots were revolutionary for their time, they had significant limitations in their ability to understand complex language inputs and adapt to dynamic conversations (Pandey et al., 2022).

As the demand for scalable mental health support increased, newer chatbots like Woebot emerged, using CBT frameworks to offer structured therapeutic interventions. Woebot, for instance, demonstrated effectiveness in providing CBT-based support, which was reflected in reduced symptoms of depression among users (Denecke et al., 2020). However, these early AI interventions often followed predefined conversational pathways, which limited their ability to provide personalized responses. Many of these systems lacked the capacity to interpret nuanced emotional states or recognize intricate contextual cues, leading to a push for more sophisticated AI tools (Omarov et al., 2023).

2.2.2 Emergence of Large Language Models (LLMs)

The introduction of LLMs, such as GPT-3 and GPT-4, marked a significant advancement in AI's capabilities for mental health applications. LLMs, with their deep learning architectures and extensive training datasets, brought substantial improvements in natural language understanding and generation. This development allowed AI models to handle complex conversations and respond more empathetically to users' emotional cues (Xu et al., 2024).

A prominent example is the development of Mental-LLM, a fine-tuned model designed to enhance mental health support through online interactions. Mental-LLM demonstrated significant improvements in predicting mental health states by using specialized datasets and fine-tuning techniques. This approach enhanced the model's performance in tasks related to mental health assessment and intervention, outperforming general-purpose models like GPT-3.5 in key accuracy metrics (Xu et al., 2024). Similarly, ChatCounselor, a model built on domain-specific datasets derived from professional counselling sessions, showcased how LLMs could be adapted to provide targeted psychological support (Liu et al., 2023). The use of domain-specific datasets and reinforcement learning techniques significantly increased the quality and relevance of the chatbot's responses in psychological contexts (Bill & Eriksson, 2023).

Despite these advancements, the deployment of LLMs in mental health contexts presents challenges related to data quality, ethical risks, and contextual understanding. To address these concerns, researchers have emphasized the importance of training LLMs with data grounded in established therapeutic techniques, such as CBT, to ensure the delivery of accurate and constructive advice (Na et al., 2023). The continued refinement of these models and their integration with professional counselling practices indicate a promising direction for AI-based mental health support (Yu & McGuinness, 2024).

2.3 Therapeutic Approaches in AI-Based Mental Health Solutions

2.3.1 Cognitive Behavioural Therapy (CBT)-Based Chatbots

CBT is a well-established form of psychotherapy that focuses on the relationship between thoughts, emotions, and behaviors. In recent years, AI-based chatbots have been designed to incorporate CBT principles, providing users with automated psychological support. For instance, Omarov et al. (2023) introduced a mobile chatbot psychologist using AI Markup Language (AIML) combined with CBT techniques. This chatbot aimed to help users recognize and challenge cognitive distortions by offering structured therapeutic interventions (Omarov et al., 2023). Another example is the development of CBT-LLM, a LLM designed specifically for CBT-based question-answering in a Chinese-speaking context. The researchers fine-tuned the model using a dataset grounded in CBT principles, resulting in more professional and structured responses compared to generic models (Na et al., 2023).

These chatbots effectively simulate therapeutic conversations, enabling users to receive immediate support and guidance. Studies have shown that such chatbots can effectively reduce symptoms of depression, anxiety, and stress by facilitating cognitive restructuring and behavioral changes (Denecke et al., 2020). However, limitations include the models' dependency on predefined scripts, which can restrict the adaptability of responses in complex or multi-faceted cases.

2.3.2 Other Therapeutic Approaches

Besides CBT, alternative therapeutic approaches such as Behavioral Activation (BA) and emotion regulation techniques have also been incorporated into chatbots. Rathnayaka et al. (2022) proposed a chatbot integrated with personalized Behavioral Activation therapy. This chatbot provided emotional support and remote health monitoring, with features like identifying key activities to uplift mood and mitigate

anxiety. The pilot study demonstrated its effectiveness in recurrent interventions, highlighting the potential of BA as an alternative to CBT in chatbot-based therapy (Rathnayaka et al., 2022).

Another example is the chatbot SERMO, which implements emotion regulation strategies to aid users in recognizing and managing their emotions. This chatbot uses natural language processing to detect the emotional state of users and suggests appropriate mindfulness exercises or activities accordingly. It was found to be beneficial in enhancing emotional awareness and improving coping skills (Denecke et al., 2020).

2.3.3 Multimodal and Hybrid Approaches

Advancements in AI have led to the emergence of multimodal and hybrid therapeutic approaches in chatbots. These chatbots combine various therapeutic principles, integrating emotion regulation with behavioral and cognitive interventions to provide comprehensive support. For instance, "Ted the Therapist" uses natural language processing and deep learning techniques to analyze and respond to user inputs. This chatbot effectively offers a blend of supportive dialogue and therapeutic exercises, catering to both cognitive restructuring and emotional regulation needs (Pandey et al., 2022).

Additionally, the hybrid chatbot developed by Yu and McGuinness (2024) integrates specialized mental health datasets with advanced language models like BERT. This model utilizes fine-tuning techniques and prompt engineering to provide contextually relevant responses that align with traditional therapeutic practices. This approach shows significant improvement in conversational quality and engagement, which is crucial for maintaining user interaction and trust in AI-based mental health tools (Yu & McGuinness, 2024).

2.4 Advancements in Large Language Models for Mental Health Applications

2.4.1 LLM-Based Chatbots for Psychological Interventions

Recent advancements in LLMs have demonstrated significant potential in enhancing chatbot-based mental health support. These LLMs, such as ChatGPT and GPT-4, have been increasingly employed to simulate conversational agents capable of providing therapeutic responses based on cognitive and emotional cues. A prominent example is the development of a Chinese large language model named CBT-LLM, which integrates CBT principles to generate structured and professional responses. Na et al. (2023) fine-tuned CBT-LLM using a CBT-specific dataset, which led to improved quality and relevance in chatbot interactions, particularly in addressing mental health concerns (Na et al., 2023).

Moreover, models like ChatCounselor have been developed using domain-specific instruction data to refine their counseling responses. Liu et al. (2023) utilized a dataset derived from real-world counseling sessions and applied fine-tuning techniques to achieve a more accurate and empathetic response generation. The evaluation framework, which incorporated metrics such as counseling relevance and conversational empathy, showed that fine-tuned models outperformed generic ones in

therapeutic settings (Liu et al., 2023). These developments highlight the importance of specialized training and fine-tuning for improving the applicability of LLMs in mental health support.

2.4.2 Predictive Models and Diagnostic Capabilities of LLMs

The use of LLMs extends beyond chatbot interactions to predictive modelling and mental health diagnostics. Xu et al. (2024) introduced Mental-LLM, a framework leveraging instruction fine-tuning to enhance mental health prediction tasks. By training on diverse online text datasets, the study demonstrated that fine-tuning significantly improved the model's performance in identifying and predicting mental health conditions. Mental-LLM achieved a 10.9% improvement in balanced accuracy compared to non-finetuned models, indicating the effectiveness of adapting LLMs for specialized mental health tasks.

Similarly, Qing Yu and McGuinness (2023) explored a hybrid model combining traditional LLMs with domain-specific datasets. Their study fine-tuned a model based on CBT principles and found that it significantly enhanced the conversational quality and relevance in mental health support. This approach highlighted the advantage of integrating LLMs with structured therapeutic methodologies to achieve better outcomes. Such advancements underscore the potential of LLMs not only as interactive chatbots but also as diagnostic tools capable of predicting and addressing mental health concerns effectively.

2.5 Datasets and Training Strategies for Mental Health AI Tools

2.5.1 Datasets for Training Mental Health Chatbots

The development of effective AI-based mental health chatbots relies significantly on the quality and relevance of datasets used for training. Various datasets have been designed to focus on different aspects of mental health interventions. For instance, the **Psych8k** dataset, constructed from over 260 in-depth counselling sessions, serves as a vital resource for fine-tuning language models for mental health counselling. This dataset enables models like **ChatCounselor** to simulate realistic therapeutic conversations (Liu et al., 2023). Similarly, the **CBT QA dataset** was developed with a focus on CBT principles, providing structured question-answer pairs that guide AI chatbots in delivering professional and relevant responses (Na et al., 2023).

Datasets such as **Ted the Therapist's database** integrate conversational data pre-processed using natural language processing (NLP) techniques and structured based on CBT principles. This enhances chatbots' ability to interact effectively with users facing mental health challenges (Pandey et al., 2022). These datasets often emphasize specialized knowledge and structured interaction to enable chatbots to respond empathetically and professionally.

2.5.2 Challenges in Dataset Diversity and Representation

One of the primary challenges in training AI-based mental health tools is the limited diversity and representation within existing datasets. Most datasets are sourced from specific cultural and linguistic contexts, limiting the models' generalizability. The CBT QA dataset, while effective for structured CBT interventions, lacks the diversity to handle varied linguistic nuances and emotional contexts (Na et al., 2023). Similarly,

datasets like Psych8k primarily cover traditional counseling settings, which may not fully encompass the range of mental health issues faced by users from different cultural and socio-economic backgrounds (Liu et al., 2023).

Moreover, biases in datasets can lead to ethical risks, particularly when AI chatbots provide advice that may not be culturally sensitive or contextually appropriate (Xu et al., 2024). Addressing these limitations requires diversifying training datasets to include different demographics, as well as refining models through continuous feedback and iterative improvements. Incorporating varied real-world data and enhancing dataset curation strategies will be crucial to creating more inclusive and reliable AI mental health tools.

2.6 Evaluation of AI-Based Mental Health Tools

2.6.1 Quantitative Evaluation Metrics

The effectiveness of AI-based mental health tools, particularly those leveraging LLMs, is assessed using several standard quantitative evaluation metrics. Perplexity is a key metric used to measure how well a model predicts a sequence of words, indicating the model's language fluency. Lower perplexity scores correlate with more coherent and contextually relevant responses (Yu and McGuinness, 2024). Another widely used metric is the BLEU score (Bilingual Evaluation Understudy), which compares generated chatbot responses against reference responses. High BLEU scores indicate that the chatbot's output closely aligns with predefined responses based on therapeutic principles, such as CBT (Na, 2024).

Additionally, models are evaluated on task-specific metrics. For instance, chatbots focusing on mental health support may be assessed based on emotion recognition accuracy or adherence to therapeutic goals. Xu et al. (2024) introduced the use of balanced accuracy in predicting mental health conditions, highlighting that LLMs fine-tuned with domain-specific data performed significantly better than generic models. These quantitative measures are vital in determining the chatbot's ability to provide therapeutic responses that are contextually relevant and clinically sound (Xu et al., 2024).

2.6.2 User and Professional Feedback

In addition to quantitative metrics, user and professional feedback is integral in evaluating AI-based mental health tools. Feedback from end-users and mental health professionals provides valuable insights into the chatbot's usability, empathy, and practical effectiveness. In the study by Liu et al. (2023), the ChatCounselor model was evaluated using feedback from patients and professionals. The results demonstrated that user feedback helped identify key improvements in chatbot responses, focusing on aspects such as empathy, clarity, and conversational flow (Liu et al., 2023).

Professional feedback is crucial for aligning chatbot interactions with therapeutic standards. For instance, the SERMO chatbot was evaluated using the **User Experience Questionnaire** (UEQ) to gauge its efficiency, perspicuity, and attractiveness. The feedback highlighted areas of improvement, emphasizing the need for balancing conversational novelty with therapeutic goals (Denecke et al., 2020). Such evaluations

are critical in refining AI-based mental health tools, ensuring they remain reliable, empathetic, and aligned with clinical standards.

2.7 Challenges and Ethical Considerations in AI for Mental Health

2.7.1 Bias and Ethical Risks in AI Models

AI models, especially LLMs, are subject to biases stemming from their training data and the algorithms themselves. In the mental health context, biases in AI can manifest in various forms, including racial, gender, and socio-economic biases, which may lead to incorrect or harmful recommendations. Xu et al. (2024) emphasized that these biases could undermine model reliability and patient safety, as LLMs may inadvertently reinforce stereotypes or provide misleading information based on biased datasets (Xu et al., 2024). Moreover, issues like overgeneralization from limited data pose significant ethical risks, especially when models lack grounding in established psychological practices (Na, 2024). Given the sensitive nature of mental health, the risk of deploying biased or unvalidated models is particularly concerning. Therefore, careful evaluation of training data and continuous monitoring of AI-based interventions are crucial to minimizing these risks (Bill and Eriksson, 2023).

2.7.2 Data Privacy and Security Concerns

Mental health applications require handling sensitive information, which raises significant privacy and security concerns. AI-based systems, particularly chatbots, often collect and process large volumes of user data to provide personalized support. Protecting this data from breaches and unauthorized access is paramount. According to Liu et al. (2023), implementing robust encryption protocols and secure data storage systems are essential to safeguard users' privacy (Liu et al., 2023). Furthermore, ethical concerns arise from the potential misuse of collected data, such as profiling or targeting vulnerable individuals. To address these concerns, mental health chatbots like CareBot prioritize data minimization practices, collecting only the information necessary for delivering effective support (Crasto et al., 2021). Adhering to stringent privacy laws and regulations, such as GDPR, is also vital to maintaining users' trust and ensuring compliance with legal requirements.

2.7.3 Cultural and Contextual Sensitivity in AI Mental Health Tools

The effectiveness of AI-driven mental health tools hinges on their ability to provide culturally and contextually relevant support. Many current models are developed and trained using datasets that may not adequately represent diverse populations, leading to limitations in their applicability (Pandey et al., 2022). For instance, chatbots like "Ted the Therapist" showed that even minor misinterpretations of cultural nuances could affect the quality of mental health interventions (Pandey et al., 2022). Addressing this requires incorporating diverse datasets and ensuring that AI systems are trained to understand different cultural contexts. Denecke et al. (2020) emphasize the need for culturally adaptive systems that can modify their recommendations based on users' cultural backgrounds and preferences (Denecke et al., 2020). Thus, developing culturally aware AI tools not only improves accessibility but also enhances the overall impact of mental health interventions.

2.8 Comparison of Related Work

Research	Model Used	Dataset	Benefits	Diagnostic Technique	Limitations	Further Improvements
Denecke et al. (2020)	SERMO (NLP Lexicon-based)	Not Specified	Daily emotion regulation support using NLP and lexicon-based techniques	CBT	Limited evidence of long-term effects; neutral evaluations for novelty	Enhanced emotion recognition and multilingual support
Crasto et al. (2021)	GPT-based Chatbot	PHQ-9, WHO-5	Provides anonymity, addresses social stigma in mental health interventions	CBT	Limited to text-based responses; needs more emotional understanding	Adding voice recognition and real-time emotional assessment
Rathnayaka et al. (2022)	Behavioral Activation Chatbot	Not Specified	Provides recurrent emotional support and personalized assistance	Behavioral Activation (BA)	Limited scope in therapy personalization; only tested in pilot studies	Expanding to include other therapeutic approaches and larger trials
Pandey et al. (2022)	CNN-based Deep Learning	Not Specified	High accuracy in appropriate response delivery, uses deep learning models	CBT	Relies heavily on training data quality; may miss nuanced interactions	Improved natural language understanding and scenario-based training
Omarov et al. (2023)	AIML-based Chatbot	Not Specified	Cost-effective and efficient solution, personalized CBT interventions	CBT	AIML lacks adaptability and flexibility for complex therapeutic interactions	Incorporating adaptive and multimodal capabilities
Liu et al. (2023)	Vicuna-7B (Fine-tuned LLM)	Psych8k	High-quality domain-specific data, improved counseling metrics	CBT	Limited generalizability due to small dataset size	Expanding dataset size and improving response diversity
Bill & Eriksson (2023)	RLHF-based LLM	Not Specified	Uses human feedback to refine responses; ethical framework considerations	CBT	No significant improvement over pre-trained models due to hardware limitations	Additional tests with enhanced datasets and hardware capabilities

Na (2024)	Fine-tuned LLM (CBT-LLM)	CBT QA Dataset	Structured and professional responses based on CBT principles	CBT	Limited to Chinese language data; lacks cross-linguistic support	Expanding to multilingual capabilities and diverse datasets
Yu & McGuinness (2024)	DialoGBT and ChatGBT 3.5	Not Specified	Blended model for better contextual awareness in mental health conversations	CBT	Limited scalability and deployment risks	Integrating improved safety measures and scalability techniques
Xu et al. (2024)	Mental-Alpaca & Mental-FLAN-T5	Online Social Media	High balanced accuracy in mental health prediction tasks	Not specified	Bias in predictions, ethical risks	Expanding ethical considerations and diverse dataset integration

Table 1 Comparison of Related Work.

2.9 Conclusion

This literature review highlights the transformative role of AI in mental health care, particularly through the use of LLMs and chatbots. The reviewed studies emphasize the effectiveness of AI-enabled systems in delivering personalized mental health support based on established therapeutic frameworks like CBT (Na, 2024). Chatbots like ChatCounselor and Ted the Therapist demonstrate the capacity of LLMs to facilitate accessible and empathetic conversations, contributing to improved mental health outcomes (Liu et al., 2023; Pandey et al., 2022).

However, significant gaps persist in the current literature. Challenges include the limited cultural adaptability of AI models and the lack of diverse, high-quality datasets (Xu et al., 2024). Moreover, while AI systems have shown initial success in emotional regulation and user engagement, their long-term efficacy in replicating professional therapeutic practices remains a critical area for further research (McGuinness and Yu, 2024).

To address these gaps, future research should focus on refining AI systems to provide more personalized and context-aware support. Incorporating interdisciplinary insights from psychology and ethics into AI training processes can enhance the adaptability and safety of these systems. Furthermore, exploring hybrid approaches that combine rule-based methods with LLMs could lead to more effective therapeutic interventions (Bill and Eriksson, 2023).

These findings offer valuable insights into my own research goals of developing a scalable and adaptive AI-based mental health support system. By addressing identified limitations and enhancing personalization, I aim to contribute to the ongoing efforts in making mental health care more accessible and effective.

References

Bill, D. and Eriksson, T., 2023. Fine-tuning a LLM using Reinforcement Learning from Human Feedback for a Therapy Chatbot Application. Royal Institute of Technology, Stockholm, Sweden.

Crasto, R., Dias, L., Miranda, D. and Kayande, D., 2021. CareBot: A Mental Health ChatBot. In: 2021 2nd International Conference for Emerging Technology (INCET). Belgaum, India: IEEE.

Denecke, K., Vaaheesan, S. and Arulnathan, A., 2021. A Mental Health Chatbot for Regulating Emotions (SERMO) - Concept and Usability Test. IEEE Transactions on Emerging Topics in Computing, 9(3), pp.1170-1179.

Liu, J.M., Li, D., Ren, T., Liao, Z. and Wu, J., 2023. ChatCounselor: A Large Language Model for Mental Health Support. In: Proceedings of PGAI CIKM 2023, Birmingham, UK.

Na, H., 2024. CBT-LLM: A Chinese Large Language Model for Cognitive Behavioral Therapy-based Mental Health Question Answering. Australian Artificial Intelligence Institute, University of Technology Sydney, Australia.

Omarov, B., Zhumanov, Z., Gumar, A. and Kuntunova, L., 2023. Artificial Intelligence Enabled Mobile Chatbot Psychologist using AIML and Cognitive Behavioral Therapy. International Journal of Advanced Computer Science and Applications, 14(6), pp.137-145.

Pandey, S., Sharma, S. and Wazir, S., 2022. Mental healthcare chatbot based on natural language processing and deep learning approaches: Ted the therapist. International Journal of Information Technology, 14(7), pp.3757-3766.

Qing Yu, H. and McGuinness, S., 2024. An Experimental Study of Integrating Fine-tuned LLMs and Prompts for Enhancing Mental Health Support Chatbot System. Journal of Medical Artificial Intelligence, pp.1-24.

Rathnayaka, P., Mills, N., Burnett, D., De Silva, D. and Gray, R., 2022. A Mental Health Chatbot with Cognitive Skills for Personalised Behavioural Activation and Remote Health Monitoring. Sensors, 22(10), 3653.

Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A.K. and Wang, D., 2024. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. Proceedings of the ACM on Interactive, Mobile, Wearable, and Ubiquitous Technologies, 8(1), pp.1-32.