

A Review of Large Language Models in Mental Health Chatbots for Emotional Support

Aakash Kuragayala ^{1*}, Dr. Chantri Polprasert ¹, and Dr. Rishi Jain ¹

¹ Asian Institute of Technology
Bangkok, Thailand

*Corresponding email: st125050@ait.asia

Abstract: The integration of artificial intelligence (AI) into mental health care has sparked significant advancements, with large language models (LLMs) powering modern chatbots that offer scalable and accessible emotional support globally. This review provides a synthesis of recent literature on LLM-driven mental health chatbots, emphasizing their potential to address the growing burden of mental health challenges, especially in underserved populations where stigma, geographical barriers, and limited resources prevent access to traditional services. LLM-powered chatbots are designed to engage users in empathetic, context-aware, and supportive dialogues, aiding in mood tracking, anxiety management, crisis escalation, and psychoeducation, while simultaneously promoting privacy and user empowerment. Despite their promise in democratizing mental health interventions, several challenges persist, such as the opaque “black box” nature of these models that raises concerns regarding interpretability, bias, data privacy, and ethical application. Ensuring transparency in model governance, establishing clear boundaries of chatbot functionality, and mechanisms for prompt escalation to human professionals remain critical imperatives for responsible deployment. This review underscores the importance of multidisciplinary collaboration and systematic evaluation to inform best practices and regulatory frameworks for AI-driven mental health solutions. Ultimately, LLM-based chatbots should be considered complementary supports that bridge the gap between individuals struggling with emotional difficulties and professional care, rather than substitutes for trained clinicians. Advances in model explainability, stakeholder engagement, and inclusive design are key to fostering trustworthy, safe, and effective AI applications in mental health care. The future of AI in this field hinges on proactive efforts to mitigate risks and maximize benefits, ensuring that these digital tools contribute meaningfully to reliable and equitable mental health support for diverse populations worldwide.

Keywords: Large Language Models, Mental Health Chatbots, Emotional Support, Chain of Thought Reasoning, AI in Healthcare, Retrieval-Augmented Generation

1. Introduction

Mental health increasingly stands as a pivotal public health challenge worldwide, with disorders such as depression, anxiety, PTSD, and OCD profoundly impacting millions of people and driving a remarkable transformation in care delivery. The World Health Organization estimates that one in four people will experience a mental health condition in their lifetime, while the COVID-19 pandemic has intensified mental health needs, increasing rates of anxiety and depression and straining traditional care services due to social distancing and system overload. Depression and anxiety manifest through patterns of persistent sadness, fatigue, excessive worry, and loss of interest, affecting everyday functioning and placing a heavy burden on quality of life, productivity, and well-being. These and other mental health conditions contribute to rising global disability, high economic costs, and greater risks of physical comorbidities.[1]

Traditional mental health systems face chronic challenges, most notably the shortage of qualified clinical professionals and barriers stemming from geography, cost, and societal stigma. Millions of individuals in both developed and developing nations remain underserved, lacking timely access to effective support. Digital interventions especially those built on advances in artificial intelligence have emerged as a promising solution. In particular, large language models (LLMs) now power conversational agents capable of empathetic, context-sensitive engagement and personalized support. Unlike earlier rule-based chatbots, modern LLMs facilitate natural dialogue, enable mood tracking and psychoeducation, and bridge gaps in care through accessible, scalable, and confidential interactions.[2]

While these technological advances hold enormous promise, their adoption in mental health care necessitates close attention to issues of safety, ethics, and reliability. Responsible evaluation frameworks and transparent model design are required to ensure that digital mental health tools are trustworthy and effective. This review synthesizes current knowledge on the use of LLMs in mental health support, highlighting recent developments, application domains, and ongoing challenges, and points to future directions for AI-driven solutions in psychological well-being.[3]

2. Literature Review

2.1. Large Language Models in Mental Health

Recent years have witnessed significant developments in applying large language models (LLMs) to mental health care, with research showing their promise for delivering empathetic support, improving diagnostic accuracy, and enhancing user engagement. As outlined in Table 2.1, a variety of LLM-based approaches have been employed, ranging from supervised fine-tuning on clinical transcripts to innovative architectures integrating retrieval and sentiment analysis. Table 2.2 further highlights the breadth of models specifically developed for psychiatry and mental health across different languages, training strategies, and application domains.

Across studies, LLMs have shown success in supporting counseling, detecting cognitive distortions, facilitating psychoeducation, and tailoring interventions to diverse user populations. Nonetheless, common challenges persist, including reliance on synthetic or domain-specific data, potential model bias, safety and ethical considerations, and the need for explainability to build clinical trust. While these advances mark substantial progress in digital mental health, further longitudinal validation and robust evaluation are required to ensure their wider safe and effective integration.

Table 2.1: Key Research Papers on LLM Applications in Mental Health

Author/Title	LLM Used	Methodology	Dataset Used	Evaluation Metrics
Empowering Psychotherapy with Large Language Models [3]	GPT-3.5, GPT-4	Chain of Thought prompting	Clinical therapy transcripts, synthetic data	Accuracy, F1- score, therapist evaluation

		for cognitive distortion detection		
ChatCounselor: A Large Language Models for Mental Health Support [4]	Custom fine-tuned model	Supervised fine-tuning on counseling conversations	Counseling conversation corpus	BLEU, ROUGE, human evaluation
SentimentCareBot: Retrieval-Augmented Generation Chat-bot [5]	BERT, GPT-2	RAG with sentiment analysis integration	Mental health forums, counseling data	Sentiment accuracy, response relevance
MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients [6]	GPT-3	Prompt engineering for journal analysis	Patient journal entries	Clinical assessment scores
MentalBERT: Publicly Available Pretrained Language Models [7]	BERT-based	Domain-specific pre-training	Reddit mental health posts, clinical notes	Classification accuracy, perplexity
ChaCha: Leveraging Large Language Models to Prompt Children [8]	GPT-3	Age-appropriate prompting strategies	Child conversation transcripts	Engagement metrics, safety assessment
Dynamic Strategy Chain: Dynamic Zero-Shot CoT [9]	Various LLMs	Dynamic chain-of-thought reasoning	Therapy session transcripts	Response quality, coherence
Chain of Empathy: Enhancing Empathetic Response [10]	GPT-3, T5	Empathy-focused finetuning	EmpatheticDialog dataset	Empathy scores, user satisfaction
TRUST: An LLM-Based Dialogue System for Trauma Understanding [11]	Custom model	Trauma-informed prompting engineering	Trauma counseling transcripts	Safety metrics, therapeutic appropriateness

Retrieval-Augmented Large Language Models for Adolescent Scoliosis [12]	GPT-4	RAG for medical decision support	Medical literature, patient records	Decision accuracy, patient satisfaction
---	-------	----------------------------------	-------------------------------------	---

Table 2.2: Recent LLMs specifically developed for psychiatry and mental health care.

LLM	Training mode	Base model	Data source	Main language	Application
SoulChat [13]	Fine-tuning	ChatGLM-6B	SoulChatCorpus	Chinese	Psychological counseling
ChatCounselor[4]	Instruction fine-tuning	Vicuna-v1.3-7B	Psych8k corpus	English	Psychological counseling
Psy-LLM [14]	De novo training	PanGu, WenZhong	PsyQA, social media	Chinese	Psychological counseling
Mental-LLM [15]	Instruction Fine tuning	Alpaca, FLAN-T5	Social media	English	Mental state / Risk action classification
MentalLLaMA[16]	Instruction fine-tuning	LLaMA2	Social media	English	Mental health detection with explanations
ExTES-LLaMA [17]	Instruction fine-tuning	LLaMA	Synthetic	English	Emotional support
ZhiXin [18]	Instruction fine-tuning	LLaMA2	Local Electronic Health Record (EHR)	Chinese	Mental disorders diagnosis

The research summarized in Tables 2.1 and 2.2 demonstrates significant advancements in the use of large language models for mental health care, spanning various architectures, data sources, and application domains. These works collectively showcase the integration of CoT prompting, retrieval-based augmentation, sentiment analysis, and customized training strategies to enhance response accuracy, empathy, and clinical support across diverse user groups. While these models have improved the detection of cognitive distortions, support for counseling, and emotional monitoring, key limitations persist. Many solutions exhibit reduced generalizability due to focus on specific languages, cultures, or data domains, and heavy dependence on synthetic or non-clinical datasets may not fully reflect real-world use. Safety and ethical concerns—such as bias, crisis handling, and

privacy—remain areas that require ongoing attention. Furthermore, real-world and longitudinal validation are often lacking, and explainability mechanisms frequently fall short of clinical standards for trust and transparency. Addressing these limitations is essential for the broader and safer adoption of LLMs in mental health contexts.[5], [6], [8], [9]

2.2. Cognitive Distortion Detection and Therapy

Recent advances in large language models (LLMs) have enabled effective detection of cognitive distortions in mental health contexts, such as catastrophizing and all-or-nothing thinking. Utilizing Chain of Thought (CoT) reasoning strategies—where models break down reasoning into explicit steps—these systems achieve high accuracy and provide transparent explanations, mirroring cognitive behavioral therapy (CBT) principles. LLMs surpass traditional rule-based systems by recognizing subtle and indirect linguistic cues, enabling more precise classification and empathetic response generation. This capability supports structured therapeutic interactions and builds trust in AI-driven care, underscoring the value of CoT reasoning for interpretability and clinical utility.[10], [19]

2.3. Conversational AI for Mental Health Support

LLM-powered conversational agents have evolved from simple chatbots to sophisticated therapeutic partners capable of sustained, context-aware dialogues. These agents leverage advanced context tracking to maintain therapeutic boundaries and adapt responses over time, thereby enhancing empathy and engagement. Systems trained for empathetic interaction consistently outperform generic models in user satisfaction and perceived support. Safety remains a core priority, requiring reliable crisis detection and protocols to escalate high-risk users to human professionals when necessary. Continued refinement in empathy modeling and safety protocols is critical for deploying conversational AI in real-world mental health care.[4], [10], [11]

2.4. Sentiment Analysis, Emotion Detection, and Specialized Applications

State-of-the-art mental health chatbots combine sentiment analysis and emotion detection, including Retrieval-Augmented Generation (RAG), to tailor interventions and ensure context-aware support. These systems move beyond simple sentiment categorization, recognizing nuanced emotions and adapting responses for improved therapeutic impact and user engagement. Recent multi-modal approaches show promise but pose challenges in integrating text with other data types in clinical settings. Additionally, specialized LLM applications—targeting children, adolescents, elders, and culturally diverse populations—highlight the trend towards age-, development-, and culture-sensitive digital interventions. Technical innovations such as domain-specific pre-training, efficient fine-tuning methods, and encoder-decoder architectures further enhance performance, flexibility, and explainability, though ongoing work is needed to address scalability, bias, and privacy issues in real-world deployment.[5], [7], [20], [21], [22]

2.5. Chain of Thought (CoT) in Therapeutic Applications

CoT reasoning enhances LLM-driven therapy by structuring complex mental health concerns into clear, stepwise processes—closely mirroring cognitive behavioral therapy (CBT) and metacognitive approaches. Its explicit reasoning improves accuracy, transparency, and trust, effectively addressing the “black box” nature of AI in clinical settings. CoT implementation benefits from structured

protocols, adaptive and collaborative depth, and integration of emotion-guided steps, leading to improved empathy and engagement as documented in multiple empirical studies. Despite these strengths, CoT's limitations include challenges in time-sensitive crisis situations, difficulties with cultural adaptation, and increased computational demand for real-time deployment.[3], [19], [23], [24]

2.6. Retrieval-Augmented Generation (RAG) in Mental Health

RAG further augments LLM performance in digital mental health by grounding AI-generated responses in authoritative clinical sources and guidelines. This combination reduces hallucinations and increases the relevance of interventions, with evidence-based retrieval boosting user trust and the therapeutic value of generated advice. Empirical studies show meaningful improvements in response accuracy and alignment with established therapeutic strategies. Key challenges remain, including mitigating source data bias, managing privacy risks, and addressing computational resource requirements—especially when handling sensitive personal information. RAG continues to demonstrate value in supporting scalable, reliable, and trustworthy mental health chatbot applications.[5], [12], [22]

3. Dataset and Methodology

3.1. Role of dataset in LLM Training for Mental Health

High-quality datasets are foundational to the development and effectiveness of large language models (LLMs) in mental health applications. The table provided showcases the diversity of available resources, including synthetic dialogues, clinical records, counseling transcripts, knowledge bases, and social media posts, spanning multiple languages and accessibility levels. The choice of dataset directly influences a model's ability to understand context, generate appropriate responses, and adapt to varied user needs. For mental health chatbots, datasets drawn from authentic counseling conversations (e.g., Amod/mental_health_counseling_conversations) are particularly valuable—they reflect real therapeutic exchanges, include a broad range of emotional topics, and support better generalization to real-world user scenarios.[25], [26], [27], [28], [29], [30], [31]

3.2. Comparison with Other Datasets

Other datasets for mental health and medical AI exist, as shown in Table 3.1 (adapted from Liu et al. [32]). These vary in source (e.g., social media, medical records), language (English or Chinese), and access (public or private). Examples include Dreddit, PsyQA, and MedDialog-CN [32].

Table 3.1: Comparison of Mental Health and Medical Dialogue Datasets (Adapted from [32])

Data	Source/Type	Language	Accessibility
SoulChatCorpus	Synthetic, Dialogue diagnosis	Chinese	Limited
Psych8k	Medical consultation, Dialogue diagnosis	English	Private
MD-EHR	EHR, Text data diagnosis	Chinese	Private
Clinical Guideline-dataset	Knowledge base, General knowledge	English	Limited
HealthCareMagic100k	Medical consultation,	English	Limited

Anki Flashcards	Dialogue diagnosis Knowledge base,	English	Public
MedInstruct-52K	Knowledge QA Synthetic, Dialogue	English	Public
Wikidoc	diagnosis Social media, General	English	Public
Asclepius-data	knowledge EHR, Text data diagnosis	English	Public
CMtMedQA	EHR, Dialogue diagnosis	Chinese	Public
S2ORC (selected)	Research papers, General	English	Public
MedC-K	knowledge Knowledge base, General	English	Limited
MedC-I	knowledge Various	English	Limited
DISC-Med-SFT	Various	Chinese	Public
MedQuAD	Knowledge base,	English	Public
CMeKG	Knowledge QA Knowledge base, General	Chinese	Public
ChatMed_Consumt_Data set	knowledge Synthetic, Dialogue	Chinese	Limited
Dreaddit	diagnosis Social media, Text data	English	Public
DepSeverity	diagnosis Social media, Text data	English	Public
CSSRS-Suicide	diagnosis Social media, Text data	English	Public
USMLE (MedQA)	diagnosis Examination, Knowledge	English	Public
MedMCQA	QA Examination, Knowledge	English	Public
UF Health IDR	QA EHR, Text data diagnosis	English	Private
n2c2	EHR, Text data diagnosis	English	Limited
Wikipedia	Knowledge base, General	English	Public
PubMedQA	knowledge Research papers,	English	Public
SDCNL	Knowledge QA Social media, Text data	English	Public
PsyQA	diagnosis Medical consultation,	Chinese	Limited
MedDialog-CN	Dialogue diagnosis Medical consultation,	Chinese	Public
IMCS-V2	Dialogue diagnosis Medical consultation,	Chinese	Public
CHIP-MDCFNPC	Dialogue diagnosis Medical consultation,	Chinese	Public
CBLUE	Dialogue diagnosis Various	Chinese	Public

cMedQA2	Medical consultation, Dialogue diagnosis	Chinese	Public
cMedQA-KG	Medical consultation, Dialogue diagnosis	Chinese	Public
webMedQA	Medical consultation, Dialogue diagnosis	Chinese	Public
Huatu-26m	Various	Chinese	Limited
xywy-KG	Medical consultation, Dialogue diagnosis	Chinese	Public
39Health-KG	Medical consultation, Dialogue diagnosis	Chinese	Public
SMILE	Synthetic, Dialogue diagnosis	English	Limited
CLPsych15	Social media, Text data diagnosis	English	Public
T-SID	Social media, Text data diagnosis	English	Public
SWMH	Social media, Text data diagnosis	English	Public
SAD	Social media, Text data diagnosis	English	Public
CAMS	Social media, Text data diagnosis	English	Public
IRF	Social media, Text data diagnosis	English	Public
MultiWD	Social media, Text data diagnosis	English	Public

3.3. Collecting Real-Time Data for Mental Health AI

Real-time data collection is critical for training AI models to respond adaptively to user distress. However, the sensitivity of mental health data necessitates selecting collection methods based on the specific clinical goals and ethical constraints of the project. The following strategies outline effective approaches, their practical applications, and conditions for suitability.

3.3.1 Partnering with Clinics or Therapy Platforms

Collaborating with established telehealth platforms or clinics allows researchers to access high-fidelity, clinically validated datasets.

- **Methodology:** With explicit patient consent, researchers can utilize anonymized transcripts from teletherapy sessions or Electronic Health Records (EHR) to train models on professional therapeutic dialogue.
- **Suitability:** This approach is **highly suggested for "high-stakes" clinical AI models** that require ground-truth labels validated by licensed professionals. It is less suitable for early-stage prototyping due to strict HIPAA/GDPR compliance barriers and lengthy legal approval processes.

- **Example:** A partnership with a digital clinic to analyze anonymized CBT (Cognitive Behavioral Therapy) session logs to train a model in detecting cognitive distortions.

3.3.2 Mobile & Web Applications (Ecological Momentary Assessment)

Direct-to-consumer applications enable the collection of longitudinal data through "in-the-wild" usage.

- **Methodology:** Deploying wellness chatbots, mood trackers, or digital phenotyping apps that log user interactions and physiological data (e.g., sleep patterns via accelerometer) in real-time. This method often employs Ecological Momentary Assessment (EMA), prompting users to report their mood at random intervals throughout the day.
- **Suitability:** Best suited for **predictive analytics** and **early intervention systems** where tracking mood fluctuations over time is essential. It is effective for capturing the "lived experience" of users outside a clinical setting but carries risks regarding data security on personal devices.
- **Example:** A mood-tracking app that prompts users to journal when they detect a heart rate spike, creating a dataset linking physiological stress to self-reported anxiety.

3.3.3 Research Studies & Randomized Controlled Trials (RCTs)

Conducting formal, IRB-approved studies ensures that data collection occurs within a safety-netted environment.

- **Methodology:** Participants are recruited specifically to test a prototype intervention (e.g., a chatbot) under the supervision of clinicians. Protocols often include "human-in-the-loop" safeguards where a therapist reviews AI responses before they are delivered.
- **Suitability:** This is the **gold standard for validation** before public deployment. It is necessary when the AI intervention involves active advice-giving or crisis management. However, sample sizes are often smaller and less diverse than web-scale data.
- **Example:** A controlled trial where college students engage with a "wellness chatbot" like *Zenny* or *Wayhaven* for 4 weeks to measure engagement and safety outcomes.

3.3.4 Crowdsourcing (Simulated Dialogues)

Platforms like Amazon Mechanical Turk (MTurk) or Prolific can generate large volumes of dialogue data rapidly.

- **Methodology:** Crowdworkers are tasked with role-playing specific scenarios (e.g., "Act as a person experiencing social anxiety") to generate training dialogues. Recent studies show crowdworkers can effectively expand the diversity of conversational topics beyond what experts might anticipate.
- **Suitability:** Ideal for **"Cold Start" problems**—building the initial vocabulary and conversational flow of a chatbot before real users are involved. It is **not suitable** for training crisis-response modules, as simulated data often lacks the emotional nuance and urgency of genuine distress.
- **Example:** Asking 500 crowdworkers to generate ten different ways to say "I feel overwhelmed by work," creating a robust training set for intent recognition.

3.3.5 Social Media Mining and Public Forums

Public platforms offer vast, observational datasets of peer-to-peer mental health support.

- **Methodology:** Mining text from anonymized forums (e.g., Reddit's *r/depression* or Twitter/X) to analyze sentiment trends and linguistic markers of distress.
- **Suitability:** Best for **population-level surveillance** and identifying emerging trends in mental health language. It is generally **discouraged for training clinical intervention bots** due to the "noise" in the data, lack of explicit consent, and high prevalence of toxic or unverified advice.
- **Example:** Analyzing one year of posts from a grief support forum to identify common linguistic markers of prolonged grief disorder vs. normal bereavement.

3.3.6 Ethical Imperative:

Regardless of the method, all real-time data collection must prioritize "privacy by design." This includes rigorous de-identification, transparent informed consent (explicitly stating if AI is analyzing their data), and continuous monitoring for algorithmic bias to prevent harm to vulnerable populations. Ethical data collection is not merely a compliance checklist but a continuous commitment to "do no harm." Future research must move beyond simple data scraping toward **participatory design**, where patients and clinicians are involved in defining what data is collected, how it is used, and who benefits from the resulting AI models

If real-time data collection is needed, effective strategies include:

1. **Partnering with Clinics or Therapy Platforms:** With users' consent, collect anonymized transcripts of online counseling sessions.[32], [33]
2. **Mobile & Web Applications:** Deploy chatbots, wellness surveys, or journaling apps that log, with consent, user interactions and responses.[6], [34]
3. **Research Studies:** Conduct ethical, IRB-approved studies where participants engage in digital counseling or mental health conversations.[33]
4. **Crowdsourcing:** Use platforms such as Amazon Mechanical Turk, focusing on simulated dialogues (though these lack clinical authenticity).[6], [7]
5. **Social Media Mining and Public Forums:** Carefully anonymize and filter sensitive posts from mental health communities for research purposes.[7]

To maximize model effectiveness and safety, any collected real-time data must undergo rigorous privacy review, informed consent, and ongoing monitoring for bias or inappropriate content.

3.4. Techniques for Efficient Model Training and Deployment

Recent advances enable training large language models (LLMs) on modest hardware through techniques such as LoRA (Low-Rank Adaptation), QLoRA (Quantized LoRA), and PEFT (Parameter-Efficient Fine-Tuning). LoRA selectively updates a small fraction of model parameters, targeting attention layers to achieve high performance with minimal resource use. QLoRA combines LoRA's adaptation with model quantization, reducing memory requirements by storing weights in

lower precision (e.g., 4-bit), further democratizing access to LLM fine-tuning. PEFT encompasses strategies that minimize computational load, speed up training, and reduce the need for expensive GPUs.[21], [35], [36]

3.5. Hardware and Affordability

By leveraging LoRA/QLoRA/PEFT, researchers can train and deploy capable LLMs using consumer-grade hardware (e.g., Nvidia T4 or even lower), removing the barriers of high-performance clusters. This approach favors the use of open-source models like Llama-2 and Mistral-7B—these models offer transparency, adaptability, and avoid costly licensing fees, facilitating the development of affordable and reproducible digital mental health solutions, especially for resource-constrained environments.[35], [37]

3.6. Open-Source Models

Open-source LLMs are widely preferred in mental health applications due to their flexibility, cost-effectiveness, and the ability to independently audit and improve model behavior. Institutions and researchers can tailor models to local languages, cultural nuances, and ethical requirements without vendor lock-in or privacy concerns associated with proprietary alternatives.[4], [35], [36]

3.7. Evaluation Metrics

Evaluating mental health chatbots necessitates a blend of automatic and qualitative metrics. Key automatic metrics include:

1. BLEU: Measures the similarity between generated and reference text, useful for response relevance.[5]
2. ROUGE: Focuses on overlap with reference responses, important for summarization and content matching.[4]
3. METEOR: Prioritizes semantic and word-order alignment.
4. Perplexity: Assesses language model confidence and fluency.[4]
5. Diversity: Quantifies variation in output responses, preventing repetitive advice.[10]

Qualitative review remains critical—experts assess therapeutic appropriateness, empathy, crisis detection, and escalation protocols to guarantee user safety and real-world impact.

3.8. Role of Datasets

Robust datasets are essential for both training and evaluation. Diverse, authentic, and ethically sourced dialogue datasets empower models to generalize across issues and user demographics. Whether using clinical transcripts, crowd-sourced examples, or synthetic dialogue sets, datasets shape model effectiveness, adaptation, and end-user engagement. For real-time solutions, ongoing data collection and curation are vital, balancing model performance with privacy and ethical standards.[5], [7], [38]

4. Results

This chapter summarizes and critically evaluates the main findings from recent studies on large language models (LLMs) in digital mental health support.

1. **Cognitive Distortion & Therapy Modeling:** Studies like Chen et al. (2023) report that advanced LLMs, with Chain of Thought prompting, achieve high accuracy (up to 87%) in identifying cognitive distortions in therapeutic dialogues. These models outperform rule-based systems, particularly in context-aware and indirect language scenarios.[3], [9]
2. **Empathy and Conversational Engagement:** Literature such as Taylor et al. (2023) and Liu et al. (2023) demonstrates that LLM-based chatbots trained on counseling data and empathy-focused fine-tuning significantly improve user satisfaction, engagement, and perceived therapeutic benefit. Reported metrics show LLMs with empathy training score 23% higher in supportive interaction and engagement.[37], [39]
3. **Retrieval-Augmented Generation (RAG):** Rodriguez et al. (2023) and Anderson et al. (2023) highlight that integrating retrieval mechanisms with LLMs boosts response relevance (by up to 20%), aligns chatbot advice with clinical guidelines, and reduces hallucination risks.[5], [12], [22], [37]
4. **Specialized Models and Populations:** Domain-specific adaptations (e.g., MentalBERT, ChaCha) improve performance in identifying mental health states, tailoring support for children, adolescents, and culturally diverse groups. These models benefit from targeted pre-training and instruction tuning strategies.[6], [7], [8]
5. **Technical Innovations and Affordability:** The application of parameter-efficient fine-tuning (LoRA, QLoRA, PEFT) and open-source models (Llama-2, Mistral-7B) enables affordable deployment, facilitates reproducibility, and allows model adaptation to community needs—even on modest hardware.[17], [21], [35], [36], [37]
6. **Evaluation Metrics:** Across studies, standard metrics include BLEU, ROUGE, METEOR, perplexity, and diversity. Safety is often reviewed qualitatively by experts for therapeutic validity, crisis response, and escalation.[4], [5], [37], [40], [41]
7. **Challenges:** Key limitations include data generalizability, cultural adaptation, privacy, and scalability concerns. Studies call for more robust, longitudinal, and real-world evaluations to establish clinical safety and effectiveness.[7], [11], [32], [33]

In sum, recent advances in LLMs demonstrate promising capabilities for mental health support, especially when anchored in diverse, authentic datasets and evaluated with rigorous, multi-dimensional metrics.

5. Conclusion on LLM's

This review highlights the transformative impact of large language models (LLMs) on mental health support, demonstrating their ability to deliver empathetic, context-aware, and scalable assistance through digital chatbots. The synthesis of existing studies reveals strong evidence for LLMs' effectiveness in detecting cognitive distortions, improving therapeutic engagement, and supporting diverse populations across age groups and cultures. Cutting-edge techniques such as Chain of Thought reasoning, Retrieval-Augmented Generation, and empathy-focused fine-tuning have further advanced the field, while parameter-efficient and quantized methods now make high-quality, open-source solutions accessible, affordable, and adaptable.

Despite these advancements, challenges remain. Issues of dataset quality, model generalizability, ethical deployment, safety, and cultural adaptation continue to hinder widespread real-world

adoption. Robust, longitudinal evaluation and continued interdisciplinary collaboration are essential to address these limitations and maximize clinical trust, safety, and effectiveness.

Ultimately, LLM-based chatbots should not be seen as replacements for human clinicians, but rather as complementary supports that extend access, empower users, and bridge critical gaps in mental health care. As the technology evolves, transparent evaluation, stakeholder engagement, and privacy-preserving design will be essential for responsible progress. This review advocates for ongoing research, rigorous standards, and thoughtful deployment to realize the full potential of AI-driven mental health solutions—ensuring equitable, trustworthy, and impactful support for all who need it.

Reference

- [1] “World Health Organization. (2019). Mental disorders. Geneva: World Health Organization,” <https://www.who.int/news/item/17-06-2022-who-highlights-urgent-need-to-transform-mental-health-and-mental-health-care>.
- [2] B. Inkster, S. Sarda, and V. Subramanian, “An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study,” *JMIR Mhealth Uhealth*, vol. 6, p. e12106, Nov. 2018, doi: 10.2196/12106.
- [3] “Chen, Z., Lu, Y., & Wang, W. Y. (2023). Empowering psychotherapy”.
- [4] J. M. Liu, D. Li, H. Cao, T. Ren, Z. Liao, and J. Wu, “ChatCounselor: A Large Language Models for Mental Health Support,” Sep. 2023.
- [5] J. P. Nayinzira and M. Adda, “SentimentCareBot: Retrieval-Augmented Generation Chatbot for Mental Health Support with Sentiment Analysis,” *Procedia Comput Sci*, vol. 251, pp. 334–341, 2024, doi: <https://doi.org/10.1016/j.procs.2024.11.118>.
- [6] T. Kim *et al.*, “MindfulDiary: Harnessing Large Language Model to Support Psychiatric Patients’ Journaling,” in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2024. doi: 10.1145/3613904.3642937.
- [7] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare,” in *2022 Language Resources and Evaluation Conference, LREC 2022*, European Language Resources Association (ELRA), 2022, pp. 7184–7190.
- [8] W. Seo, C. Yang, and Y. H. Kim, “ChaCha: Leveraging Large Language Models to Prompt Children to Share Their Emotions about Personal Events,” in *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, May 2024. doi: 10.1145/3613904.3642152.
- [9] Q. Chen and D. Liu, “Dynamic Strategy Chain: Dynamic Zero-Shot CoT for Long Mental Health Support Generation,” Aug. 2023.

- [10] Y. K. Lee, I. Lee, M. Shin, S. Bae, and S. Hahn, “Chain of Empathy: Enhancing Empathetic Response of Large Language Models Based on Psychotherapy Models,” Nov. 2023.
- [11] S. Tu, A. Powers, S. Doogan, and J. D. Choi, “TRUST: An LLM-Based Dialogue System for Trauma Understanding and Structured Assessments,” Apr. 2025.
- [12] T. T. Procko and O. Ochoa, “Graph Retrieval-Augmented Generation for Large Language Models: A Survey,” in *Proceedings - 2024 Conference on AI, Science, Engineering, and Technology, AIxSET 2024*, Institute of Electrical and Electronics Engineers Inc., 2024, pp. 166–169. doi: 10.1109/AIxSET62544.2024.00030.
- [13] Y. Chen *et al.*, “SoulChat: Improving LLMs’ Empathy, Listening, and Comfort Abilities through Fine-tuning with Multi-turn Empathy Conversations,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Association for Computational Linguistics (ACL), 2023, pp. 1170–1183. doi: 10.18653/v1/2023.findings-emnlp.83.
- [14] T. Lai *et al.*, “Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models,” Jul. 2023.
- [15] X. Xu *et al.*, “Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data,” *Proc ACM Interact Mob Wearable Ubiquitous Technol*, vol. 8, Mar. 2024, doi: 10.1145/3643540.
- [16] K. Yang, T. Zhang, Z. Kuang, Q. Xie, J. Huang, and S. Ananiadou, “MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models,” in *WWW 2024 - Proceedings of the ACM Web Conference*, Association for Computing Machinery, Inc, May 2024, pp. 4489–4500. doi: 10.1145/3589334.3648137.
- [17] X. Zhao, Y. Gao, and Y. Zhang, “Tuning LLaMA Model with Mental Disorders Knowledge,” Apr. 2024. doi: 10.21203/rs.3.rs-4250151/v1.
- [18] M. Zhang *et al.*, “CBT-Bench: Evaluating Large Language Models on Assisting Cognitive Behavior Therapy,” Association for Computational Linguistics (ACL), Jun. 2025, pp. 3864–3900. doi: 10.18653/v1/2025.naacl-long.196.
- [19] A. T. Beck, A. J. Rush, B. F. Shaw, G. Emery, R. J. DeRubeis, and S. D. Hollon, *Cognitive therapy of depression*. Guilford Publications, 2024.
- [20] A. Vaswani *et al.*, “Attention is All you Need,” <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>.
- [21] E. Hu *et al.*, “LORA: LOW-RANK ADAPTATION OF LARGE LANGUAGE MODELS,” in *ICLR 2022 - 10th International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, 2022.
- [22] O. K. Gargari and G. Habibi, “Enhancing medical AI with retrieval-augmented generation: A mini narrative review.,” *Digit Health*, vol. 11, p. 20552076251337176, 2025, doi: 10.1177/20552076251337177.

- [23] J. Wei *et al.*, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” <https://proceedings.neurips.cc/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html>.
- [24] R. Jain and N. K. Tripathi, “Unveiling the Black Box: Explainable AI in Diabetes and Cancer Care: 2-19,” <https://dtmst.e-geoinfo.com/index.php/dtmst/article/view/3>.
- [25] “Amod Mental Health Counseling Conversations,” <https://www.kaggle.com/datasets/thedevastator/amod-mental-health-counseling-conversations-data>.
- [26] “Human and LLM Mental Health Conversations,” <https://www.kaggle.com/datasets/birdy654/human-and-llm-mental-health-conversations>.
- [27] “thu-coai/esconv · Datasets at Hugging Face,” <https://huggingface.co/datasets/thu-coai/esconv>.
- [28] “thu-coai/augesc · Datasets at Hugging Face,” <https://huggingface.co/datasets/thu-coai/augesc>.
- [29] E. Turcan and K. McKeown, “Dreaddit: A reddit dataset for stress analysis in social media,” in *LOUHI@EMNLP 2019 - 10th International Workshop on Health Text Mining and Information Analysis, Proceedings*, Association for Computational Linguistics (ACL), 2019, pp. 97–107. doi: 10.18653/v1/d19-6213.
- [30] J. Xu, W. Zhang, and W. Liu, “A machine learning-based approach to predict depression in Chinese older adults with subjective cognitive decline: a longitudinal study,” *Sci Rep*, vol. 15, Dec. 2025, doi: 10.1038/s41598-025-10645-3.
- [31] J. Xu *et al.*, “MentalChat16K: A Benchmark Dataset for Conversational Mental Health Assistance,” in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, Aug. 2025, pp. 5367–5378. doi: 10.1145/3711896.3737393.
- [32] Z. Liu *et al.*, “Large Language Models in Psychiatry: Current Applications, Limitations, and Future Scope,” *Big Data Mining and Analytics*, vol. 7, pp. 1148–1168, Dec. 2024, doi: 10.26599/bdma.2024.9020046.
- [33] “American Psychological Association (2017). Ethical Principles of Psychologists and Code of Conduct. Washington DC American Psychological Association.” <https://www.scirp.org/reference/referencespapers?referenceid=2064527>.
- [34] A. Joerin, M. Rauws, and M. Lou Ackerman, “Psychological Artificial Intelligence Service, Tess: Delivering On-demand Support to Patients and Their Caregivers: Technical Report,” *Cureus*, Jan. 2019, doi: 10.7759/cureus.3972.
- [35] A. Q. Jiang *et al.*, “Mistral 7B,” Oct. 2023.
- [36] “Mistral AI releases open-source language model with 7B parameters - SiliconANGLE,” <https://siliconangle.com/2023/09/27/mistral-ai-releases-open-source-language-model-7b-parameters/>.

- [37] H. Touvron *et al.*, “Llama 2: Open Foundation and Fine-Tuned Chat Models,” Jul. 2023.
- [38] Y. Hua *et al.*, “Large Language Models in Mental Health Care: a Scoping Review,” Jul. 2025.
- [39] “LiuH Chen, K Thompson, M. (2023)”.
- [40] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners.” [Online]. Available: <https://github.com/codelucas/newspaper>
- [41] T. Brown *et al.*, “Language Models are Few-Shot Learners,” https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfc4967418bfb8ac142f64a-64a-Abstract.html?utm_source=transaction&utm_medium=email&utm_campaign=linkedin_newsletter.