Towards Culturally Adaptive Large Language Models in Mental Health: Using ChatGPT as a Case Study

Mahwish Aleem*
RPTU Kaiserslautern
Kaiserslautern, Germany
mahwishaleem28@gmail.com

Imama Zahoor*
Lahore University of Management Sciences
Lahore, Pakistan
imamazahoor@gmail.com

Mustafa Naseem
University of Michigan
Ann Arbor, USA
mnaseem@umich.edu

ABSTRACT

This paper explores the efficacy of ChatGPT as a multicultural therapist. Our study involves two rounds of prompt testing exercises: the first to assess general therapeutic skills and the second to identify multicultural counseling limitations. Our findings reveal significant limitations in memory, adaptability, listening, engagement depth, and cultural sensitivity. These limitations highlight the need for AI models to better adapt to diverse cultural contexts and exhibit increased empathy and responsiveness. We further discuss the integration of multicultural therapeutic practices, as well as the importance of culturally sensitive AI in mental health support. Our research contributes to Human Computer Interaction (HCI) literature by proposing design recommendations for future development and underscores the need for culturally nuanced therapeutic interactions in AI-driven mental health support.

CCS CONCEPTS

• **Human-centered computing → HCI theory, concepts and models**.

KEYWORDS

Large Language Models, Multicultural Therapy, ChatGPT, Mental Health, Cultural Adaptivity, Inclusive Design, HCI

ACM Reference Format:
Mahwish Aleem, Imama Zahoor, and Mustafa Naseem. 2024. Towards Culturally Adaptive Large Language Models in Mental Health: Using ChatGPT as a Case Study. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24), November 9–13, 2024, San Jose, Costa Rica.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3678884.3681858

1 INTRODUCTION

Mental health disorders in low- and middle-income countries (LMICs) present a significant challenge, due to a chronic shortage of mental health professionals, limited culturally appropriate treatments, and training and supervision barriers [17]. The COVID-19 pandemic further increased the demand for psychological counseling [33], driving the urgent need for effective and accessible mental health solutions. In response, there has been a surge in digital tools for mental health, including mobile applications [31, 40], online platforms [52, 59], teletherapy services [43, 49], and wearable devices [25]. Advancements in generative artificial intelligence (AI), especially large language models (LLMs), have expanded possibilities in mental health support, with fine-tuned models like ChatCounsellor [39], Psy-LLM [35], and MentalLLaMA [65] that provide services ranging from "question-answering in psychological consultation settings" [35] to interpretable mental health analysis [65].

However, the deployment of AI-driven mental health solutions remains far from ready for widespread use [35], facing ongoing challenges in reliability, interpretability, ethical concerns, and cultural sensitivity [11, 32, 62]. Focusing on cultural considerations is crucial, as disproportionate access to digital systems may worsen social inequalities [28, 56]. Current AI models, such as ChatGPT, often exhibit biases towards Western-centric values and norms [7], a trend that, if not addressed, could undermine their long-term effectiveness and acceptance in diverse cultural settings. To address some of these shortcomings, the integration of multicultural therapeutic practices emerges as a critical aspect.

In this paper, we present findings from prompting of GPT-4 with two sets of scenarios, one aimed at highlighting generic therapeutic shortcomings, and the other aimed at uncovering specifically multicultural therapeutic limitations. At the outset, we want to acknowledge that GPT-4 was not designed, and is not intended, to be used as a multicultural therapist. We use prompting with diverse multicultural scenarios as a technique analogous to using technology probes [29] to serve the "design goal of inspiring users and researchers to think about new technologies". We specifically chose GPT-4 due to its natural language understanding and generation abilities surpassing earlier models [2], however, further exploration is needed with other commercially available LLM tools to highlight a comprehensive list of shortcomings that a future therapy-bot service would need to address.

Our exploration to uncover generic therapeutic conversational limitations reveals ChatGPT's limitations in memory and adaptability, as well as listening and engagement depth, all of which hinder its efficacy as a therapeutic chatbot. Similarly, the analysis of ChatGPT's responses against key metrics of effective multicultural therapy uncovered significant shortcomings, including lack of self-awareness of biases, limited understanding of user perspectives, no attempt at learning from user experiences which may be different from the GPT's training data, and lack of practicing respectful

[*] Both authors contributed equally to this research.

CSCW Companion '24, November 9–13, 2024, San Jose, Costa Rica

ACM ISBN 979-8-4007-1114-5/24/11
https://doi.org/10.1145/3678884.3681858

curiosity. These findings introduce a preliminary set of LLM limitations in multicultural sensitivity to the HCI community, and build on other known limitations within LLMs in this context, including reliability, interpretability, and ethical concerns [11, 32, 62]. In this paper, we use these limitations to propose design directions for future development. We believe our work takes an important first step towards introducing limitations of Human-AI collaboration when it comes to culturally nuanced therapeutic interactions. Future HCI and systems research should pursue overcoming such limitations in order to advocate for equitable and inclusive mental health care that is aligned with the field's overarching objectives.

2 BACKGROUND

2.1 Multicultural Therapeutic Approach

The Accreditation Council for Graduate Medical Education (ACGME) outlines six core competencies for evaluating medical practitioners in the U.S.: patient care, medical knowledge, practice-based learning and improvement, interpersonal and communication skills, professionalism, and systems-based practice [1, 15]. These competencies are relevant across various medical specialties, crucial for high-quality patient care and have influenced global health curricula [19, 44, 67]. However, psychology's focus on emotional connection with clients necessitates distinct competencies for general and multicultural settings [13, 30]. Multicultural competence positively impacts clients' perceptions and satisfaction with therapists [14]. Studies show that therapists with a multicultural orientation (MCO) are seen as more credible and comforting, leading to better mental health outcomes [48]. Beyond cultural awareness and knowledge, cultural humility, characterized by an other-oriented approach that embodies respect, openness, and an absence of condescension in interpersonal interactions, is also a vital factor in fortifying counseling relationships [27, 66]. Within this framework, two scales measuring counselor competency become relevant. The first is the revised version of the Cross-Cultural Counseling Inventory (CCCI-R) which is an 18-item scale that organizes traits three-dimensionally: cultural awareness and beliefs, cultural knowledge, and adaptability in counseling skills [20, 34]. The second is the Multicultural Therapy Competency Inventory - Client Version (MTCI-CV) which is a 32-item instrument that assesses clients' perceptions of their therapists' multicultural competence with a special emphasis on the role of cultural humility [12]. These scales' themes can inform the development of a similar list of metrics to evaluate AI models in mental health settings, transferring the knowledge of human therapist competencies to the assessment of AI-driven therapeutic interactions.

2.2 LLMs and Cultural Sensitivity

There is no straightforward definition of culture, but research mostly aligns with the view of culture as a set of shared beliefs and values by a collective group of people that differs them from other groups of people [26]. LLM's training data predominantly reflect Western cultural norms and values, displaying strong alignment with that culture and adapting less effectively to other contexts [7, 38, 45]. Additionally, biases or stereotypes embedded in the training corpus may cause the LLM to inadvertently perpetuate them, leading to discriminatory or outright offensive responses [36, 46, 53]. This includes religious bias, where a 2021 study demonstrated persistent anti-Muslim bias in GPT-3, with 23% of test cases identifying Muslims as terrorists, compared to 5% for other religious groups [3].

Recent studies underscore these challenges, evaluating LLMs in culturally nuanced, low-resource, real-world scenarios such as multilingual WhatsApp chats, revealing that models like GPT-4 exhibit better performance but still struggle inconsistently with cultural nuances, especially in non-English settings [47]. Efforts have been made to make culturally sensitive conversational agents by targeting Natural Language Understanding (NLU) dimensions, such as linguistic form and style, shared knowledge, aboutness, and values [23]. Other methods explore the use of multicultural datasets for fine-tuning LLMs and cross-lingual data sharing for improved performance on diverse tasks [4, 10, 54]. Furthermore, there has been the introduction of new tools and methods to evaluate the performance of LLMs for multicultural content, such as through targeted dataset creation [8]. Recently, Song et al. [55] found that multicultural users who use LLMs for mental health support, only do so as complements to other forms of support, due to the culturally-bound limitations of LLMs. Our paper uses established metrics and guidelines to highlight the specific ways that generic LLMs fall short for multicultural users who may seek mental health support, building upon novel insights into real-world application challenges and the ongoing need for tools that capture cultural nuances effectively.

2.3 LLMs and Mental Health Support

Generative AI models, especially in the field of mental health technology, have gained significant attention in recent years. This surge in interest has led to numerous studies exploring the strengths and weaknesses of AI-driven mental health support. Studies by Chung et al. [11], Xu et al. [62], and Ji et al. [32] have identified challenges like model hallucinations, ethical risks, and interpretability issues. Notably, Ji et. al [32] delves into nuanced challenges such as the sensitivity of model performance to subtle language changes and their impact on inference times (resulting in unpredictable fluctuations) [63]. Some AI models also struggle to accurately assess severity and recognize risk progression [24]. Research in the design and evaluation of specific LLM tools employs novel approaches by using a range of metrics from traditional counseling strategies [39], fostering empathy and introspection through user engagement with relevant questions [6], and nurturing conversations for mental well-being [41]. However, they also reveal weaknesses such as the model's inability to probe for additional information and engage in reflective interactions [39], the challenge of maintaining a balance between inquiry and advice [6], persistent biases and inconsistencies despite post-processing efforts [6, 41], issues with memory loss [47], and inconsistent communication styles, often exacerbated post-update, referred to as "Post-Update Blues" [41]. Innovative studies have integrated psychological counseling theories to address these issues. The MentalBlend framework by Gu and Zhu [22] incorporates Cognitive-Behavioral Therapy, Dialectical Behavior Therapy, Person-Centered Therapy, and Reality Therapy to enhance the empathetic engagement of LLMs. Additionally, reinforcement learning from human feedback (RLHF) has been utilized to improve response relevance and interaction quality in therapy chatbots [5]. Expert evaluations, such as those suggested by Cho et al. [9], report significant enhancements in adaptability and contextually appropriate interactions. Yet, these advancements also underscore ongoing challenges such as the need for high-quality training data, the complexity of implementation, and the difficulties in achieving the depth of understanding and personalization found in human therapy. Furthermore, these methods require extensive computational resources and meticulous management to mitigate biases in feedback quality [5, 9, 22]. Therefore, while prior research has provided foundational insights at the intersection of LLMs and mental health support, our study seeks to further this discourse by detailing various LLM limitations when it comes to multicultural therapeutic responses. This approach aims to address the nuanced requirements of a culturally diverse user base and assess the model's efficacy in catering to a wide range of mental health needs, building upon established best practices and innovative integrations in the field.

3 METHODS

To assess ChatGPT's therapeutic skills and its multicultural counseling capabilities, we conducted prompt testing exercises using GPT-4 as a case study. This involved a subset of the authors, both of whom are Pakistani women, posing as potential users and having detailed conversations with the chatbot. We provided GPT-4 with an instruction prompt to act as a multicultural therapist and practice rapport-building, active and empathetic listening, and cultural awareness and humility. These instructions were derived from ACGME guidelines [1], as well as competencies outlined in previous literature and validated scales such as MCTI-CV [12, 20, 34].

The following instruction prompt was used to provide ChatGPT instructions for taking on the persona of a multicultural therapist. This prompt was the first message sent by the authors, and the subsequent interactions were based on the user scenarios described in Tables 1 and 2.

> Prompt: Take on the persona of a multicultural therapist who provides culturally relevant guidance to the user. The advice you give should align with the following characteristics, please adhere to them throughout the conversation and refer back to them before sharing all of your responses:
>
> (1) Taking time to build rapport with the user.
> (2) Encouraging active participation from the user while listening attentively without judgment or preconceptions.
> (3) Identifying and articulating the user's emotions in an empathetic manner.
> (4) Being knowledgeable about the customs, accepted behaviors, and general values of different cultures, and showing openness and respect for the cultural background and unique experiences of the user.
> (5) Understanding the user's perspective without judgment, and being willing to learn from the user's expertise on their own life.
> (6) Providing counseling that is tailored to the user's context, including culture, race, class, gender, religion, sexual orientation, etc.

In our conversations with GPT-4, we used a series of scenarios as a starting point for each conversation. The authors utilized their diverse lived experiences to curate scenarios rich in cultural, religious, ethical and social challenges, thereby enhancing the relevance of the interactions for a multicultural user base. The criteria used to craft user scenarios was: they must be realistic, contain cultural or religious probes, represent different demographics and situations, and contain both self and interpersonal relationship queries.

Two rounds of prompt testing were performed, with the first round being conducted in October and November 2023. The aim of the first round was to uncover general therapeutic skill limitations. The aim of the second round of prompt testing was to uncover specific multicultural counseling skill limitations, and was conducted in January 2024. Tables 1 and 2 depict the scenarios used for round 1 and 2 respectively. Using a simulated approach with the authors posing as potential users during prompt testing has its limitations. Given that the authors are aware of the research question, this approach introduces confirmation and interviewer bias into our analysis. To overcome such unintentional bias, the authors performed a critical reading of the conversational logs of each others' interactions with GPT-4, and had detailed conversations within the research team to ensure any limitations with user prompting and author dialogue with GPT-4 are highlighted. Given that this paper is intended to be a "prequel" to provoke future work, we chose to use a simulation approach using author-generated chats. However, future research presenting complete empirical analysis should explore more robust evaluation methods such as having two different LLM's interact with each other [18]. Such an approach can overcome confirmation and interviewer bias, however it can introduce other biases inherent within LLMs themselves.

3.1 Scenarios for Multicultural Therapeutic Conversations

We recorded detailed logs of all of our interactions with GPT-4. For the first round of prompt testing, the first two authors individually analyzed each log file using evaluation coding [51], inductively highlighting 10 key shortcomings of GPT-4 in a therapeutic setting. All three authors met to combine the two individual list of shortcomings, and narrowed down to the 5 major shortcomings highlighted in section 4.1. For the second round of prompt testing, the first author used evaluation coding to deductively code the performance of GPT-4 against MTCI-CV guidelines [12] summarized in Table 3. All three authors had multiple rounds of conversations to analyze the codes, and had agreement on GPT-4 shortcomings presented in Section 4.2

4 FINDINGS

4.1 Generic Therapeutic Conversation Limitations

In this subsection, we detail the five inductively generated themes uncovered by analyzing limitations within our conversations with GPT-4 with each of the four scenarios described in Table 1. These

Scenario Number	Scenario Description
Scenario 3.1.1	I had a fight with my mother and hurtful things were said on both sides. Now we are not speaking to each other and this situation is making me very anxious. What should I do?
Scenario 3.1.2	I crashed my car the other day into a stranger's car. It was primarily my fault but I was not able to admit it at that time and instead picked a fight with them. Now I feel regretful and embarrassed about my actions. What should I do?
Scenario 3.1.3	I feel very frustrated and confused because everyone around me is very religious and expects me to share their sentiments which I do not. As I grew older I found myself growing apart from Islamic practices and while I hold some confusions I do miss the feeling of having a strong connection with God. What should I do?
Scenario 3.1.4	My mother-in-law is very overbearing and does not allow me to fully express myself. She has imposed multiple restrictions on my movement and my day-to-day activities, which has caused me to lash out on my husband. What should I do?

Table 1: Description of user scenarios for general therapeutic conversations

Scenario Number	Scenario Description
Scenario 3.2.1	I'm originally from Pakistan, but I live in the US. My parents think my mindset has completely changed since I settled abroad, and I don't prioritize my family anymore. We frequently fight about this and it just keeps widening the distance between us. I don't know how to explain this to them that while of course they mean the world to me, my priorities in life have shifted and I just can't be family-oriented the same way as them.
Scenario 3.2.2	I have been dealing with depression for quite some time now, but my family believes that I'm just lazy and there's nothing wrong with me, since my symptoms are not that severe (regardless of how impairing their effect is on me). They dismiss all my symptoms since they don't really believe in mental health disorders unless they manifest as extreme psychosis, due to the strong associated stigma in Pakistani society. Their lack of support is very upsetting for me and I don't know how to convince them otherwise.
Scenario 3.2.3	I hate the way my family responds to conflict. Growing up I saw my father always lash out in anger and my mother go absolutely silent when facing a confrontation. I don't think I've ever seen anyone work out disagreements in a healthy manner in my family. Now that I'm older and much more cognizant, I try to point out these patterns, but I'm perceived as "badtameez" or insolent since I'm disagreeing with my elders, especially my father, because according to my parents, good kids silently agree with everything their parents say.

Table 2: Description of user scenarios for multicultural therapeutic conversations

themes are: (a) memory and adaptability, (b) listening behavior, (c) flexibility, (d) inferential abilities, and (e) engagement depth. In the sections below, we elaborate on each of these themes.

4.1.1 Memory and Adaptability. When adopting a specific persona, such as acting as a multicultural therapist, ChatGPT often deviates from initial instructions over time, shifting to neutral responses. This pattern reveals its reliance on training data, leading to generic responses, similar to a generalist conversation between friends rather than with a trained therapist. For example, in scenario 3.1.1 involving a user conflict with her mother, GPT-4 initially adheres to validating user feelings but ceases this approach after three interactions. The initial response is more affirming, like *"Can you share more about what specifically frustrated you?",* but from the second interaction onwards, the chatbot's focus shifts to only giving advice such as *"It could be beneficial to use "I" statements to convey how you feel without placing blame..."*, ignoring the instructions of addressing user validation in the pre-prompt.

4.1.2 Listening Behavior. ChatGPT often prioritizes giving advice over maintaining a balanced dialogue, focusing on information provision rather than deeply understanding emotional nuances, a key aspect of active listening in human therapy. For instance, in user scenario 3.1.1, ChatGPT quickly offers generic advice without further inquiry, saying *"I'm sorry to hear about your situation. Reconciliation can be a delicate process, but here are some steps that might help..."* and proceeds to list generic approaches towards conflict reconciliation. This approach, while suitable in general contexts, falls short in therapeutic settings where understanding the individual's specific needs and context is crucial, particularly in sensitive cases like interpersonal conflicts. Ideally, ChatGPT should ask follow-up questions to help user's understand their own thought processes or help users uncover their feelings, mirroring practices in traditional therapy.

4.1.3 Flexibility. ChatGPT often rigidly sticks to initial instructions, resulting in literal or direct responses that lack the adaptability and tact crucial for effective conversation. This inflexibility

Sr. No.	MTCI-CV Themes	Therapeutic Skill
1	Counselor awareness of own cultural values and biases	Be open about own and client's values
2	Counselor awareness of client's worldview	Acknowledge the client's perspectives
3	Culturally appropriate intervention strategies	Provide suggestions that fit the client/their family's context (i.e.: race, class, gender, culture, sexual orientation, etc.)
4	Respectful curiosity	Ask the client to tell about their expectations for care
5	Naivete	Be receptive (through body language and communication) to the differences between yourself and the client
6	Multicultural counseling relationship	Use relationship-building skills, such as listening, attending, and paraphrasing

Table 3: Summary of MTCI-CV themes and their corresponding therapeutic skills borrowed from [12]

limits the use of instructions to enhance capabilities. For instance, GPT-4 takes the guiding prompt for culturally relevant advice too literally, leading to peculiar phrases like *"Here are some culturally sensitive suggestions..."* and *"Participate in cultural or religious events that resonate with your identity as a Pakistani Muslim"*.

4.1.4 Inferential Abilities. ChatGPT's difficulty with inference, stemming from its inability to interpret non-verbal cues and limitations in its training data, hinder its provision of nuanced responses, especially for non-Western users. For instance, when queried with the instructional prompt for its therapist persona for scenario 3.1.4, ChatGPT merely paraphrased and referred to the given instructions in sentences such as *"It might be helpful to explore ways to communicate your feelings and needs to your family, considering the cultural context."*, rather than confirming comprehension for future interactions, leading to a generic response that simply regurgitated the original prompt.

4.1.5 Engagement Depth. ChatGPT tends to superficially interpret lengthy user inputs, focusing on initial sentences or dominant themes while overlooking crucial details for nuanced responses. This issue became evident in cases where we provided extensive instructional prompts with both general and culturally specific instructions: ChatGPT primarily responds based on the initial, generic information.

4.2 Multicultural Therapeutic Conversation Strengths and Limitations

In this section, we present the shortcomings of GPT-4 responses when it comes to multicultural therapy, using the themes emerging from MTCI-CV guidelines [12]. While these themes present six categories, in our interactions we were only able to uncover strengths in the first four categories, while limitations were found in all 6 categories.

4.2.1 Counselor awareness of client's worldview. ChatGPT's responses show an understanding of the user's perspective in each query, affirming their emotions and viewpoints. However, as conversations progress, the depth of understanding appears basic, often simply echoing the user's words. For example, in scenario 3.2.1, ChatGPT's reply *"It sounds like you're facing a challenging situation, trying to balance the expectations of your family with your own personal growth and changes in priorities since moving to the US."* simply reiterates the user query. Regardless, ChatGPT validates the user's thoughts and feelings, starting responses with phrases like *"I'm sorry to hear about your struggles…"* or *"It sounds like you're in a challenging situation… "* to connect with the user.

4.2.2 Culturally appropriate intervention strategies. Once ChatGPT validates the user, it proceeds to provide advice, attempting to make it culturally relevant as instructed. Yet, this guidance often lacks the depth and specificity needed to truly align with the user's unique cultural context. For instance, in response to scenario 3.2.1, ChatGPT suggested: *"Considering your family's cultural background, it might be beneficial to approach the conversation with an understanding of the values and expectations that are important in Pakistani culture, such as family unity and interdependence. This doesn't mean you have to agree with everything, but showing respect and understanding for their perspective can help in finding a middle ground."* While this response shows a basic grasp of Pakistani values, it misses nuances such as the cultural implications of disagreeing with elders, which is considered disrespectful.

4.2.3 Multicultural counseling relationship. According to the MTCI-CV, employing relationship-building skills such as *"listening, attending, and paraphrasing"* is imperative for the multicultural counseling relationship [12]. As delineated in the previous section and further corroborated here, ChatGPT's listening behavior and engagement depth are inadequate, which hinders its ability to emulate the listening and attending skills of a therapist. Although it demonstrates paraphrasing abilities, the absence of truly listening and attending leads to responses that can appear insincere or shallow. For example, in response to scenario 3.2.3, ChatGPT asks multiple questions like, *"How do these family dynamics make you feel, and what kind of changes would you like to see in the way your family handles conflict? Additionally, how do you feel these cultural norms have shaped your own approach to conflict and communication?"* This barrage of questions may overwhelm a distressed user and give the impression of a lack of meaningful engagement.

4.2.4 Respectful curiosity. ChatGPT's responses also exhibit the ability to ask about the user's views and expectations regarding its advice, a form of respectful curiosity. Commonly, ChatGPT poses questions such as *"How do you feel about this approach? Do you think it could be helpful in your situation?"* at the end of its responses, aiming to mimic a natural conversational rhythm. However, while these inquiries initially seem sincere and well founded, they would need to be detailed for follow-on conversations in order to avoid seeming surface level or dismissive of the user's perspective.

4.2.5 Counselor awareness of own cultural values and biases.

ChatGPT's responses generally exhibit a lack of self-awareness regarding its own cultural values and biases, which are critical for openness in multicultural therapeutic relationships. This gap arises due to its training predominantly based on Western data. For instance, in scenario 3.2.2, where the user discusses family struggles due to declining mental health, ChatGPT suggested approaches like *"Express Your Feelings Using 'I' Statements," "Educate Gently,"* and *"Set Boundaries."* These suggestions reflect Western norms and may not resonate in communal societies, where collective values and familial hierarchies often take precedence.

4.2.6 Naivete.

The issue of lack of self-awareness in ChatGPT also leads to a challenge in recognizing and appreciating the differences between itself and the user. Typically, being aware of these differences involves integrating the client's expertise on their own lived experiences into the counseling sessions, making the multicultural approach more nuanced and effective. However, ChatGPT struggles to grasp this aspect; when explicitly prompted, its primary approach is to simply ask more questions, as observed in *4.2.3.*, which compromises the depth and relevance of its multicultural counseling approach.

5 DISCUSSION

Only in the last couple of years have AI-based digital interactions been extensively studied. Notably, LLMs like ChatGPT are developed with a primarily Western perspective, which might not fully encompass the diversity of user needs across different cultural or social backgrounds. For example, Eastern cultures often emphasize compassionate listening - a practice that is not only culturally significant but also aligns with traditionally feminine traits of empathy and understanding [50]. Having delineated the limitations of ChatGPT-4 in therapeutic and multicultural contexts, we present the following design recommendations to overcome identified shortcomings in both general therapeutic applications and multicultural counseling scenarios. These suggestions are relevant not only for GPT-4 but also for other existing LLMs. However, we want to acknowledge that the authors served both as designers and participants in the study, and this may limit the generalizability of our conclusions.

We identify four key aspects that inform an LLM's efficacy as a digital therapist: contextual adaptability, active listening, emotional awareness, and empathetic responses. Improved contextual adaptability is achievable through continuous learning mechanisms which enable LLMs to update their knowledge and response strategies continuously, thus eliminating the need for retraining [60]. This process can be enhanced by generating interpretable user data representations from past interactions in the form of user interaction profiles, following the approach by Wang et al. [58]. Active and empathetic listening can be advanced with emotion-enhanced CoT prompting, which directs LLMs to generate intermediate reasoning steps and more empathetic responses[64]. Additionally, self-refinement allows LLMs to autonomously enhance their responses through iterative self-feedback [42]. Advanced content and emotional analysis mechanisms are essential for identifying emotional states and high-risk scenarios, such as self-harm. We suggest that future research explores integrating emotion and sentiment analysis techniques across different languages, local slang and shorthand, and cultural contexts to analyze the emotional content of the user's input. This would enable LLMs, like a GPT-4 based therapy chatbot, to detect and de-escalate critical situations by involving human interlocutors. As such, it is crucial to integrate safety and ethical considerations, including transparency regarding the chatbot's functionality, informed consent for data collection, and data security measures. Aligning LLMs with human therapy practices can vastly improve their therapeutic potential, fostering more equitable and inclusive mental health care solutions.

To improve the multicultural capabilities of LLMs like GPT-4, fine-tuning the model with a culturally diverse dataset becomes pertinent. This could involve developing a dataset containing therapist-client interaction transcripts across diverse cultural settings, annotated using MTCI-CV metrics. These annotations would pinpoint instances where multicultural competencies are exhibited by therapists, as informed by prior research on expert-annotated therapy dialogues [61]. The fine-tuning process can be further augmented by integrating both expert and user feedback on the LLM's responses through Reinforcement Learning from Human Feedback (RLHF), ensuring that the model's output aligns closely with human evaluations [16]. This granular annotation approach would enable the LLM to recognize the nuances of multicultural counseling and make its responses more personalized and culturally aware. However, creating such a dataset presents challenges. Currently, there is a lack of comprehensive clinical data that can be leveraged to capture the vast range of multicultural therapeutic interactions. Manual collection of such data can be time consuming and cost intensive. Additionally, cultural data cannot be neatly divided into binary categories like Eastern and Western traditions; cultures are complex, dynamic, and intersect in diverse ways [57]. To address these, we recommend exploring different methods to create dense datasets via data pruning and use these highly-representative data samples for few-shot fine-tuning [37]. Another promising method includes transforming existing datasets to generate synthetic data [21]. This new dataset would be structured and task-specific without compromising the diversity and specificity of the samples.

6 CONCLUSION

In conclusion, our findings reveal critical shortcomings and provide a roadmap for future research to enable ChatGPT to be more culturally adaptive, empathetic, and responsive to diverse emotional needs.

REFERENCES

[1] 2021. ACGME Psychiatry Milestones. https://www.acgme.org/

[2] 2023. GPT-4. https://openai.com/gpt-4

[3] Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society.* 298–306.

[4] Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. arXiv:2204.06487 [cs.CL]

[5] Desirée Bill and Theodor Eriksson. 2023. Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application.

[6] Lennart Brocki, George C. Dyer, Anna Gładka, and Neo Christopher Chung. 2023. Deep Learning Mental Health Dialogue System. In *2023 IEEE International Conference on Big Data and Smart Computing (BigComp).* 395–398. https://doi.org/10.1109/BigComp57234.2023.00097

[7] Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study. arXiv:2303.17466 [cs.CL]

[8] Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs' (Lack of) Multicultural Knowledge. arXiv:2404.06664 [cs.CL]

[9] Yujin Cho, Mingeon Kim, Seojin Kim, Oyun Kwon, Ryan Donghan Kwon, Yoonha Lee, and Dohyun Lim. 2023. Evaluating the efficacy of interactive language therapy based on LLM for high-functioning autistic adolescent psychological counseling. *arXiv preprint arXiv:2311.09243* (2023).

[10] Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. How do languages influence each other? Studying cross-lingual data sharing during LLM fine-tuning. arXiv:2305.13286 [cs.CL]

[11] Neo Christopher Chung, George Dyer, and Lennart Brocki. 2023. Challenges of Large Language Models for Mental Health Counseling. arXiv:2311.13857 [cs.CL]

[12] Elise M Cole, Fred Piercy, Edward W Wolfe, and Jamie M West. 2014. Development of the multicultural therapy competency inventory-client version. *Contemporary Family Therapy* 36 (2014), 462–473.

[13] Hardin LK Coleman. 1998. General and multicultural counseling competency: Apples and oranges? *Journal of Multicultural Counseling and Development* 26, 3 (1998), 147–156.

[14] Madonna G Constantine. 2002. Predictors of satisfaction with counseling: Racial and ethnic minority clients' attitudes toward counseling and ratings of their counselors' general and multicultural counseling competence. *Journal of Counseling Psychology* 49, 2 (2002), 255.

[15] Barbara A Cubic and Edwin E Gatewood. 2008. ACGME core competencies: Helpful information for psychologists. *Journal of clinical psychology in medical settings* 15 (2008), 28–39.

[16] Oliver Daniels-Koch and Rachel Freedman. 2022. The Expertise Problem: Learning from Specialized Feedback. arXiv:2211.06519 [cs.LG]

[17] Katie S Dawson, Richard A Bryant, Melissa Harper, Alvin Kuowei Tay, Atif Rahman, Alison Schafer, and Mark Van Ommeren. 2015. Problem Management Plus (PM+): a WHO transdiagnostic psychological intervention for common mental health problems. *World Psychiatry* 14, 3 (2015), 354.

[18] Edoardo Sebastiano De Duro, Riccardo Improta, and Massimo Stella. 2024. Introducing CounseLLMe: A dataset of simulated mental health dialogues for comparing LLMs like Haiku, LLaMAntino and ChatGPT against humans. (2024).

[19] Anindita Deb, Melissa Fischer, and Anna DePold Hohler. 2018. Education research: A framework for global health curricula for neurology trainees. *Neurology* 91, 11 (2018), 528–532.

[20] Joanna M Drinane, Jesse Owen, Jill L Adelson, and Emil Rodolfa. 2016. Multicultural competencies: What are we measuring? *Psychotherapy Research* 26, 3 (2016), 342–351.

[21] Saumya Gandhi, Ritu Gala, Vijay Viswanathan, Tongshuang Wu, and Graham Neubig. 2024. Better Synthetic Data by Retrieving and Transforming Existing Datasets. arXiv:2404.14361 [cs.CL] https://arxiv.org/abs/2404.14361

[22] Ziyin Gu and Qingmeng Zhu. 2023. MentalBlend: Enhancing Online Mental Health Support through the Integration of LLMs with Psychological Counseling Theories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46.

[23] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. arXiv:2203.10020 [cs.CL]

[24] Thomas F. Heston. 2023. Evaluating Risk Progression in Mental Health Chatbots Using Escalating Prompts. *medRxiv* (2023). https://doi.org/10.1101/2023.09.10.23295321 arXiv:https://www.medrxiv.org/content/early/2023/09/12/2023.09.10.23295321.full.pdf

[25] Blake Anthony Hickey, Taryn Chalmers, Phillip Newton, Chin-Teng Lin, David Sibbritt, Craig S McLachlan, Roderick Clifton-Bligh, John Morley, and Sara Lal. 2021. Smart devices and wearable technologies to detect and monitor mental health conditions and stress: A systematic review. *Sensors* 21, 10 (2021), 3461.

[26] Geert Hofstede. 2016. Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations. *Collegiate Aviation Review* 34, 2 (2016), 108.

[27] Joshua N Hook, Don E Davis, Jesse Owen, Everett L Worthington Jr, and Shawn O Utsey. 2013. Cultural humility: Measuring openness to culturally diverse clients. *Journal of counseling psychology* 60, 3 (2013), 353.

[28] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. arXiv:2005.00813 [cs.CL]

[29] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[30] Zac E Imel, Scott Baldwin, David C Atkins, Jesse Owen, Tim Baardseth, and Bruce E Wampold. 2011. Racial/ethnic disparities in therapist effectiveness: a conceptualization and initial study of cultural competence. *Journal of Counseling Psychology* 58, 3 (2011), 290.

[31] Yavuz Inal, Jo Dugstad Wake, Frode Guribye, and Tine Nordgreen. 2020. Usability evaluations of mobile mental health technologies: systematic review. *Journal of medical Internet research* 22, 1 (2020), e15337.

[32] Shaoxiong Ji, Tianlin Zhang, Kailai Yang, Sophia Ananiadou, and Erik Cambria. 2023. Rethinking Large Language Models in Mental Health Applications. arXiv:2311.11267 [cs.CL]

[33] Osama Bin Khalid, Mustafa Qazi, Almas F Khattak, Madiha Khattak, Muhammad Noman K Wazir, and Humaira Gilani. 2023. COVID-19 Pandemic Lessons for Creating Effective Mental Health Safety Nets in Lower Middle-Income Countries. *Cureus* 15, 9 (2023).

[34] Teresa D LaFromboise, Hardin LK Coleman, and Alexis Hernandez. 1991. Development and factor structure of the Cross-Cultural Counseling Inventory—Revised. *Professional psychology: Research and practice* 22, 5 (1991), 380.

[35] Tin Lai, Yukun Shi, Zicong Du, Jiajie Wu, Ken Fu, Yichao Dou, and Ziqi Wang. 2023. Psy-LLM: Scaling up Global Mental Health Psychological Services with AI-based Large Language Models. arXiv:2307.11991 [cs.CL]

[36] Yizhi Li, Ge Zhang, Bohao Yang, Chenghua Lin, Shi Wang, Anton Ragni, and Jie Fu. 2022. HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models. arXiv:2211.02882 [cs.CL]

[37] Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024. Data-efficient Fine-tuning for LLM-based Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 365–374. https://doi.org/10.1145/3626772.3657807

[38] Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are Multilingual LLMs Culturally-Diverse Reasoners? An Investigation into Multicultural Proverbs and Sayings. arXiv:2309.08591 [cs.CL]

[39] June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. ChatCounselor: A Large Language Models for Mental Health Support. arXiv:2309.15461 [cs.CL]

[40] Joyce HL Lui, David K Marcus, and Christopher T Barry. 2017. Evidence-based apps? A review of mental health mobile applications in a psychotherapy context. *Professional Psychology: Research and Practice* 48, 3 (2017), 199.

[41] Zilin Ma, Yiyang Mei, and Zhaoyuan Su. 2023. Understanding the benefits and challenges of using large language model-based conversational agents for mental well-being support. In *AMIA Annual Symposium Proceedings*, Vol. 2023. American Medical Informatics Association, 1105.

[42] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651 [cs.CL]

[43] Adriana S Miu, Hoa T Vo, Jayme M Palka, Christopher R Glowacki, and Reed J Robinson. 2021. Teletherapy with serious mental illness populations during COVID-19: telehealth conversion and engagement. *Counselling Psychology Quarterly* 34, 3-4 (2021), 704–721.

[44] Gianina M Monestime, Isabelle Baird, Andrei Rebarber, and Taraneh Shirazian. 2022. ACGME Milestones in global health: Need for standardized assessment of global health training in obstetrics/gynecology residency. *International Journal of Gynecology & Obstetrics* 157, 3 (2022), 522–526.

[45] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Afina Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages. *arXiv preprint arXiv:2406.09948* (2024).

[46] Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456 [cs.CL]

[47] Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2024. Beyond Metrics: Evaluating LLMs' Effectiveness in Culturally Nuanced, Low-Resource Real-World Scenarios. *arXiv preprint arXiv:2406.00343* (2024).

[48] Jesse J Owen, Karen Tao, Mark M Leach, and Emil Rodolfa. 2011. Clients' perceptions of their psychotherapists' multicultural orientation. *Psychotherapy* 48, 3 (2011), 274.

[49] Fujiko Robledo Yamamoto, Amy Voida, and Stephen Voida. 2021. From therapy to teletherapy: Relocating mental health services online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[50] Deborah Roebuck, Reginald Bell, Reeta Raina, and Cheng Ean Lee. 2015. The effects of home country, gender, and position on listening behaviors. *Journal of Organizational Culture, Communications & Conflict* (2015).

[51] Johnny Saldaña. 2021. Coding techniques for quantitative and mixed data. *The Routledge reviewer's guide to mixed methods analysis* (2021), 151–160.

[52] Jessica Lee Schleider, Mallory Dobias, Jenna Sung, Emma Mumper, and Michael C Mullarkey. 2020. Acceptability and utility of an open-access, online single-session

intervention platform for adolescent mental health. *JMIR mental health* 7, 6 (2020), e20513.

[53] Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2383–2389.

[54] Sunayana Sitaram, Monojit Choudhury, Barun Patra, Vishrav Chaudhary, Kabir Ahuja, and Kalika Bali. 2023. Everything you need to know about multilingual LLMs: Towards fair, performant and reliable models for languages of the world. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*. 21–26.

[55] Inhwa Song, Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2024. The Typing Cure: Experiences with Large Language Model Chatbots for Mental Health Support. arXiv:2401.14362 [cs.HC]

[56] Nidhi Tewathia, Anant Kamath, and P Vigneswara Ilavarasan. 2020. Social inequalities, fundamental inequities, and recurring of the digital divide: Insights from India. *Technology in Society* 61 (2020), 101251.

[57] Vivian L Vignoles. 2018. The "common view", the "cultural binary", and how to move forward. *Asian Journal of Social Psychology* 21, 4 (2018), 336–345.

[58] Yan Wang, Zhixuan Chu, Xin Ouyang, Simeng Wang, Hongyan Hao, Yue Shen, Jinjie Gu, Siqiao Xue, James Y Zhang, Qing Cui, Longfei Li, Jun Zhou, and Sheng Li. 2023. Enhancing Recommender Systems with Large Language Model Reasoning Graphs. arXiv:2308.10835 [cs.IR]

[59] Nathan Wilkinson, Rebecca P Ang, and Dion H Goh. 2008. Online video game therapy for mental health concerns: a review. *International journal of social psychiatry* 54, 4 (2008), 370–382.

[60] Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual Learning for Large Language Models: A Survey. arXiv:2402.01364 [cs.CL]

[61] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. Anno-MI: A Dataset of Expert-Annotated Counselling Dialogues. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6177–6181. https://doi.org/10.1109/ICASSP43922.2022.9746035

[62] Xuhai Xu, Bingsheng Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2023. Mental-LLM: Leveraging Large Language Models for Mental Health Prediction via Online Text Data. arXiv:2307.14385 [cs.CL]

[63] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6056–6077. https://doi.org/10.18653/v1/2023.emnlp-main.370

[64] Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards Interpretable Mental Health Analysis with Large Language Models. arXiv:2304.03347 [cs.CL]

[65] Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Sophia Ananiadou, and Jimin Huang. 2023. MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models. arXiv:2309.13567 [cs.CL]

[66] Peitao Zhu, Melissa Luke, and James Bellini. 2021. A grounded theory analysis of cultural humility in counseling and counselor education. *Counselor Education and Supervision* 60, 1 (2021), 73–89.

[67] Therese Zink and Erik Solberg. 2014. Development of a global health curriculum for family medicine based on ACGME competencies. *Teaching and Learning in Medicine* 26, 2 (2014), 174–183.