

Cultural Adaptation and Evaluation of LLM-Driven Mental Health Conversational Agents

Serena Jinchun Xie

A dissertation

submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2025

Reading Committee:

Trevor Cohen, Chair

Weichao Yuwen

Magaly Ramirez

Andrea Hartzler

Program Authorized to Offer Degree:

Biomedical Informatics and Medical Education

©Copyright 2025

Serena Jinchun Xie

University of Washington

Abstract

Cultural Adaptation and Evaluation of LLM-Driven Mental Health Conversational Agents

Serena Jinchen Xie

Chair of the Supervisory Committee:

Trevor Cohen

Department of Biomedical Informatics and Medical Education

Mental health disparities disproportionately affect underserved family caregivers, due in part to the limited availability of culturally responsive interventions. While large language model (LLM)-driven conversational agents hold promise for scalable mental health support, their responses often lack cultural responsiveness, undermining empathy, therapeutic alliance, and engagement among diverse populations.

This dissertation develops and evaluates a novel approach for dynamic cultural adaptation of LLM-based mental health agents leveraging context engineering. The process began with stakeholder engagement to identify culturally salient caregiving challenges among Chinese American family caregivers (Aim 1). In collaboration with domain experts, a cultural context database was developed to capture these challenges alongside culturally responsive response examples, enabling real-time retrieval and integration of relevant context during agent interactions.

To evaluate this approach, I conducted a contextualized, multiphase evaluation through a series of user studies with Chinese American and Latino American family caregivers. Findings show that the context-engineered system consistently generated responses rated as more culturally responsive and empathic than both a prompt-based adaptation strategy and a baseline non-adapted agent (Aim 2). In a randomized user study with Chinese American caregivers (Aim 3), the culturally adapted agent significantly improved near-term emotional well-being and received higher ratings for cultural competence and therapeutic alliance compared to the non-adapted agent. Notably, participants indicated greater willingness to recommend the adapted agent within their communities.

Together, this work contributes: (1) a workflow for adapting LLM-based conversational agents to diverse populations, (2) empirical evidence that cultural adaptation enhances user outcomes in mental health agents, and (3) a human-centered evaluation procedure for assessing cultural responsiveness of AI in mental health contexts. While future research is needed to refine and validate this approach across additional cultural groups and care domains, the findings represent a significant step toward advancing equitable, culturally responsive digital mental health support and reducing persistent disparities in access and quality of care.

Acknowledgements

First and foremost, I would like to thank my supervisory committee, Dr. Trevor Cohen, Dr. Weichao Yuwen, Dr. Andrea Hartzler, Dr. Magaly Ramirez, and my GSRs, Dr. Clara Berridge and Dr. Tim Althoff. I am profoundly grateful for such a supportive committee and will always remember how everyone attended my proposal talk the morning after the bomb cyclone hit Seattle, quite literally overcoming obstacles such as fallen tree branches on the streets.

I would like to express my deepest gratitude to Dr. Trevor Cohen, who has been a wonderful mentor throughout this journey. His insightful guidance, drawn from his extensive expertise, has been instrumental in helping me navigate complex challenges and identify critical gaps in my work. Beyond his academic mentorship, he has provided invaluable guidance for my research and career development, always encouraging me to reach my full potential. I will forever be grateful.

I am also profoundly thankful to Dr. Weichao Yuwen, an exceptional mentor, team leader, and dear friend. I often feel I must have saved the world in a past life to have her as my mentor. Her unwavering support, visionary guidance, and dedication have shaped not only this dissertation but also my outlook on life and research. From her, I learned what it means to be a compassionate leader and a mentor who empowers others. She showed me that with persistence and dedication, it is possible to secure the support for research endeavors that are close to our hearts—work that serves our communities and the people we care about.

I would like to especially extend my heartfelt thanks to Dr. Andrea Hartzler, who played a formative role during the earliest stages of my PhD journey. I am forever grateful for her patience and mentorship as she guided me through many of my “firsts” in academia. Andrea encouraged me to explore my research interests with openness and curiosity, and I continue to carry that spirit with me.

I would also like to thank Dr. Maggie Ramirez for her invaluable support and thoughtful guidance. Her insights helped me recognize and address important gaps in my work. Maggie's passion for her research and her encouragement were deeply inspiring.

I am grateful to my collaborators and colleagues on the COCO team—Jingyi Li, Yanjing Liang, Liying Wang, Daniil Filienko, Shumenghui Zhai, Xuehong Fan, Priscilla Carmiol, Genevieve Aguilar, Yinzhou Wang, Amadalyne Paullada, Jingduo Zhou, Xiaolong Wang, Cindy Dong, Jingjing Xu, Rebekah Lee Baik, William Kearns, Oliver Li, Caroline El Jazmi, and Shubhi Jain. This work was never a solo endeavor; I was fortunate to have such a dedicated and supportive team by my side. Thank you all for your contributions, support, and encouragement. I am also grateful to my friends in the Biomedical and Health Informatics program, including alumni, who have generously offered timely advice and support whenever I needed it.

A special thank you is also due to Dr. Rema Padman at Carnegie Mellon University, who introduced me to the world of research and inspired me to pursue this path.

I owe my deepest gratitude to my parents, Jing Jin and Jianping Xie, for their unwavering love and belief in me. My mother always reminded me to be detail-oriented, while my father encouraged me to reflect on the broader implications of my work—together teaching me to hold myself to the highest standards. I also wish to honor the memory of my grandfather, Yanying Jin, who passed away shortly before I began my PhD. His stories as a scientist inspired me from a young age to pursue my own path in research.

Finally, I would like to thank my partner, Jingwei Li, who has been my greatest supporter and cheerleader. He believed in me unfailingly, lifted me up during difficult moments, and celebrated every small success along the way. This journey would not have been possible without his love and encouragement.

This work was partly funded by the National Institutes of Health (NIH) [No.1OT2OD032581 and R21NR020634] and the University of Washington Population Health Initiative Pilot Grants.

Table of Contents

CHAPTER 1. Introduction	1
1.1 The Need for Mental Health Conversational Agent	1
1.2 Challenges for AI-Based Mental Health Support	2
1.3 Making AI Useful in Mental Health Care	5
1.4 The Case of Family Caregivers and COCO Health	7
1.5 Specific Aims	9
1.6 Overview of Dissertation	12
CHAPTER 2. Background	16
2.1 The Potential of AI-Powered Conversational Agents the Gap in Mental Health Care Services	16
2.2 The Cultural Responsiveness Imperative	19
2.3 The Gap Between AI Capabilities and Cultural Adaptation Methods	23
2.4 Summary of Key Contributions in Addressing Gaps in Culturally Responsive Mental Health AI	28
CHAPTER 3. Understanding the Needs and Challenges of Chinese American Caregivers: Implications for Culturally Tailored Digital Mental Health Tools	31
3.1 Background and Objectives	33
3.2 Study Design and Methods	35
3.3 Results	38
3.4 Discussion	47
3.5 Conclusion	52
CHAPTER 4. Cultural Prompting Improves the Empathy and Cultural Responsiveness of GPT-Generated Therapy Responses	53
4.1 Chinese Version	54
4.1.1 Introduction	55
4.1.2 Methods	59
4.1.3 Results	63
4.1.4 Discussion	68
4.1.5 Conclusion	70
4.2 Spanish Version	71

4.2.1 Methods	71
4.2.2 Results	73
4.2.3 Discussion	77
CHAPTER 5. Design and Implementation of a Retrieval-Augmented LLM Workflow for Culturally Responsive Mental Health Chatbots: A Context Engineering Approach	79
5.1 Development of cultural context database	80
5.2 Design and Experiment with LLMs workflows and prompts - Context Engineering	83
CHAPTER 6. Evaluation of Prompt-Based vs Context-Engineering-Based Cultural Adaptation Approaches	91
6.1 Methods	92
6.2 Results	97
6.3 Discussion	103
6.4 Conclusion	105
CHAPTER 7. Randomized User Study: Impact of Cultural Adaptation on Emotional Well-being and User Experience	107
7.1 Methods	108
7.2 Results	113
7.3 Discussion	116
7.4 Conclusion	120
CHAPTER 8. Discussion & Conclusion	122
8.1 Discussion	122
8.2 Contributions	124
8.3 Limitations and Future Directions	127
8.4 Generalizability and Application	132
8.5 Conclusion	133
References	134
Appendices	166
Appendix 1. Interview Questions for Chapter 3	166
Appendix 2. LLMs' System Prompts (Chapter 5)	171
Appendix 3. Evaluation Questions (Chapter 6)	176

List of Abbreviations

Abbreviation	Definition
AAPI	Asian American and Pacific Islanders
AI	Artificial Intelligence
CBT	Cognitive Behavioral Therapy
CCM	Cultural Competence Measure
CRQ	Cultural Relevance Questionnaire
GPT	Generative Pre-trained Transformer
COCO	Care for Caregivers Platform
HRSA	Health Resources & Services Administration
NLP	Natural Language Processing
LLM	Large Language Models
NAMI	National Alliance on Mental Illness
NPS	Net Promoter Score
PANAS	Positive and Negative Affect Schedule
PHQ-2	Patient Health Questionnaire-2
PST	Problem Solving Therapy
PsyQA	Psychological Question-Answer dataset

CHAPTER 1. Introduction

1.1 The Need for Mental Health Conversational Agent

Mental health conditions affect approximately 22.8% of U.S. adults, roughly one in five adults in the U.S. (National Alliance on Mental Illness [NAMI], n.d.). Yet, less than half of those individuals receive any form of treatment (NAMI, n.d.). Access to mental health care remains a challenge. A limited availability of licensed mental health professionals has left many areas designated as Mental Health Professional Shortages Areas, affecting about 160 million people in the U.S., with rural communities especially impacted (Health Resources & Services Administration [HRSA], 2023). The gap between mental health care need and access is particularly stark among racial and ethnic minorities. For instance, only 25.4% of Non-Hispanic Asian adults and 36.1% of Hispanic or Latino adults who need mental health treatment receive any form of care (National Alliance on Mental Illness [NAMI], n.d.). For many individuals, costs and lack of insurance coverage further impede access to traditional therapy (Han et al., 2017). Additionally, cultural stigma around mental health can discourage help-seeking, especially in certain ethnic communities (Cabassa et al., 2007; Kung, 2004).

Artificial Intelligence (AI) systems, especially conversational AI agents or “chatbots,” have the potential to help bridge these mental health care gaps. Recent advances in large language models (LLMs) have enabled AI systems to simulate therapists to provide support or deliver psychotherapy (Lawrence et al., 2024; Li et al., 2023). Unlike traditional human therapists, these AI conversational agents can be accessed on-demand 24/7 (Inkster et al., 2018; Fulmer et al., 2018), potentially at lower costs (Vaidyam et al., 2019), and with lower fear of judgement (Hoermann et al., 2017). Research suggests some users may feel more comfortable disclosing information freely to AI agents than to human clinicians (Gale et al., 2014). Unlike the

previous generations of rule-based chatbots that relied on prewritten responses, LLM-based conversational agents are able to engage users in natural, context-aware dialogues. Thakkar et al. (2024) reported that AI-powered mental health tools can facilitate emotional regulation through tailored interventions, such as guiding users in relaxation exercises, deep breathing, and personalized stress management strategies. Lawrence et al. (2024) further highlighted the emerging applications of LLM-powered conversational agents in mental health education, assessment, and intervention, highlighting their potential to deliver novel, large-scale solutions to address growing public mental health needs. Building on these foundations, Heinz et al. (2025) conducted one of the first large-scale randomized controlled trials of a fully generative AI-based chatbot, Therabot. In this study, 210 U.S. adults who screened positive for major depressive disorder, generalized anxiety disorder, or clinically high-risk feeding and eating disorders were randomized to either a four-week Therabot intervention or waitlist control. Compared with controls, Therabot users demonstrated large, statistically significant symptom reductions across all diagnostic groups (Cohen's $d = 0.63\text{--}0.90$), with high engagement (averaging over 6 hours total use) and therapeutic alliance ratings comparable to human-delivered therapy. Collectively, these insights support the role of AI as a scalable, front-line component in mental health systems, with the potential to bridge the gap between public need and professional resources.

1.2 Challenges for AI-Based Mental Health Support

Conversational AI has demonstrated impressive capabilities in providing mental health support and general well-being coaching (Li et al., 2023). Yet, the context of mental health poses unique challenges not encountered in more task-oriented domains (such as customer service chatbots). LLMs are known to have hallucinations and sycophancy issues —factually incorrect or misleading responses—which can erode user trust and introduce harm in emotionally vulnerable

contexts, for example, encouraging clients' delusional thinking (Moore et al., 2025). Privacy is also a critical issue, as users may not fully understand how their sensitive disclosures are stored, used, or shared. AI models trained on mental health data face risks of inadvertently encoding or revealing sensitive user information (Mandal et al., 2025). Additionally, these AI models often reflect the sociocultural assumptions embedded in their training data, which can lead to biased or culturally incongruent interactions. Mental health conversations are deeply personal and emotional, and an individual's cultural, social, and psychological background profoundly influences how they express distress, frame coping strategies, and receive support (Song et al., 2024; Chu & Leino, 2017). Cultural beliefs, for example, shape how symptoms of emotional distress are recognized and labeled (Okazaki & Tanaka-Matsumi, 2006), and determine which coping strategies are considered acceptable or helpful. In some East Asian cultural contexts, coping through quiet endurance or family-centered activities (e.g. taking a restorative bath or hosting a family gathering) may be more valued than openly discussing one's emotions (Tsai et al., 2006). Furthermore, communication styles vary – Latino and East Asian cultural norms often favor indirect communication and high-context conversational cues (e.g., relying on tone of voice, pauses, shared cultural references, and nonverbal signals such as eye contact or gestures to convey meaning without explicit statements), which can add complexity to achieving mutual understanding in a therapeutic dialogue (Santiago-Rivera & Altarriba, 2002).

Current LLM-powered conversational agents (i.e., AI agents) may be poorly responsive to these cultural nuances. Studies have shown that state-of-the-art LLMs are trained predominantly on data reflecting Western norms and values (Johnson et al., 2022; Arora et al., 2023). As a result, their default language patterns and elicited emotional responses align more closely with Western individualistic perspectives, potentially overlooking or misinterpreting non-Western cultural expressions (Havaladar et al., 2023; Keleg et al., 2023). This cultural asymmetry in training data means that important nuances – such as collectivist values, family

dynamics, or cultural idioms of distress – may not be adequately represented in the AI agent’s responses. Indeed, recent work has found that LLM outputs tend to align more strongly with Western values and viewpoints, even in multilingual models, indicating that linguistic diversity alone does not guarantee multicultural understanding (Benkler et al., 2023; Rystrøm et al., 2025). This lack of cultural alignment can undermine user trust and reduce the perceived empathy of the AI agents’ responses (Song et al., 2024; Ullrich, 2019). If an AI agent produces advice or emotional reflections that feel irrelevant – or worse, insensitive – to a user’s lived experience, the user may quickly disengage (Song et al., 2024). In therapeutic terms, such misalignment can erode the therapeutic alliance – the collaborative, trusting bond between client and therapist that is critical for positive outcomes (Ullrich, 2019; Wampold, 2015).

Given that AI-based mental health tools are often proposed as a way to expand access for underserved populations, any inability to account for cultural differences poses a serious concern: rather than narrowing mental health disparities, generic AI tools could inadvertently amplify them. Therefore, this dissertation tackles the crucial question of how we can systematically and scalably culturally adapt AI-driven mental health interventions to better serve culturally diverse users. In particular, we focus on Chinese American and Latino American family caregivers—populations that face substantial caregiving burdens while remaining traditionally underserved in mental health care (NAMI, n.d.). These groups also represent speakers of the two most commonly used non-English languages in the U.S., contexts where AI interventions can significantly reduce language barriers and improve accessibility.

However, while the AI tools offer considerable promise, one challenge remains in ensuring this potential is realized equitably. In particular, this dissertation argues, AI-powered mental health support must move beyond a one-size-fits-all approach and toward cultural responsiveness if it is to effectively reach and engage diverse populations.

1.3 Making AI Useful in Mental Health Care

For AI-powered tools to be useful across diverse populations, it is not enough for them to demonstrate technical accuracy or provide generic expressions of empathy. Mental health support is deeply relational and culturally embedded. In human-delivered psychotherapy, therapists tailor their language, tone, communication style, and therapeutic strategies to align with the cultural backgrounds and lived experiences of their clients (Tseng, 2004; Hwang et al., 2006; Rathod et al., 2019). This individualized cultural responsiveness is mirrored at the level of intervention design through a process known as cultural adaptation—a critical strategy for improving engagement and treatment efficacy for diverse clients from different cultural backgrounds (Bernal et al., 1995; Perera et al., 2020; Naeem et al., 2019; Ramirez et al., 2023). This process is known as cultural adaptation. Cultural adaptation goes beyond surface-level translation of language; it is a systematic, theory-informed process of aligning intervention content with the values, norms, idioms, practices, and worldviews of a specific cultural group (Bernal et al., 1995; Chu & Leino, 2017).

Empirical research has consistently demonstrated that culturally adapted mental health interventions yield better outcomes for marginalized groups (Ennis et al., 2020; Degnan et al., 2018; Griner & Smith, 2006). Meta-analytic evidence shows that interventions enhanced with cultural tailoring are more effective in improving clinical outcomes than their unadapted counterparts (Griner & Smith, 2006)^[OBJ]. Recent work has also shown that culturally tailored AI tools—such as chatbots that integrate culture-specific knowledge into the AI system—can lead to higher user satisfaction and engagement (Chang et al., 2024). In practice, communities ranging from Latino families to Chinese American caregivers have been shown to experience higher engagement, greater satisfaction, and improved outcomes when services are delivered in a culturally congruent manner (Beasley et al., 2021; Hong et al., 2022; Cardona et al., 2012^[OBJ]).

These findings underscore that cultural relevance is not a mere “nice-to-have,” but rather a determinant of an intervention’s effectiveness and user acceptance.

In the realm of AI and LLMs, researchers have begun exploring methods to imbue models with cultural knowledge. One approach involves fine-tuning or continued training on culturally specific datasets to better align model outputs with a target culture’s language use, norms, and values (Bai et al., 2025; Huang et al., 2024; Masala et al., 2024). While promising, these weight-tuning approaches require substantial technical expertise, computational resources, and large volumes of high-quality, culturally specific data. Even when community organizations partner with technical developers or researchers to create culturally adapted AI tools, the resource demands of fine-tuning approaches often exceed what these partnerships can sustain long-term or scale across diverse populations. Moreover, for many cultural groups—particularly those underrepresented in digital spaces—relevant training data may be scarce, inaccessible, or simply nonexistent.

This work responds to these questions by experimenting with LLM-facilitated cultural tailoring techniques that do not require retraining the model from scratch. Specifically, we investigate approaches that leverage prompt tailoring and retrieval-augmented generation (RAG) within LLM workflows to dynamically deliver culturally responsive support. Grounded in evidence-based psychotherapy practices and co-designed with community stakeholder input, our approach aims to move beyond proof-of-concept toward practical, adoptable methods for real-world mental health interventions that promote equity. In essence, we seek to demonstrate that cultural tailoring can be achieved “on the fly” – during AI-generated conversations – through intelligent prompting and context injection, rather than through data-intensive training or manual editing.

1.4 The Case of Family Caregivers and COCO Health

To explore the real-world design, development, and evaluation of culturally responsive conversational AI, this dissertation centers on the context of family caregivers and leverages an existing platform called COCO Health. COCO Health is a conversational AI system originally developed by colleagues as a digital mental health support tool for family caregivers (Kearns et al., 2020; Wang et al., 2025; Filienko et al., 2024; COCO - Care for Caregivers., n.d.). It provides an ideal testbed for studying cultural tailoring of conversational agents for several reasons: it targets a high-need population, it is built on evidence-based therapy techniques, and it was designed with flexibility for research-driven modifications.

Family caregivers experience significant physical, emotional, and financial burdens as they care for relatives with illnesses or disabilities (AARP, 2020). Asian American caregivers – and Chinese American family caregivers in particular – have been found to experience higher stress and depression levels on average than White American caregivers (AARP, 2020; Pinquart & Sörensen, 2003; Semere et al., 2019). Despite this elevated burden, they also report more barriers to accessing and completing mental health services and often have a lower perceived need for formal support (Green et al., 2020). This population faces unique challenges when balancing caregiving and self-care. For example, in Chinese culture the value of filial piety creates a strong sense of obligation to care for aging parents, which can lead caregivers to neglect their own well-being out of duty (Guo et al., 2019; Miyawaki, 2015). Language barriers and limited English proficiency among immigrant caregivers further hinder access to support resources (Chi & Boot, 2019), potentially exacerbating stress (Hong et al., 2024). Caregivers often struggle to incorporate self-care practices into their lives due to these cultural and practical barriers (Hong et al., 2024; Wang et al., 2021), despite an unmet need for self-care support

documented in this community (Xu et al., 2024; Yu et al., 2024; Chi et al., 2020). AI systems offer promising scalable solutions to help support family caregivers (Wang et al., 2024).

COCO Health was created to support caregiver mental health using evidence-based methods, specifically Problem-Solving Therapy (PST) – a structured therapeutic approach that helps individuals manage stress and address everyday problems. The key feature of COCO Health for our purposes is its conversational AI module. Rather than relying on rigid scripts or simple rule-based dialogues, COCO’s chatbot utilizes a semi-structured conversation flow powered by generative LLMs. It is guided by carefully designed prompts (including few-shot examples) authored by clinical experts to ensure the LLM-driven conversational agent’s responses follow therapeutic principles (Filienko et al., 2024). This design allows the AI system to respond flexibly and empathetically to users, within the boundaries of PST’s framework. The system was not intended to replace human therapists, but to provide low-intensity, easily accessible support with an emphasis on empathy and emotional validation. In prior work, the COCO Health AI has been evaluated to refine its prompting strategies and assess its capacity to build a therapeutic alliance with users. For example, Filienko et al. (2024) identified optimal prompt configurations for eliciting emotionally intelligent, tailored support responses. Wang et al. (2025) demonstrated that carefully crafted prompts (in English) enabled the AI to foster a reasonable level of therapeutic alliance with caregivers, indicating potential for users to feel heard and supported by the chatbot.

This dissertation builds upon COCO Health by extending it to support culturally responsive dialogue for specific caregiver communities. In doing so, we transform COCO Health into a platform for investigating how cultural context can be integrated into AI-driven mental health support. Our focus is on two caregiver groups: (1) Chinese American caregivers of older adults who prefer to speak Chinese, and (2) Latino American caregivers of children with special needs (specifically, autism) who prefer to speak Spanish. These groups were chosen because

they represent speakers of the two most commonly used non-English languages in the U.S. and underserved populations with well-documented cultural influences on caregiving and mental health, allowing us to study the impact of adaptation in contrasting cultural contexts.

Ultimately, COCO Health serves both as a practical application – aiming to deliver real support to caregivers – and as a research vehicle for examining the processes and outcomes of culturally tailored conversational agents.

1.5 Specific Aims

The overarching goal of this dissertation is to develop and evaluate culturally responsive LLM-based conversational agents for mental health support, with a focus on addressing the needs of underserved cultural communities. In pursuit of this goal, the work is organized around three specific aims:

Aim 1: Identify culturally salient caregiving challenges and self-care support needs to guide adaptation of LLM-based mental health tools. For this formative aim, we engaged directly with stakeholders to understand the cultural and contextual needs of the target populations. We conducted qualitative interviews and focus groups with two key stakeholder groups: family caregivers from the target cultural communities and staff from community-based organizations that serve these caregivers. The objectives were: (1) to establish the need for cultural adaptation in digital mental health interventions for caregivers, and (2) to identify caregiving-related stressors and self-care challenges that are particularly salient within the cultural context of the target populations.

Findings from Aim 1 yielded a set of culturally relevant themes and scenarios that formed the foundation for constructing a cultural context database. This database was used in Aims 2 and 3 to enable retrieval-augmented cultural adaptation, allowing the AI system to

integrate culture-specific knowledge into its responses through context-aware prompt construction.

Aim 2: Design and implement LLM-based cultural adaptation techniques to improve the cultural responsiveness of conversational agents. Based on insights from Aim 1, this aim explores and experiments with two different technical approaches for making an LLM-driven mental health agent more culturally responsive:

Aim 2.1 Prompt-based adaptation. We developed prompt tailoring methods, where culturally relevant context or instructions are embedded directly into the prompts given to the LLM. This prompt-based strategy called ‘cultural prompting’ was used in cultural alignment research in the field of natural language processing. We evaluated this prompt-based strategy in a controlled setting (offline, static evaluation) with family caregivers to assess their impact on cultural responsiveness and perceived empathy of the LLM’s responses. We conducted separate evaluations for Chinese American caregivers of older adults and Spanish-speaking Latino caregivers of children with autism, to examine the effectiveness of prompt-based cultural adaptation in each group. We also analyzed the relationship between the LLMs’ cultural responsiveness and users’ perceptions of the LLMs’ empathy.

Aim 2.2 Dynamic retrieval-based adaptation. We designed a dynamic adaptation strategy using context engineering with Retrieval-Augmented Generation (RAG). In this approach, the agent can retrieve culturally relevant information on the fly and incorporate it into its responses. We implemented this as an LLM workflow (multi-step LLM process) and evaluated it in an interactive setting. Family caregivers and community organization staff from the Chinese American caregiver community participated in real-time, hands-on sessions with three versions of the conversational agent: (1) a baseline agent without cultural adaptation, (2) a culturally adapted agent using the prompt-based method, which used the pre-existing strategy

described in Aim 2.1, and (3) a novel culturally adapted agent using the dynamic context engineering method. This within-subject comparative evaluation allowed us to directly measure improvements in cultural responsiveness, empathy, and therapeutic alliance due to the adaptive techniques. We also identified which of the two adaptation methods (prompt-based vs. context engineered) performed better, to determine which method would be used to power the adapted agent in Aim 3.

Aim 3: Evaluate the impact of cultural adaptation on perceived empathy, therapeutic alliance, and preliminary health outcomes in a real-world usage scenario. For this aim, we conducted a randomized experiment with end-users to rigorously assess whether cultural adaptation affects user experience and outcomes. Chinese American family caregivers were randomly assigned to use either: (1) the culturally adapted conversational agent (as developed through Aim 2 and refined based on the comparative results), or (2) a baseline non-adapted version of the agent. After a full interactive session with the AI, we measured key outcomes including the caregiver's perceived empathy from the agent, the therapeutic alliance formed, their satisfaction with the interaction, and some preliminary mental health outcome indicators, specifically immediate changes in emotional states. In doing so, we examined the practical impact of cultural adaptation on user experience and short-term mental health outcomes. The results provide insight into whether the benefits observed in controlled comparisons translate into meaningful improvements in real-world usage.

For both Aim 1 and Aim 2.1, the studies were first conducted with Chinese-speaking Chinese American family caregivers, followed by parallel studies with Spanish-speaking Latino family caregivers. The parallel interview study in Aim 1 was conducted by bilingual Spanish-speaking collaborators, whereas the parallel study in Aim 2.1 was carried out collaboratively by the author and Spanish–English bilingual researchers. Accordingly, the Aim 1

parallel study with Latino caregivers is not presented in this dissertation, whereas the Aim 2.1 parallel study with Latino caregivers is included.

Following completion of the first draft of this dissertation, I used ChatGPT for copyediting assistance to identify grammatical errors and enhance the clarity and flow of the writing.

1.6 Overview of Dissertation

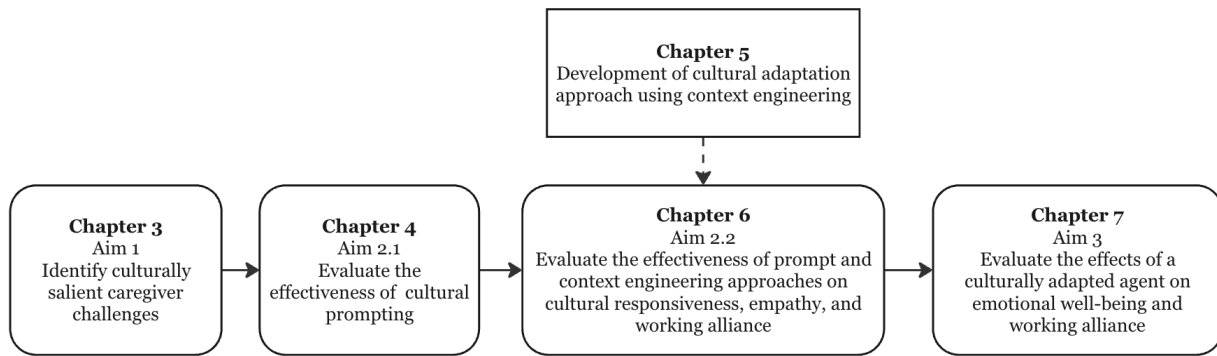


Figure 1.1. Overview of the dissertation structure.

This dissertation is structured as follows. Chapter 2 (Background) frames the research problem by reviewing relevant literature and highlighting the key gaps that this work addresses. In particular, Chapter 2 discusses: (1) the lack of scalable, real-time cultural adaptation methods that are compatible with current generative AI technologies, and (2) the lack of contextually grounded, user-centered evaluation frameworks for cultural responsiveness in mental health AI systems. These gaps motivate the aims and approach of this dissertation.

Chapter 3 (Aim 1) describes the qualitative study conducted to understand stakeholder needs. Through interviews and focus groups with Chinese American caregivers and community staff, this chapter identifies the culturally salient caregiving, self-care, and mental health care challenges of this population. Examples of challenges uncovered include communication

barriers between adult children and their elderly parents, the social isolation of older adults resulting in caregiver guilt, and difficulties navigating healthcare and social service systems in a culturally unfamiliar environment. These insights directly informed the development of a cultural context database designed to support culturally responsive responses from the mental health conversational agent. The themes identified in this chapter lay the foundation for the AI adaptation process detailed in Chapter 5.

Chapter 4 (Aim 2.1) details the development and evaluation of the prompt-based cultural adaptation technique, referred to as cultural prompting (Tao et al., 2024). In this chapter, two studies (one with Chinese caregivers, one with Latino caregivers) are presented. Caregivers rated the cultural competence, cultural relevance, and empathy of AI-generated mental health support responses with vs. without cultural prompting. The results showed that adding cultural context into the prompts significantly improved GPT-4o's performance on all measured dimensions. Moreover, a mediation analysis suggested that the enhancements in perceived empathy were driven by the agent's improved cultural responsiveness, highlighting the importance of cultural responsiveness in building empathy.

Chapter 5 (Aim 2.2) then describes the development of a dynamic cultural adaptation system using context engineering. In this approach, instead of pre-defining cultural content in the prompt, the system can intelligently determine when to retrieve relevant cultural context from the database (constructed from Chapter 3's findings) and incorporate it into its responses in real time. Chapter 5 details the design decisions and implementations involved in building this LLM-based workflow, including how we developed the cultural context database and integrated retrieval-augmented generation (RAG) into the conversational agent. The chapter explains how the system decides when to fetch cultural context, how the retrieved information is filtered and used within the final response generation, and how we balance the contributions of human-authored cultural knowledge with the generative model's capabilities.

Chapter 6 (Aim 2.2, continued) presents a with-in subjects evaluation study with three prototype agents: (1) an unadapted baseline agent, (2) a prompt-based culturally adapted agent (from Chapter 4), and (3) a dynamically adapted agent using the context-engineering approach (from Chapter 5). In this study, which involved both Chinese American caregivers and community staff as participants, each participant interacted with all three versions (in a randomized order) using the same caregiving scenario. We collected quantitative evaluations using validated instruments – such as the Session Rating Scale (SRS) for therapeutic alliance, and custom cultural competence and relevance questionnaires – to compare empathy, alliance, and cultural responsiveness across the three conditions. Findings demonstrate that the dynamic context-engineered agent (Prototype 3) outperformed both the prompt-based and non-adapted agents on most metrics, indicating the added value of real-time cultural adaptation with human-authored use-case specific cultural knowledge.

Chapter 7 (Aim 3) reports on the randomized user study with Chinese American caregivers as end-users. In this chapter, the best-performing culturally adapted agent from Chapter 6 is compared against a non-adapted baseline in a between-subjects design. Whereas Chapter 6’s evaluation involved each participant trying all versions, Chapter 7 focuses on real end-users each using one version in a realistic usage scenario. Outcome measures include preliminary health indicators (e.g., mood changes measured by the PANAS), as well as perceived empathy, therapeutic alliance, and user satisfaction. This evaluation enables us to examine the practical impact of cultural adaptation on user experience and short-term mental health outcomes. The results provide insight into whether the benefits observed in controlled comparisons (Chapter 6) translate into meaningful improvements in real-world scenarios.

Finally, Chapter 8 provides a synthesis of the contributions and findings of the dissertation. We summarize the end-to-end workflow we developed for culturally adapting LLM-driven mental health conversational agents, compare the effectiveness of the two

adaptation strategies (prompt-based vs. dynamic retrieval-based), and discuss the implications for generalizability. We also address the limitations of our approach and outline directions for future research, such as extending the methods to other cultural groups, exploring long-term outcomes, and ensuring ethical, safe deployment of culturally adaptive AI in real-world care settings.

CHAPTER 2. Background

2.1 The Potential of AI-Powered Conversational Agents the Gap in Mental Health Care Services

2.1.1 Mental Health Gap and Equity

The United States is facing an ongoing mental health crisis. More than one in five American adults, approximately 57.8 million people, currently live with a mental illness (National Institute of Mental Health, 2023). Additionally, 27% of Americans report stress levels that significantly impair their daily functioning (American Psychological Association, 2022). Unfortunately, these substantial mental health needs are not being met with adequate services. As of 2024, over 120 million Americans (around 37% of the population) reside in areas officially designated as having a shortage of mental health professionals (HRSA, 2023). In practical terms, this means tens of millions of people cannot readily access a therapist or counselor in their community.

Crucially, the gap between mental health needs and available care is not uniform across society – it disproportionately affects marginalized and minority communities. Asian Americans, for example, have the lowest rate of mental health service utilization among major ethnic groups (about 25% usage among those with a need), followed by Hispanic/Latino Americans (around 36% usage), despite reporting high levels of psychological distress (NAMI, n.d.). Within the context of caregiving, research shows that Asian American and Latino family caregivers experience elevated stress levels compared to their White counterparts, yet also face more barriers to completing treatment and often report a lower perceived need for mental health services (AARP, 2020; Green et al., 2020; Semere et al., 2019).

Multiple factors contribute to these disparities in mental health care access and outcomes. Structural issues include inadequate insurance coverage and an uneven geographic distribution of providers – rural areas and some urban neighborhoods simply lack mental health clinics or professionals. There is also a scarcity of culturally and linguistically appropriate resources; many mental health services are not tailored to non-English speakers or do not account for cultural norms, which discourages use by minority groups. Moreover, cultural stigma around mental illness and help-seeking remains prevalent in many communities, further suppressing utilization of care (Ahad et al., 2023; Mongelli et al., 2020). Historical distrust of healthcare systems and a lack of exposure to behavioral health concepts can also lead to lower perceived need for services in some groups (The Commonwealth Fund, 2003). These factors intersect to create a substantial mental health equity gap – even as overall awareness of mental health grows, those in marginalized communities often do not benefit equally from advances in care.

2.1.2 AI-Powered Mental Health Solutions: Current Landscape and Potential

To address the mismatch between widespread mental health needs and scarce human resources, a range of AI-based solutions have emerged in recent years (Torous et al., 2023; Li et al., 2023). In particular, AI-driven mental health conversational agents (chatbots) are seen as a promising approach to provide scalable, cost-effective support that transcends some limitations of traditional care (De Choudhury et al., 2023). These AI agents can operate without being bound to office hours or physical locations, thus sidestepping many infrastructure constraints of the healthcare system. Importantly, they also offer an anonymous, stigma-reducing access point for individuals who may be hesitant to seek help due to fear of judgment or cultural stigma, making communicating with a chatbot can feel more private and accessible (Corrigan & Matthews, 2003; Shi et al., 2025). Moreover, AI-driven agents have personalization capabilities to provide tailored responses (Shi et al., 2025).

A number of AI mental health platforms and chatbots have shown encouraging results in research studies and pilot deployments. For example, Woebot, a fully automated chatbot delivering principles of Cognitive Behavioral Therapy (CBT) (Beck, 1976), demonstrated effectiveness in reducing symptoms of depression and anxiety among college students and young adults over several weeks of use (Fitzpatrick et al., 2017; Fulmer et al., 2018; Ponzo et al., 2020). Similarly, Wysa, another popular CBT-based mental health chatbot, has reported measurable reductions in depression and anxiety severity across diverse user populations in both self-directed use and clinical evaluations (Inkster et al., 2018; MacNeill et al., 2024). Beyond individual products, systematic research reviews have highlighted the broader potential of AI interventions: studies suggest that mental health chatbots can improve adherence to therapeutic exercises or self-management practices (Yasukawa et al., 2024) and can promote overall well-being by providing accessible, on-demand emotional support (Li et al., 2023). These findings collectively point to AI-powered conversational agents as viable tools to extend the reach of mental health care.

Despite this promise, it must be noted that the evidence base is still developing. Many of these studies have been conducted in English-speaking contexts and with relatively homogenous user samples. The success stories of Woebot, Wysa, and others largely reflect outcomes in Western cultural settings. To date, the promise of AI-powered mental health support remains underexplored among culturally and linguistically diverse populations, such as recent immigrants who prefer mental health services in languages other than English, including Chinese and Spanish. In other words, while early trials and applications show that chatbots can help users in general, we have far less understanding of how well they work for users from non-Western backgrounds, or how to make them work better for those groups. The design and content of current systems tend to be Western-centric, which may not resonate with users from different cultural traditions. In the following sections, we delve into why cultural responsiveness

is imperative for mental health interventions, and how current AI tools measure up (or fall short) in this regard.

2.2 The Cultural Responsiveness Imperative

In the fields of psychology and medicine, the importance of delivering care in a culturally responsive manner has been well established through decades of research and clinical experience (Hofmann, 2006). Over time, several terms have been used to describe the integration of cultural context into clinical care. Culturally responsive therapy refers to a deliberate approach in which therapists actively (1) demonstrate cultural awareness, (2) demonstrate knowledge of their clients' cultural backgrounds, and (3) tailor therapy practices to the unique cultural context and life experiences of clients accordingly (Eustis et al., 2023; Jones-Smith, 2019; Center for Substance Abuse Treatment, n.d.; Sue et al., 1991). This concept builds upon cultural competence, a widely adopted concept defined as “a set of congruent behaviors, attitudes, and policies that come together in a system, agency, or among professionals that enables effective work in cross-cultural situations” (SAMHSA, 2019; Chu et al., 2022). However, cultural competence has been criticised that it tends to generalize behaviors of a group and perpetuate cultural group stereotypes (Kirmayer, 2012; Jones et al., 1996). Closely related to it is cultural humility, which emphasizes an ongoing commitment to self-reflection of cultural awareness, and open to learning new cultural information from clients, particularly in acknowledging and addressing power imbalances within therapeutic relationships (Delfish et al., 2025; Mosher et al., 2017). Cultural sensitivity, often linked with cultural awareness, focuses more on recognizing and respecting cultural differences (Sue et al., 2008; Graham et al., 2013). Regardless of the terminology, the consensus is that acknowledging and integrating a client's cultural context enhances therapeutic trust and understanding between provider and client (Sue & Zane, 1987; Sue et al., 2008). In this work, I adopt the term

cultural responsiveness, as it represents the broader end-goal of cultural competence and humility and it emphasizes not only awareness and respect but also proactively demonstrating awareness and knowledge of clients' cultural context and life experiences throughout the therapeutic process.

2.2.1 Cultural Bias in Large Language Models

Turning to AI systems, we face the reality that large language models (LLMs), which form the foundation of modern conversational agents, come with significant cultural biases. The effectiveness of AI-driven mental health interventions is fundamentally constrained by the cultural limitations embedded in these underlying models. LLMs (such as the GPT series, etc.) have been trained on vast internet text corpora that are predominantly in English and thus largely reflect Western norms, values, and viewpoints, which refers to the cultural orientations commonly associated with North America and Western Europe, such as individualism, including individualism, low power distance (i.e., egalitarian social structures), and a strong emphasis on personal rights and freedoms (Hofstede et al., 2010; Johnson et al., 2022; Tao et al., 2024; Havaladar et al., 2023). As a result, these models tend to mirror Western cultural patterns more readily than those of non-Western cultures.

Research has consistently shown that popular LLMs exhibit systematic biases aligning with Western values. For instance, one study found that an advanced LLM's responses demonstrated stronger alignment with individualistic, Western cultural perspectives compared to viewpoints rooted in non-Western cultures (Benkler et al., 2023). In another study, while pride is more openly expressed in Western cultures and more often inhibited in Eastern cultures such as Japanese culture, Havaladar et al. (2023) found that LLMs failed to reproduce this well-established cultural difference. Similarly, Tao et al. (2024) reported that in culturally sensitive scenarios, such as attitudes toward family caregiving or guest etiquette, LLM responses

frequently reflected Western individualist values (e.g., prioritizing personal boundaries). This Western-centric orientation persists even in models that are multilingual or have seen diverse data: being able to generate text in multiple languages is not the same as possessing a deep understanding of multiple cultures. A model might generate responses in Chinese or Spanish, yet still implicitly follow Western patterns of reasoning or counseling. This distinction is critical in sensitive domains like mental health^[69]. Linguistic diversity in training data does not automatically equate to cultural diversity of thought in the model's responses.

The implications of these biases are far beyond research – they have real-world consequences for users seeking help. Users from non-Western backgrounds might find LLM-generated mental health advice to be subtly off-target, culturally insensitive, or simply not relatable. They may express that “the bot doesn’t get where I’m coming from,” which reflects a failure of the AI to align with the user’s cultural frame of reference (Hershovich et al., 2022; Tschuschke et al., 2022). In another study, the users from Black/African American communities expressed strong preference for culturally relevant content in a generative AI powered App (Bosco et al., 2025). Such misalignment could undermine the therapeutic potential of these AI interventions. In the worst case, an insensitive or irrelevant response might even exacerbate a user’s feelings of alienation or distress, especially among those who already feel marginalized. Recent advances in personalization of LLM outputs raise the possibility of more tailored support (Dingler et al., 2021; Lawrence et al., 2024), but as Pawar et al. (2024) argue, personalization without explicit cultural awareness may just reinforce dominant norms and exclude marginalized perspectives. In other words, a system might become very adept at individual personalization (e.g. using a person’s name, remembering facts about them) yet still fail at cultural personalization (understanding the cultural meaning behind the person’s experiences). This points to the need for approaches that explicitly inject cultural knowledge and sensitivity into AI systems.

2.2.2 The Clinical Importance of Cultural Responsiveness in Mental Health Care

It is valuable to reiterate why cultural responsiveness is so crucial in a mental health context, whether the provider is human or AI. Decades of clinical research confirm that when mental health care is delivered in a way that respects and incorporates the patient's cultural background, outcomes improve (Sue & Zane, 1987; Sue et al., 2008; Chu et al., 2022). For instance, encouraging a coping strategy that involves socializing with friends might be ineffective (or even stressful) advice for a caregiver in a community where privacy and family honor dictate keeping problems within the family. Therapists are trained to pick up on such cultural cues and adjust their approach accordingly (Delfish & Chadha, 2025; Mosher et al., 2017). An AI agent needs a way to do the same. In the context of digital mental health tools, cultural responsiveness could mean the difference between a user sticking with an app or abandoning it after one try. If a Latina mother using a parenting stress chatbot never sees her cultural values or daily realities acknowledged in the conversation, she may conclude the tool “just isn’t meant for people like me.” On the other hand, if the chatbot’s responses make her feel seen – for example, validating the importance of *familismo* (i.e., strong family interconnectedness) while helping her set personal boundaries – she may be more likely to continue engaging and benefit from the program.

Recognizing the critical role of cultural responsiveness in mental health care, researchers have developed cultural adaptation frameworks to guide the cultural adaptation of evidence-based interventions (Bernal et al., 1995; Barrera & Castro, 2006; Domenech Rodríguez et al., 2011). For instance, Barrera and Castro (2006) outlined a stepwise framework for cultural adaptation that involves gathering information to identify potential areas for modification, implementing preliminary adaptations based on these insights, testing the adaptations in pilot studies, and refining them in response to pilot findings. Interventions adapted through such structured processes have consistently demonstrated greater engagement and improved

outcomes for target communities compared with unadapted interventions. These benefits have been observed across a range of health conditions and cultural groups—for instance, tailoring diabetes management programs to reflect dietary practices of specific ethnic communities, or modifying therapeutic modalities to better align with community communication styles (Resnicow et al., 1999; Castro et al., 2010).

In summary, both the technical characteristics of LLMs and the clinical principles of mental health care point to cultural adaptation as an essential frontier. AI mental health agents must account for cultural differences to truly fulfill their promise of broadening access and improving outcomes. Failing to do so risks leaving entire groups of users behind and potentially widening the disparities we set out to reduce.

2.3 The Gap Between AI Capabilities and Cultural Adaptation Methods

The intersection of AI-powered mental health tools and the need for cultural adaptation reveals two major gaps that this dissertation seeks to address:

1. **The lack of real-time, dynamic cultural adaptation methods for LLM-driven conversational agents.** In other words, current strategies for cultural adaptation have not caught up with the capabilities (and challenges) of generative AI systems that produce content in real-time.
2. **The lack of contextually grounded, human-centered evaluation frameworks for cultural responsiveness in AI outputs.** That is, there is a lack of evaluation to measure whether an AI conversational agent is culturally responsive in the context of an actual therapeutic conversation, as opposed to abstract, questionnaire style evaluations.

2.3.1 Gap 1: Dynamic Cultural Adaptation in Real-Time LLM Systems

Incompatibility of Traditional Adaptation Methods with LLM-Generated Content.

Traditional approaches to cultural adaptation in mental health interventions assume a relatively static content delivery. For example, a cognitive-behavioral therapy workbook might be translated and adapted by a team of experts for use in a new cultural group, or a scripted therapy program might be modified to include culturally relevant metaphors (Spanhel et al., 2021). These processes involve expert review and revision of fixed materials to better align with the target population's beliefs, values, communication styles, and lived experiences (Barrera & Castro, 2006; Castro et al., 2010; Spanhel et al., 2021). For instance, the team behind the Wysa mental health app recently created a culturally adapted Spanish-language version by having Spanish-speaking clinicians translate and tailor Wysa's pre-written English responses and content for the new audience (Dinesh et al., 2024). In another example, researchers adapted a physical activity promotion chatbot for Chinese users by applying Hofstede's cultural dimensions theory and manually tailoring the agent's dialogue to better match Chinese cultural values, such as emphasizing collective family benefits of exercise rather than individual goals (Zhou et al., 2017).

However, these conventional methods cannot be directly applied to LLM-based agents, which generate content dynamically and non-deterministically in response to each user's unique inputs. Unlike traditional systems, there are no static, pre-written materials or finite message sets that can be manually adapted in advance. Instead, LLMs generate content in real-time, guided by their training data and prompts. This dynamic, generative nature makes it technically infeasible to rely on human experts to review and tailor materials beforehand. Cultural adaptation workflows for LLM-based conversational agents must instead embed cultural understanding and human experts' input directly into the AI system itself.

Resource and Scalability Constraints

A further complication in adapting LLM-based conversational agents culturally is the limited availability of data and resources for many cultures, particularly for underserved populations whose languages and cultural contexts are underrepresented in foundational model training datasets. Approaches to cultural alignment in NLP generally fall into two broad categories:

- Weight-changing methods: e.g., fine-tuning or continued pre-training the model on culturally relevant data (Bai et al., 2025; Huang et al., 2024; Masala et al., 2024). This alters the model’s internal parameters to better reflect the linguistic patterns and knowledge of the target culture.
- Non-weight-changing methods: e.g., prompt-based techniques that provide cultural context or instructions to steer the model’s outputs without changing its weights (Cao et al., 2023; Tao et al., 2024; Mukherjee et al., 2024).

Weight-changing methods, while potentially very effective, demand extensive resources – not only do they require technical expertise and computational power, they also depend on having a substantial corpus of culturally specific data to fine-tune on. For many underserved languages or cultural groups, such large datasets simply do not exist (or are proprietary and not freely available). Even if they did, smaller-scale organizations or research teams typically lack the resources to train large models. Community-based organizations are not generally engaged in developing large AI systems today, but with the growing accessibility of new AI tools in recent years, they may begin to adopt and adapt lighter-weight approaches for their communities. In this way, our work helps lower the barrier for these organizations to engage with AI in ways that are aligned with their cultural and service contexts. On the other hand, prompt-based methods are lighter weight and easier to implement, making them more accessible to community organizations or smaller research teams. While prompt engineering alone may have limitations

and mixed results in deeply aligning outputs with cultural nuances (Tao et al., 2024; Cao et al., 2023), recent approaches that augment prompts with retrieval-enhanced cultural context could offer a promising middle ground (Chang et al., 2024). Our approach builds on this latter direction, aiming to retain the accessibility of prompt-based methods while improving cultural responsiveness through dynamic retrieval-enhanced cultural context injection.

Compounding the issue, there is a scarcity of culturally relevant conversational datasets to even evaluate or train LLMs for cultural responsiveness. To our knowledge, one of the few datasets of this kind is PsyQA, which contains mental health Q&A pairs with culturally tailored answers for Chinese users (Sun et al., 2021). But even PsyQA is limited in scope – it covers certain topics and a general Chinese cultural context, which may not directly translate to the caregiver context or to other groups. We have not identified comparable dataset exists for Spanish-speaking or Latino cultural contexts in therapy. This data scarcity motivates methods that do not require large training corpora and that can instead leverage expert knowledge in more efficient ways.

In summary, Gap 1 is that we need methods for cultural adaptation that work with LLM-generated contents (i.e., dynamic generation) and are practical given resource constraints. This dissertation’s approach to closing Gap 1 is to use LLM workflows and context engineering (i.e., “building dynamic systems to provide the right information and tools in the right format such that the LLM can plausibly accomplish the task” (Chase, 2025)): by orchestrating multiple LLM and retrieval steps to inject cultural context into responses as needed, rather than altering the model or relying on static prompts alone.

2.3.2 Gap 2: Contextually Grounded Evaluation of Cultural Responsiveness

The second major gap lies in how we evaluate an AI system’s cultural responsiveness and its impact on the user. Existing research on cultural awareness in language models has mostly

relied on decontextualized evaluation metrics. For example, some studies pose multiple-choice questions to an LLM about cultural facts or values and see if it chooses the correct answer that a culturally knowledgeable person would (Arora et al., 2023; Zhao et al., 2024). Others test an LLM’s understanding of proverbs or idioms from a certain culture (Liu et al., 2024). Some evaluation schemes involve asking the model to generate responses and then using either humans or another AI to judge if those responses show cultural knowledge (Shi et al., 2024; Chiu et al., 2024; Zhou et al., 2025; Ramezani et al., 2023). While these approaches provide useful diagnostics, they fall short of capturing the interactive nature in the therapeutic conversation context.

In a real counseling session (or a chatbot-based session), cultural responsiveness is evidenced in nuanced ways: the therapist or AI might pick up on a client’s indirect reference to a family expectation and gently probe it, or they might frame advice in terms of culturally resonant values. These subtleties are hard to measure with a multiple-choice survey or a single-turn judgment. As Pawar et al. (2024) argue, real-world human-AI interactions require models to demonstrate cultural empathy and contextual appropriateness, not just regurgitate facts about a culture. For instance, knowing the definition of “filial piety” (a cultural value) is different from appropriately responding to a caregiver who feels guilty because of that value.

What is needed is a human-centered, contextualized evaluation framework. This means involving target users (or culturally informed human evaluators) to assess AI responses in the context of full dialogues or scenarios, focusing on whether the agent’s behavior feels culturally sensitive, respectful, and helpful. It also means considering outcomes that are indicators of therapeutic effectiveness (e.g., Does cultural adaptation improve the user’s therapeutic alliance with the agent?). These kinds of questions connect cultural responsiveness to tangible impacts on user experience and health outcomes.

To the best of our knowledge, no prior study has systematically evaluated if or how LLM-driven mental health conversational agents demonstrate cultural responsiveness in the context of a therapeutic encounter. Our work addresses this gap by designing evaluations where real users from the target communities interact with the AI systems and then provide feedback about the AI agent’s cultural responsiveness through surveys such as the Cultural Relevance Questionnaire and Cultural Competence Measure. This approach ensures that evaluation is grounded in the lived experiences and perceptions of the very users we aim to serve, rather than abstract metrics alone. As Blodgett et al. (2020) emphasize, AI evaluation, especially for bias or fairness, should center on the perspectives of the affected communities. In our case, this translates to designing studies where Chinese and Latino caregivers directly judge the cultural responsiveness of the AI agent’s support for them.

In summary, Gap 2 is the lack of contextualized, human-centered evaluation to measure cultural responsiveness in mental health conversational agents. This dissertation addresses this by directly eliciting user feedback on empathy, therapeutic alliance, and cultural relevance, and by comparing adapted vs. non-adapted systems in user studies, thereby generating evidence on the value of cultural adaptation in LLM-based therapeutic conversation context.

2.4 Summary of Key Contributions in Addressing Gaps in Culturally Responsive Mental Health AI

Health Informatics Contribution. This dissertation contributes a novel, generalizable workflow for operationalizing cultural adaptation in LLM-based mental health interventions. In contrast to traditional adaptation methods, the proposed workflow emphasizes real-time adaptation that aligns with the interactive nature of conversational agents. I integrated culturally grounded insights from end-users and community experts as contexts for AI-generated responses. Concretely, I implemented and evaluated two scalable, low-resource

techniques: (1) prompt-based cultural tailoring, and (2) dynamic context engineering. The context engineering method enables the delivery of culturally tailored mental health support incorporating human expertise (e.g., interview findings, culturally relevant context) and without requiring model fine-tuning or large-scale cultural datasets, making it accessible in low-resource settings (e.g., community clinics). We provide a framework that health informatics practitioners can adopt for other contexts where cultural adaptation is needed.

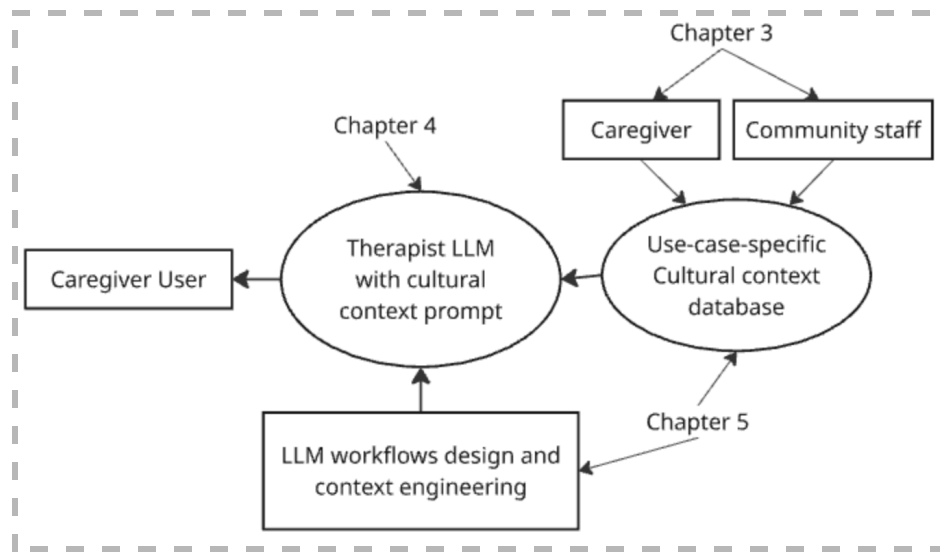


Figure 2.1. Overview of how each chapter contributes to the development of a prompt-based and a context engineering approach for culturally adapting mental health conversational agents.

Contextualized and human-centered evaluation contribution. This dissertation also explores evaluating mental health conversational agents through contextual, human-centered evaluation of cultural responsiveness. It describes work in which I designed studies that put the conversational agents in mental health use cases with actual stakeholders, family caregivers and community staff, and assessed outcomes such as empathy and working alliance that serve as strong indicators for health outcomes. This evaluation procedure represents an informatics contribution that evaluates not only whether LLMs possess cultural knowledge, but whether they can appropriately apply this knowledge in their responses to serve

diverse users. I adopted evaluation instruments such as Cultural Relevance Questionnaire and Cultural Competence Measure, for use with this new context with AI conversational agents. By linking improved cultural responsiveness to improved user-reported outcomes (e.g., higher empathy and alliance ratings), this work provides evidence of the value of cultural adaptation in digital health tools. Such evidence is crucial for justifying the integration of cultural considerations into design processes and, ultimately, for persuading healthcare providers and policymakers to support culturally adapted technologies that can help reduce health disparities.

CHAPTER 3. Understanding the Needs and Challenges of Chinese American Caregivers: Implications for Culturally Tailored Digital Mental Health Tools

In the work described in this chapter, I aimed at understanding the caregiving and self-care experiences of Chinese American family caregivers of older adults, with a focus on identifying culturally specific challenges and opportunities for digital mental health interventions. As part of the foundation for developing culturally adapted AI-powered mental health tools, this study explores how cultural values, caregiving norms, and systemic barriers shape caregivers' stress, help-seeking behavior, and openness to AI-driven support. The study involved semi-structured interviews with Chinese American family caregivers and focus groups with staff members from community organizations that support these families. Through inductive thematic analysis, the study surfaced key insights related to caregiving burden, cultural barriers to self-care, the potential role of digital tools, and perceptions of chatbots in mental health support. These findings provide important findings for the development work presented in Chapter 6.

In addition to the work presented below, my colleagues in the team who speak Spanish also conducted a parallel study with Spanish-speaking Latino family caregivers.

The following manuscript is submitted to the journal *The Gerontologist* and is currently under review. Identifying Opportunities for Technologies to Support the Well-being of Chinese American Caregivers Caring for Older Adults. Serena Jinchun Xie, MS, Jingyi Li, RN, PhD, Yanjing Liang, MS, Andrea Hartzler, PhD, Magaly Ramirez, PhD, Trevor Cohen, MBChB, PhD, and Weichao Yuwen, RN, PhD.

Abstract

Background and Objectives: Chinese American caregivers face unique challenges influenced by cultural values, yet few culturally tailored digital solutions exist. This study aimed to explore the challenges Chinese American caregivers face in caregiving, self-care, and mental health, identify opportunities for digital solutions, and understand their perspectives on using AI-driven tools for mental health support.

Research Design and Methods: We conducted 13 semi-structured interviews with Chinese American family caregivers and four focus groups with a total of 21 staff members from community organizations serving Chinese American families. Interviews were conducted in participants' preferred languages (English, Mandarin, or Cantonese). Participants shared caregiving challenges, self-care practices, and perspectives on digital tools. Transcripts were analyzed using inductive content analysis.

Results: Four themes emerged: (1) Caregiving challenges, including communication difficulties influenced by filial piety, older adult social isolation, and complex system navigation; (2) Caregivers' self-care, highlighting practices and barriers to prioritizing self-care and seeking mental health support; (3) Opportunities for digital solutions, including fostering connections among caregivers, providing trustworthy information, and enabling organizations to provide services; and (4) Perspectives on chatbots, appreciating privacy and non-judgment while expressing concerns about human connection and cultural understanding.

Discussion and Implications: Findings highlight the need for culturally sensitive digital tools that consider engagement strategies, incorporate practical resources, and align with cultural values and norms. Digital solutions, including AI-driven chatbots, hold promise for supporting Chinese American caregivers' mental health but should be designed with cultural adaptations that address unique barriers and incorporate cultural contexts.

3.1 Background and Objectives

The Asian American and Pacific Islanders (AAPI) population in the United States is rapidly growing and aging. Between 2010 and 2030, the older adult population among this group is projected to increase by 145% (Lim et al., 2020). Among AAPI, Chinese Americans are the largest subgroup, making up 22% of the Asian American population (USAFacts, 2023). As the Chinese American population ages, the need for caregiving is expected to rise. Studies have shown that about 50.4% of Chinese American older adults experience some levels of functional decline and depressive symptoms (Dong, Chang, et al., 2014; Dong, Chen, et al., 2014). As older adults experience declining health and independence, family caregivers play a vital role in providing care for older adults at home, offering physical, emotional, and social support (Schumacher et al., 2006).

Because of these barriers and gaps, culturally and linguistically appropriate interventions for Chinese American caregivers are needed. However, there has been a lack of culturally appropriate resources for this population (Li, 2004; Miyawaki, 2015). Linguistically appropriate, culturally responsive, and technology-enabled interventions could be the solutions to provide accessible and tailored support to this population. Studies have shown that technology-enabled interventions can improve care for the older adult population and their family caregivers by reducing caregiving stress, optimizing care time, and improving the quality of life for the care dyads (de-Moraes-Ribeiro et al., 2024). Despite the growing variety of mobile health applications, few are designed for Chinese American caregivers. Caregivers experience significant emotional, physical, and mental exhaustion, often leading to burnout. Yet, only a small number of mobile health applications include functions that address these critical needs, particularly self-care support, which is essential for maintaining caregivers' well-being (Grossman et al., 2018). To the best of our knowledge, there are only a couple of mobile health applications designed for Chinese immigrant caregivers with limited English proficiency to

engage in self-care (Hong et al., 2022; Liu et al., 2020; Yu et al., 2024), and none of them leverage the capability of artificial intelligence.

With the enhancement of technology, artificial intelligence (AI) has shown remarkable potential in areas such as tailoring treatments to individual patients (Alowais et al., 2023). The ability of AI to process vast amounts of data quickly and accurately is transforming diagnostics and treatment, offering new possibilities for individualized care. The potential of AI extends beyond diagnostics and treatment planning. AI-powered applications have several advantages over traditional mobile health applications, such as personalization (Esmailzadeh, 2020), including tailored language and cultural preferences (Dinesh et al., 2024). While AI-powered tools have demonstrated success in patient care (Alowais et al., 2023), little research has focused on how AI can be used to support caregivers in their roles and facilitate self-care, including in Chinese American caregivers of older adults.

This project aims to explore the needs of families caring for Chinese American older adults and identify opportunities to leverage digital health tools, including AI-driven technologies, to address the unmet needs. The findings will provide insights into unique cultural factors the older adults and caregivers experience that influence the successful adoption of digital health interventions and inform the design, development, and delivery of culturally sensitive digital health tools, such as chatbots, to support caregivers. Additionally, the implications of this study could guide future digital health tools tailored to culturally diverse caregiving populations. We aim to answer the following research questions:

Research question 1: What unique challenges do Chinese American caregivers of older adults face in caregiving, self-care, and mental health?

Research question 2: What opportunities exist for digital solutions to support Chinese American caregivers of older adults?

Research question 3: What are Chinese American caregivers' perspectives on using AI-driven tools, such as chatbots, for mental health and well-being support?

3.2 Study Design and Methods

Design

We conducted semi-structured individual interviews with Chinese American family caregivers caring for older adults and focus group interviews with staff members who have worked closely with these families. Including both groups allowed us to capture complementary perspectives: family caregivers first-person accounts of their lived caregiving experiences, including encountered challenges, self-care practices, and unmet needs, while professional staff provided observations of and insights into salient and recurrent challenges, common barriers, and current services gap informed by their work with multiple Chinese American families. Both stakeholder groups provided perspectives on the potential of digital health technologies, particularly chatbot interventions, for addressing caregiver support needs. This dual perspective enabled a more comprehensive understanding of the caregiving context and challenges, which is critical for designing culturally responsive interventions. The study was approved by the University of Washington's Institutional Review Board.

Data Collection

From October 2023 to February 2024, we worked closely with four local community organizations in the greater Seattle metropolitan area where the university resides, serving Chinese American families to recruit study participants. To recruit family caregivers, we (1) posted study flyers at community centers, (2) shared study information at community events, (3) social media groups organized by our community partners, and (4) had community partners refer potential participants to the study. Caregivers interested in participating contacted the

study lead, who screened for eligibility and scheduled interviews. Eligible family caregiver participants were at least 18 years of age, spoke English, Mandarin, or Cantonese, were living in the United States, self-identified as Chinese or Chinese American, and self-reported having current or previous experience caring for an older adult family member who had one or more chronic conditions. We also recruited staff members working with Chinese American families with older adults directly through the four community partners serving Chinese American families in the Greater Seattle area.

Following each interview, the research team member conducting interviews verbally administered a structured survey to participants to obtain sociodemographic characteristics about both themselves and their care recipients. The survey collected information on participants' and their care recipients' ages, length of time participants had lived in the United States, their English literacy (Karliner et al., 2008), health literacy (Chew et al., 2008; Wallace et al., 2006), and digital literacy (Roque & Boot, 2018), and self-reported frequency of depressed mood assessed by PHQ-2 (*Patient Health Questionnaire-2 (PHQ-2)*, n.d.). Semi-structured interviews were conducted virtually via Zoom or phone call in participants' preferred language. Family caregivers were asked about their caregiving challenges, current self-care practices, perspectives on using digital tools to address the caregiving challenges, and, more specifically, perspectives on using chatbots to help with self-care. To develop the interview guide, we reviewed the Cultural Treatment Adaptation Framework (Chu & Leino, 2017) and existing literature on Chinese American caregiver experiences (Miyawaki, 2020; Han et al., 2019; Spitzer et al., 2003; Lai, 2010; Tang, 2011; Halland et al., 2010). The research team, including an expert in cultural adaptation of evidence-based health interventions (Ramirez et al., 2023), collaborated with bilingual-bicultural research team members to develop the initial questions. The interview guide is in Appendix 1. Bilingual researchers later translated the guide into Chinese. The surveys and interviews were conducted in English or Chinese (Mandarin or Cantonese) based on the participant's preferred language.

Prior to each focus group, each staff participant attending the focus group completed a brief survey that included questions about their age, current role in the community organization, how long they have worked in their current role, and their level of education. Surveys were distributed and collected in person before the session began. During the in-person focus groups, staff member participants were asked about their observations on the unique challenges faced by Chinese American families with older adults, family caregivers' self-care practices and challenges, potential opportunities and barriers for mobile-based applications and specifically conversational agents to help Chinese American families with older adults, and how the unique challenges related to some Chinese cultural values. The focus group interview guide is in Appendix 1. The focus groups were conducted in English and Chinese based on the staff participants' preferred language.

Data Analysis

The interviews and focus groups were recorded and transcribed using the Notta.ai transcription platform for the following qualitative analysis. We conducted an inductive content analysis (Elo & Kyngäs, 2008), analyzing combined interview and focus group transcripts in their original language (Chinese or English) to preserve cultural nuances inherent in participants' expressions. First, three bilingual-bicultural research team members independently analyzed a subset of data comprising six individual interview transcripts and four focus group transcripts. Each researcher conducted in vivo coding (Saldaña, 2023), extracting participants' exact words and phrases as preliminary codes to preserve cultural nuances. Next, two team members collaboratively applied the initial codes to all remaining transcripts, iteratively refining codes and definitions and developing a shared understanding of emerging themes and an initial codebook. Then, four team members then engaged in a consensus validation process during weekly meetings over a two-month period, focusing on coding consistency and reliability,

thematic coherence, and the relationships among codes and themes. The consensus process continued until consensus was reached. Finally, the first author translated all themes and codes, which were reviewed and verified by the bilingual team members.

3.3 Results

Participant Characteristics

The study included 34 participants, including 13 family caregivers (P1-P13) and 21 staff participants (S1-S21) recruited from four local community organizations. Among the caregiver interviews, nine were conducted in Mandarin, three in English, and one in Cantonese. We conducted four focus groups with the staff participants. Most staff participants were bilingual (95.2%), and focus groups were conducted in English and Chinese, depending on the staff participants' preferred language, to ensure clear communication. The average age of caregiver participants was 58.1 (std=13.9), ranging from 43 to 82 years old. The average age of the care recipients is 80.3 (std=4.5), ranging from 74 to 85 years old. The majority of caregivers were adult children of the care recipients and were currently homemakers. There was a high variation in the number of years the caregivers lived in the U.S., with an average of 32.0 years (std=23.3, 4-65 yrs). Caregivers had various levels of English proficiency and health literacy, relatively high digital literacy, and high education and socioeconomic status. On average, caregivers reported very low levels of depressive symptoms with an average PHQ-2 of 0.58 (std=0.71, 0-2). Staff participants held various roles in their organizations, with work experiences in their current roles ranging from 1 to 31 years (mean=7.8, std=9.5). The majority of staff participants had a Bachelor's degree in social work. More details can be found in Table 1.

Table 1. *Participants characteristics*

Caregivers characteristics		Staff characteristics	
N = 13 (n, %)		N=21 (n, %)	
Relationship with care recipients		Current position in the organization	
Adult children	10, 76.9%	Program directors/managers	5, 23.8%
Spouse	2, 15.4%	Case manager/Social worker	9, 42.9%
Other (neighbor)	1, 7.7%	Home care specialists	2, 9.5%
How well do you speak English?		Registered nurses	4, 19.0%
Very well	4, 30.8%	Advanced practice provider	1, 4.8%
Well	5, 38.5%	Education	
Not well	1, 7.7%	Bachelor's degree	14, 66.7%
Not at all	3, 23.1%	Master's degree	4, 19.0%
Health literacy: How confident are you in filling out medical forms by yourself?		Doctorate degree	3, 14.3%
Extremely	1, 7.7%		
Quit a bit	4, 30.8%		
Somewhat	3, 23.1%		
A little bit	1, 7.7%		
Not at all	4, 30.8%		
Digital literacy Q1: Using a mobile device, I can use the onscreen keyboard to type			
Very easily	7, 53.8%		
Somewhat easily	3, 23.1%		
Not very easily	1, 7.7%		
Not at all	2, 15.4%		
Digital literacy Q2: Using a mobile device, I can find information I need on the Internet			
Very easily	6, 46.2%		
Somewhat easily	4, 30.8%		
Not very easily	1, 7.7%		
Not at all	2, 15.4%		
Education			
Some high school	1, 7.7%		
High school or GED	1, 7.7%		
Bachelor's degree	9, 69.2%		
Master's degree	2, 15.4%		
Annual household income			
Less than \$19,999	1, 7.7%		
\$20,000-\$39,999	2, 15.4%		
\$60,000-\$79,999	1, 7.7%		
Over \$100,000	9, 69.2%		
Employment status			
Full-time	2, 15.4%		
Part-time	3, 23.1%		
Stay-at-home	8, 61.5%		

Caregiver participants described caring for older adults with a diverse spectrum of health conditions, including chronic pain (e.g., chronic back pain, knee pain), cardiovascular conditions (e.g., hypertension), post-surgical recovery needs, and dementia. The majority of the participants described caring for older adults managing multiple concurrent health conditions. Although systematic health condition data collection was not incorporated into the survey, participant disclosures during interviews revealed substantial heterogeneity in care complexity, resulting in varied caregiving demands and support requirements across the participants. Another characteristic that was not captured in the survey but came up in the interview was the caregiver participants' residential locations. Although recruitment was conducted through local community organizations within the Seattle metropolitan area, three participants accessed study information through informal social networks (family and friends) and were subsequently identified as residing outside Washington State. The remaining participants resided within the Seattle metropolitan area.

Themes and Subthemes Identified

We identified four themes, as shown in Table 2. Under the theme “Caregiving Challenges,” we outlined several challenges caregivers face. The theme “Caregivers’ Self-Care” highlighted current self-care practices among caregivers. It also detailed significant barriers to prioritizing self-care as well as seeking mental health support. Under the theme “Opportunities for Digital Solutions,” we explored how caregivers and staff envisioned digital tools addressing their needs. Lastly, the theme “Perspectives on Chatbots to Support Mental Health” examined caregivers’ openness to using chatbots to support their mental health and well-being. While participants appreciated chatbots’ private and non-judgmental nature, they raised concerns about the lack of human connection and trust and the importance of cultural understanding in fostering meaningful support.

Table 2. Themes and subthemes.

Themes	Subthemes
Caregiving challenges	<i>Communication challenges in filial caregiving</i>
	<i>Older adult social isolation</i>
	<i>Navigating healthcare and social service systems</i>
Caregivers' self-care	<i>Current self-care practices</i>
	<i>Barriers to prioritizing self-care</i>
	<i>Barriers to seeking mental health support</i>
Opportunities for digital solutions to support caregivers	<i>Fostering connections with other Chinese American caregivers</i>
	<i>Providing trustworthy information</i>
	<i>Enabling organizations to provide services</i>
Perspectives on chatbots to support mental health	<i>Providing private and non-judgmental support</i>
	<i>The need for human connection and trust</i>
	<i>The importance of cultural understanding</i>

Theme 1. Caregiving challenges

Communication Challenges in Filial Caregiving

A prominent theme among caregivers was the psychological burden of navigating communication with their older adult parents. Many caregivers found it challenging to discuss important caregiving topics such as healthy lifestyles and caregiving options due to the deeply ingrained filial piety and family hierarchy, which shaped caregiving dynamics and placed caregivers in a difficult position, balancing respect for their parents with the need to address practical caregiving concerns. As one caregiver expressed, "The main thing I feel about caring for my mom is that it's psychologically stressful... she doesn't listen to you... we have a strong tradition of filial piety, which means it is hard to be disobedient to our parents." (P6)

Participants highlighted that cultural norms rooted in filial piety, such as the expectation that children take care of their aging parents, also influenced older adults' rejection of services like professional caregiving and senior living facilities. As one caregiver shared, "The traditional

concept of three or even four generations under one roof is something that Chinese people aspire to... It's hard for them [older adult parents] to accept the idea of living in a senior living." (P10)

Older Adult Social Isolation

Social isolation among Chinese older adults emerged as a significant concern for caregivers. Factors such as living in neighborhoods with few Chinese families, language barriers, and health conditions limited opportunities for older adults to engage socially. This isolation not only affected the older adults' emotional well-being but also strained caregivers. As one participant explained, "In terms of mental and emotional aspects, they [older adult parents] have no friends or other elderly people to talk to... In this situation, they feel bored and pessimistic, which inevitably affects me." (P6) Many caregivers expressed feelings of guilt, feeling responsible for their parents' isolation since they had facilitated their relocation to the U.S. They also voiced frustration at being unable to find meaningful ways to improve their parents' social connectedness and well-being.

Navigating Healthcare and Social Service Systems

Caregivers and staff participants identified significant challenges in navigating the U.S. healthcare and social service systems. Even caregivers fluent in English reported feeling overwhelmed by the complexity of these systems and unaware of available resources. Staff participants noted that one of the most frequent requests from family caregivers was assistance with navigating these systems and applying for services. For Chinese caregivers, these challenges were further compounded by cultural barriers.

Theme 2. Caregivers' self-care

Current Self-Care Practices

Caregivers reported engaging in various self-care activities, including physical, leisure, social, and spiritual practices. A few participants recognized the psychological benefits of physical activities, such as Tai Chi, golf, and hiking. Leisure activities, such as reading, shopping, crafting, and cooking, were frequently mentioned, along with social activities, such as playing Mahjong with friends and volunteering at community centers. As one caregiver stated, "I think... just being able to find out where those connections are has given me a lot of pleasure in appreciating my life." (P4)

Spiritual practices were less commonly reported. Two caregivers mentioned participating in a spiritual study group and attending church, highlighting the benefits of positive energy and mindset. Professional mental health support was rarely sought. Only one caregiver participant shared seeing a therapist and found it helpful. Another caregiver reported attending a weekly well-being talk series tailored to the Chinese American community.

Barriers to Prioritizing Self-Care

Almost all caregivers linked their well-being to solving immediate problems, believing that addressing immediate challenges would naturally improve their mental health. For example, one caregiver who cared for her father with dementia noted, "Improving my mood feels useless since the reality would still be there." (P7)

Many caregivers found it difficult to practice self-care as the cultural value of "self-restraint" led them to view self-care almost as indulgent, especially when caregiving responsibilities loomed large. Staff participants noted that Chinese caregivers frequently experience guilt, feeling it is their responsibility to take on all caregiving tasks alone.

Additionally, staff participants noted that immigrant caregivers face unique challenges. Navigating a new society without a support network often leaves them little capacity for self-care. A staff participant, who was also a family caregiver, shared, “As an immigrant and a caregiver, I don’t have any relatives in the U.S. who can assist me. I have to do everything alone with no helping hands.” (S20)

Barriers to seeking mental health support

Awareness of mental health issues was limited among Chinese American caregivers. One staff participant with over three decades of experience providing consultation and care planning for Chinese American families shared an example, “Sometimes they [caregivers] are depressed, but they don’t recognize it at all. Once, a client of mine [a caregiver] filled out a survey, which showed that she had severe depression, but she had no idea that she was depressed and she had no plan to seek help.” (S7) Many other staff participants echoed this observation.

From caregivers’ perspectives, several mentioned that they often questioned whether their mental health concerns warranted action. As one caregiver stated, “I definitely realize there are times when I feel depressed... but it doesn’t seem to reach a point where I feel strongly enough to take action.” (P2) Caregivers and staff participants recognized that this barrier might stem from “the culture of endurance,” a cultural narrative that valorizes resilience and endurance through adversity and hardship, contributing to viewing seeking mental health help as a sign of weakness.

Theme 3. Opportunities for digital solutions to support caregivers

Fostering connections with other Chinese American caregivers

Caregivers frequently discussed the opportunity for digital solutions to help them connect with other caregivers within the Chinese American community. For those living in areas with small Chinese-speaking communities, digital solutions were seen as promising tools to connect with

other caregivers who share similar cultural values and challenges. One caregiver caring for her father with dementia highlighted the importance of culturally specific networks, saying she wished for a platform where she could learn from other dementia caregivers within the Chinese American community and exchange advice. Such solutions could address the psychological strain and foster emotional and practical support among caregivers.

Providing informational support

Both family caregivers and staff members identified a significant need for immediate, trustworthy, and comprehensive information tailored to the challenges of caregiving. These include navigating policies, insurance, social services, public assistance programs, and caregiving knowledge. Participants expressed interest in digital solutions, such as apps or platforms, that could offer accurate information in real time. Features like a Q&A function and culturally tailored content were particularly appealing. These tools could help caregivers overcome challenges by providing accessible, on-demand guidance.

Enabling organizations to provide services

From the perspective of community organizations, staff participants shared that digital solutions could be integrated into their current workflow. Tools like apps with translation functions could streamline their daily tasks, such as filling out forms and reducing administrative burdens for both staff and caregivers. Moreover, staff participants noted the gap between in-person visits. A digital platform could facilitate more frequent check-ins with the families, addressing their concerns in real time and mitigating the gaps in service delivery. These features could also open pathways for introducing caregivers to digital mental health interventions, such as chatbots, to support their emotional well-being.

Theme 4. Perspectives on chatbots to support caregivers' mental health and well-being

Providing private and non-judgmental support

Both caregivers and staff participants highlighted the advantage of chatbots to provide a form of mental health support free of potential judgment from a human. Many family caregivers noted that they would feel comfortable discussing personal challenges with a chatbot, particularly when concerns might be judged by others as failing to meet expectations of filial piety. As one caregiver said, "Sometimes you can't say something to others. You don't want others to know. This [i.e., talking to a chatbot] will definitely help." (P5) Several staff participants echoed this, noting that chatbots could serve as a discreet and private option for Chinese caregivers who might otherwise feel hesitant to discuss their struggles openly, which could potentially encourage more caregivers to seek emotional support.

The need for human connection and trust

Despite the advantage of reducing the fear of judgment, participants voiced a few concerns about the lack of human connection and emotional nuance. Staff participants, particularly those with extensive experience supporting Chinese American caregivers, stressed the importance of human connection. One participant noted, "It's not merely having a listener. It's about having a facial expression reaction. They [i.e., caregivers] need some reactions beyond words, which is difficult for a machine." (S2)

Additionally, staff participants highlighted the need for building trust, suggesting that caregivers might hesitate to open up emotionally without a pre-established trusting relationship, which is typically built through multiple human interactions. Caregivers also expressed skepticism about whether a chatbot could replicate the sense of care provided by a human. As

one caregiver remarked, “It is just a feeling... I would not feel good if I knew it was a robot on the other side doing the caring work.” (P10)

The importance of cultural understanding

A recurring concern among caregivers was whether chatbots could demonstrate “cultural understanding” and provide empathetic responses grounded in shared cultural contexts. Participants emphasized that connecting with someone who shares their cultural background, without needing to explain cultural nuances, was essential for fostering a sense of understanding and emotional connection. One caregiver shared her experience with her husband, who was not a Chinese American: “For example, if I have to tell my husband, he doesn’t understand why I can’t talk to my mom directly. I feel that it’s important when we chat with someone, if we have the same background, I don’t need to explain, and they understand. I don’t know what background or context a chatbot will have.” (P6) Many caregivers stressed that cultural context plays a vital role in feeling emotionally connected and supported, raising doubts about whether chatbots could replicate the depth of understanding required for meaningful interactions.

3.4 Discussion

Our findings highlight the challenges, some unique to Chinese American caregivers, and the potential of digital solutions to address their needs. Cultural values such as “filial piety” and “endurance” significantly shaped caregivers’ experiences, influencing communication dynamics, caregiving decisions, and willingness to seek support. These cultural norms often created psychological strain, particularly when caregivers felt obligated to meet traditional expectations while navigating the complexities of caregiving in the U.S. system. Social isolation among older adults and lack of support networks further compounded these challenges, leaving caregivers

with limited capacity for self-care. Regarding these difficulties, caregivers and staff identified opportunities for digital tools to provide informational support, facilitate connections within culturally specific communities, and assist community organizations in delivering services. While participants appreciate the private and non-judgemental nature of using chatbots to get support for mental health and well-being, they expressed concerns about the human connection, trust-building, and cultural sensitivity of chatbot solutions, underscoring the need for culturally tailored approaches to digital caregiving support.

While previous research identified the language barrier as a critical obstacle for caregivers to access and utilize resources (Knipping et al., 2023), our findings indicate that even caregivers proficient in English often face difficulties navigating available resources due to the complexity of these systems. This underscores the need for information and system navigation support to improve service utilization among Chinese American caregivers. The challenge is further compounded by cultural expectations that influence the acceptability of certain services, such as senior living, within Chinese American families. Prior studies consistently pointed to the significant role of filial piety in caregiving among Chinese and other East Asian cultures (Guo et al., 2019; Miyawaki, 2020). While this value can serve as a driving force for caregivers to deliver exceptional care (Kim, 2022), it can also result in communication challenges for adult children caregivers caring for their parents.

Additionally, the cultural value of filial piety results in feelings of guilt and inadequacy when caregivers are unable to meet the emotional and social needs of their parents. Contributing to this caregiving challenge is the prevalence of social isolation and loneliness among Chinese older adults (Simon et al., 2014). While having certain health conditions (Mayo et al., 2021) and limited English proficiency (Jang et al., 2022) contribute to being more prone to social isolation, our study highlights the additional barrier for older adults living in neighborhoods with a relatively small Chinese American population, where connecting with

individuals with the same cultural background presents a challenge. The isolation of older adults also takes an emotional toll on their caregivers, which is supported by previous research (Mayo et al., 2021). Digital tools could serve as a platform to facilitate peer support and foster social connections for caregivers and older adults with similar cultural backgrounds. Moreover, given the interest of community partners in using digital tools to better serve their clients, future digital tools could also consider including features that facilitate communication between caregivers and local community organizations.

In the study, participants shared their experiences and observations with self-care and their perspectives on using chatbots for mental health and well-being support. Although this study is not about adapting a specific evidence-based intervention, prior cultural adaptation frameworks (e.g., Chu & Leino, 2017) offer useful concepts for thinking about engagement, modifications to key elements, and delivery considerations.

Engagement considerations. Our findings revealed that Chinese American caregiver participants often struggle with limited awareness of mental health issues, influenced by cultural narratives that emphasize resilience and endurance. This is also discussed in previous literature as a strong sense of self-managing (Li, 2004; Napoles et al., 2010). Taken together these may limit the likelihood of engagement with an intervention intended to support mental health. Overcoming these barriers to mental health support may need increased effort in outreach and psychoeducation, such as establishing referral channels through trusted local community organizations and delivering culturally sensitive psychoeducation in Chinese to educate families in the community about mental health and destigmatize mental health care. Additionally, caregivers shared their view of mental health challenges as personal weaknesses rather than as conditions requiring professional support. Addressing the tendency toward self-judgment could be a crucial step in engaging caregivers in mental health and well-being interventions. However, it is also true that the participants recognized reducing fear of judgment as a major advantage of

using chatbots. This suggests that chatbots can support mental health and well-being in this population without the fear of being judged or viewed as ‘weak’ which may manifest in an encounter with a human clinician or peer.

Adapting core components. It became clear that the caregivers in this study viewed accessing tangible services and resolving practical issues as closely intertwined with their mental health, prioritizing them over their self-care and well-being. This highlights the necessity of incorporating information about available resources as part of the digital mental health and well-being tool for Chinese American caregivers, especially for first-generation immigrants, who may be unaware of available resources.

Treatment delivery considerations. In our findings, Chinese American caregivers engaged in various self-care practices, including all five structures of the self-care framework (Lee & Miller, 2013): physical, social, leisure, psychological and emotional, and spiritual. It is important to recognize that some activities (e.g., Tai-Chi and Mahjong) are more culturally salient and include them in the digital mental health solution to improve the cultural relevance of the intervention while also recognizing the diversity of self-care practices and not being confined by a few stereotypical ones. In our study, spiritual practices were seldom mentioned, and psychological and emotional self-care practices were only mentioned by one caregiver who found seeing mental health therapists significantly beneficial to improving her psychological well-being. There is a gap in this population in engaging in spiritual and psychological self-care practices despite their being proven to be effective in promoting well-being (Manning et al., 2019). This presents an opportunity for the incorporation of such practices within components of digital mental health solutions.

Additionally, the study participants revealed that cultural norms and values are deeply embedded in their caregiving challenges and self-care barriers; therefore, they were skeptical about conversational agents’ capacity to incorporate cultural contexts to comprehend their

struggles and offer a level of empathetic understanding comparable to humans. For example, in our study, participants shared that Chinese American caregivers tend to view self-care activities as an indulgence. This mindset is particularly prevalent among immigrant caregivers who often lack a support system. It would be important for a mental health support chatbot to recognize this lack of support and understand the feeling of indulgence. This concern of lack of cultural sensitivity is supported by studies, which have found that AI-based conversational agents can leave the impression of reflecting predominant Western perspectives (Havaladar et al., 2023; Johnson et al., 2022), lacking understanding of cultural contexts (Song et al., 2024) and being inaccurate in culture-specific values (Arora et al., 2023). Therefore, it's crucial to develop culturally tailored mental health chatbots that incorporate culturally relevant examples and themes and use tailored language (e.g., using cultural sayings). This can potentially improve the therapeutic alliance between the chatbot and users, which is crucial to a successful therapeutic session (Stern et al., 2024).

Limitations. This study has several limitations. One limitation is that the majority of the caregivers interviewed were Mandarin-speaking, and only one caregiver was Cantonese-speaking. This may limit the generalizability of the findings to the Cantonese-speaking population. Another limitation is the high heterogeneity of chronic conditions among the care recipients that the caregivers care for, leading to high variability in the caregiving experiences. While dementia requires intensive caregiving and high dependency on the caregiver, some chronic conditions, such as hypertension, allow older adults to maintain high levels of independence. Additionally, the majority of caregiver participants were recruited through local community organizations in the Seattle metropolitan area, potentially introducing a selection bias, as these individuals already have actively looked for and have access to community resources. This recruitment method may have contributed to the inclusion of a participant group with relatively high socioeconomic status and high education level, which may

not fully represent the experiences of caregivers with lower socioeconomic status, who might face additional caregiving challenges.

3.5 Conclusion

The study identifies both challenges faced by Chinese American caregivers regarding their caregiving and self-care and opportunities for digital solutions, such as chatbots, to support this population. The findings underscore the need for culturally sensitive mental health and well-being solutions that account for necessary cultural adaptations in user engagement, the inclusion of information on tangible resources, and the delivery of interventions that incorporate cultural contexts.

CHAPTER 4. Cultural Prompting Improves the Empathy and Cultural Responsiveness of GPT-Generated Therapy Responses

This chapter presents the first empirical evaluation of a lightweight, prompt-based cultural adaptation technique for large language model (LLM)-driven therapeutic responses. Addressing the need for scalable and culturally responsive AI mental health tools, this study investigates whether cultural prompting—a method of embedding culturally relevant instructions in the system prompt—can enhance the cultural responsiveness and perceived empathy of AI-generated responses. The evaluation focuses on two LLMs, GPT-4o and DeepSeek-V3, prompted in Chinese and assessed by Chinese-speaking Chinese American caregivers. The results presented in this chapter inform both the technical feasibility and design considerations for developing culturally adaptive AI mental health tools.

In addition, Section 4.2 presents a parallel study applying the same prompt-based cultural adaptation approach in Spanish for Spanish-speaking Latino family caregivers in the U.S. In this evaluation, two LLMs, GPT-4o and DeepSeek-V3, were prompted in Spanish and assessed by Spanish-speaking Latino caregivers. As in the Chinese-language study, the analysis examines whether embedding culturally specific cues enhances both the cultural responsiveness of the responses and perceived empathy. The Spanish-language study is ongoing, and this dissertation presents the data collected to date. Together, the two studies offer complementary evidence on the effectiveness of this lightweight, prompt-based cultural adaptation approach for AI mental health support.

4.1 Chinese Version

A version of the following work is accepted by American Association for Medical Informatics (AMIA) as an open-access article. © AMIA.

Xie, S. J., Zhai, S., Liang, Y., Li, J., Fan, X., Cohen, T., & Yuwen, W. (2025). Cultural prompting improves the empathy and cultural responsiveness of GPT-generated therapy responses. In Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2025). (In press).

Abstract

Large Language Model (LLM)-based conversational agents offer promising solutions for mental health support, but lack cultural responsiveness for diverse populations. This study evaluated the effectiveness of cultural prompting in improving cultural responsiveness and perceived empathy of LLM-generated therapeutic responses for Chinese American family caregivers. Using a randomized controlled experiment, we compared GPT-4o and Deepseek-V3 responses with and without cultural prompting. Thirty-six participants evaluated input-response pairs on cultural responsiveness (competence and relevance) and perceived empathy. Results showed that cultural prompting significantly enhanced GPT-4o's performance across all dimensions, with GPT-4o with cultural prompting being the most preferred, while improvements in DeepSeek-V3 responses were not significant. Mediation analysis revealed that cultural prompting improved empathy through improving cultural responsiveness. This study demonstrated that prompt-based techniques can effectively enhance the cultural responsiveness of LLM-generated therapeutic responses, highlighting the importance of cultural responsiveness in delivering empathetic AI-based therapeutic interventions to culturally and linguistically diverse populations.

4.1.1 Introduction

Large language model (LLM)-based conversational agents (chatbots) have emerged as promising tools to bridge mental health service gaps by providing cost-effective and on-demand therapeutic support (Yasukawa et al., 2024), offering immediate emotional support, stress management strategies, and culturally relevant coping techniques (Borna et al., 2024). However, research shows that many LLMs face critical challenges in cultural alignment, often misinterpreting culturally specific expressions or providing advice that does not align with the traditions, beliefs, and family dynamics of diverse communities (Dinesh et al., 2024). In particular, studies have demonstrated that LLM-generated responses tend to align more closely with Western cultural values than those from non-Western backgrounds, which can result in responses that feel culturally insensitive or irrelevant to users from minority communities (Hershcovich et al., 2022), making it difficult to be effectively used in mental health care for diverse populations (Tao et al., 2024). There is a critical need for culturally responsive LLM-based mental health interventions.

Culturally responsive therapy is the approach that considers the interaction between a client's presenting problem and their culture (Jones-Smith, 2018). Cultural responsiveness goes beyond the cultural sensitivity of therapists. It consists of understanding and respecting cultural differences and actively adapting and tailoring one's approach to ensure relevancy, acceptability, and effectiveness for the client's cultural context (Eustis et al., 2023). The importance of cultural responsiveness in mental health care has been emphasized for decades (Sue & Zane, 1987), with evidence indicating that cultural insensitivity in therapy can lead to a lack of trust between patients and providers and alliance ruptures (Tschuschke et al., 2022). Empathy, a core mechanism of building therapeutic alliances and subsequent changes, requires cultural contextualization to be effectively conveyed in clinical interactions. Elliott et al. (2018) highlighted that empathy strongly predicted positive therapy outcomes. However, without

cultural responsiveness, any therapeutic interaction may fail to recognize a client's cultural background, and empathy can be misunderstood or perceived as a superficial approach, leading to reduced engagement and less effective treatment (Ullrich, 2019). We believe that cultural responsiveness has the potential to deepen empathy by making therapeutic interactions more contextually relevant and resonant with clients, deepening the sense that clients feel heard and supported.

Cultural adaptation of psychosocial interventions originated with the application of the ecological validity theory (Bernal et al., 1995) to adapt interventions to specific cultural patterns, meanings, and values. Traditional adaptation methods involve an iterative and collaborative process that includes input from individuals within the target populations (Dinesh et al., 2024; Spanhel et al., 2021). However, this presents challenges for AI-driven mental health interventions. Unlike conventional interventions with static and manually written content, LLM-driven chatbots generate dynamic responses in real-time, rendering traditional expert-guided cultural adaptation approaches inapplicable (Spanhel et al., 2021). To date, there is a gap in the literature on methodologies for culturally adapting LLM-driven mental health chatbots. Given the dynamic nature of LLM responses, new methodologies are required to ensure that AI-generated content is culturally responsive.

Previous research has shown that merely prompting in the language of the intended culture is not sufficient for cultural adaptation (Arora et al., 2023). Subsequent approaches have attempted to modify model behavior using more explicit methods. There are two general directions of model adaptation: prompt-based methods (e.g., prompt tailoring) and weight-changing methods (e.g., fine-tuning). For example, Solaiman and Dennison fine-tuned LLMs on a value-targeted dataset (Solaiman & Dennison, 2023), and Nguyen et al. (2024) continued pre-trained LLMs to capture more Southeast Asia cultural norms, customs, and stylistic preferences. Non-weight-changing methods rely on prompts to modify model behavior

and require fewer resources compared to weight-changing approaches. One example is the cultural prompting technique, which incorporates cultural contexts as part of prompts (e.g., “You are an average human being born in [country/territory] and living in [country/territory] responding to the following survey question.”) (Tao et al., 2024). Tao et al. (2024) demonstrated that cultural prompting can effectively reduce cultural bias in GPT. In another work that quantified the cultural alignment of LLMs, GPT-4 showed the capability to adapt to cultural nuances in response to an explicit cultural context prompt, particularly in Chinese settings (Masoud et al., 2025). However, other research yielded mixed results on the effectiveness of cultural prompting (Cao et al., 2023). Beck et al. (2024) questioned the readiness of prompt-based methods by highlighting the lack of robustness, being highly sensitive to prompt formulation, model choice, and specific tasks performed. Mukherjee et al. (2024) observed that model responses varied significantly for both culturally sensitive and non-sensitive prompts, suggesting that cultural conditioning may be a ‘placebo effect’ rather than actual cultural adaptation.

While prior studies have investigated the cultural biases embedded within LLMs—including their alignment with cultural values (Arora et al., 2023; Masoud et al., 2025), interpretations of cultural proverbs (Liu et al., 2024), and knowledge of cultural norms (Shi et al., 2024; Rao et al., 2025; Zhou et al., 2025)—these works have primarily focused on probing model characteristics rather than evaluating their effectiveness in real-world therapeutic interactions. The most commonly used tools for probing cultural alignment were Hofstede’s Value Survey Model (VSM) (Hofstede et al., 2010) and World Value Survey (WVS) (Ronald et al., 2002) for quantifiable evaluation. VSM and WVS are two influential frameworks for cross-cultural analysis, where VSM categorizes national cultures in six dimensions and WVS is a global research effort that maps cultural values through representative population sampling. Studies then probed the cultural alignment of LLMs by reformulating the survey questions from WVS and VSM into cloze-style (fill-in-the-blank) questions, analyzing the models’ predicted

word probabilities, and comparing the predictions with established cultural value survey results (Arora et al., 2023; Masoud et al., 2025). A recent survey provided a comprehensive overview of cultural alignment methodologies, benchmarks, and evaluation approaches (Pawar et al., 2024). However, for healthcare, cultural responsiveness should be situated within the context of treatment delivery (Bernal et al., 1995). Yet, to our knowledge, no study has systematically evaluated how an AI-driven mental health chatbot's cultural responsiveness affects user experiences in treatment settings or with human-centered evaluation approaches. As Blodgett et al. (2024) highlighted, NLP models are ultimately integrated into real-world applications that directly impact individuals' lives. In the context of mental health care, evaluation methods should adopt a human-centered perspective, prioritizing users' needs, values, and interaction behaviors to ensure AI-driven interventions have meaningful and desirable real-world effects.

The current work addresses these gaps in the literature by improving and evaluating the cultural responsiveness of LLM-generated responses in a therapy-delivery context. To this end, we used a prompt-based technique, cultural prompting, to adapt the responses of an AI therapist for Chinese American family caregivers. Asian Americans are the fastest-growing racial and ethnic minority group in the United States, with approximately 67% of Asian adults being immigrants (Budiman et al., 2024). Among Asian American subgroups, Chinese Americans represent the largest part, making up 24% of the Asian American population (Ruiz & NG, 2021). Studies indicate that Chinese American caregivers experience higher levels of psychological distress but remain underserved by mental health services due to reasons such as a lack of linguistically appropriate services (Hong et al., 2024). Our objectives were to conduct a comparative assessment of culturally adapted versus baseline LLM-generated interactions and to assess (1) how effective cultural prompting is in improving the cultural responsiveness of AI therapist responses and (2) whether improving cultural responsiveness of these responses also improves perceived empathy. Our study aims to evaluate the hypothesis that a culturally responsive AI therapist will also be perceived as more empathetic, which could lead to better

clinical outcomes (Figure 1). This research addresses a critical gap by introducing and evaluating LLM-based methodologies for real-time cultural adaptation in AI-driven mental health applications. By promoting the cultural responsiveness of an AI therapist, we seek to make mental health care more inclusive, engaging, and effective. Findings from this study will help advance the development of culturally responsive AI-driven mental health interventions and improve their accessibility and effectiveness for diverse caregiving populations.

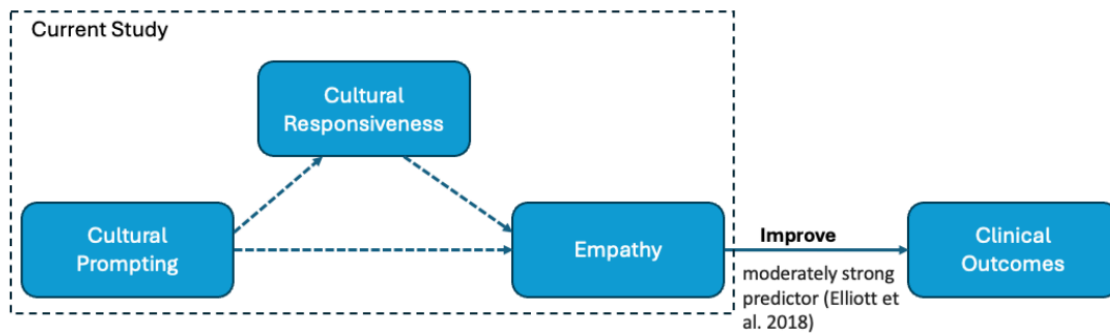


Figure 4.1. Illustration of the hypothesis tested in the current study and the relationship with clinical outcomes.

4.1.2 Methods

We conducted a within-subject experiment in the following steps: (1) curating high-quality, culturally relevant example user inputs, (2) generating LLM responses using two LLMs with and without cultural prompting techniques (2×2 design), and (3) a blinded, randomized mixed-methods assessment of input-response pairs across four model configurations.

Cultural Representative Input Development and Response Generation

We curated a set of example family caregiver inputs by working with six community partners who have rich experience working with Chinese American family caregivers and are familiar with the challenges caregivers face. Through weekly meetings, we developed six personas of Chinese caregivers. Each community partner generated four example client inputs in simplified or traditional Chinese per persona, following detailed guidelines. In total, we collected 144

examples of caregiver inputs in Chinese. After that, the research team selected 45 inputs with rich cultural contexts for response generation. Inputs were excluded if they were too general (e.g., “I am too stressed”), not culturally salient (e.g., “Every time I see the hospital bills, I’m getting a headache”), or similar to an already selected input.

We utilized GPT-4o and DeepSeek-V3 to generate responses to the selected caregiver inputs, using system prompts tested in a previous study (Filienko et al., 2024). Our model selection was motivated by variations in model characteristics (e.g., open-source vs. closed-source), performance metrics, and potential cultural competencies. GPT-4o, a closed-source model developed by OpenAI, represents the current benchmark in general language capabilities (OpenAI, 2024) and has demonstrated robust performance across multilingual contexts (Busch et al., 2025). For comparison, we selected DeepSeek-V3 as the open-source model. We chose DeepSeek-V3 because it is trained on a multilingual corpus, predominantly consisting of English and Chinese text. DeepSeek-V3 outperforms other open-source models, such as LLaMA-3, across most benchmarks while achieving comparable performance to leading closed-source models like GPT-4o and Claude-Sonnet-3.5. Notably, DeepSeek-V3 exceeds both GPT-4o and Claude-Sonnet-3.5 in Chinese factual knowledge benchmarks, particularly Chinese SimpleQA (DeepSeek-AI et al., 2025). This deliberate comparison of models enables examination of whether models with different training paradigms yield varying degrees of cultural responsiveness with and without explicit cultural prompting.

Moreover, previous research has demonstrated that using English prompts could homogenize cultural differences in the responses from LLMs (Cao et al., 2023), suggesting the importance of constructing prompts directly in the target language. Therefore, all system prompts were translated into Chinese by the authors. For each model, we used two prompt versions: with and without cultural prompting. The prompts with cultural prompting included the following sentence (translated into Chinese) as the first sentence in the system prompt:

“When responding to caregivers, please fully consider their Chinese cultural background. Use language, values, and communication styles that align with Chinese culture to ensure your message resonates with them.” The rest of the prompt was the same for both versions. This design yielded four distinct response versions (Table 1) for each user input, allowing for a systematic comparison of cultural competency across prompting techniques and models.

Table 4.1. Four model configurations using two different LLMs with and without cultural prompting.

Model id	LLM	With Cultural Prompting Techniques?
1	GPT-4o	No
2	GPT-4o	Yes
3	Deepseek-V3	No
4	Deepseek-V3	Yes

Recruitment and Data Collection

We employed a combination of snowball sampling and targeted outreach to recruit participants. We distributed study flyers via social media platforms and recruited participants at two adult daycare centers primarily serving Chinese Americans. All evaluations were conducted through the Qualtrics online surveys. Each participant was instructed to assess four sets of five input-response pairs, randomly selected and with presentation order randomized to mitigate sequence effects. Participants evaluated each set on cultural responsiveness and perceived empathy.

We used two complementary measures to assess cultural responsiveness:

1. The **Healthcare Provider Cultural Competency Measure (CCM)** is a validated patient-reported instrument that captures patients’ perceptions of their providers’ cultural competency (Lucas et al., 2008). This 9-item evaluation assesses three critical

domains—cultural knowledge, awareness, and skill. This scale has established reliability (Cronbach’s alpha 0.68–0.98) in previous studies (Dawadi et al., 2022; Alizadeh & Chavan, 2020; Nápoles et al., 2012). To minimize participant burden while maintaining construct validity, we selected a representative item from each domain plus one overall assessment item, preserving the theoretical structure while optimizing completion time. The internal consistency reliability of the four CCM items was assessed using Cronbach’s alpha (Cronbach, 1951), showing excellent internal consistency for the selected CCM items ($\alpha = 0.96$).

2. The **Cultural Relevance Questionnaire (CRQ)** was informed by the literature on cultural sensitivity (Resnicow et al., 1999) and ecological validity theory (Bernal et al., 1995). The CRQ has demonstrated utility in assessing internet-delivered therapeutic interventions (Salamanca-Sanabria et al., 2018) and low-intensity psychological approaches (Perera et al., 2020), making it particularly appropriate for our low-intensity therapy intervention context. Prior studies report good internal consistency (Cronbach’s $\alpha > 0.70$) (Perera et al., 2020; Salamanca-Sanabria et al., 2019). The Cronbach’s alpha for CRQ in our study is 0.94. By employing both CCM and CRQ, we aimed to probe cultural responsiveness comprehensively.

Empathy was measured by a 3-item instrument specifically used to evaluate the empathy expressed in text-based mental health support (Sharma et al., 2023). Specifically, the instrument assesses three aspects of empathy: emotional reactions (expressing emotions such as warmth, compassion, and concern), interpretations (communicating an understanding of feelings and experiences inferred from the user input), and explorations (exploring the feelings and experiences not stated in the user input).

Following quantitative ratings, participants selected their most and least preferred response versions and provided rationales through an open-ended prompt: “Can you share why

you like/dislike the responses?” Demographic data collected included age, gender, years living in the U.S., education level, and self-reported English proficiency level.

Data Analysis

We conducted a comprehensive analysis combining descriptive and statistical methods. For descriptive analysis, we computed means, standard deviations, and ranges for each evaluation measure (cultural competence (CCM), cultural relevance (CRQ), and perceived empathy) across four model configurations. Inter-rater reliability was assessed using intraclass correlation coefficients. We examined preference patterns across the four configurations using Pearson’s chi-square test against the null hypothesis of random distribution, followed by a post hoc analysis to identify specific model configurations contributing statistically significantly to the chi-square value.

We analyzed correlations between ratings and participant demographics using Spearman’s Rank Correlation for numerical variables and the Mann-Whitney U test for categorical variables. We conducted statistical comparisons using Wilcoxon signed-rank tests to compare ratings between (1) responses generated by different LLMs and (2) responses generated with and without cultural prompting. To investigate the relationship between cultural responsiveness and perceived empathy, we conducted an Analysis of Covariance (ANCOVA) test using differences in CRQ and CCM ratings as covariates. Statistical significance was considered established at $p < .05$.

4.1.3 Results

In total, 36 participants participated in the study. Participants were predominantly female ($n=28$, 77.8%) with a mean age of 50.8 years ($SD=9.8$, range=28-70). The sample represented a well-established immigrant population with an average U.S. residency duration of 20.9 years ($SD=8.0$, range=8-35). Education levels were notably high, with the majority holding graduate

or professional degrees (n=22, 61.1%), followed by Bachelor's degrees (n=10, 27.8%), some college education (n=3, 8.3%), and one participant declining to specify (2.8%). Regarding English language proficiency, most participants self-reported high proficiency levels (n=19, 52.8%), with the rest indicating moderate (n=12, 33.3%) or limited proficiency (n=5, 13.8%).

Cultural responsiveness and empathy among different models

The descriptive statistics of ratings for the three evaluation measures across different model configurations are shown in Table 2. GPT-4o with cultural prompting was consistently rated the highest across all three assessment dimensions. The inter-rater reliability analysis yielded an intraclass correlation coefficient of 0.86 (95% CI [0.79, 0.91]), indicating good reliability of the ratings across the 36 evaluators.

Table 4.2. Descriptive statistics of the ratings for responses generated by each model configuration.

Model Configuration	CCM (7-point Likert Scale, mean (SD, range))	CRQ (5-point Likert Scale, mean (SD, range))	Empathy (5-point Likert Scale, mean (SD, range))
GPT-4o without cultural prompting	4.01 (1.65, 1-6.75)	2.96 (1.19, 1-5)	3.52 (1.02, 1-5)
GPT-4o with cultural prompting	5.00 (1.61, 1-7)	3.63 (1.09, 1-5)	3.94 (0.94, 1.67-5)
Deepseek without cultural prompting	4.08 (1.72, 1-6.75)	2.98 (1.22, 1-5)	3.39 (0.98, 1-5)
Deepseek with cultural prompting	4.26 (1.53, 1-7)	3.22 (1.00, 1-5)	3.41 (0.88, 1.67-5)

The Wilcoxon Signed-Rank test showed that cultural prompting significantly enhanced performance across all measured dimensions for GPT-4o with relatively large effect sizes ($r > 0.5$) but not for Deepseek-V3. With cultural prompting, the GPT model showed statistically significantly better performance compared to Deepseek with cultural prompting across all measured dimensions with large effect sizes ($r > 0.5$).

Table 4.3. Wilcoxon Signed-Rank Test Results for Between-Condition Comparisons.

Comparison	Cultural Competency (CCM)	Cultural Relevance (CRQ)	Empathy
GPT-4o with cultural prompting vs. without	W=55.0, p<.001, r=0.74*	W=68.5, p<.001, r=0.71*	W=91.5, p=.01, r=0.66*
DeepSeek-V3 with cultural prompting vs. without	W=204.0, p=.11, r=0.39	W=177.0, p=.16, r=0.46	W=188.0, p=.98, r=0.43
GPT-4o vs. DeepSeek-V3 (both without cultural prompting)	W=149.5, p=.09, r=0.52	W=172.5, p=.49, r=0.47	W=139.5, p=.36, r=0.54
GPT-4o vs. DeepSeek-V3 (both with cultural prompting)	W=117.5, p=.01, r=0.60*	W=148.5, p=.02, r=0.52*	W=110.5, p=.01, r=0.61*

* indicates statistically significant differences at p<.05.

We employed Spearman's Rank Correlation and the Mann-Whitney U test to examine associations between ratings and participant demographics (age, gender, duration of time living in the U.S., education level, self-reported English proficiency level). Analysis revealed no statistically significant correlations between any demographic characteristics and rating patterns across responses generated by different model configurations.

Model preferences

The chi-square analysis on the relationship between model configuration and participants' preferences revealed a statistically significant association between model configuration and participants' preferences ($\chi^2(3) = 23.78$, $p < .001$). Post-hoc analysis of the residuals indicated that responses generated by GPT-4o with cultural prompting were significantly more preferred than expected under the null hypothesis (randomly distributed preference) with a standardized residual equal to 2.21. Qualitative feedback showed that users

felt GPT-4o with cultural prompting provided a balanced mix of emotional resonance, practical recommendations, and encouragement. They noted that culturally tuned responses did a better job of "expressing care, providing concrete suggestions, and offering thoughtful reasoning." On the other hand, responses generated by GPT-4o without cultural prompting demonstrated a trend toward being the least preferred, though this did not reach statistical significance based on the post-hoc analysis. Relevant qualitative feedback includes "too official" and "sounded like ChatGPT." Others described them as "empty, repetitive, and indirect" and expressed frustration that the chatbot "avoided giving direct responses" or was "too many words without getting to the point." Although quantitative ratings and the preference votes did not show Deepseek with cultural prompting was preferred, some participants noted Deepseek with cultural prompting as a "culturally focused and individualized conversation compared to the other groups" and had "a strong sense of empathy and understanding of Chinese culture." On the contrary, responses from Deepseek without cultural prompting were described as "(the responses) not specifically targeting the cultural background of Chinese users." These responses suggested that responses generated by LLMs with cultural prompting enabled a deeper emotional connection and made participants feel more understood.

Table 4.4. Counts of each model's responses are selected as the most and least preferred models.

Model Configuration	Most preferred Counts	Least Preferred Counts
GPT-4o without cultural prompting	5	15
GPT-4o with cultural prompting	19	4
Deepseek without cultural prompting	4	10
Deepseek with cultural prompting	8	7

Cultural responsiveness mediates the relationship between cultural prompting and empathy

To test the hypothesis that perceived empathy is primarily mediated by cultural relevance rather than direct cultural prompting, we conducted an analysis of covariance (ANCOVA), examining empathy rating differences between GPT-4o's responses with and without cultural prompting, controlling for differences in cultural relevance ratings. Results demonstrated the strong explanatory power of adjusted $R^2 = 0.41$ ($F(1,34) = 24.94$, $p < .0001$). The regression coefficient for cultural relevance difference ($\beta = 0.54$, $SE = 0.108$, $t = 4.90$, $p < .0001$, 95% CI [0.32, 0.76]) indicates that with a 1 point increase in cultural relevance rating difference, the empathy rating difference increases about 0.54 point in the empathy rating scale. When controlling for cultural relevance as a covariate, the independent effect of adding cultural prompting on empathy rating difference became non-significant ($\beta = 0.05$, $p = 0.70$). This result provided evidence that differences in perceived empathy between the configurations were predominantly mediated by differences in cultural relevance rather than resulting directly from prompt modifications.

Similarly, we tested the hypothesis that cultural competency mediated the perceived empathy in the GPT-4o responses. A similar mediation pattern emerged for cultural competence as a mediator for the change in empathy. The ANCOVA model demonstrated strong statistical power ($F(1, 34) = 34.94$, $p < .0001$, adjusted $R^2 = 0.49$), with a non-significant intercept term ($\beta = -0.068$, $SE = 0.13$, $t = -0.510$, $p = .61$), indicating that when controlling for cultural competence as a covariate, the independent effect of adding cultural prompting on empathy rating difference became non-significant. The regression coefficient of cultural competence difference ($\beta = 0.49$, $SE = 0.083$, $t = 5.91$, $p < .001$, 95% CI [0.32, 0.66]) revealed a strong positive relationship between the improvements in cultural competence and the corresponding improvement in perceived empathy. Specifically, a 1-point increase in cultural competence rating difference corresponds to about 0.49 points in empathy rating difference.

4.1.4 Discussion

This study evaluated the cultural responsiveness (relevance and competence) and empathy of AI-generated therapeutic responses for Chinese American family caregivers through a randomized controlled comparison of two LLMs (GPT-4o and DeepSeek-V3) with and without a cultural prompting technique. Our findings demonstrate that cultural prompting enhances GPT-generated responses' perceived cultural responsiveness and empathy. Although the two LLMs showed comparable performance without cultural prompting, GPT-4o showed greater improvement across the three evaluation dimensions when a cultural context prompt was added. Moreover, our mediation analyses revealed that cultural prompting improved perceived empathy in the therapeutic responses through increased cultural responsiveness. This indicates that cultural adaptation can improve the perceived empathy of therapeutic communication.

Our findings extend previous research evaluating the efficacy of cultural prompting on cultural alignment using cultural value survey data (Tao et al., 2024), by demonstrating its effectiveness in improving cultural responsiveness in the context of a healthcare application. In a previous study that used the CRQ for evaluating culturally adapted psychological interventions, a score of 3.5 on a scale of 1–5 was considered “average to good” and adequate to proceed with implementing the adapted intervention (Perera et al., 2020). In our findings, only the GPT model with cultural prompting achieved an average CRQ rating of 3.6, which would be considered an average-good cultural adaptation by the standards of this prior work. While GPT-4o with cultural prompting emerged as the most preferred, GPT-4o without cultural prompting was the least preferred. This highlights the importance of adopting cultural prompting techniques when using LLMs such as GPT-4o for therapeutic intervention delivery.

The effectiveness of this relatively straightforward cultural prompting technique suggests that meaningful improvements in cultural responsiveness and perceived empathy can be

achieved without extensive system design, additional computing, or further human resources. Whereas traditional cultural adaptation approaches require significant pre-implementation resources (Spanhel et al., 2021; Zhou et al., 2017), our results demonstrate that real-time adaptation can be achieved through appropriate prompt engineering.

The observed difference in the effectiveness of cultural prompting between GPT-4o and DeepSeek-V3 suggests that the efficacy of such techniques is model-dependent, aligning with previous findings in cultural alignment research (Masoud et al., 2025; Beck et al., 2024). We observed that responses generated by GPT-4o with cultural prompting contained more tokens than those from DeepSeek-V3, indicating that GPT-4o tends to produce longer, more elaborative responses that may incorporate more cultural nuance. This difference may stem from architectural variation or distinct training protocols. Alternatively, it could reflect the fact that our prompt design was optimized for GPT models (Filenko et al., 2024) and not further tailored for DeepSeek.

The results of the mediation analysis provide compelling evidence that cultural responsiveness served as a **mediator** through which cultural prompting influences perceived empathy. This finding suggests that perceived empathy in AI-delivered therapeutic interventions for culturally specific populations can be significantly improved by embedding cultural context. Our analysis showed that 40–50% of the perceived empathy improvement could be attributed to enhanced cultural responsiveness. This finding has major implications for LLM-based therapeutic applications: including cultural context in prompt design should not be treated as an optional enhancement, but rather as a core component for delivering empathetic and effective interventions. Without addressing cultural context, LLM-based interventions may fall short in serving diverse populations effectively.

Several limitations of the current study present opportunities for future work. First, participants were generally highly educated—predominantly holding a bachelor’s or more advanced degree—and had immigrated to the U.S. on average two decades ago. Future research should aim to include participants from broader educational backgrounds, immigration experiences, and cultural-linguistic contexts to enhance generalizability. In this study, culturally adapting LLM responses was achieved through system prompt augmentation, relying heavily on pre-existing cultural knowledge embedded within the models. Future studies should explore strategies that integrate **external cultural knowledge**, such as culturally specific sayings (Liu et al., 2024). In addition, more sophisticated adaptation techniques should be explored, including multi-agent workflows that account for specific cultural dimensions such as values, burdens, and communication styles. Finally, this study evaluated responses within **single-turn conversation instances**, rather than full **multi-turn therapeutic sessions**. Future work should explore the effectiveness of cultural prompting in multi-turn contexts, which would allow for evaluation of additional clinical outcomes and proxies such as emotional state and therapeutic alliance.

4.1.5 Conclusion

Our study contributes to the literature on culturally responsive AI applications in healthcare by presenting a human-centered evaluation of the cultural responsiveness of LLM-generated responses in a therapeutic context. This study provides evidence that cultural prompting techniques can significantly improve the cultural responsiveness (competence and relevance) and perceived empathy of GPT-4o-generated therapeutic responses in supporting Chinese American family caregivers. Parallel evaluations of similar approaches on different languages and cultures are warranted. Furthermore, our findings revealed that cultural responsiveness mediated cultural prompting’s improvements in perceived empathy, suggesting that cultural adaptation is not merely an additional feature but a core component for delivering

empathetic AI-based health interventions to culturally and linguistically diverse populations. Our study also highlights that although the prompt-based technique offers a practical solution for cultural adaptation without requiring additional resources for model modification, the efficacy of cultural prompting is model-dependent. Future work should continue exploring additional techniques to improve cultural responsiveness and evaluate their effects on LLM-delivered therapy chatbots among diverse populations.

4.2 Spanish Version

This is a parallel study to the Chinese version study presented above, we conducted the evaluation study with Spanish-speaking family caregivers.

4.2.1 Methods

Recruitment

We recruited Spanish-speaking Latino family caregivers of children with chronic health conditions using a multi-pronged approach. Recruitment took place between June and July, 2025. Initial recruitment was through the University of Washington Autism Center and participants from a previous study our team conducted. We then used snowball sampling, leveraging existing participants to identify new ones. Eligible caregivers were over 18, self-identified as Latino/Hispanic, spoke Spanish, and cared for an underage relative with a chronic health condition such as Autism. Upon meeting eligibility, caregivers were invited to

participate via email with an online survey link. This study was determined exempt by the University of Washington Institutional Review Board.

Model Configuration and Responses Generation

The simulated user inputs used to generate the input–response pairs for evaluation were extracted from a qualitative interview study with Latino caregivers by two bilingual researchers. User inputs with rich cultural contexts were selected, and in total, 39 inputs were used for response generation. For generating the responses, we utilized GPT-4o and Llama-3. Our model selection was motivated by differences in model characteristics (open-source vs. closed-source). GPT-4o, a state-of-the-art closed-source model, demonstrated strong general language capabilities and robust multilingual performance at the time of the study (Busch et al., 2025). Llama-3, one of the best-performing open-source models at the time (Grattafiori et al., 2024), was trained with expanded high-quality instruction-tuning data across several languages, including Spanish, making it particularly well-suited for Spanish use cases. By including both models, we aimed to investigate whether variations in training paradigms influence the degree of cultural responsiveness, both with and without explicit cultural prompting.

Table 4.5. Four model configurations using two different LLMs with and without cultural prompting.

Model id	LLM	With Cultural Prompting Techniques?
v1	GPT-4o	No
v2	GPT-4o	Yes
v3	Llama 3	No
v4	Llama 3	Yes

Data Analysis

We followed the same data analysis procedures as described in the Chinese version of this evaluation study. In addition, we implemented a data quality control step to ensure the validity of responses. Specifically, we identified and removed outliers exhibiting straight-lining behavior, defined as providing identical ratings across all items within a measure (e.g., all 5s on CRQ items) across all model configurations. Sixteen responses were excluded based on this criterion. Sixteen responses were excluded based on this criterion. We chose this approach over conventional distribution-based methods such as the interquartile range (IQR) because our concern was not extreme numerical values, since some items were expected to receive legitimately high or low ratings, but rather response patterns indicative of low engagement or invalid data quality, which distribution-based trimming would not reliably capture.

4.2.2 Results

In total, there were 41 responses collected through the online platform. After removing responses with straight-lining behavior (i.e., identical ratings across all items of a measure across all versions), data from 25 participants were retained for analysis. The participants had a mean age of 35.5 years ($SD = 10.2$, range = 18–66). Most participants were female (76%) and married or living with a partner (76%). All participants reported communicating in Spanish daily and identified as Hispanic, Latino/a, or Spanish origin, with highly varied levels of English proficiency. Most of the participants (84%) agreed or strongly agreed that they had a strong sense of belonging to their cultural group.

Table 4.6. Participant demographic characteristics

Age (Mean, SD, range)	35.5 (10.2, 18-66)	Marital Status (n, %)	
Gender		Married	15 (60%)
Female	19 (76.0%)	Living with my partner	4 (16%)
Male	6 (24.0%)	Separated	2 (8%)
Years living in the U.S.	12.7 years (10.2, 1-36) (5 not reported)	Single, never married	4 (16%)
Are you of Hispanic, Latino, or Spanish origin?		English proficiency	
Mexican, Mexican American, or Chicano(a)	11 (44%)	Cannot communicate or understand English	4 (16.0%)
Hispanic or Latino(a) from another place, or of Spanish origin	12 (48.0%)	Beginner level	7 (28.0%)
Cuban	2 (8%)	Intermediate level	3 (12.0%)
Puerto Rican	0 (0%)	Advanced level	3 (12.0%)
Not of Hispanic, Latino, or Spanish origin	0 (0%)	Bilingual / Native speaker	7 (28.0%)
Communicate in Spanish on a daily basis		Prefer not to answer	1 (4.0%)
Yes	25 (100%)		
No	0 (0%)		
I have a strong sense of belonging to my ethnic, racial, or cultural group			
Strongly Agree	12 (48.0%)		
Agree	9 (36%)		
Neutral	4 (16%)		
Disagree	0 (0%)		
Strongly disagree	0 (0%)		

Table 4.7 shows the mean ratings for cultural competence (CCM), cultural relevance (CRQ), and empathy across the four model configurations. Across all three evaluation domains, models with cultural prompting (GPT-4o+cultural prompting; Llama 3+cultural prompting) achieved slightly higher mean scores than their respective non-prompted counterparts (GPT-4o; Llama 3). Among all configurations, the highest mean scores were observed for Llama 3 with cultural prompting on CCM (Mean = 5.76, SD = 1.24), CRQ (Mean = 4.39, SD = 0.82), and

Empathy (Mean = 4.47, SD = 0.66), suggesting a potential benefit of cultural prompting, particularly with the Llama 3 model.

Table 4.7. Descriptive statistics of the ratings for responses generated by each model configuration.

Model Configuration	CCM (7-point Likert), M (SD, range)	CRQ (5-point Likert), M (SD, range)	Empathy (5-point Likert), M (SD, range)
GPT-4o without cultural prompting	5.52 (1.44, 2-7)	4.19 (0.92, 1-5)	4.43 (0.77, 2-5)
GPT-4o with cultural prompting	5.69 (1.32, 2-7)	4.33 (0.83, 2-5)	4.47 (0.68, 2-5)
Llama 3 without cultural prompting	5.36 (1.40, 2-7)	4.13 (1.08, 1-5)	4.45 (0.77, 2-5)
Llama 3 with cultural prompting	5.76 (1.24, 3-7)	4.39 (0.82, 2-5)	4.47 (0.66, 3-5)

For GPT-4o, no statistically significant differences were found between versions with and without cultural prompting for any of the three domains. For Llama 3, cultural prompting significantly improved CCM ratings ($W = 30.5$, $p = .02$, $r = 0.71$), and marginally improved CRQ ratings ($W = 71.0$, $p = .07$, $r = 0.49$). No significant difference was observed in Empathy ratings. Comparing the two LLMs (GPT-4o vs Llama 3), without prompting, GPT-4o outperformed Llama 3 on CCM with approaching statistical significance ($W = 56.0$, $p = .06$, $r = 0.57$), but no significant differences were observed on CRQ or Empathy. With prompting, GPT-4o and Llama 3 did not differ significantly in any evaluation domain.

Table 4.8. Wilcoxon Signed-Rank Test Results for Between-Condition Comparisons.

Comparison	Cultural Competency (CCM)	Cultural Relevance (CRQ)	Empathy
GPT-4o with cultural prompting vs. without	W=74.5, p=.25, r=0.47	W=62.5, p=.18, r=0.54	W=61.0, p=.71, r=0.55
Llama 3 with cultural prompting vs. without	W=30.5, p=.02, r=0.71**	W=71.0, p=.07, r=0.49*	W=48.0, p=.78, r=0.62
GPT-4o vs. Llama 3 (both without cultural prompting)	W=56.0, p=.06, r=0.57*	W=80.5, p=.55, r=0.44	W=38.0, p=.78, r=0.62
GPT-4o vs. Llama 3 (both with cultural prompting)	W=67.5, p=.67, r=0.51	W=72.5, p=.57, r=0.48	W=52.0, p=.97, r=0.59

*indicating statistical significance at the $p < .10$ level, and **indicating significance at the $p < .05$ level.

We further examined the impact of cultural prompting at the item level. As shown in Table 4.9, GPT-4o with cultural prompting received significantly higher ratings than without prompting on a CRQ item assessing goal alignment (“The goals suggested in the response are tailored to work with Latino caregivers”). Additionally, both GPT-4o and Llama 3 with cultural prompting were rated significantly higher on a CCM item evaluating recommendation likelihood (“Would you recommend the AI therapist to someone with Latino cultural background?”). These findings indicate that cultural prompting could have enhanced users’ perceptions of the agent’s relevance and appropriateness for Latino caregivers.

Table 4.9. Evaluation items that reached statistical significance across the pair-wise comparisons.

GPT-4o	without cultural prompting M (SD)	with cultural prompting M (SD)	W, p-value, r
CRQ - “The goals suggested in the response are tailored to work with Latino caregivers.”	4.24 (0.88)	4.60 (0.50)	W = 18.0, p = .04, r = 0.78
CCM - “Would you recommend the AI therapist to someone with Latino cultural background?”	5.60 (1.41)	5.96 (1.24)	W = 9.0 p = .047, r = 0.83
Llama 3	without cultural prompting M (SD)	with cultural prompting M (SD)	W, p-value, r
CCM - “Would you recommend the AI therapist to someone with Latino cultural background?”	5.64 (1.29)	6.16 (1.14)	W = 6.5, p = .002, r = 0.84

Ongoing Data Collection

The current sample size limits statistical power, and many comparisons did not reach significance despite medium-to-large effect sizes. Data collection is ongoing, and we anticipate that with additional participants, we may achieve sufficient power to detect statistically significant differences in empathy and cultural relevance across more conditions.

4.2.3 Discussion

These findings suggest that cultural prompting improved users’ perceptions of cultural competence and relevance more effectively for the open-source LLM Llama 3 than for the proprietary LLM GPT-4o. In particular, Llama 3 with prompting showed statistically significant improvement in perceived cultural competence and marginal improvements in cultural relevance compared to its non-prompted version. The lack of consistent significant differences

across empathy ratings suggests that cultural prompting may more directly impact perceived cultural alignment than general emotional understanding.

One limitation that may help explain discrepancies in results between Chinese and Latino family caregivers is the length of responses generated by the LLMs (GPT-4o, DeepSeek-V3, and Llama-3), which may align differently with the preferred communication styles of the two cultural groups. We observed that Llama-3 generally produces more verbose responses than GPT-4o. Future work should examine how variations in response length across different model configurations influence user perceptions, and how these differences may contribute to cross-cultural variation in caregiver ratings.

CHAPTER 5. Design and Implementation of a Retrieval-Augmented LLM Workflow for Culturally Responsive Mental Health Chatbots: A Context Engineering Approach

In this chapter, I describe the development of an LLM-facilitated approach for dynamically culturally adapting a mental health conversational agent. The goal in pursuing this work of cultural adaptation was to devise a method that is compatible with the dynamic, real-time content generation of an LLM, while incorporating human expert insights and minimizing the need for extensive manual data curation. We build upon the findings of Chapter 3, which identified culturally salient user needs, and the prompt-based techniques from Chapter 4, moving toward a more automated, scalable adaptation system.

To develop the cultural context database and the conversational agent, six bilingual-bicultural researchers engaged in an iterative assessment and refinement process through weekly meetings over two months, continuing until consensus was reached that the adapted responses were determined as more culturally responsive than the non-adapted GPT-4o responses through manual review. Prior to each meeting, I generated input-response pairs using the culturally relevant inputs described in Chapter 4, producing responses using both the non-adapted GPT-4o and the context-engineered agent at that stage. Each researcher independently reviewed these two sets of input-response pairs, after which the team convened to provide feedback and discuss strategies for improvement. Based on these discussions, I implemented modifications and conducted further experimentations. This iterative process resulted in several major refinements to both the database and the LLM workflow, including (1) the addition of culturally responsive response examples to the database, (2) modifications to the prompt for incorporating extracted data as context to avoid “context confusion”, (3) empirical

selection of a similarity threshold, and (4) introduction of two auxiliary modules—a gatekeeper cultural relevance detection LLM agent and a validation LLM agent.

5.1 Development of cultural context database

As a first step, I created a structured cultural context database to serve as a knowledge repository of culturally tailored content. As described in Chapter 3, I learned from interviews and focus groups about numerous culturally salient caregiving and self-care challenges faced by Chinese American family caregivers of older adults. These insights provided the themes around which cultural adaptation was focused. I structured the context database using these challenge themes including, communication barriers due to culturally rooted family dynamics, feelings of inadequacy in meeting traditional caregiving expectations, guilt related to older adult parents' social isolation, difficulties navigating healthcare systems, a sense that improving one's own well-being is futile, and the perception of self-care as an "indulgence" in conflict with duty (Table 5.1).

To enrich the database with culturally relevant examples, six bilingual Chinese American researchers on our team (including myself)—all native Chinese speakers with expertise in health and caregiving research—collaborated in this curation process. Each researcher selected two of the identified challenge themes to focus on. Drawing on the interview transcripts from Aim 1 and relevant cultural literature, team members independently generated underlying reasons or contextual factors that may contribute to each challenge. The team then convened to review and discuss these factors collaboratively until consensus was reached. This interpretive process aimed to unpack the implicit cultural norms, values, and common situations underlying a caregiver's experiences. For example, for the challenge "Communication Barriers in Filial Caregiving," an underlying reason might be "emotional restraint in the family due to cultural emphasis on respecting elders, leading adult children to hide stress." Through this process, we

uncovered a total of 44 unique underlying reasons across all challenges, ensuring a rich coverage of cultural context relevant to this use case.

We initially included only “challenges” and “underlying reasons” in the database. However, we observed that the model frequently integrated these reasons into responses in an overly literal and awkward manner (e.g., “Chinese culture emphasizes filial piety, and therefore you might feel...”). While technically accurate, these insertions reduced the authenticity and empathy of the responses. To address this, we expanded the database to also include culturally responsive responses, which served as examples for the LLM to emulate more nuanced and implicit ways of incorporating underlying reasons into its outputs. Based on each underlying reason, the researchers crafted one or more culturally responsive response snippets that were intended to address the caregiver’s challenge in a way that reflects cultural understanding and emotional support. These were written in Chinese and designed as empathetic, supportive messages that show a deep understanding of the cultural context and that would feel familiar to Chinese American family caregivers.

Each entry in the database thus consists of a triad: a challenge, an underlying reason, and an example culturally responsive message. Table 5.1 presents the set of identified challenge themes along with brief descriptions, as derived from the Aim 1 study. The examples below (translated into English) show how the underlying reasons and responses capture culturally specific context that a generic system would likely miss. This database effectively encodes community-informed cultural knowledge in a form that can be later retrieved to augment the AI agent’s replies.

Table 5.1. The cultural salient challenges identified in Aim 1.

Culturally Salient Challenges Identified in Aim 1	Description
Communication Barriers in Filial Caregiving	Difficulty expressing needs and emotions within culturally rooted family dynamics.
Stress Over Unmet Caregiving Expectations	Persistent feelings of inadequacy in fulfilling traditional caregiving roles.
Burden of Older Adult Social Isolation	Caregivers feel guilt and pressure to compensate for their older adult parents' loneliness.
Healthcare and Social Service System Navigation	Overwhelmed by the complexity of healthcare and social service systems.
Improving Emotional Well-being Feels Useless	Efforts to improve emotional well-being feel futile.
Self-care as Indulgence	Caring for oneself is viewed as indulgent or selfish.

Example 1:

- Challenge: Stress Over Unmet Caregiving Expectations
- Underlying reason: Fast-paced life limits caregiving capacity
- Culturally responsive response: Even if you want to do more, you are often stretched thin. You are afraid your parents may feel, “My children are too busy for me.” You may feel “I really want to spend more time with them, but I’m just so exhausted.” You’re not a superhero. You’re already holding up your family, your career, and your parents—all at once. You have the right to feel tired. You have the right to pause. Taking care of yourself isn’t selfish—it’s the most reliable protection you can offer your loved ones.”

Example 2:

- Challenge: Healthcare and Social Service System Navigation
- Underlying reason: Language and medical jargon barriers
- Culturally responsive response: You’ve acted as a “bridge” between two languages countless times—listening to technical terms and translating them into something your parents can understand. Often, you’re learning and explaining at the same time, under immense pressure. Please don’t doubt yourself. You’re not a professional interpreter or a doctor—you’re a devoted child trying to find a way forward for your parents. Every translation, every explanation, every doctor’s visit you accompany them to—that’s an act of love. They may say, “I didn’t understand,” but they can feel that you’re doing everything in your power to guide them through a foreign world.

5.2 Design and Experiment with LLMs workflows and prompts - Context Engineering

Having built the cultural context database, I next developed an LLM workflow that enables the conversational agent to utilize this database during live interactions. To do so, I employed an approach known as *context engineering* (Chase, 2025), which is broadly defined as “building dynamic systems to provide the right information and tools in the right format such that the LLM can plausibly accomplish the task.” In our case, context engineering involved designing when and how the conversational agent retrieves and incorporates culturally relevant content from the cultural context database into its responses in real-time to enhance cultural responsiveness. This process relies on thoughtful LLM workflow design and the use of retrieval augmented generation (RAG), an approach that combines the model’s pre-trained parametric memory with non-parametric memory sources such as external databases (Lewis et al., 2020).

Context engineering is an approach that has recently emerged in AI system design, defined as the process of building dynamic systems that provide an LLM with the right information and tools, in the right format, so that it can accomplish a task effectively (Chase, 2025; Mei et al., 2025). Unlike basic prompt engineering, which focuses on phrasing a single prompt, context engineering is about constructing an entire workflow that can bring in relevant data as context in response to the situation. In our case, the “task” is to conduct an empathetic therapeutic dialogue that is culturally responsive. The external data or tool is a vector database of snippets describing cultural context, and the dynamic information from relevant cultural snippets might help an LLM to generate more culturally responsive responses. Integrating external knowledge via retrieval has been identified as a key technique in context engineering frameworks, and the quality of context is crucial to shape the quality of the conversational agent response (Chase, 2025). By applying RAG and context engineering to our problem, we aim to bridge the gap between the LLMs’ limitations due to their Western-oriented training data and the need for culturally rich responses informed by stakeholder interviews and expert-crafted responses.

5.2.1 Overview of the Dynamic Cultural Adaptation System

Our system augments a base therapist LLM (Filienko et al., 2024; Wang et al., 2025) with a workflow of auxiliary LLM modules and a retrieval mechanism tapping into a developed cultural context database described in section 5.1. Figure 5.1 illustrates the architecture, and full prompts can be found in Appendix 2.

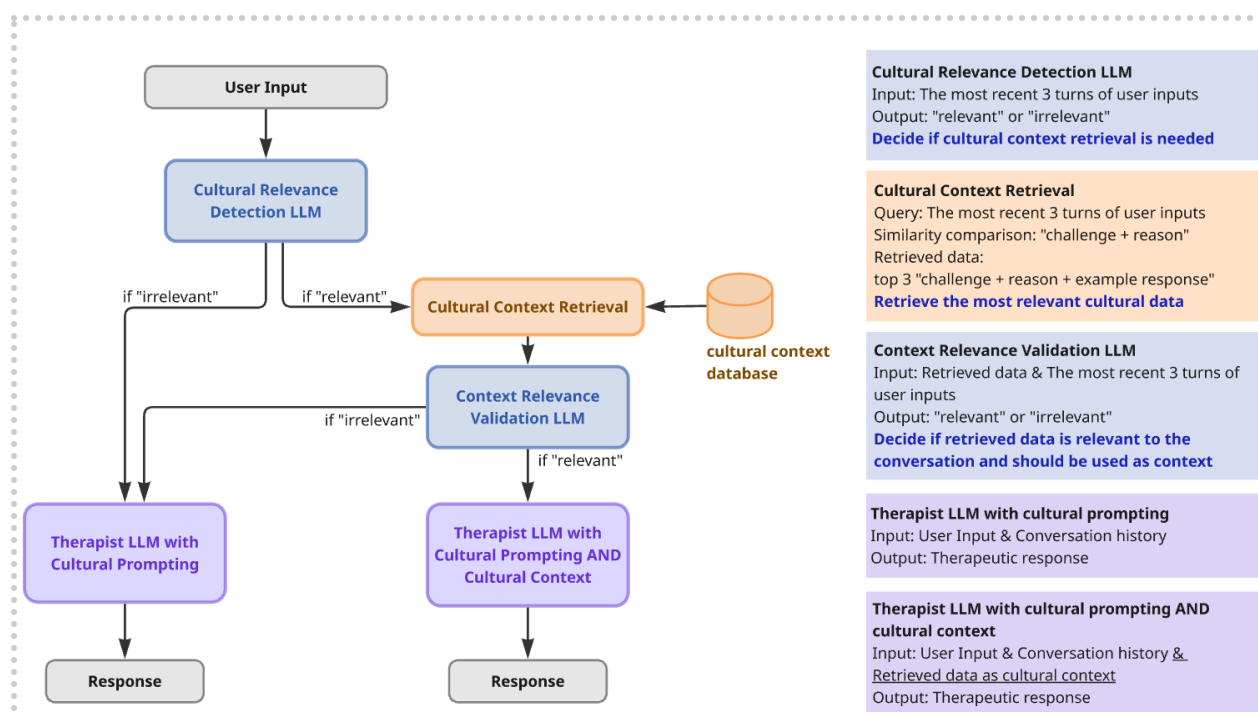


Figure 5.1. Components in the context-engineering-based cultural adaptation LLM workflow.

5.2.2 Retrieval Component Design and Implementation

To implement the retrieval component, I utilized Chroma (chroma, n.d.), an open-source embedding database, to store vector representations of structured data. I used Chroma as the vector store for its good integration with LangChain, which enabled efficient experimentation and deployment (i.e., via *langchain_community.vectorstores import Chroma*). Embeddings were generated using OpenAI's *text-embedding-3-large* model. The retrieval database consisted of a structured index with three fields per entry: (1) challenge, (2) underlying reason, and (3) culturally responsive response examples. Although users may not explicitly articulate the underlying reasons, we included them because the “challenge” labels alone were often too high-level or abstract to meaningfully match the nuanced ways users described their experiences. In contrast, the underlying reasons—derived from qualitative analysis—tended to capture semantically richer and more specific contextual cues that more closely aligned with users’ real-time input. During setup, each challenge+reason pair texts was concatenated,

embedded, and indexed. This approach enabled the retrieval system to leverage both the general topic (challenge) and its specific context (reason), improving semantic similarity with user inputs. At runtime, the user’s recent messages are concatenated (the most recent three turns, with a delimiter indicating separate turns) and embedded in real-time. This design decision was empirically supported during iterative experimentations, where retrieval relevance was informally evaluated by examining retrieved ranked results and determining whether top matches accurately aligned with the user’s intent using simulated user inputs. We observed qualitatively improved alignment between retrieved content and user-described scenarios when “reason” was included, compared to using “challenge” alone.

We retrieved the top three matches to allow some breadth – often multiple entries might be relevant to one user statement. For example, if a user says, “I feel so guilty taking a break; my father raised me to always put family first,” the system might retrieve an entry about self-care as indulgence and another about filial obligation causing guilt. To filter out irrelevant results, we applied a cosine similarity threshold of 0.6. This threshold was chosen through iterative prototyping and testing during development and experimentation phases, where I qualitatively observed that lower thresholds often retrieve content that was irrelevant to the simulated user’s input and therefore reduced response quality. If nothing cleared the threshold, the system would end up with an empty retrieval result, and thus it would just proceed without context injection. Once the top entries were identified, all three fields—including challenge, underlying reason, and the culturally responsive response examples—were retrieved and concatenated to form the contextual input for the language model.

5.2.3 Integrating Retrieved Context into LLM Responses

Integrating the retrieved information into the LLM’s response was one of the most critical design aspects. Our initial intuitive approach was to insert the retrieved data as examples for the therapist LLM to learn from. However, during our iterative assessment process, we observed

that this design caused a context confusion issue as the model mixed context with user input. The concept of context confusion is noted in LangChain’s documentation and by other practitioners as a common pitfall “when superfluous content in the context is used by the model to generate a low-quality response” (Breunig, 2025). Upon analysis, we realized the model treated the retrieved information as additional messages or facts that the user provided, which was not our intention.

To improve this, we refined the system prompt given to the therapist LLM. We clarified the role of the provided context as: “Leverage the provided cultural context information to develop a deeper understanding of the client’s cultural background and circumstances.” This practice is akin to how a human therapist might recall something they learned about the client’s culture and use it to inform their next response, rather than abruptly blurting, for example, “*As you said, taking a break is selfish,*” when the client never said that.

At this stage, the therapist LLM produces a final response that ideally blends therapeutic empathy with culturally specific context and culturally responsive response examples from the retrieved data. The following is an example of how the system behaved in practice based on a user input. When a user said, “*I’m so tired but I feel like taking time for myself is selfish,*” the system retrieved a relevant entry (translated into English) to use as context with the following content: “*Challenge: Self-care feels like indulgence; Underlying reason: Influenced by the cultural value of filial piety, there’s a belief that sacrificing one’s own life is simply what one is expected to do; Culturally responsive response: You might feel that caring for your parents is something you’re ‘supposed to do,’ but the dedication and commitment you’ve shown go far beyond obligation, and that deserves to be seen and recognized.*” Given this retrieved context and the user’s input, the therapist LLM generated the following response: “*You’ve been carrying so much for so long. It’s completely understandable to feel that taking a break might seem selfish—especially when you might deeply value self-sacrifice for the family. But you have the*

right to pause. Caring for yourself isn't selfish—it's one of the most important ways you can continue caring for the people you love." This illustrates how dynamic cultural context injection shaped the AI agent's output in practice.

Despite these improvements, we identified additional needs. Not every user message about family or stress necessarily warrants a cultural context insertion each time. We wanted the system to be selective in retrieval usage, which is why we added the "Cultural Relevance Detection" module (described next). Moreover, we noticed that sometimes the retrieved snippet, though thematically related, wasn't a good fit for the specific conversation moment – hence we added a validation step as well (described next).

5.2.4 Cultural Relevance Detection Module

The Cultural Relevance Detection LLM is a classifier module that acts as a gatekeeper for the retrieval process. Its purpose is to prevent the system from performing needless retrieval when the conversation topic doesn't pertain to our cultural database. This is important for efficiency and cost-effectiveness (each retrieval has a cost and latency) and for avoiding injecting off-base content that could confuse the dialogue.

We implemented this detector using a prompt-based approach with an LLM (GPT-4o). The prompt for this module is *"Based on the user input below, determine whether it involves topics related to Chinese culture (such as communication with family members, language, family expectations, parents' feelings of loneliness, lack of understanding about health and social services, viewing self-care and relaxation as indulgent, lack of support from relatives or friends, values, family, immigration, etc.). Respond with only one word: 'Relevant' or 'Not Relevant.'"* During the conversation, if the user's recent messages (i.e., user inputs from the most recent three turns) are classified as "Not Relevant" by this module, the system will skip the retrieval step and default to using the prompt-based only cultural prompting approach from

Chapter 4. If classified as “Relevant,” the retrieval workflow proceeds. This mechanism ensures that we leverage the cultural database only when likely to be helpful. For example, if a user’s problem is about family expectations or immigration stress, it will trigger retrieval of “filial piety” and “immigration” related data. But if the user is talking about a very general issue (e.g., trouble sleeping), the system doesn’t try to force a cultural angle.

5.2.5 Context Relevance Validation Module

Even after retrieving context, we found a need for a second filter: sometimes the retrieved data, though related to the general theme, did not actually apply well to the specific user situation at that moment. Including them could clutter the prompt or lead the therapist LLM to address an aspect that the user isn’t concerned with. To handle this, we introduced a Context Relevance Validation LLM, which assesses whether the retrieved cultural context is truly relevant to the user’s last few turns. This component essentially provides a quality control on the retrieval.

The prompt for this validator is *“Please determine whether the following cultural context information is highly relevant to the user’s three most recent inputs. Respond ‘Relevant’ or ‘Not Relevant’.”* Recent user inputs (i.e., user inputs from the most recent three turns) and retrieved data are provided and the LLM must judge if they match well. If the output is “Not Relevant,” the system will discard the retrieved context and not feed it into the therapist LLM, to avoid context confusion.

5.2.6 Summary of the Context Engineering Workflow and Link to Overall Goals

In summary, the dynamic cultural adaptation system we developed consists of multiple coordinated components: a therapist LLM, a cultural context knowledge base, and two LLM modules for decision-making (when to retrieve, whether to use retrieved info). Through context

engineering, we created a workflow where the LLM workflow can inject culturally relevant knowledge into the conversation in a controlled, contextually appropriate manner.

This design addresses Gap 1 identified in Chapter 2 by providing a method for real-time cultural adaptation that does not rely on pre-written scripts or data-intensive fine-tuning. It leverages expert knowledge (from the database) in a scalable way. Once this LLM workflow is set up, adding a new cultural scenario (i.e., with its own context database) or updating information is straightforward, and the conversational agent can switch contexts dynamically. If a developer wants to adapt this system to another culture or language, they can follow a similar procedure—gather cultural challenges, create a context database, and plug it into the workflow—without needing to collect hundreds of training examples. This offers a practical path for community organizations or clinics or other low-resource settings to adapt a generic chatbot by leveraging human experts’ insights with the cultural group/community and the use case.

In the following chapter (Chapter 6), I describe the evaluation on the effectiveness of this LLM workflow in comparison to other versions, examining whether this context engineered system indeed translates to measurable improvements in cultural responsiveness, empathy, and working alliance in practice.

CHAPTER 6. Evaluation of Prompt-Based vs Context-Engineering-Based Cultural Adaptation Approaches

This chapter presents the study where I directly compared three prototypes of a mental health conversational agent: a baseline agent without cultural adaptation (V1), culturally adapted agent using a prompt-based method (V2, as described in Chapter 4), and a dynamically adapted agent using a context-engineered retrieval-augmented in the previous study (Chapter 4) in Chinese. The goal of this study was to determine whether embedding culturally relevant information via dynamic retrieval in context would enhance the agent's cultural responsiveness, empathy, and therapeutic alliance as perceived by users. In addition to the outcomes examined in the prior prompt-based culturally adapted agent study, this evaluation also measured therapeutic alliance (i.e., working alliance), which was enabled by the interactive evaluation design. Therapeutic alliance is a well-established predictor of clinical outcomes and has shown strong association with treatment success (Bordin, 1994; Wampold, 2015; Anderson et al., 2012; Orlinsky et al., 2004). This relationship is particularly critical for clients from diverse cultural backgrounds, as research shows that racial and ethnic minority clients often experience weaker alliances in therapy, which can contribute to higher dropout rates (Thomas, 2016; Owen et al., 2011). I also chose this outcome as it is also highly related to the cultural backgrounds of the client and therapist (Ullrich, 2019; Okazaki & Tanaka-Matsumi, 2006; Shonfeld-Ringel, 2001). Moreover, previous research has found that digital text-based interventions were able to establish a comparable alliance to face-to-face therapy (Darcy et al., 2021; Prochaska et al., 2021), and appear to be related to user engagement (Hillier, 2018). We hypothesized that culturally adapting the AI agent (through either prompts or RAG-based context injection) would

strengthen users' sense of alliance with the agent, in addition to improving perceived empathy and cultural relevance of the agent's responses.

We conducted a within-subjects experiment in which each participant interacted with all three agent versions (V1, V2, V3) in a randomized order. This design allowed each participant to serve as their own control, enabling sensitive detection of differences across the adaptation conditions. After each interaction, participants provided evaluations of the agent's therapeutic alliance (using the Session Rating Scale), empathy, cultural competence, cultural relevance, and overall satisfaction. By comparing these metrics across V1, V2, and V3, we aimed to isolate the effects of the two cultural adaptation strategies and identify whether the more complex context engineering method yields measurable improvements over the simpler prompt-based technique.

6.1 Methods

The three agent versions were developed to deliver supportive responses to users discussing caregiving challenges. Agent V1 was an unaltered baseline agent with no specific cultural adaptation. Agent V2 was adapted using the cultural prompting method described in Chapter 4 – i.e., the system prompt included culturally relevant instructions and context (in this case, Chinese-American cultural cues) to guide the LLM's responses. Agent V3 was powered by a dynamic context-engineering workflow (Chapter 5).

Three Versions of the Mental Health Conversational Agents

All three versions of the agents were powered by GPT-4o, based on findings from Chapter 4, where GPT-4o was rated higher than DeepSeek-V3. The system prompts were written in Chinese and reviewed by a bilingual clinical psychologist to ensure clinical appropriateness. The V1 agent was configured to deliver a standard problem-solving therapy session without any cultural adaptation. The V1 agent was only instructed to conduct a problem-solving therapy. The V2 agent used the same base prompt as V1 but added an explicit cultural cue in its system prompt:

“When responding to caregivers, please fully consider their Chinese cultural background. Use language, values, and communication styles that align with Chinese culture to ensure your message resonates with them.” This instruction directed the agent to take the user’s cultural background into account when generating responses, representing a prompt-based cultural adaptation approach.

The V3 agent incorporated a dynamic cultural adaptation system built on a context-engineering framework. This system integrates a retrieval-augmented generation (RAG) workflow that enables the therapeutic LLM to incorporate culturally specific context in real time. The approach draws on prior qualitative research (Chapter 3) identifying culturally salient challenges faced by Chinese American family caregivers. These insights informed the creation of a structured cultural context database containing expert-derived underlying reasons for caregiving challenges and culturally responsive response examples. Six bilingual Chinese American researchers with expertise in caregiving and health collaboratively curated content across six thematic challenge areas. Each entry included: (1) a caregiving challenge, (2) an underlying cultural reason, and (3) a Chinese-language example message designed to convey empathy and understanding while reflecting cultural values such as filial piety and emotional restraint. The final database comprised 44 unique underlying reasons and over 100 tailored messages. All entries were stored in a searchable vector database using OpenAI’s text-embedding-3-large model and Chroma for retrieval. The cultural adaptation workflow begins with a cultural relevance detection module, implemented as a GPT-4o classifier, which determines whether the user’s recent input (last three turns) is culturally pertinent. If relevant, the system embeds the input, performs a similarity search against the vector database, and retrieves the top three matching entries. These are then screened by a context validation module—another LLM classifier—to confirm alignment with the user’s current concerns. Validated entries are appended to the system prompt for the therapeutic LLM, which then generates culturally responsive responses. To address context confusion—a known issue where

LLMs misinterpret retrieved snippets as user input—the system prompt explicitly instructs the agent to “leverage the provided cultural context information to develop a deeper understanding of the client’s background and circumstances,” ensuring that the LLM behaves like a culturally informed therapist rather than simply repeating retrieved text. This modular, scalable architecture enables real-time personalization without fine-tuning, offering a practical, low-resource pathway to delivering culturally responsive AI-based mental health support. A detailed description of this development process is provided in Chapter 5.

Recruitment and Data Collection

We recruited participants through two local community organizations. We distributed study flyers to recruit professional staff members who had experiences working with Chinese family caregivers for older adults. The sample size was determined based on recruitment feasibility, consistent with our previous study using a similar evaluation framework. In the previous study (Chapter 4), a sample of 36 participants was sufficient to detect statistically significant differences across conditions (with and without prompt-based cultural adaptation) with good effect sizes (e.g., $r = .76$ for cultural relevance, $r = .57$ for cultural competency, $r = .85$ for empathy). Given the comparable design and analytic strategy in the present study, we anticipated that a similar sample size would be adequate to detect meaningful within-subject differences across prototypes.

Each participant was given a link to Qualtrics online survey and instructed to interact with all three mental health conversational agent prototypes. To control for order effects, we randomized the sequence in which participants interacted with the three agents, ensuring that each prototype appeared equally often in each position across participants. This approach minimized potential biases related to fatigue and learning effects. To ensure comparability across prototypes while allowing for personalized relevance, participants were instructed to imagine a consistent caregiving scenario—either drawn from prior experience supporting family

caregivers or from their personal experience as a family caregiver. They were asked to maintain the same imagined scenario across all three interactions with three agents. Following interacting with each mental health conversational agent, participants were asked to complete an evaluation of the agent on its therapeutic alliance, empathy, cultural responsiveness, and their satisfaction.

Therapeutic Alliance. Therapeutic alliance (or working alliance) describes the collaborative and trusting relationship between a client and their therapist. We measured the therapeutic alliance using the 4-item Session Rating Scale (SRS) (Duncan et al., 2003), which is a tool commonly used in psychotherapy to assess clients' subjective experiences of a single therapy session. It is a session-level measurement. For instance, a recent study has used SRS to assess therapeutic alliance of an Empathic LLM (EmLLM) chatbot (Dongre, 2024; Dongre et al., 2024). The SRS was also examined cross-culturally and deemed largely consistent with the English studies (Hafkenschied et al., 2010; She et al., 2023). SRS covers the three cores of the alliance according to the highly influential model of the working alliance by Bordin (1994): agreement on therapy goals, agreement on tasks, and the bond between therapist (i.e., conversational agent) and client (i.e., participant). Specifically, it measures four dimensions including relationship (to what extent you felt heard, understood, and respected by the chatbot), goals & topics (to what extent the session focused on what you wanted to work on and talk about), approach or method (to what extent the therapy approach is a good fit for you), and overall satisfaction (overall rating of interaction) (Duncan et al., 2003). The SRS is a four-item visual analogue instrument on a 0–10 scale. Higher scores indicate better performance in that dimension.

Empathy. We measured perceived empathy using a 3-item computational empathy measure developed by Sharma et al. (Sharma et al., 2020). Specifically, the instrument assesses three aspects of empathy: emotional reactions (expressing emotions such as warmth, compassion, and concern), interpretations (communicating an understanding of feelings and

experiences inferred from the user input), and explorations (exploring the feelings and experiences not stated in the user input).

Cultural Competence. In addition, we used two complementary measures to assess cultural responsiveness: The Healthcare Provider Cultural Competency Measure (CCM) is a validated patient-reported instrument that captures patients' perceptions of their providers' cultural competency (Lucas et al., 2008). This 9-item evaluation assesses three critical domains—cultural knowledge, awareness, and skill. This scale has established reliability (Cronbach's alpha 0.68–0.98)(Cronbach, 1951) in previous studies (Dawadi et al., 2022; Alizadeh & Chavan, 2020; Nápoles et al., 2012). To minimize participant burden while maintaining construct validity, we selected a representative item from each domain plus one overall assessment item, preserving the theoretical structure while optimizing completion time. The consistency reliability of the four selected CCM items was assessed in our previous study (Chapter 4), showing excellent internal consistency (Cronbach's $\alpha = 0.96$).

Cultural Relevance. The Cultural Relevance Questionnaire (CRQ) was informed by the literature on cultural sensitivity (Resnicow et al., 1999) and ecological validity theory (Bernal et al., 1995). The CRQ has demonstrated utility in assessing internet-delivered therapeutic interventions (Salamanca-Sanabria et al., 2018) and low-intensity psychological approaches (Perera et al., 2020), making it particularly appropriate for our low-intensity therapy intervention context. Prior studies report good internal consistency (Cronbach's $\alpha > 0.70$) (Perera et al., 2020; Salamanca-Sanabria et al., 2019). The consistency reliability of the CRQ items was assessed in our previous study (Chapter 4), showing excellent internal consistency (Cronbach's $\alpha = 0.94$). By employing both CCM and CRQ, we aimed to probe cultural responsiveness comprehensively.

In addition, we included the Net Promoter Score (NPS) as a single-item measure of overall user satisfaction, rated on a scale from 0 (not at all likely) to 10 (extremely likely).

Participants were asked, “How likely are you to recommend this AI therapist to a family member or friend?”—a widely used item for assessing general satisfaction and likelihood of endorsement. All evaluation items are provided in Appendix 3.

Data Analysis

We conducted a comprehensive analysis using both descriptive and statistical methods. Descriptively, we calculated the mean and standard deviation for each evaluation measure—cultural competence (CCM), cultural relevance (CRQ), perceived empathy, therapeutic alliance (SRS), and user satisfaction (NPS)—across the three conversational agent prototypes. We first examined the distribution of scores for each measure to determine the appropriateness of parametric versus nonparametric tests. Because the data did not consistently meet normality assumptions, we performed Wilcoxon signed-rank tests for each pairwise comparison (V1 vs. V2, V2 vs. V3, V1 vs. V3) across all five evaluation dimensions. In addition, we conducted subscale-level analyses to gain deeper insights into each construct. For therapeutic alliance, we analyzed the four subscales of the Session Rating Scale (SRS): relationship, goals and topics, approach, and overall rating. For perceived empathy, we examined the three dimensions outlined by Sharma et al. (2020): emotional reactions, interpretations, and explorations. For cultural competence, we assessed three domains: cultural knowledge, awareness, and skill. For cultural relevance, each individual item in the Cultural Relevance Questionnaire (CRQ) was analyzed. Wilcoxon signed-rank tests were also conducted for all subscales, including pairwise comparisons between each version of the conversational agent. A p-value of less than 0.05 was considered statistically significant.

6.2 Results

A total of 36 participants were included in the study. Age data were provided by 31 participants, with a mean age of 37.9 years (SD = 8.1, range = 24-55). Gender was reported by all participants,

with 21 (58.3%) identifying as female and 15 (41.7%) as male. Most participants held a bachelor's degree ($n = 18$, 50.0%) or a graduate/professional degree ($n = 7$, 19.4%). A small number reported having some high school or less ($n = 1$, 2.8%), a high school diploma ($n = 1$, 2.8%), or preferred not to answer ($n = 2$, 5.6%). For English proficiency, the majority of participants indicated they spoke English "well" ($n = 18$) or "very well" ($n = 16$), 2 (5.6%) selected "not well" and 1 selected "not at all". Thirty-four participants reported their length of residence in the U.S., with an average of 12.5 years ($SD = 6.87$, range = 3-30 years). Sixteen (44.4%) out of the participants were caring for or had experience caring for their older adult family members.

We evaluated three prototypes of mental health conversational agents developed to support Chinese family caregivers: (V1) a baseline agent with no cultural adaptation, (V2) an agent using a prompt-based cultural adaptation method, and (V3) an LLM workflow-powered agent utilizing a cultural context database. As shown in Table 6.1, the LLM workflow-powered agent (V3) consistently outperformed both the prompt-based agent (V2) and the baseline agent with no cultural adaptation (V1) across all measures. Table 6.2 further shows that only the context-engineered model (V3) achieved consistent, statistically significant improvements over the baseline (all $p < .05$) with relatively large effect sizes ($r = 0.54-0.65$). Although the prompt-based culturally adapted agent (V2) demonstrated improvements over V1 in the expected direction, it reached statistical significance only for user satisfaction ($p = .02$, $r = 0.58$). This distinction highlights the superior effectiveness of the context-engineered approach (V3) compared with prompt-based adaptation (V2). Direct comparisons between V2 and V3 revealed a trend favoring the LLM workflow context engineering method, including a trend toward a significant difference in empathy ($p = .09$). These findings support the added value of context engineering and retrieval-augmented generation utilizing the cultural context database in delivering culturally responsive and empathetic support.

Table 6.1. Comparison of conversational agent prototypes across five evaluation dimensions. V1 = no cultural adaptation; V2 = culturally adapted using prompt-based method (Chapter 4); V3 = LLM workflow with cultural context database (Chapter 5).

Dimension	V1 Mean (SD)	V2 Mean (SD)	V3 Mean (SD)
Therapeutic Alliance (SRS)	5.69 (0.53)	6.15 (0.39)	6.53 (0.49) [†]
Empathy	3.22 (0.21)	3.37 (0.20)	3.62 (0.21) [†]
Cultural Competence (CCM)	3.72 (0.09)	4.03 (0.17)	4.25 (0.15) [†]
Cultural Relevance (CRQ)	2.84 (0.24)	2.99 (0.21)	3.19 (0.17) [†]
User Satisfaction (NPS)	5.37 (2.56)	5.94 (2.44) [†]	6.44 (2.27) [†]

[†] Indicates statistically significant improvement over the baseline (V1) at $p < .05$.

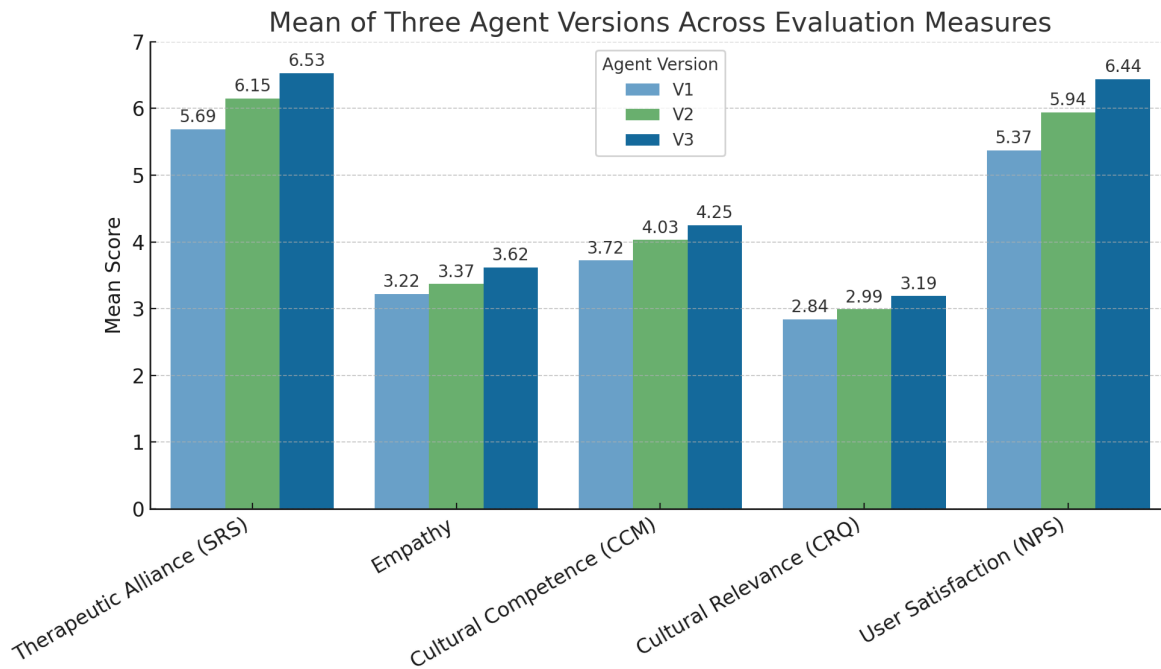


Figure 6.1. Mean scores across five evaluation measures for each of the three conversational agent prototypes.

Table 6.2. Pairwise comparisons of conversational agent prototypes across evaluation dimensions. Wilcoxon signed-rank test results are reported for each comparison, including the W statistic, p-value, and effect size (r).

	SRS average	Empathy average	CCM average	CRQ average	NPS average
V1 vs V2	W=227.0, p=.34, r=0.28	W=171.0, p=.47, r=0.42	W=137.5, p=.13, r=0.51	W=135.5, p=.19, r=0.52	W=113.0, p=.02, r=0.58*
V2 vs V3	W=155.5, p=.11, r=0.46	W=109.0, p=.09, r=0.59	W=171.5, p=.47, r=0.42	W=160.0, p=.213, r=0.45	W=151.5, p<.05, r=0.48*
V1 vs V3	W=124.5, p=.02, r=0.55*	W=91.5, p=.01, r=0.63*	W=125.5, p=.03, r=0.54*	W=82.5, p=.003, r=0.66*	W=83.5, p=.01, r=0.65*

* indicates statistically significant differences at $p < .05$.

To further examine therapeutic alliance, we analyzed the four subscales of the Session Rating Scale (SRS): relationship, goals and topics, approach, and overall rating. As shown in Table 6.3, the LLM workflow and RAG-enabled culturally adapted agent (V3) demonstrated statistically significant improvements over the baseline (V1) on all four subscales (all $p < .05$). The prompt-based agent (V2) did not show significant improvements over baseline (V1) on any subscale. For perceived empathy, we analyzed three subscales: emotional reactions, interpretations, and explorations. As shown in Table 6.4, V3 significantly outperformed V1 on all three dimensions—emotional reactions ($W = 34.0$, $p = .01$), interpretations ($W = 63.0$, $p = .02$), and explorations ($W = 57.5$, $p = .02$). V2 again showed no statistically significant improvements over V1. These results underscore that the most consistent and substantial gains from baseline were achieved with the context-engineered V3, particularly in fostering stronger therapeutic alliance and richer expressions of empathy.

Table 6.3. Pairwise comparisons of **therapeutic alliance** subscales (Session Rating Scale, SRS) across conversational agent versions.

	SRS - Relationship	SRS- Goals & Topics	SRS - Approach	SRS - Overall
V1 vs V2	W=122.0, p=.16, r=0.55	W=68.5, p=.10, r=0.69	W=193.5, p=.60, r=0.37	W=83.0, p=.03, r=0.65*
V2 vs V3	W=170.0, p=.12, r=0.43	W=163.5, p=.36, r=0.44	W=112.0, p=.03, r=0.58*	W=98.5, p=.36, r=0.61
V1 vs V3	W=113.0, p=.02, r=0.58*	W=88.0, p=.04, r=0.64*	W=47.0, p=.01, r=0.75*	W=113.0, p=.01, r=0.58*

* indicates statistically significant differences at $p < .05$.

Table 6.4. Pairwise comparisons of **empathy** dimensions across conversational agent versions.

	Empathy- Emotional Reactions	Empathy- Interpretations	Empathy- Explorations
V1 vs V2	W=91.0, p=.58, r=0.63	W=70.5, p=.30, r=0.69	W=72.5, p=.34, r=0.68
V2 vs V3	W=37.5, p<.05, r=0.77*	W=72.0, p=.31, r=0.68	W=78.5, p=.17, r=0.67
V1 vs V3	W=34.0, p=.01, r=0.78*	W=63.0, p=.02, r=0.71*	W=57.5, p=.02, r=0.72*

* indicates statistically significant differences at $p < .05$.

To more closely examine cultural responsiveness, we analyzed the subscales of cultural competence, which capture cultural knowledge, cultural awareness, and skillfulness in providing care that respects the client’s cultural values and traditions. As shown in Table 6.5, the LLM workflow–powered agent (V3) achieved significantly higher ratings than the baseline agent (V1) for both overall cultural competence ($W = 77.5$, $p = .01$) and the “well-equipped with skills” subscale ($W = 78.5$, $p = .04$). The prompt-based agent (V2) showed only a marginally significant improvement over V1 in overall cultural competence ($W = 70.0$, $p < .05$) and did not differ significantly from baseline on other subscales.

We also assessed cultural relevance using the Cultural Relevance Questionnaire (CRQ). As shown in Table 6.6, V3 scored significantly higher than V1 on CRQ-2 (tailoring goals to the

user's cultural context, e.g., using relevant examples or personal stories; $W = 30.0$, $p = .02$) and CRQ-3 (use of dialects, sayings, or slang relevant to the cultural context; $W = 20.0$, $p < .001$). The prompt-based agent (V2) showed a significant improvement over V1 only on CRQ-3 ($W = 34.0$, $p < .001$). Across all cultural competence and cultural relevance measures, V3 consistently achieved the highest mean scores, indicating that the most substantial gains from baseline were realized with the LLM workflow approach.

Table 6.5. Pairwise comparisons of **cultural competence** (Cultural Competence Measure, CCM) dimensions across conversational agent versions.

	CCM- Cultural Knowledge	CCM- Cultural Awareness	CCM- Well-equipped with Skills	CCM- Overall
V1 vs V2	$W=61.5$, $p=.10$, $r=0.71$	$W=75.5$, $p=.42$, $r=0.66$	$W=59.0$, $p=.23$, $r=0.72$	$W=70.0$, $p=.05$, $r=0.69^*$
V2 vs V3	$W=114.0$, $p=.96$, $r=0.57$	$W=107.5$, $p=.53$, $r=0.57$	$W=129.5$, $p=.35$, $r=0.53$	$W=113.0$, $p=.27$, $r=0.58$
V1 vs V3	$W=83.0$, $p=.09$, $r=0.65$	$W=102.0$, $p=.06$, $r=0.60$	$W=78.5$, $p=.04$, $r=0.67^*$	$W=77.5$, $p=.01$, $r=0.67^*$

* indicates statistically significant differences at $p < .05$.

Table 6.6. Pairwise comparisons of **cultural relevance** (Cultural Relevance Questionnaire, CRQ) items across conversational agent versions. CRQ 1 - reflect the people and broader cultural context, including social, political, economic, ethnic, and historical factors; CRQ 2 - tailoring goals to the user's cultural context, e.g., using relevant examples or personal stories; CRQ 3 - use of dialects, sayings, or slang relevant to the cultural context; CRQ 4 - uses behavioral or emotional expressions familiar to the target cultural group; CRQ 5 - incorporates culturally shared symbols, concepts, or analogies, into its responses.)

	CRQ 1	CRQ 2	CRQ 3	CRQ 4	CRQ 5
V1 vs V2	$W=78.0$, $p=.74$, $r=0.66$	$W=90.0$, $p=.13$, $r=0.62$	$W=34.0$, $p=.003$, $r=0.78^*$	$W=92.5$, $p=.41$, $r=0.63$	$W=76.5$, $p=.67$, $r=0.67$
V2 vs V3	$W=42.0$, $p=.15$, $r=0.76$	$W=68.0$, $p=.67$, $r=0.69$	$W=61.5$, $p=.45$, $r=0.71$	$W=56.0$, $p=.18$, $r=0.73$	$W=53.5$, $p=.14$, $r=0.72$
V1 vs V3	$W=32.5$, $p=.10$, $r=0.78$	$W=30.0$, $p=.02$, $r=0.79^*$	$W=20.0$, $p<.001$, $r=0.82^*$	$W=158.5$, $p=.65$, $r=0.46$	$W=34.0$, $p=.07$, $r=0.78$

* indicates statistically significant differences at $p < .05$.

6.3 Discussion

This study evaluated and compared three versions of a mental health conversational agent designed to support Chinese family caregivers, comparing a baseline agent (V1), an agent that is culturally adapted using prompt-based method (V2), and an agent enhanced through a context-engineered LLM workflow that dynamically retrieved culturally relevant context from a database (V3). Our findings demonstrate that the context-engineered LLM workflow (V3) consistently and significantly improved user perceptions of therapeutic alliance, empathy, and cultural responsiveness when compared to the baseline agent (V1), and in some cases, the prompt-based adaptation, while the prompt-based adapted agent (V2) only significantly improved baseline performance on overall user satisfaction.

In particular, the V3 agent outperformed both V1 on all four subscales of the Session Rating Scale with statistical significance, indicating stronger perceived relationship, better alignment on goals and topics, greater fit of the therapeutic approach, and higher overall satisfaction with the interaction. These improvements in alliance are especially notable given prior research showing that racial and ethnic minority clients often experience weaker therapeutic relationships in psychotherapy settings (Thomas, 2016; Owen et al., 2011). These findings suggest that incorporating dynamically retrieved, culturally grounded content can strengthen therapeutic alliance in AI-mediated interventions.

The LLM workflow-powered agent (V3) also demonstrated significantly higher ratings on all three subscales of empathy compared to non-culturally adapted V1. These results suggest that the integration of culturally grounded examples—retrieved in real time from a curated knowledge base—enabled the agent to respond in ways that were more attuned to users' emotional needs. Similarly, the V3 agent achieved higher ratings in both cultural competence and cultural relevance measures, indicating that users perceived the agent as better equipped to understand, respect, and reflect their cultural contexts. While all CRQ items showed directional

improvements, only CRQ-2 and CRQ-3 reached statistical significance. This suggests that the retrieval-augmented LLM workflow was particularly effective in enhancing the agent’s ability to incorporate concrete, context-specific elements of cultural communication, such as culturally relevant goals, sayings, and expressions.

This study makes several contributions to the field of culturally responsive AI for mental health support. First, comparing to prior work that has explored context-engineered chatbots for cultural preservation and education (Chang et al., 2024), our work uniquely applies RAG to improve cultural responsiveness within a mental health context, explicitly targeting outcomes that are clinically meaningful—such as therapeutic alliance, empathy, and cultural competence. These constructs, drawn from psychotherapy research, offer a more outcome-oriented lens than general usability or engagement measures. Importantly, we conducted subscale-level analyses across all evaluation measures, offering deeper insights into the specific dimensions where cultural adaptation was effective. Second, we conducted a within-subject comparison of three adaptation methods—baseline, prompt-based, and context-engineered agents—allowing us to isolate the benefits of context engineering with statistical evidence. By directly evaluating both prompt-based and context engineering methods, we demonstrate that an agent that dynamically retrieved cultural context was able to achieve better outcomes. Third, our system is supported by a carefully curated cultural context database, informed by prior qualitative research with Chinese family caregivers and community staff and developed by bilingual Chinese American researchers. This database embeds culturally grounded examples and nuanced expressions aligned with values such as filial piety and emotional restraint. Fourth, we incorporated safeguards within the LLM workflow to mitigate contextual confusion—explicitly guiding the LLM to treat retrieved content as background cultural information rather than user input. Collectively, these contributions represent a substantial step forward in designing and evaluating culturally responsive AI systems tailored for mental health care delivery.

Despite these promising findings, several limitations should be noted. First, while our sample was community-based, participants were primarily staff members from local community organizations. As such, their experiences may not fully reflect those of Chinese-speaking family caregivers actively engaged in caregiving roles. Second, although the use of validated measures allowed us to capture key dimensions of perceived effectiveness, future studies should incorporate behavioral or longitudinal outcomes—such as user engagement over time, adherence to self-care recommendations, and mental health outcomes—to evaluate the practical impact of cultural adaptation. Additional work is also needed to assess real-world health outcomes of culturally adapted mental health AI agents. Third, because the adapted agent was implemented as a multi-component system, we were not able to determine which component, such as the cultural context database or the retrieval-and-integration mechanism, contributed most to the observed effects. Moreover, we observed that the responses generated across the three versions were different. The context engineering approach tends to generate longer responses. Finally, we observed systematic differences in response styles across models: the context-engineered agent tended to generate longer responses. Future work should examine how response length influences user perceptions of empathy and therapeutic alliance, through post-hoc or controlled analyses.

6.4 Conclusion

Taken together, the findings of this study underscore that embedding culturally grounded, cultural context-specific information into mental health conversational agents through a RAG-enabled context engineering workflow can substantially enhance users' perceptions of therapeutic alliance, empathy, and cultural responsiveness. For family caregivers, such improvements may not only foster more relatable interactions but may also translate into greater engagement and adherence to self-care strategies. More broadly, the work demonstrated a scalable framework for operationalizing cultural adaptation in LLM-driven mental health

interventions, which could be extended to other underserved communities and caregiving contexts. By bridging cultural gaps in communication, this approach holds promise for reducing disparities in access to and quality of mental health care, advancing more personalized digital mental health support for caregivers.

CHAPTER 7. Randomized User Study: Impact of Cultural Adaptation on Emotional Well-being and User Experience

In this chapter, I describe a user study (Aim 3) that examines the real-world impact of cultural adaptation by testing the chatbot with target end-users: Chinese American family caregivers caring for older adults. Based on the findings from Chapter 6, we selected the best-performing culturally adapted agent (the dynamic cultural context-engineered V3) and compared it against the non-adapted baseline agent (V1) in a randomized between-subjects experiment. This study was designed to evaluate whether the advantages of cultural adaptation observed in controlled comparisons would translate into improved user experience and short-term emotion state improvement in a practical use scenario. In particular, we focused on outcomes immediately following a single chat session, including the user's emotional states, perceived cultural competence, empathy and therapeutic alliance of the agent, and satisfaction. Our hypothesis was that caregivers who interacted with the culturally adapted agent would report greater improvements in their emotions, and higher ratings of the agent's empathy, alliance, and cultural competence, compared to those who interacted with the non-adapted agent.

This study builds upon the previous evaluation in three key ways to enhance ecological validity and strengthen the assessment of cultural adaptation: (1) participants engaged with only one version of the agent, (2) actual caregivers were recruited and asked to interact with the agent using their current real-life experiences rather than hypothetical scenarios, and (3) an additional measure of change in emotional states (PANAS) was included to provide a more comprehensive evaluation of cultural adaptation. These design choices aimed to more closely mirror the real-world use of a mental health support conversational agent for caregivers and preliminarily explore the effect on preliminary health outcomes. By measuring preliminary

health outcomes (i.e., immediate change of affective states), this user study addresses the call for more outcome-oriented evaluation of culturally adapted AI conversational agents in the context of mental health care delivery^[66]. While the intervention was limited to a single, brief session, immediate mood improvements observed may indicate the potential for sustained benefits with repeated use (Di Pompeo et al., 2023).

7.1 Methods

We conducted a two-group randomized user study. Participants were randomly assigned to interact with one of two agent prototypes: Non-culturally adapted agent – interaction with the baseline (GPT-4o), or the culturally adapted agent that uses the context-engineered LLM workflow. Randomization was implemented via the Qualtrics survey platform. Participants were not explicitly told which version (adapted or not) they were using and were instructed that they would be interacting with a “caregiver support AI therapist” and to engage with it as if seeking help with their caregiving stress.

Participants & Recruitment

We targeted Chinese-speaking family caregivers of older adults residing in the United States or Canada as the intended end-user population. Inclusion criteria were: (1) age 18 or older; (2) current experience providing care to an older family member; and (3) the ability to read and type in Chinese. We recruited participants through multiple community channels. Our partner organizations (the same ones from earlier studies) helped distribute study flyers in their network. Additionally, we posted study advertisements on social media platforms popular in Chinese American communities (e.g., WeChat groups). Snowball sampling was encouraged (caregivers could share the study link with others). We aimed to recruit approximately 80 participants based on a power analysis using results from a prior study examining pre- and post-session PANAS differences, as well as user study feasibility considerations. Given the

preliminary nature of this user study, we acknowledged that our sample might not be powered to detect small effects. However, we anticipated observing clear directional trends and gathering feasibility data to inform future trials.

Data Collection

The study was conducted online. All evaluations were conducted through the Qualtrics online surveys. Demographic data collected included age, gender, years living in the U.S., education level, and self-reported English proficiency level. Each participant was assigned to engage with either the non-culturally adapted prototype or the culturally adapted prototype. Each participant was assessed on their change in emotional states, and asked to evaluate the prototype that they interacted with on therapeutic alliance, perceived empathy, and cultural competency of the agent.

Both versions of the agent (adapted and non-adapted) were configured to deliver a brief Problem-Solving Therapy (PST) initial session. PST has been shown to be effective in reducing anxiety and emotional distress (Ferré-Grau et al., 2014). Specifically, the agent would greet the user, ask them about their current problem or stressor regarding their caregiving, and then follow the PST framework focusing on *symptom assessment*. We chose to focus on the symptom assessment phase because it emphasizes eliciting the user’s caregiving stressors, which would align with our cultural context database, which contains caregiving challenges and underlying reasons. Both chatbot versions had identical PST prompts, except that the adapted version had the added cultural context retrieval capability. The culturally adapted agent would, whenever relevant, pull in culturally specific understanding or examples (from the Chinese caregiver database) as part of its responses (described in detail in Chapter 5). The baseline agent, on the other hand, responded based on generic PST prompts without any cultural context injection. After a complete PST first session, the agent will signaled a wrap up and instruct participants to return to the survey to complete post-session evaluations. These included the same mood scale

as pre-session and questionnaires on the agent's empathy, therapeutic alliance, and cultural competence, identical to those used in Chapter 6.

Emotional Well-being (PANAS-X). To quantify changes in participants' emotional states after chatting with the conversational agent, we used subscales from the Positive and Negative Affect Schedule - Expanded Scale (PANAS-X) (David & Clark, 2011). PANAS (Watson et al., 1988) is a tool that has been used by therapists to track changes in positive and negative emotions for clients. It can be used to measure the immediate effects of therapy sessions and outcomes associated with psychological interventions (Saxon et al., 2017). We used a Chinese translation of PANAS-X that has been validated in prior research (Guo & Gan, 2010). The PANAS-X contains 11 specific emotional subscales, in this study, specifically, we selected four subscales corresponding to four affective states, **fatigue**, **guilt**, **sadness**, and **serenity**, that we hypothesized to be relevant to stressed caregivers and responsive to a therapeutic session based on a prior study (Divina et al., 2021). The survey included 14 items of 4 subscales from PANAS-X in total: Fatigue (3-item), Guilt (6-item), Sadness (3-item), and Serenity (2-item). The items were presented in Chinese. Participants rated the extent to which they felt each emotion "right now" on a 1 (not at all) to 5 (extremely) scale. The PANAS survey was administered twice: once right before the chat with the agent, and once immediately after the chat. We then scored each affect by summing its items, yielding four scores for pre-session and four for post-session for each participant. An effective intervention should reduce the scores for negative affects (fatigue, guilt, sadness) and increase the score for positive affect (serenity).

Therapeutic Alliance. As with the work described in Chapter 6, to assess the strength of the therapeutic relationship between participants and the AI agent, we utilized the 4-item Session Rating Scale (SRS) (Duncan et al., 2003). The SRS is a widely used session-level instrument designed to capture clients' perceptions of the therapeutic process in a single session. It aligns with Bordin's (1994) foundational model of working alliance, which includes

three core components: shared goals, agreement on tasks, and the emotional bond. The four subscales of the SRS reflect these dimensions: (1) Relationship (feeling heard, understood, and respected), (2) Goals & Topics (alignment with what the user wished to discuss), (3) Approach or Method (fit of the agent’s approach), and (4) Overall Satisfaction. Participants rated each item on a 0–10 visual analogue scale, with higher scores indicating stronger therapeutic alliance. Prior work has used the SRS to evaluate human and AI-delivered therapy, including LLM-based agents (Dongre, 2024; Dongre et al., 2024), and cross-cultural validation studies have shown consistent psychometric properties across languages and populations (Hafkenscheid et al., 2010; She et al., 2023).

Empathy. As with the previous study, we evaluated perceived empathy using a three-item instrument developed by Sharma et al. (2020), specifically designed to capture the emotional and cognitive elements of empathetic responses in computational systems. The tool measures three key aspects: (1) Emotional Reactions (e.g., displays of warmth or concern), (2) Interpretations (inferring and acknowledging users’ emotions or experiences), and (3) Explorations (prompting reflection on unspoken feelings). These dimensions collectively provide insight into the depth and quality of the AI agent’s empathic communication.

Cultural Competence. Consistent with the previous study, to assess perceptions of cultural competency, we adapted four items from the Healthcare Provider Cultural Competency Measure (CCM) (Lucas et al., 2008), a validated tool developed to capture patient-reported experiences of provider cultural sensitivity. The full CCM includes nine items spanning three key domains—cultural knowledge, awareness, and skills. We selected one item from each domain, plus an overall item, in order to reduce respondent burden while preserving conceptual coverage. This reduced set demonstrated excellent internal consistency in prior work (Chapter 4; Cronbach’s $\alpha = 0.96$), aligning with previously reported reliability estimates in the literature (Cronbach, 1951; Dawadi et al., 2022; Alizadeh & Chavan, 2020; Nápoles et al., 2012).

Data Analysis

First, to ensure data quality, we set a threshold that surveys completed under five minutes were flagged and excluded, as internal testing within the research team concluded that it would be not feasible to thoughtfully complete both the chat interaction and evaluations in such a short time. Four data entries were excluded using this criterion. We used ANOVA and Chi-square to test if there was any major difference between the two groups among demographic variables (age, education level, English proficiency, length of residence in the U.S.) and baseline PANAS score. Within each group (non-adapted and adapted), Wilcoxon signed-rank tests were used to assess whether the within-subject pre-post changes in emotional states were statistically significant.

Next, we conducted independent samples t-tests to examine differences between groups (adapted vs. non-adapted) on change of emotional states (post- minus pre-session scores), therapeutic alliance, perceived empathy, and cultural competence. Prior to analysis, data were assessed for normality. Given the sample sizes, the Central Limit Theorem was deemed sufficient to support the robustness of the t-tests despite minor deviations from normality. Effect sizes were calculated using Cohen's d. Additionally, given the observed gender distribution differences between two groups, multiple linear regression analyses were conducted to examine whether group assignment (adapted vs. non-adapted) predicted each outcome variable while controlling for gender. Post-hoc exploratory analyses were also conducted on individual measured items to examine item-level difference between groups. Because this is a user study with limited sample size, constrained by recruitment feasibility, and the goal was to identify potential areas for further investigation, we also presented results that were trending significance ($p < 0.10$). However, any significant findings should be interpreted with caution given the limited sample size.

7.2 Results

A total of 58 family caregivers participated in this study: 26 interacted with the non-adapted agent, and 32 interacted with the culturally adapted agent. Table 7.1 summarizes the demographic characteristics of each group. No statistically significant differences were found between groups on key variables including age, education level, English proficiency, or years living in North America. However, there was a significant gender imbalance: the adapted group had a higher proportion of female participants (90.3% vs. 64.0%, $\chi^2 = 5.70$, $p = .02$). To account for this imbalance, subsequent analysis of group comparisons controlled for gender.

Table 7.1. Participant demographic characteristics by group assignment.

	Non-adapted agent group (n=26)	Adapted agent group (n=32)	t/χ^2	p-value
Age (years) (mean, SD, min-max)	50.5 (14.9, 19-79)	54.1 (19.2, 18-86)	0.77	.45
Gender (n, %)				
Male	9 (36.0%)	3 (9.7%)	5.70	.02*
Female	16 (64%)	28 (90.3%)		
Education (n, %)				
Some high school or less	1 (3.8%)	3 (9.4%)	4.80	.44
High school diploma or GED	2 (7.7%)	5 (15.6%)		
Some college, but no degree	3 (11.5%)	5 (15.6%)		
Bachelor's degree	10 (38.5%)	10 (31.3%)		
Graduate or professional degree	9 (34.6%)	6 (18.8%)		
Prefer not to say	1 (3.8%)	3 (9.4%)		
English proficiency (n, %)				
Not at all	2 (7.7%)	3 (9.4%)	0.06	.98
Not well	6 (23.1%)	7 (21.9%)		
Well	7 (26.9%)	9 (28.1%)		
Very well	9 (34.6%)	11 (34.4%)		
Prefer not to say	2 (7.7%)	2 (6.3%)		
Years living in the U.S. or Canada (mean, SD, min-max)	15.1 (11.7, 1-42)	16.1 (8.4, 1-39)	0.39	.70

Tables 7.2 and 7.3 show the results of Wilcoxon signed-rank tests to assess pre- and post-differences in emotional states (PANAS-X) within each group. Participants interacting with both the non-adapted and adapted agents showed significant improvements in fatigue, guilt, and sadness after the session ($p < .05$ for all). However, for positive affect, no significant change in serenity was observed ($p = .90$) for the non-adapted agent group while the adapted agent group showed a trend toward a significant improvement in serenity ($p = .07$, $r=0.74$). These findings suggest that both conversational agents led to immediate reductions in negative emotions, and that the culturally adapted agent may have conferred additional benefits in enhancing positive affect.

Table 7.2. Within-group pre–post changes in emotional states for the non-adapted agent group.

Affect	Change in Score M(SD)	Wilcoxon Stats (W)	Effect size (r)	p-value
Negative Affect				
Fatigue	-1.85 (2.43)	35.5	0.70*	.002
Guilt	-3.65 (4.53)	10.5	0.82*	.001
Sadness	-0.81 (1.86)	19.0	0.78*	.03
Positive Affect				
Serenity	+0.15 (2.07)	74.0	0.51	.90

Table 7.3. Within-group pre–post changes in emotional states for the culturally adapted group.

Affect	Change in Score M(SD)	Wilcoxon Stats	Effect size (r)	p-value
Negative Affect				
Fatigue	-2.12 (2.08)	4.0	0.86*	<.001
Guilt	-2.69 (3.60)	10.0	0.84*	<.001
Sadness	-1.12 (2.49)	31.5	0.77*	.01
Positive Affect				
Serenity	+0.56 (1.72)	39.0	0.74*	.07

Independent samples t-tests revealed no statistically significant differences between groups on any outcome variable (all $p > 0.10$). Effect sizes were small, ranging from $d = 0.10$ to $d = 0.35$. Post-hoc power analysis indicated that the current sample size ($n = 58$) only had adequate power ($>80\%$) to detect medium to large effect sizes ($d \geq 0.5$) but limited power to detect small effects ($d < 0.5$). Power estimates further suggested that more than 200 participants would be required in a future study to reliably detect small effects with statistical significance. Although group differences did not reach statistical significance on these overall composite measures (p 's $> .10$), participants in the adapted agent group reported consistently higher scores across all three domains including therapeutic alliance, empathy and cultural competence (Table 7.5).

Table 7.4. Between-group comparisons of outcome measures using independent samples t-tests.

Variable	Non-adapted version M (SD)	Adapted version M (SD)	t-statistic	df	p-value	Cohen's d
Fatigue (pre–post diff)	-1.85 (2.43)	-2.12 (2.08)	-0.46	49.4	0.64	-0.13
Guilt (pre–post diff)	-3.65 (4.53)	-2.69 (3.60)	0.88	47.2	0.38	0.24
Sadness (pre–post diff)	-0.81 (1.86)	-1.12 (2.49)	-0.56	55.6	0.58	-0.14
Serenity (pre–post diff)	0.15 (2.07)	0.56 (1.72)	0.80	48.6	0.42	0.22
Therapeutic Alliance (SRS)	6.29 (2.82)	6.62 (2.57)	0.45	50.8	0.66	0.12
Empathy	3.60 (1.16)	3.71 (0.96)	0.37	48.6	0.71	0.10
Cultural Competence (CCM)	4.12 (1.82)	4.72 (1.62)	1.32	50.7	0.19	0.35

As shown in Table 7.5, after controlling for gender ($\text{Outcome} = \beta_0 + \beta_1(\text{Group}) + \beta_2(\text{Gender}) + \epsilon$), group assignment was still not a significant predictor of any outcome variable (all $p > .10$). This is consistent with and strengthens the results shown in Table 7.4 that adapted and non-adapted versions have no significant difference in effects.

Table 7.5. Multivariate regression results: effect of group assignment on outcome measures controlling for gender.

Variable	Group Coefficient (β_1)	p-value
Fatigue (pre–post diff)	-0.19	0.70
Guilt (pre–post diff)	1.02	0.40
Serenity (pre–post diff)	0.33	0.50
Sadness (pre–post diff)	-0.24	0.70
Therapeutic Alliance (SRS)	-0.11	0.80
Empathy	0.04	0.80
Cultural Competence (CCM)	0.52	0.20

Exploratory post-hoc analysis of individual CCM items revealed a significant group difference ($t = 2.13$, $p = 0.046$, $d = 0.56$) on the item: "Would you recommend the AI therapist to someone with a Chinese cultural background? 1=Not at all, 7=Highly recommend." Participants in the adapted group reported significantly higher recommendation scores ($M = 5.03$, $SD = 1.79$) compared to the non-adapted group ($M = 4.00$, $SD = 1.88$). This medium effect size ($d = 0.56$) suggests that while overall cultural competence scores did not differ significantly between groups, cultural adaptation specifically enhanced participants' willingness to endorse the agent within their cultural community.

7.3 Discussion

This user study aims to evaluate whether the culturally adapted AI agent, adapted through a context-engineering approach using a cultural context database, would yield better short-term emotional outcomes, alliance, and empathy for Chinese-speaking caregivers living in North America, compared to a generic, non-adapted baseline agent. Participants in both groups reported significant reductions in negative affects (fatigue, guilt, sadness) following a single-session interaction. Notably, only the culturally adapted agent led to a statistically significant improvement in positive affect (serenity). These results suggest that caregivers were

able to get measurable emotional benefits from even a single-session interaction with a therapeutic conversational AI agent, which is particularly promising for populations under high stress, such as family caregivers. In contrast to prior studies of non-LLM-based therapy chatbots, which did not observe significant changes in PANAS scores (e.g., Liu et al., 2022; Fitzpatrick et al., 2017), our results point to the promise of LLMs in facilitating emotionally supportive therapeutic conversations.

Despite the overall positive impact, between-group comparisons did not reveal statistically significant differences across emotional states improvements and most post-session evaluation metrics, including therapeutic alliance, empathy, and cultural competence. However, the culturally adapted agent consistently achieved higher mean scores across these dimensions, indicating directional improvements. Importantly, participants who interacted with the adapted agent were significantly more likely to recommend it to someone with a similar cultural background, suggesting a potentially higher perceived cultural alignment and acceptability. Even though differences in measured outcomes are modest, enhanced cultural alignment and acceptability may influence user endorsement and long-term engagement, which will potentially affect adoption and dissemination in real-world mental health interventions.

Several factors may help explain the absence of statistically significant between-group differences. First, individual differences in participants' characteristics and therapeutic preferences may have introduced substantial variability that obscured the cultural adaptation effects in our between-subjects design. Future studies might benefit from statistically controlling for key variables such as gender, age, acculturation level, therapeutic expectations, and baseline mental health status. Second, the interaction with the agents in this study is limited to only one single session and the agent focused solely on symptom assessment, a relatively structured phase of problem-solving therapy in which even non-adapted agents may perform competently. In this context, the added value of cultural adaptation may not be fully realized, as the session

primarily involved prompting users to describe their stressors—an interaction that does not necessarily require culturally grounded responses. Third, the study’s modest sample size limited statistical power. With only 25–30 participants per group, the study was underpowered to detect small-to-moderate effects. A post-hoc power analysis indicated that over 200 participants per group would be required to detect a medium effect size ($d \sim 0.4$) on alliance with 80% power at a conventional alpha level of .05. We did observe a consistent trend (adapted > non-adapted baseline), which could indicate small but meaningful effects that may become evident in a larger trial. Further research is needed to determine the differences and impact of cultural adaptation on long-term, real-world effectiveness of therapeutic AI agents.

The higher willingness to recommend the adapted agent to someone of similar background is a key finding. It might suggest that the culturally adapted version had higher perceived fit from the participants. Our adapted chatbot, by demonstrating knowledge of Chinese cultural nuances, may have appeared more “made for” Chinese Americans, hence participants felt more confident recommending it to someone also with Chinese cultural background. The practical implication is that even if immediate health outcomes were not significantly different, cultural adaptation might improve uptake and dissemination. As such, future evaluations might look at engagement metrics, for example, one may measure caregivers assigned to a culturally adapted agent have lower dropout rates and if use is sustained for longer durations than those with a general agent.

Limitations

This study has several limitations. First, the sample size was small, so results should be considered preliminary. Second, we looked at only immediate, short-term emotional changes after a single session. It is unclear whether any emotional state improvements would last in the long-term and yield long-term benefits. Moreover, because of the single session format, we could not assess engagement or adherence rate, which could also be the value of cultural

adaptation. Third, the perceived empathy measure (Sharma et al., 2020) we used in this study was developed in the context of peer support online rather than therapy sessions. This contextual mismatch may have limited the instrument's sensitivity to detect empathy differences between the two groups in therapeutic interactions with AI conversational agents. Last but not the least, our cultural context database covers limited underlying reasons and therefore, there were possibilities that the specific content of conversations that users in this study discussed, even though culturally relevant, didn't always trigger the cultural adaptation workflow. We didn't strictly control the conversation topics beyond instructing them to discuss a caregiving problem. It could be that some participants talked about issues where culture was less obviously involved (e.g., a participant talked about difficulties to find parking), resulting in the adapted bot not always retrieving extra cultural context. For those participants, the two agents would have behaved almost identically, thus diluting differences. In future work, we should continue to expand the cultural context database to be more comprehensive.

Future work

This study naturally points to several next steps. First, a more robust trial with a larger sample and extended follow-up period (e.g., several weeks) is needed to provide stronger evidence regarding the effects of cultural adaptation on both mental health outcomes and user engagement indicators such as usage frequency and dropout rates. In particular, future studies should further investigate the impact of cultural adaptation on willingness to recommend the agent to peers within the same cultural group, a significant finding in this study, and examine how this factor influences adoption and sustained engagement, which remain critical gaps.

Second, future research should explore qualitative differences in how users interact with and perceive culturally adapted versus non-adapted agents. For example, chat transcripts could be analyzed to determine whether users disclose more emotions when the agent demonstrates cultural understanding. One testable hypothesis is that cultural responsiveness facilitates deeper

and more meaningful conversations. Complementary to this, exit interviews should be conducted to capture participants' qualitative feedback, providing a feedback loop for refining and expanding the cultural context database.

Third, while this study was designed to compare adapted and non-adapted agents, we did not systematically track how often the cultural adaptation workflow was triggered or which cultural context entries were retrieved during participants' conversations. Future studies should quantify the frequency and circumstances under which cultural adaptation is invoked, providing greater clarity on how augmented context is used and how it contributes to outcomes. In addition, qualitative analyses are needed to assess whether the retrieved data were contextually relevant and genuinely enhanced cultural responsiveness within each conversation. Finally, the cultural context database could be expanded to include challenges faced by more diverse user populations beyond Chinese family caregivers of older adults, thereby extending the culturally adapted agent to broader and more varied communities while testing the generalizability of these findings.

7.4 Conclusion

This study demonstrates that culturally adapted AI therapeutic agents can achieve meaningful immediate emotional improvements among Chinese American family caregivers, despite limited statistical power to detect between-group differences. While both culturally adapted and non-adapted agent versions produced significant improvements in negative emotions, only the adapted agent improved positive emotional states and generated significantly higher willingness for the users to recommend it to their cultural community. Limited sample size combined with relatively high variation of individual differences in participants characteristics may have obscured the between-group treatment effects, suggesting future studies should employ larger samples and statistical control of key covariates. These

findings provide preliminary evidence that cultural adaptation of AI therapeutic agents enhances immediate positive emotional states outcome and willingness to endorse the agent within users' cultural community, warranting larger-scale investigation to establish the effectiveness of culturally responsive therapeutic AI agents on improving health outcomes and user engagement.

CHAPTER 8. Discussion & Conclusion

8.1 Discussion

This dissertation investigated how large language model (LLM)-driven conversational agents can be culturally adapted and evaluated to better support mental health in underserved communities (Chinese-speaking and Spanish-speaking caregivers in the U.S.). Addressing a critical gap in digital health equity, this work demonstrates that integrating cultural context into AI systems, through the development of a cultural context database and a tailored LLM workflow for retrieval and integration, enhances users' perceptions of empathy, cultural responsiveness, and therapeutic alliance. Below, I synthesize the key findings across the three aims and discuss their implications for culturally responsive digital mental health interventions.

In Aim 1 (Chapter 3), through qualitative research, I identified culturally salient challenges among our primary target population—Chinese American family caregivers. These insights informed the construction of a structured cultural context database containing caregiving challenges and example responses. By grounding this resource in community-informed needs, I laid a foundation for cultural adaptation that covers the culturally salient topics and reflects users' lived experiences and that resonate with the specific populations. A parallel study was also conducted to draw insights from Spanish-speaking Latino family caregivers and provide foundation for the cultural adaptation of the conversational AI agent for this population.

In Aim 2, I developed and evaluated two methods for culturally adapting an LLM-powered therapeutic agent: (1) a prompt-based technique, and (2) a context engineering and retrieval-augmented generation (RAG)-enabled method that enables real-time context retrieval and integration using the cultural context database developed based on Aim 1 insights. To evaluate these techniques, I conducted two within-subject comparison studies.

The first evaluation study examined the effect of cultural prompting. Participants compared LLM-generated responses with and without cultural prompting. Results showed that the culturally prompted responses were perceived as significantly more empathetic and culturally responsive, validating the utility of this lightweight adaptation technique and supporting its inclusion in later comparisons.

The second evaluation study directly compared three agents: (1) a baseline LLM without adaptation, (2) an agent using cultural prompting, and (3) a culturally adapted agent using the cultural context engineering workflow. In contrast to the first evaluation study in which participants evaluated static input-response pairs, participants interacted with each version of the agent in a live conversational setting. Findings indicated that the dynamic context-engineered agent consistently received the highest ratings across all outcomes, including perceived empathy, cultural relevance, cultural competence, and therapeutic alliance, with statistically significant improvements over the baseline. These results support the effectiveness of the context engineering adaptation over the static cultural prompting method.

In Aim 3, a two-arm user study evaluated the culturally adapted agent compared to the baseline agent in real-world interactions with Chinese American family caregivers. Both agents demonstrated significant improvements in negative emotional states (guilt, fatigue, sadness) while the culturally adapted agent also showed trends toward improving positive affect (serenity). Although between-group differences did not reach statistical significance due to limited sample size, the adapted agent consistently trended toward better outcomes across cultural competence, empathy, and therapeutic alliance measures. The adapted agent significantly improves willingness to recommend to peers with the same cultural background. These findings suggest that LLM-driven conversational agents provide immediate emotional benefits for caregivers, with cultural adaptation offering additional advantages in positive emotional outcomes and strengthening community endorsement, which may enhance adoption

within the Chinese American community and address the critical gap of low utilization of mental health services. Larger-scale, longitudinal studies are needed to more robustly assess the sustained impact of cultural adaptation on mental health outcomes and user engagement.

In Aim 3, a two-arm user study evaluated the culturally adapted agent compared to the baseline agent in real-world interactions with Chinese American family caregivers. Both agents demonstrated significant improvements in negative emotional states (guilt, fatigue, sadness), while the culturally adapted agent also showed trends toward improving positive affect (serenity). Although between-group differences did not reach statistical significance due to limited sample size, the adapted agent consistently trended toward better outcomes across cultural competence, empathy, and therapeutic alliance measures. The adapted agent significantly improved participants' willingness to recommend the system to peers with the same cultural background. These findings suggest that LLM-driven conversational agents provide immediate emotional benefits for caregivers, with cultural adaptation potentially offering additional advantages in positive emotional outcomes and community acceptability. These findings suggest the need for larger-scale, longitudinal studies to more robustly assess the impact of cultural adaptation on mental health outcomes and user experience.

8.2 Contributions

First, findings from this work demonstrate that the dynamic context engineering approach for adapting LLM-driven conversational agents to mental health support outperformed the prompt-based adaptation approach in cultural responsiveness, empathy, and therapeutic alliance. Both adapted versions outperformed the non-adapted agent. Moreover, while both adapted and non-adapted agents produced significant immediate emotional benefits for stressed caregivers, the culturally adapted agents using the context engineering approach

received significantly higher user endorsement, as measured by caregivers' willingness to recommend the system to others in their community.

Second, this work makes a contribution through the development of a **structured cultural context database** specifically designed for context engineering-based cultural adaptation of mental health conversational agents. Grounded in comprehensive qualitative research with Chinese American family caregivers and community stakeholders, the database is organized in three aspects: “challenges, underlying reasons, and culturally responsive response examples.” This structure was developed and refined through iterative rapid prototyping and assessment with a multidisciplinary team of bilingual and bicultural researchers. The database represents the first structured repository for LLM-based cultural adaptation in mental health support contexts. Beyond its application in this dissertation work, the database holds broader potential for application in other contexts serving this population, such as informing culturally tailored caregiver support programs. Moreover, following the same structure, the database can be extended to encompass additional caregiving contexts beyond caring for older adults and adapted for other cultural groups, providing a blueprint for developing knowledge resources to inform future cultural adaptation of mental health agents.

Third, this dissertation contributes a **generalizable workflow for culturally adapting LLM-driven mental health conversational agents**. Specifically, it introduces a community-informed, dynamic context-engineering workflow that extends beyond static cultural prompting by enabling real-time context integration. The design of this workflow is inspired by Barrera and Castro's (2006) heuristic framework and grounded in principles of human-centered design, which emphasize incorporating the perspectives and lived experiences of end users and stakeholders throughout the development process. The workflow begins with the elicitation of population-specific caregiving challenges and stressors, gathered through qualitative research with community stakeholders. These insights are then used to construct a

structured cultural context database containing key caregiving challenges, their underlying cultural interpretations, and example supportive responses. During real-time interaction, the LLM workflow detects user disclosures that reflect culturally salient concerns, retrieves relevant content from the database, and integrates this information into the model's context. This process ensures that responses are dynamically generated in a manner directly informed by community knowledge and expert insights. Importantly, this workflow is generalizable as the database is community informed and can be easily extended with salient challenges, underlying interpretations, and example responses that correspond to other subgroups (e.g., parents of children with Autism, Latino family caregivers).

Besides technical development, this dissertation advances the evaluation of culturally adapted mental health AI agents through the design and implementation of a contextualized, multiphase evaluation procedure. Aim 1 drew on human-centered design methods to ground the system in user needs (Maguire, 2001), using participatory and community-based approaches to identify caregiver priorities and culturally specific requirements. Aim 2 applied principles of iterative system prototyping and evaluation from health informatics (Kushniruk & Patel, 2004) to refine prototypes. The overall framing drew on Shneiderman's Human-Centered AI framework (2020), which emphasizes the centrality of human needs, values, and oversight in AI evaluation. Unlike existing approaches for assessing cultural alignment in LLMs (Masoud et al., 2025; Tao et al., 2024), which rely largely on computational metrics, this dissertation expanded evaluation to methods that directly engaged users in reflecting on their perceptions and experiences. The multiphase evaluation procedure incorporated constructs from mental health and cultural adaptation research, including measures used in adapting evidence-based interventions, indicators of therapeutic experiences such as empathy and alliance, and preliminary outcomes in emotional well-being. The procedure also reflects principles of progressive, staged evaluation for real-world deployment (Mohr et al., 2017) and parallels the pursuit of higher ecological validity in cultural adaptation research (Bernal et al., 1995).

Specifically, these phases map directly onto the dissertation’s aims. Phase 1 (Aim 2.1) involved static evaluation in a controlled setting to test whether cultural adaptations produced detectably different outputs; Phase 2 (Aim 2.2) introduced interactive evaluation to assess subjective constructs critical to therapeutic efficacy and provide proof-of-concept for real-time adaptation; and Phase 3 (Aim 3) extended evaluation to end-users in real-world settings, exploring short-term effects on emotional well-being. Taken together, these three phases demonstrate how a human-centered (i.e., Phase 1-3) and contextualized evaluation strategy (i.e., Phase 2 and 3) can generate a holistic understanding of cultural adaptation in AI systems. Together, this dissertation work contributes an example for future researchers seeking to evaluate AI system’s cultural alignment and responsiveness beyond computational evaluations, demonstrating how iterative, staged, and culturally grounded assessments can be systematically operationalized in practice.

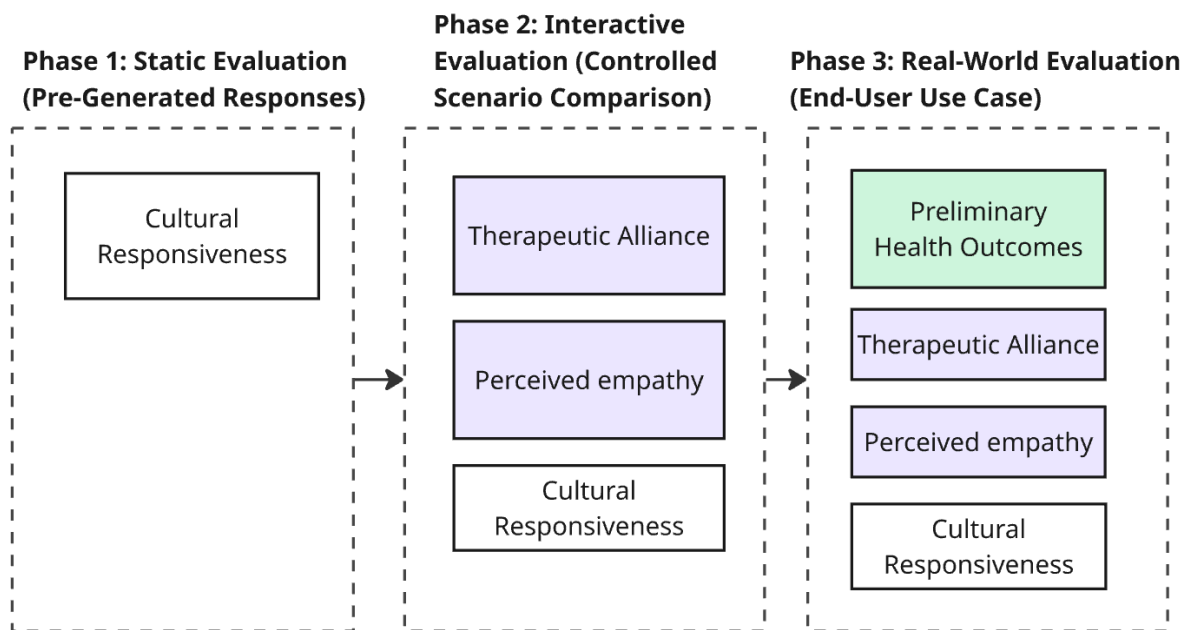


Figure 8.1. Multiphase evaluation procedure for cultural adaptation of LLM-based mental health agents. Each phase builds on the previous one by adding new evaluation outcomes. Phase 1 (static evaluation) assesses cultural responsiveness (white). Phase 2 (interactive evaluation under controlled scenarios) includes therapeutic alliance and perceived empathy (purple), alongside cultural responsiveness. Phase 3 (real-world evaluation) additionally incorporates preliminary health outcomes (green).

8.3 Limitations and Future Directions

While this dissertation advances both the implementation and evaluation of culturally adapted LLM-driven therapeutic conversational agents, several limitations should be acknowledged to contextualize the findings and inform directions for future research.

Single-Session Intervention Design

This study was limited to a single-session interaction with the AI agent, focusing primarily on immediate, short-term outcomes. While this design enabled controlled comparisons across versions and efficient evaluation of the intervention's initial effects, it did not capture patterns of long-term engagement or sustained mental health benefits. In reality, psychological support—particularly for caregivers coping with chronic stress—often requires ongoing, repeated interactions to meaningfully influence emotional well-being, coping strategies, and mood regulation. It remains uncertain whether the observed short-term improvements (e.g., increased positive affect or reduced negative emotions) would persist, diminish, or evolve with continued use. Future longitudinal research is needed to assess user behaviors such as retention and adherence, as well as to examine changes in emotional well-being over time resulting from repeated interactions with culturally adapted AI agents in real-world caregiving contexts.

Limited Cultural Coverage and Retrieval Sensitivity

While the cultural context database was developed through qualitative research with bilingual experts and encompasses a variety of common caregiving scenarios, it remains in its early stages and does not yet reflect the full diversity and complexity of cultural experiences among Chinese American caregivers—or among individuals from other cultural backgrounds. As a result, not all user inputs triggered culturally relevant retrievals, either due to gaps in the database's content or limitations of the gatekeeping LLM in recognizing implicit cultural cues.

Moreover, because this study did not systematically track how often the cultural adaptation workflow was triggered or which entries were retrieved, it remains unclear how frequently and in what circumstances adaptation was actually utilized in practice. This lack of visibility may have led to underutilization of the adaptation workflow and diluted observable differences between adapted and non-adapted agents.

To enhance inclusivity and effectiveness, future work should expand the cultural context database to more comprehensively represent diverse subpopulations and caregiving experiences, ensuring a broader range of user inputs activate culturally relevant retrievals. Additionally, improving the LLM's ability to detect implicit cultural cues is needed; future research should explore combining embedding-based similarity with LLM-based gatekeeping to more accurately determine when to trigger the RAG workflow. Systematically logging trigger events and retrieved content will also be essential to assess how augmented context is used, its relevance to user concerns, and its contribution to outcomes.

Currently, the database is centered on challenges faced by Chinese family caregivers and reflects Chinese cultural norms and values. Its applicability to other populations, such as Latino or Indigenous caregivers, has yet to be systematically evaluated. Ongoing efforts are extending this workflow to Spanish-speaking Latino family caregivers, using a similar methodology to build and assess a parallel cultural context database. Evaluating this approach with additional populations will be essential for determining the generalizability and scalability of the overall workflow.

Lack of Qualitative Data Analysis

Although this study collected voluntary open-ended feedback and session transcripts, a detailed qualitative analysis of participant feedback and conversational content was beyond the scope of the current work. As a result, we were unable to fully explore where and why the

adapted agent may have fallen short in specific scenarios. Rich qualitative data, such as moments of emotional disconnection, could offer valuable insight into the limitations of the current adaptation and highlight areas for refinement. In particular, qualitative analysis could help assess whether retrieved cultural content was contextually relevant and perceived as useful by participants, thereby clarifying how and when cultural adaptation contributes to responsiveness. Such analysis could also reveal patterns in how cultural nuances are missed or mishandled by the agent, as well as identify missing themes, overlooked norms and values, or new caregiving scenarios that warrant inclusion. Future work should supplement quantitative measures with qualitative findings, driving iterative improvement of the cultural context database and the overall design of culturally responsive AI agents.

Safety Integration and Integration Over Replacement

As LLM-driven agents become more prevalent in mental health contexts, safeguarding users must remain a top priority. Future development should incorporate real-time risk detection systems capable of identifying high-risk disclosures—such as suicidal ideation—and triggering appropriate escalation protocols. This includes establishing clear boundaries for agent autonomy, implementing robust guardrails, and ensuring compliance with emerging regulatory requirements. For example, Utah’s recently enacted legislation (HB0452, 2025) mandates that licensed mental health providers be involved in the development of mental health apps and that such apps include in-app safety mechanisms. Equally important is transparency about how risks are detected and managed, as this can help build user trust in AI-based mental health interventions. Additionally, AI conversational agents in mental health should function as supportive tools rather than replace traditional therapeutic treatment. The American Psychological Association emphasizes that no AI chatbot has received FDA approval for diagnosing or treating mental health disorders, and there are significant risks when vulnerable

populations mistake chatbots for actual therapists (Khawaja & Belisle-Pepon, 2023; Abrams, 2025).

Psychoeducation and Engagement

Situated within Chu and Leino's (2017) cultural adaptation framework, this dissertation focused primarily on the Treatment Delivery component and did not address the Engagement component, which encompasses treatment access, entry, and psychoeducation. For populations such as Chinese Americans, psychoeducation is particularly critical. As noted by Hwang (2009) in his empirical work on the Formative Method for Adapting Psychotherapy (FMAP), it is important to provide education to clients about the course of therapy and set clear expectations at the outset. These early engagement efforts can help reduce stigma, increase understanding of therapeutic processes, and improve retention—factors that are often culturally mediated. While this work demonstrated the value of cultural adaptation in AI-driven treatment delivery, future research should also explore how psychoeducational elements can be integrated into LLM-powered mental health agents to support culturally aligned engagement from the very beginning of the therapeutic interaction.

Personalization Beyond Culture

Although this dissertation focused on cultural adaptation based on users' ethnic backgrounds, it is important to acknowledge that cultural identity extends beyond ethnicity. For example, groups such as military veterans may share distinct values, norms, and communication styles that constitute a cultural identity. Recognizing these diverse cultural dimensions opens pathways for expanding adaptation frameworks to other non-ethnic cultural groups.

In addition, future work should explore broader forms of personalization that account for individual-level differences. Tailoring AI interactions based on users' coping styles (e.g., taking a bath vs. engaging in physical exercise; Clobert et al., 2020), emotional states, or personality traits (e.g., extraverted vs. introverted) could enhance the responsiveness and

therapeutic potential of LLM-driven agents. Integrating these individual dimensions alongside cultural context may allow AI systems to deliver more nuanced, adaptive, and user-centered support that fully leverages the generative capacity of large language models. Advances in user modeling, affective computing, and real-time context sensing offer promising avenues to support this direction, enabling AI agents to adapt dynamically to both situational cues and personal characteristics. This line of research could further advance the broader goal of delivering equitable, personalized, and culturally responsive mental health support at scale.

8.4 Generalizability and Application

Although this work focused on Chinese-speaking caregivers in North America, the underlying methods are designed to be generalizable to other cultural groups and caregiving contexts. The dynamic context-engineering adaptation mechanism supports extensibility to additional populations—such as Latino or Indigenous family caregivers—without requiring substantial modifications to the overall LLM workflow. Moreover, the design is well-suited for low-resource environments, such as community health centers, where culturally adapted agents can provide meaningful support to clients. Its reliance on context engineering rather than model retraining makes the approach particularly attractive for equity-oriented health informatics applications.

That said, findings from Chapter 6 indicate that prompt-based adaptation had a more limited impact compared to the dynamic context-engineering approach. This raises important questions about the trade-offs between implementation simplicity and contextual richness in real-world settings. While the context engineering workflow offers more dynamic and personalized responses, its added complexity—such as the need for a curated cultural context database and real-time retrieval—may not always be warranted. If real-world deployments (e.g., Chapter 7) reveal minimal differences in user outcomes between context-engineered and

prompt-based approaches, one might reasonably argue that prompting methods are “good enough” for certain practical applications, especially in settings where resources for database construction are constrained.

To better inform these trade-offs, future research should conduct comparative deployments of prompt-adapted and dynamic context-engineered agents in similar pilot environments. Such evaluations would yield valuable insights into the cost-benefit dynamics of each approach and help guide scalable deployment strategies in diverse contexts.

8.5 Conclusion

This dissertation advances the field of culturally responsive AI for mental health by introducing a scalable method for adapting LLM-driven conversational agents to provide culturally responsive support. The approach is designed to incorporate community stakeholders’ insights and is aligned with the emerging generation of GenAI-powered health technologies. While broader application across diverse cultural groups and health contexts will be essential to further refine, validate, and generalize its utility, this work represents an important step toward reducing disparities in access to and quality of mental health care, advancing more personalized digital mental health support for caregivers, and fostering the future adoption and effective utilization of digital mental health technologies to address the persistent gap in mental health service utilization among marginalized and minority communities.

References

- AARP. (2020, May 14). Caregiving in the United States 2020. AARP.
<https://doi.org/10.26419/ppi.00103.001>
- Abbasi MA, Ghafouri A, Firouzmandi M, et al. PersianLLaMA: Towards building first Persian large language model. arXiv [cs.CL]. 2023.
- Abrams, Z. (2025). Using generic AI chatbots for mental health support: A dangerous trend. American Psychological Association.
<https://www.apaservices.org/practice/business/technology/artificial-intelligence-chatbots-therapists>
- Ahad AA, Sanchez-Gonzalez M, Junquera P. Understanding and addressing mental health stigma across cultures for improving psychiatric care: A narrative review. *Cureus*. 2023;15:e39549.
- Alizadeh S, Chavan M. Perceived Cultural Distance in Healthcare in Immigrant Intercultural Medical Encounters. *Int Migr* 2020; 58: 231–254.
- Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K., Badreldin, H. A., Al Yami, M. S., Al Harbi, S., & Albekairy, A. M. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. *BMC Medical Education*, 23(1), 689.
- American Psychological Association. More than a quarter of U.S. adults say they're so stressed they can't function. (2022, October 19).
<https://www.apa.org/news/press/releases/2022/10/multiple-stressors-no-function>
- Anderson REE, Spence SH, Donovan CL, et al. Working alliance in online cognitive behavior

- therapy for anxiety disorders in youth: comparison with clinic delivery and its role in predicting outcome. *J Med Internet Res.* 2012;14:e88.
- Arora, A., Kaffee, L.-A., & Augenstein, I. (2023). Probing Pre-Trained Language Models for Cross-Cultural Differences in Values. In S. Dev, V. Prabhakaran, D. Adelani, D. Hovy, & L. Benotti (Eds.), *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)* (pp. 114–130). Association for Computational Linguistics.
- Bai, Y., Du, X., Liang, Y., Jin, L., Zhou, J., Liu, Z., Fang, F., Chang, M., Zheng, T., Zhang, X., Ma, N., Wang, Z. M., Yuan, R., Wu, H., Lin, H., Huang, W., Zhang, J., Lin, C., Fu, J., ... Zhang, G. (2025). COIG-CQIA: Quality is all you need for Chinese instruction fine-tuning. *Findings of the Association for Computational Linguistics: NAACL 2025*, 8190–8205.
- Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., Do, Q. V., Xu, Y., & Fung, P. (2023). A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. In *arXiv [cs.CL]*. arXiv. <http://arxiv.org/abs/2302.04023>
- Barrera, M., & Castro, F. G. (2006). A heuristic framework for the cultural adaptation of interventions. *Clinical Psychology: A Publication of the Division of Clinical Psychology of the American Psychological Association*, 13(4), 311–316.
- Beasley, L. O., King, C., Esparza, I., Harnden, A., Robinson, L. R., So, M., Morris, A., & Silovsky, J. F. (2021). Understanding initial and sustained engagement of Spanish-speaking Latina mothers in the Legacy for children program™: A qualitative examination of a group-based parenting program. *Early Childhood Research Quarterly*, 54, 99–109.

- Beck T, Schuff H, Lauscher A, et al. Sensitivity, Performance, Robustness: Deconstructing the Effect of Sociodemographic Prompting. Epub ahead of print 8 February 2024. DOI: 10.48550/arXiv.2309.07034.
- Beck, AT. (1976). *Cognitive Therapy and the Emotional Disorders*. International Universities Press, New York.
- Benkler, N., Mosaphir, D., Friedman, S., Smart, A., & Schmer-Galunder, S. (2023). Assessing LLMs for Moral Value Pluralism. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2312.10075>
- Blodgett SL, Cheung JCK, Liao V, et al. Human-Centered Evaluation of Language Technologies. In: Li J, Liu F (eds) *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*. Miami, Florida, USA: Association for Computational Linguistics, pp. 39–43.
- Bordin ES. Theory and research on the therapeutic working alliance: New directions. *The working alliance: Theory, research, and practice*. 1994;304:13–37.
- Borna S, Maniaci MJ, Haider CR, et al. Artificial Intelligence Support for Informal Patient Caregivers: A Systematic Review. *Bioengineering* 2024; 11: 483.
- Bosco, C., Shojaei, F., Theisz, A. A., Nguyen, V., Song, H., Han, R., Osorio Torres, J. A., Chheda, D., Lin, J., Peng, X., Waseem, N. Z., Simpkins, C., Cureton, B., Himes, A. K., Jessup, N. M., Lu, Y., Hendrie, H. C., Barnes, P. A., Hill, C. V., & Shih, P. C. (2025). “I don’t see anything specifically about Black/African Americans.” Testing an Alzheimer-specific generative AI tool tailored for African American/Black communities. *ACM Transactions on Computing for Healthcare*. <https://doi.org/10.1145/3712295>
- Breunig, D. (2025, June 22). How Long Contexts Fail. Drew Breunig.

<https://www.dbreunig.com/2025/06/22/how-contexts-fail-and-how-to-fix-them.html>

Budiman ZT, Carolyn Im, Sahana Mukherjee and Abby. Why Asian Immigrants Come to the U.S. and How They View Life Here. Pew Research Center, <https://www.pewresearch.org/race-and-ethnicity/2024/10/09/why-asian-immigrants-come-to-the-u-s-and-how-they-view-life-here/> (2024, accessed 23 March 2025).

Busch F, Prucker P, Komenda A, et al. Multilingual feasibility of GPT-4o for automated Voice-to-Text CT and MRI report transcription. *Eur J Radiol* 2025; 182: 111827.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Lagos, M., Diez-Medrano, J., Norris, P., Ponarin, E., & Puranen, B. (2022). World values survey: Round seven – Country-pooled datafile (Version 3.0) [Data set]. JD Systems Institute & WWSA Secretariat.

Cabassa LJ, Lester R, Zayas LH. “It’s like being in a labyrinth:” Hispanic immigrants’ perceptions of depression and attitudes toward treatments. *J Immigr Minor Health*. 2007 Jan;9(1):1-16. doi: 10.1007/s10903-006-9010-1. PMID: 17001516; PMCID: PMC3160757.

Cao, Y., Zhou, L., Lee, S., Cabello, L., Chen, M., & Hershcovich, D. (2023). Assessing cross-cultural alignment between ChatGPT and human societies: An empirical study. *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, 53–67.

Castro, F. G., Barrera, M., Jr, & Holleran Steiker, L. K. (2010). Issues and challenges in the design of culturally adapted evidence-based interventions. *Annual Review of Clinical Psychology*, 6(1), 213–239.

Center for Substance Abuse Treatment (US). Improving Cultural Competence. Rockville (MD):

- Substance Abuse and Mental Health Services Administration (US); 2014. (Treatment Improvement Protocol (TIP) Series, No. 59.) 3, Culturally Responsive Evaluation and Treatment Planning. Available from:
https://www.ncbi.nlm.nih.gov/books/NBK248423/?utm_source=chatgpt.com
- Chang, C.-C., Chang, H.-P., & Lee, H.-S. (2024). Leveraging RetrievalAugmented Generation for Culturally Inclusive Hakka Chatbots: Design Insights and User Perceptions. arXiv.
- Chase, H. (2025, June 23). The rise of “context engineering.” LangChain Blog.
<https://blog.langchain.com/the-rise-of-context-engineering/>
- Chew, L. D., Griffin, J. M., Partin, M. R., Noorbaloochi, S., Grill, J. P., Snyder, A., Bradley, K. A., Nugent, S. M., Baines, A. D., & Vanryn, M. (2008). Validation of screening questions for limited health literacy in a large VA outpatient population. *Journal of General Internal Medicine*, 23(5), 561–566.
- Chi, I., & Boot, W. (2019). Interest group session—Chinese gerontology studies: Developing a Body-Mind-Spirit Caregiver Self-Management Program and a mobile application for Chinese Americans. *Innovation in Aging*, 3(Supplement_1), S546–S546.
- Chi, I., Liu, M., & Wang, S. (2020). Developing the Asian pacific islander caregiver train-the-trainer program in the United States. *Journal of Ethnic & Cultural Diversity in Social Work*, 29(6), 490–507.
- Chiu YY, Jiang L, Lin BY, et al. CulturalBench: A robust, diverse and challenging benchmark on measuring the (lack of) cultural knowledge of LLMs. arXiv [cs.CL]. 2024.
- Cho, Y.-M., Rai, S., Ungar, L., Sedoc, J., & Guntuku, S. C. (2023). An integrative survey on mental health conversational agents to bridge computer science and medical perspectives. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

- Conference on Empirical Methods in Natural Language Processing, 2023, 11346–11369.
- chroma: the AI-native open-source embedding database. (n.d.). Github. Retrieved July 5, 2025, from <https://github.com/chroma-core/chroma>
- Chu, W., Wippold, G., & Becker, K. D. (2022). A Systematic Review of Cultural Competence Trainings for Mental Health Providers. *Professional Psychology, Research and Practice*, 53(4), 362. <https://doi.org/10.1037/pro0000469>
- Chu, J., & Leino, A. (2017). Advancement in the maturing science of cultural adaptations of evidence-based interventions. *Journal of Consulting and Clinical Psychology*, 85(1), 45–57.
- Clobert, M., Sims, T. L., Yoo, J., Miyamoto, Y., Markus, H. R., Karasawa, M., & Levine, C. S. (2020). Feeling excited or taking a bath: Do distinct pathways underlie the positive affect-health link in the U.S. and Japan? *Emotion*, 20(2), 164–178.
- COCO - Care for Caregivers. (n.d.). Retrieved July 23, 2025, from <https://coco.health/>
- Corrigan, P., & Matthews, A. (2003). Stigma and disclosure: Implications for coming out of the closet. *Journal of Mental Health (Abingdon, England)*, 12(3), 235–248.
- Cronbach LJ. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 1951; 16: 297–334.
- Dai, X., Zhou, L., Wang, B., & Li, H. (2025). From word to world: Evaluate and mitigate culture bias via word association test. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2505.18562>
- Darcy A, Daniels J, Salinger D, et al. Evidence of human-level bonds established with a digital conversational agent: Cross-sectional, retrospective observational study. *JMIR Form Res*. 2021;5:e27868.
- David, W., & Clark, L. A. (2011). Positive and negative affect schedule—expanded version [Data

- set]. In PsycTESTS Dataset. American Psychological Association (APA).
<https://doi.org/10.1037/t04754-000>
- Dawadi A, Lucas ,Todd, Drolet ,Caroline E., et al. Healthcare provider cultural competency and receptivity to colorectal cancer screening among African Americans. *Psychol Health Med* 2022; 27: 2073–2084.
- De Choudhury, M., Pendse, S. R., & Kumar, N. (2023). Benefits and harms of large language models in digital mental health. In arXiv [cs.CL]. arXiv.
<https://doi.org/10.31234/osf.io/y8ax9>
- de-Moraes-Ribeiro, F. E., Moreno-Cámara, S., da-Silva-Domingues, H., Palomino-Moral, P. Á., & Del-Pino-Casado, R. (2024). Effectiveness of internet-based or mobile app interventions for family caregivers of older adults with dementia: A systematic review. *Healthcare (Basel, Switzerland)*, 12(15), 1494.
- DeepSeek-AI, Liu A, Feng B, et al. DeepSeek-V3 Technical Report. Epub ahead of print 18 February 2025. DOI: 10.48550/arXiv.2412.19437.
- Degnan, A., Baker, S., Edge, D., Nottidge, W., Noke, M., Press, C. J., Husain, N., Rathod, S., & Drake, R. J. (2018). The nature and efficacy of culturally-adapted psychosocial interventions for schizophrenia: a systematic review and meta-analysis. *Psychological Medicine*, 48(5), 714–727.
- Delfish, M., & Chadha, P. (2025). From Competence to Humility: Advancing Culturally Responsive Mental Health Care. *Psychiatric News*, 60(1).
- Denecke, K., Abd-Alrazaq, A., & Househ, M. (2021). Artificial intelligence for chatbots in mental health: Opportunities and challenges. In *Multiple Perspectives on Artificial Intelligence in Healthcare* (pp. 115–128). Springer International Publishing.

- Di Pompeo, I., D'Aurizio, G., Burattini, C., Bisegna, F., & Curcio, G. (2023). Positive mood induction to promote well-being and health: A systematic review from real settings to virtual reality. *Journal of Environmental Psychology*, 91(102095), 102095.
- Dinesh, D. N., Rao, M. N., & Sinha, C. (2024). Language adaptations of mental health interventions: User interaction comparisons with an AI-enabled conversational agent (Wysa) in English and Spanish. *Digital Health*, 10, 20552076241255616.
- Divina, M., Fan, X., Mikhail, N., Cheng, S.C., Ward, T.M., Hartzler, A.L., & Yuwen, W. (June 2021). Reducing negative affect through problem solving therapy with a conversational agent for family caregivers. Oral presentation, 15th International Family Nursing Conference. (Virtual).
- Dong, X., Chang, E.-S., & Simon, M. A. (2014). Physical function assessment in a community-dwelling population of U.S. Chinese older adults. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, 69 Suppl 2(Suppl 2), S31–S38.
- Dong, X., Chen, R., Li, C., & Simon, M. A. (2014). Understanding depressive symptoms among community-dwelling Chinese older adults in the Greater Chicago area. *Journal of Aging and Health*, 26(7), 1155–1171.
- Dongre P. Physiology-driven empathic large language models (EmLLMs) for mental health support. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM 2024:1–5.
- Dongre P, Behravan M, Gupta K, et al. Integrating physiological data with Large Language Models for empathic human-AI interaction. arXiv [eess.SP]. 2024.
- Duncan BL, Miller SD, Sparks JA, et al. The Session Rating Scale: Preliminary psychometric

- properties of a “working” alliance measure. *Journal of brief Therapy*. 2003;3(1):3-12.
- Elliott R, Bohart AC, Watson JC, et al. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy* 2018; 55: 399–410.
- Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1), 107–115. <https://doi.org/10.1111/j.1365-2648.2007.04569.x>
- Ennis, N., Shorer, S., Shoval-Zuckerman, Y., Freedman, S., Monson, C. M., & Dekel, R. (2020). Treating posttraumatic stress disorder across cultures: A systematic review of cultural adaptations of trauma-focused cognitive behavioral therapies. *Journal of Clinical Psychology*, 76(4), 587–611.
- Esmailzadeh, P. (2020). Use of AI-based tools for healthcare purposes: a survey study from consumers’ perspectives. *BMC Medical Informatics and Decision Making*, 20(1), 170.
- Eustis, E. H., LoPresti, J., Aguilera, A., & Schueller, S. M. (2023). Cultural responsiveness in technology-enabled services: Integrating culture into technology and service components. *Journal of Medical Internet Research*, 25, e45409.
- Ferré-Grau, C., Casado, M. S., Cid-Buera, D., LLeixà-Fortuño, M., Monteso-Curto, P., & Berenguer-Poblet, M. (2014). Caring for family caregivers: An analysis of a family-centered intervention. *Revista Da Escola de Enfermagem Da U S P*, 48 Spec No(spe), 87–94.
- Filienko D, Wang Y, Jazmi CE, et al. Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. Epub ahead of print 27 August 2024. DOI: 10.48550/arXiv.2409.00112.
- Filienko, D., Wang, Y., Jazmi, C. E., Xie, S., Cohen, T., De Cock, M., & Yuwen, W. (2024). Toward Large Language Models as a therapeutic tool: Comparing prompting techniques

- to improve GPT-delivered Problem-Solving Therapy. AMIA Annual Symposium Proceedings, 2024, 417–426.
- Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Mental Health*, 4(2), e19.
- Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial. *JMIR Ment Health* 2018;5(4):e64
- Gale M. Lucas, Jonathan Gratch, Aisha King, Louis-Philippe Morency, It's only a computer: Virtual humans increase willingness to disclose, *Computers in Human Behavior*, Volume 37, 2014, Pages 94-100, ISSN 0747-5632, <https://doi.org/10.1016/j.chb.2014.04.043>
- Graham JR, Sorenson S, Hayes-Skelton SA. Enhancing the Cultural Sensitivity of Cognitive Behavioral Interventions for Anxiety in Diverse Populations. *The Behavior Therapist*. 2013 Jun;36(5):101-108. PMID: 25392598; PMCID: PMC4225557.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). The Llama 3 herd of models. In arXiv [cs.AI]. arXiv. <http://arxiv.org/abs/2407.21783>
- Green, J. G., McLaughlin, K. A., Fillbrunn, M., Fukuda, M., Jackson, J. S., Kessler, R. C., Sadikova, E., Sampson, N. A., Vilsaint, C., Williams, D. R., Cruz-Gonzalez, M., & Alegría, M. (2020). Barriers to mental health service use and predictors of treatment drop out: Racial/ethnic variation in a population-based study. *Administration and Policy in*

- Mental Health, 47(4), 606–616.
- Griner, D., & Smith, T. B. (2006). Culturally adapted mental health intervention: A meta-analytic review. *Psychotherapy (Chicago, Ill.)*, 43(4), 531–548.
- Guo F, Hanley T. Adapting cognitive behavioral therapy to meet the needs of Chinese clients: Opportunities and challenges. *Psych J*. 2015;4:55–65.
- Guo, M., & Gan Y. (2010). Reliability and validity of the Chinese version of Positive and Negative Affect Schedule-Expanded in 660 college students. *Chinese Mental Health Journal*, 24(7), 524–528.
- Guo, M., Kim, S., & Dong, X. (2019). Sense of Filial Obligation and Caregiving Burdens Among Chinese Immigrants in the United States. *Journal of the American Geriatrics Society*, 67(S3), S564–S570.
- Gupta, A., & Jaffer, M. (2025, July 9). Prompt Engineering in 2025: The Latest Best Practices. Product Growth. <https://www.news.aakashg.com/p/prompt-engineering>
- Hafkenscheid A, Duncan B, Miller SD. The Outcome and Session Rating Scales: A Cross-Cultural Examination of the Psychometric Properties of the Dutch Translation. Published Online First: 2010.
- Han B, Compton WM, Blanco C, Colpe LJ. Prevalence, Treatment, And Unmet Treatment Needs Of US Adults With Mental Health And Substance Use Disorders. *Health Aff (Millwood)*. 2017 Oct 1;36(10):1739-1747. doi: 10.1377/hlthaff.2017.0584. PMID: 28971918.
- Han, M., Diwan, S., & Sun, K. (2019). Exploring caregiving-related experiences among Chinese American and European American family caregivers of persons with mental illness. *Transcultural Psychiatry*, 56(3), 491–509.

Havaladar S, Singhal B, Rai S, et al. Multilingual Language Models are not Multicultural: A Case Study in Emotion. In: Barnes J, De Clercq O, Klinger R, eds. Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis. Toronto, Canada: Association for Computational Linguistics 2023:202–14.

Havaladar, S., Singhal, B., Rai, S., Liu, L., Guntuku, S. C., & Ungar, L. (2023). Multilingual Language Models are not Multicultural: A Case Study in Emotion. In J. Barnes, O. De Clercq, & R. Klinger (Eds.), Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis (pp. 202–214). Association for Computational Linguistics.

HB0452. (n.d.). Retrieved August 13, 2025, from
<https://le.utah.gov/~2025/bills/static/HB0452.html>

Health Resources & Services Administration. (2023). Designated Health Professional Shortage Areas Statistics: First Quarter of Fiscal Year 2023 Designated HPSA Quarterly Summary. U.S. Department of Health & Human Services.
<https://data.hrsa.gov/Default/GenerateHPSAQuarterlyReport>

Heinz, M. V., Mackin, D. M., Trudeau, B. M., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A. J., Griffin, T. Z., & Jacobson, N. C. (2025). Randomized trial of a generative AI chatbot for mental health treatment. *NEJM AI*, 2(4).
<https://doi.org/10.1056/aioa2400802>

Hello GPT-4o, <https://openai.com/index/hello-gpt-4o/> (accessed 23 March 2025).

Helms JE. An examination of the evidence in culturally adapted evidence-based or empirically supported interventions. *Transcult Psychiatry* 2015; 52: 174–197.

Herscovich D, Frank S, Lent H, et al. Challenges and Strategies in Cross-Cultural NLP. Epub

ahead of print 18 March 2022. DOI: 10.48550/arXiv.2203.10020.

Hillier L. Exploring the Nature of the Therapeutic Alliance in Technology-Based Interventions for Mental Health Problems. 2018.

Hispanic Patients' Double Burden: Lack of Health Insurance and Limited English. (2003, February).

<https://www.commonwealthfund.org/publications/fund-reports/2003/feb/hispanic-patients-double-burden-lack-health-insurance-and>

Hoermann S, McCabe KL, Milne DN, Calvo RA. Application of Synchronous Text-Based Dialogue Systems in Mental Health Interventions: Systematic Review. *J Med Internet Res*. 2017 Jul 21;19(8):e267. doi: 10.2196/jmir.7023. PMID: 28784594; PMCID: PMC5595406.

Hofmann, S. G. (2006). The importance of culture in cognitive and behavioral practice. *Cognitive and Behavioral Practice*, 13(4), 243–245.

Hofstede G, Hofstede GJ, Minkov M. *Cultures and Organizations: Software of the Mind*, Third Edition. McGraw Hill Professional, 2010.

Holland, J. M., Thompson, L. W., Tzuang, M., & Gallagher-Thompson, D. (2010). Psychosocial factors among Chinese American women dementia caregivers and their association with salivary cortisol: Results of an exploratory study. *Ageing International*, 35(2), 109–127.

Hong, Y. A., Shen, K., Lu, H. K., Chen, H., Gong, Y., Ta Park, V., & Han, H.-R. (2022). A Social Media-Based Intervention for Chinese American Caregivers of Persons With Dementia: Protocol Development. *JMIR Aging*, 5(3), e40171.

Hong, Y. A., Shen, K., Han, H.-R., Ta Park, V., Lu, H. K., & Cleaveland, C. (2024). "It's a lonely journey": caregiving experiences and psychosocial distress among Chinese

- American dementia family caregivers. *Aging & Mental Health*, 28(3), 466–472.
- Hsieh, H.-F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.
- <https://www.nami.org/about-mental-illness/mental-health-by-the-numbers/#:~:text=22.8%25%20of%20U.S.%20adults%20experienced,represents%201%20in%2020%20adults>
- Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Dingjie, S., Chen, Z., Alharthi, M., An, B., He, J., Liu, Z., Chen, J., Li, J., Wang, B., Zhang, L., Sun, R., Wan, X., Li, H., & Xu, J. (2024). AceGPT, localizing large language models in Arabic. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 8139–8163.
- Hwang, W.-C., Wood, J. J., Lin, K.-M., & Cheung, F. (2006). *Cognitive-Behavioral Therapy With Chinese Americans: Research, Theory, and Clinical Practice*. *Cognitive and Behavioral Practice*, 13(4), 293–303.
- Hwang, W.-C. (2006). The psychotherapy adaptation and modification framework: application to Asian Americans. *The American Psychologist*, 61(7), 702–715.
- Hwang, W.-C. (2009). The Formative Method for Adapting Psychotherapy (FMAP): A community-based developmental approach to culturally adapting therapy. *Professional Psychology, Research and Practice*, 40(4), 369–377.
- Inkster, B., Sarda, S., & Subramanian, V. (2018). An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth and uHealth*, 6(11), e12106.
- Jang, Y., Park, J., Choi, E. Y., Cho, Y. J., Park, N. S., & Chiriboga, D. A. (2022). Social isolation in Asian Americans: risks associated with socio-demographic, health, and immigration

- factors. *Ethnicity & Health*, 27(6), 1428–1441.
- Jiang, L., Sorensen, T., Levine, S., & Choi, Y. (2024). Can language models reason about individualistic human values and preferences? In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2410.03868>
- Johnson, R. L., Pistilli, G., Menéndez-González, N., Duran, L. D. D., Panai, E., Kalpokiene, J., & Bertulfo, D. J. (2022). The Ghost in the Machine has an American accent: value conflict in GPT-3. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2203.07785>
- Jones JM, Carter RT. Racism and white racial identity: merging realities. In: Bowser BP, Hunt RG, editors. *Impacts of racism on White Americans*. 2nd edition. Thousand Oaks, CA: Sage Publications; 1996:1-23
- Jones-Smith, E. (2019). *Culturally diverse counseling: Theory and practice*. SAGE Publications, Inc.
- Karliner, L. S., Napoles-Springer, A. M., Schillinger, D., Bibbins-Domingo, K., & Pérez-Stable, E. J. (2008). Identification of limited English proficient patients in clinical care. *Journal of General Internal Medicine*, 23(10), 1555–1560.
- Kasirzadeh, A., & Gabriel, I. (2023). In conversation with artificial intelligence: Aligning language models with human values. *Philosophy & Technology*, 36(2), 1–24.
- Kearns, W. R., Kaura, N., Divina, M., Vo, C., Si, D., Ward, T., & Yuwen, W. (2020). A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health Counseling Dialog. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9.
- Keleg A, Magdy W. DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models. arXiv [cs.CL]. 2023.

- Khawaja, Z., & Bélisle-Pipon, J.-C. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, 5, 1278186.
- Kim, J. H. (2022). Grandparenting, filial piety, and well-being of Chinese-American older adults. *Current Sociology. La Sociologie Contemporaine*, 70(6), 860–879.
- Kirmayer LJ. Rethinking cultural competence. *Transcult Psychiatry* 2012 Apr 16;49(2):149-164 [doi:10.1177/1363461512444673] [Medline: 22508634]
- Knipping, D., Garnett, A., & Jiang, B. B. (2023). Access and use of services by caregivers of older adults: A scoping review of cultural and linguistic diversity. *Journal of Applied Gerontology: The Official Journal of the Southern Gerontological Society*, 42(7), 1672–1686.
- Kung, W. V. (2004). Cultural and Practical Barriers to Seeking Mental Health Treatment for Chinese Americans. *Journal of Community Psychology*, 32(1), 27–43.
<https://doi.org/10.1002/jcop.10077>
- Kushniruk, A. W., & Patel, V. L. (2004). Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of Biomedical Informatics*, 37(1), 56–76.
- Lai, D. W. L. (2010). Filial piety, caregiving appraisal, and caregiving burden. *Research on Aging*, 32(2), 200–223.
- Lawrence HR, Schneider RA, Rubin SB, et al. The opportunities and risks of large language models in mental health. *JMIR Ment Health*. 2024;11:e59479.
- Lee Anna Clark and David Watson. 1994. The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form. Retrieved September 12, 2024 from

<https://iro.uiowa.edu/esploro/outputs/other/9983557488402771>.

- Lee EE, Torous J, De Choudhury M, Depp CA, Graham SA, Kim HC, Paulus MP, Krystal JH, Jeste DV. Artificial Intelligence for Mental Health Care: Clinical Applications, Barriers, Facilitators, and Artificial Wisdom. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2021 Sep;6(9):856-864. doi: 10.1016/j.bpsc.2021.02.001. Epub 2021 Feb 8. PMID: 33571718; PMCID: PMC8349367.
- Lee HS, Wright C, Ferranto J, Buttimer J, Palmer CE, Welchman A, Mazor KM, Fisher KA, Smelson D, O'Connor L, Fahey N and Soni A (2025) Artificial intelligence conversational agents in mental health: Patients see potential, but prefer humans in the loop. *Front. Psychiatry* 15:1505024. doi: 10.3389/fpsyt.2024.1505024
- Lee, J. J., & Miller, S. E. (2013). A Self-Care Framework for Social Workers: Building a Strong Foundation for Practice. *Families in Society: The Journal of Contemporary Human Services*, 94(2), 96–103.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *arXiv [cs.CL]*. arXiv.
<http://arxiv.org/abs/2005.11401>
- Li H, Zhang R, Lee Y-C, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit Med*. 2023;6:236.
- Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024a). CultureLLM: Incorporating cultural differences into large language models. *arXiv.Org*.
<https://doi.org/10.48550/ARXIV.2402.10946>
- Li, C., Teney, D., Yang, L., Wen, Q., Xie, X., & Wang, J. (2024b). CulturePark: Boosting

- cross-cultural understanding in large language models. In arXiv [cs.AI]. arXiv.
<http://arxiv.org/abs/2405.15145>
- Li, H. (2004). Barriers to and unmet needs for supportive services: experiences of Asian-American caregivers. *Journal of Cross-Cultural Gerontology*, 19(3), 241–260.
- Lim, S., Mohaimin, S., Min, D., Roberts, T., Sohn, Y.-J., Wong, J., Sivanesathurai, R., Kwon, S. C., & Trinh-Shevrin, C. (2020). Alzheimer’s disease and its related dementias among Asian Americans, native Hawaiians, and pacific islanders: A scoping review. *Journal of Alzheimer’s Disease: JAD*, 77(2), 523–537.
- Liu, C., Koto, F., Baldwin, T., & Gurevych, I. (2024). Are multilingual LLMs culturally-diverse reasoners? An investigation into multicultural proverbs and sayings. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2016–2039.
- Liu, H., Peng, H., Song, X., Xu, C., & Zhang, M. (2022). Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness. *Internet Interventions*, 27(100495), 100495.
- Liu, M., Jiang, T., Yu, K., Wu, S., Jordan-Marsh, M., & Chi, I. (2020). Care Me Too, a Mobile App for Engaging Chinese Immigrant Caregivers in Self-Care: Qualitative Usability Study. *JMIR Formative Research*, 4(12), e20325.
- Lucas, T., Michalopoulou, G., Falzarano, P., Menon, S., & Cunningham, W. (2008). Healthcare provider cultural competency: development and initial validation of a patient report measure. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, 27(2), 185–193.
- MacNeill AL, Doucet S, Luke A. Effectiveness of a mental health chatbot for people with

- chronic diseases: Randomized controlled trial. *JMIR Form Res.* 2024;8:e50025.
- Maguire, M. (2001). Methods to support human-centred design. *International Journal of Human-Computer Studies*, 55(4), 587–634.
- Mandal, A., Chakraborty, T., & Gurevych, I. (2025). Towards privacy-aware mental health AI models: Advances, challenges, and opportunities. In *arXiv [cs.CL]*. *arXiv*.
<http://arxiv.org/abs/2502.00451>
- Manning, L., Ferris, M., Rosario, C. N., Prues, M., & Bouchard, L. (2019). Spiritual resilience: Understanding the protection and promotion of well-being in the later life. *Journal of Religion, Spirituality & Aging*, 31(2), 168–186.
- Marko, J. G. O., Neagu, C. D., & Anand, P. B. (2025). Examining inclusivity: the use of AI and diverse populations in health and social care: a systematic review. *BMC Medical Informatics and Decision Making*, 25(1), 57.
- Masala, M., Ilie-Ablachim, D., Dima, A., Corlatescu, D. G., Zavelca, M.-A., Olaru, O., Terian, S.-M., Terian, A., Leordeanu, M., Velicu, H., Popescu, M., Dascalu, M., & Rebedea, T. (2024). “vorbești românește?” A recipe to train powerful Romanian LLMs with English instructions. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 11632–11647.
- Masoud, R., Liu, Z., Ferienc, M., Treleaven, P. C., & Rodrigues, M. R. (2025). Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede’s Cultural Dimensions. *Proceedings of the 31st International Conference on Computational Linguistics*, 8474–8503.
- Mayo, C. D., Kenny, R., Scarapicchia, V., Ohlhauser, L., Syme, R., & Gawryluk, J. R. (2021). Aging in place: Challenges of older adults with self-reported cognitive decline. *Canadian*

- Geriatrics Journal: CGJ, 24(2), 138–143.
- Mei, L., Yao, J., Ge, Y., Wang, Y., Bi, B., Cai, Y., Liu, J., Li, M., Li, Z.-Z., Zhang, D., Zhou, C., Mao, J., Xia, T., Guo, J., & Liu, S. (2025). A survey of Context Engineering for Large Language Models. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2507.13334>
- Miyawaki, C. E. (2015). A Review of Ethnicity, Culture, and Acculturation Among Asian Caregivers of Older Adults (2000-2012). *SAGE Open*, 5(1).
<https://doi.org/10.1177/2158244014566365>
- Miyawaki, C. E. (2020). Caregiving Attitudes and Needs of Later-generation Chinese-American Family Caregivers of Older Adults. *Journal of Family Issues*, 41(12), 2377–2399.
- Mohr, D. C., Schueller, S. M., Riley, W. T., Brown, C. H., Cuijpers, P., Duan, N., Kwasny, M. J., Stiles-Shields, C., & Cheung, K. (2015). Trials of intervention principles: Evaluation methods for evolving behavioral intervention technologies. *Journal of Medical Internet Research*, 17(7), e166.
- Mongelli, F., Georgakopoulos, P., & Pato, M. T. (2020). Challenges and opportunities to meet the mental health needs of underserved and disenfranchised populations in the United States. *Focus (American Psychiatric Publishing)*, 18(1), 16–24.
- Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 599–627.
- Mosher, D. K., Hook, J. N., Captari, L. E., Davis, D. E., DeBlaere, C., & Owen, J. (2017). Cultural humility: A therapeutic framework for engaging diverse clients. *Practice Innovations (Washington, D.C.)*, 2(4), 221–233.

- Mukherjee, S., Adilazuarda, M. F., Sitaram, S., Bali, K., Aji, A. F., & Choudhury, M. (2024). Cultural conditioning or placebo? On the effectiveness of Socio-demographic prompting. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15811–15837.
- Munder, T., Wilmers, F., Leonhart, R., Linster, H. W., & Barth, J. (2010). Working Alliance Inventory-Short Revised (WAI-SR): psychometric properties in outpatients and inpatients. *Clinical Psychology & Psychotherapy*, 17(3), 231–239.
- Naeem F, Phiri P, Rathod S, et al. Cultural adaptation of cognitive–behavioural therapy. *BJPsych Advances*. 2019;25:387–95.
- Nagoudi EMB, Abdul-Mageed M, Elmadany A, et al. JASMINE: Arabic GPT models for few-shot learning. *arXiv [cs.CL]*. 2022.
- Napoles, A. M., Chadiha, L., Eversley, R., & Moreno-John, G. (2010). Reviews: developing culturally sensitive dementia caregiver interventions: are we there yet? *American Journal of Alzheimer's Disease and Other Dementias*, 25(5), 389–406.
- National Institute of Mental Health. Mental illness. National Institute of Mental Health. Published March 2023. <https://www.nimh.nih.gov/health/statistics/mental-illness>.
- Nguyen, X.-P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., Chia, Y. K., Li, X., Wang, J., Tan, Q., Cheng, L., Chen, G., Deng, Y., Yang, S., Liu, C., Zhang, H., & Bing, L. (2024). SeaLLMs - large language models for southeast Asia. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 294–304.
- Nápoles AM, Santoyo-Olsson J, Farren G, et al. The patient-reported Clinicians' Cultural Sensitivity Survey: a field test among older Latino primary care patients. *Health Expect*

2012; 15: 63–77.

Okazaki S, Tanaka-Matsumi J. Cultural considerations in cognitive-behavioral assessment.

Culturally responsive cognitive-behavioral therapy: Assessment, practice, and supervision. Washington: American Psychological Association 2006:247–66.

On-demand support to patients and their caregivers through Cass. (n.d.). Retrieved August 14, 2025, from

https://www.cass.ai/blog/psychological-artificial-intelligence-service-cass-delivering-on-demand-support-to-patients-and-their-caregivers-technical-report?utm_source=chatgpt.com

Orlinsky, D. E., Ronnestad, M. H., & Willutzki, U. (2004) Fifty years of psychotherapy process-outcome research: Continuity and change. Bergin and Garfield's handbook of psychotherapy and behavior change. ;307–89.

Owen L, Tripathi V, Kumar A, et al. Komodo: A linguistic expedition into Indonesia's regional languages. arXiv [cs.CL]. 2024.

Owen, J., Imel, Z., Tao, K. W., Wampold, B., Smith, A., & Rodolfa, E. (2011). Cultural ruptures in short-term therapy: Working alliance as a mediator between clients' perceptions of microaggressions and therapy outcomes. *Counselling and Psychotherapy Research*, 11(3), 204–212.

Palmer CE, Marshall E, Millgate E, Warren G, Ewbank M, Cooper E, Lawes S, Smith A, Hutchins-Joss C, Young J, Bouazzaoui M, Margoum M, Healey S, Marshall L, Mehew S, Cummins R, Tablan V, Catarino A, Welchman AE, Blackwell AD. Combining Artificial Intelligence and Human Support in Mental Health: Digital Intervention With Comparable Effectiveness to Human-Delivered Care. *J Med Internet Res* 2025;27:e69351

Parra Cardona, J. R., Domenech-Rodriguez, M., Forgatch, M., Sullivan, C., Bybee, D., Holtrop, K., Escobar-Chew, A. R., Tams, L., Dates, B., & Bernal, G. (2012). Culturally adapting an evidence-based parenting intervention for Latino immigrants: the need to integrate fidelity and cultural relevance. *Family Process*, 51(1), 56–72.

Patient Health Questionnaire-2 (PHQ-2). (n.d.). Retrieved January 20, 2025, from <https://www.hiv.uw.edu/page/mental-health-screening/phq-2>

Pawar, S., Park, J., Jin, J., Arora, A., Myung, J., Yadav, S., Haznitrama, F. G., Song, I., Oh, A., & Augenstein, I. (2024). Survey of cultural awareness in language models: Text and beyond. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2411.00860>

Perera C, Salamanca-Sanabria A, Caballero-Bernal J, et al. No implementation without cultural adaptation: a process for culturally adapting low-intensity psychological interventions in humanitarian settings. *Confl Health* 2020; 14: 46.

Pinquart, M., & Sörensen, S. (2003). Differences between caregivers and noncaregivers in psychological health and physical health: a meta-analysis. *Psychology and Aging*, 18(2), 250–267.

Ponzo S, Morelli D, Kawadler JM, et al. Efficacy of the digital therapeutic mobile app BioBase to reduce stress and improve mental well-being among university students: Randomized controlled trial. *JMIR MHealth UHealth*. 2020;8:e17767.

Prochaska JJ, Vogel EA, Chieng A, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug Alcohol Depend*. 2021;227:108986.

Ramezani A, Xu Y. Knowledge of cultural moral norms in large language models. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

- Long Papers). Toronto, Canada: Association for Computational Linguistics 2023:428–46.
- Ramirez M, Duran MC, Penfold RB, et al. STAR-Caregivers Virtual Training and Follow-up: a cultural adaptation for Latino caregivers of people with dementia. *Transl Behav Med*. Published Online First: 2 May 2023. doi: 10.1093/tbm/ibad028
- Rao A, Yerukola A, Shah V, et al. NormAd: A Framework for Measuring the Cultural Adaptability of Large Language Models. Epub ahead of print 6 March 2025. DOI: 10.48550/arXiv.2404.12464.
- Rathod S, Phiri P, Naeem F. An evidence-based framework to culturally adapt cognitive behaviour therapy. *the Cognitive Behaviour Therapist*. 2019;12:e10.
- Resnicow K, Baranowski T, Ahluwalia JS, et al. Cultural Sensitivity in Public Health: Defined and Demystified. *Ethn Dis* 1999; 9: 10–21.
- Ronald I, Basáñez M, Jaime D-M, et al. Human beliefs and values: A cross-cultural sourcebook based on the 1999-2002 value surveys.
- Roque, N. A., & Boot, W. R. (2018). A new tool for assessing mobile device proficiency in older adults: The Mobile Device Proficiency Questionnaire. *Journal of Applied Gerontology: The Official Journal of the Southern Gerontological Society*, 37(2), 131–156.
- Ruiz AB and NG. Key facts about Asian origin groups in the U.S. Pew Research Center, <https://www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-origin-groups-in-the-u-s/> (2021, accessed 23 March 2025).
- Rystrøm, J., Kirk, H. R., & Hale, S. (2025). Multilingual != multicultural: Evaluating gaps between multilingual capabilities and cultural alignment in LLMs. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2502.16534>
- Salamanca-Sanabria A, Richards D, Timulak L. Adapting an internet-delivered intervention for

depression for a Colombian college student population: An illustration of an integrative empirical approach. *Internet Interv* 2019; 15: 76–85.

Saldana, J. (2023, September 28). *The Coding Manual for Qualitative Researchers*. SAGE Publications Inc.

<https://us.sagepub.com/en-us/nam/the-coding-manual-for-qualitative-researchers/book273583>

Santiago-Rivera AL, Altarriba J. The role of language in therapy with the Spanish-English bilingual client. *Prof Psychol Res Pr*. 2002;33:30–8.

Saxon, L., Henriksson, S., Kvarnström, A., & Hiltunen, A. J. (2017). Affective changes during Cognitive Behavioural Therapy-as measured by PANAS. *Clinical Practice and Epidemiology in Mental Health: CP & EMH*, 13(1), 115–124.

Scholich T, Barr M, Wiltsey Stirman S, Raj S. A Comparison of Responses from Human Therapists and Large Language Model–Based Chatbots to Assess Therapeutic Communication: Mixed Methods Study. *JMIR Ment Health* 2025;12:e69709

Schumacher, K., Beck, C. A., & Marren, J. M. (2006). FAMILY CAREGIVERS: Caring for older adults, working with their families. *The American Journal of Nursing*, 106(8), 40.

Selina Meyer, David Elswailer. LLM-based conversational agents for behaviour change support: A randomised controlled trial examining efficacy, safety, and the role of user behaviour, *International Journal of Human-Computer Studies*, Volume 200, 2025, 103514,ISSN 1071-5819.

Semere, W., Nápoles, A. M., Gregorich, S., Livaudais-Toman, J., & Karliner, L. (2019). Caregiving for Older Adults with Limited English Proficiency: Transitioning from Hospital to Home. *Journal of General Internal Medicine*, 34(9), 1744–1750.

- Sharma, A., Miner, A., Atkins, D., & Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276.
- She Z, Shi Y, Duncan BL, et al. Psychometric properties and longitudinal invariance of the session rating scale in Chinese clinical samples. *Curr Psychol*. 2023;42:3650–7.
- Shi, J. M., Yoo, D. W., Wang, K., Rodriguez, V. J., Karkar, R., & Saha, K. (2025). Mapping caregiver needs to AI chatbot design: Strengths and gaps in mental health support for Alzheimer’s and dementia caregivers. In *arXiv [cs.HC]*. arXiv. <http://arxiv.org/abs/2506.15047>
- Shi, W., Li, R., Zhang, Y., Ziemis, C., Yu, S., Horesh, R., Paula, R. A. D., & Yang, D. (2024). CultureBank: An online community-driven knowledge base towards culturally aware language technologies. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 4996–5025.
- Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, 36(6), 495–504.
- Shonfeld-Ringel S. The effect of culture on the working alliance between Asian American clients and Western therapists. 2001.
- Simon, M. A., Chang, E.-S., Zhang, M., Ruan, J., & Dong, X. (2014). The prevalence of loneliness among U.S. Chinese older adults. *Journal of Aging and Health*, 26(7), 1172–1188.
- Solaiman I, Dennison C. Process for Adapting Language Models to Society (PALMS) with Values-Targeted Datasets. In: *Advances in Neural Information Processing Systems*.

- Curran Associates, Inc., pp. 5861–5873.
- Song, I., Pendse, S. R., Kumar, N., & De Choudhury, M. (2024). The typing cure: Experiences with Large Language Model chatbots for mental health support. In arXiv [cs.HC]. arXiv. <http://arxiv.org/abs/2401.14362>
- Spanhel K, Balci S, Feldhahn F, et al. Cultural adaptation of internet- and mobile-based interventions for mental disorders: a systematic review. *Npj Digit Med* 2021; 4: 1–18.
- Spitzer, D., Neufeld, A., Harrison, M., Hughes, K., & Stewart, M. (2003). Caregiving in transnational context: “my wings have been cut; Where can I fly?” *Gender & Society: Official Publication of Sociologists for Women in Society*, 17(2), 267–286.
- Stern, E., Breton, Z., Alexaline, M., Geoffroy, P. A., & Bungener, C. (2024). Redefining the relationship in digital care: A qualitative study of the Digital Therapeutic Alliance. *L’Encephale*. <https://doi.org/10.1016/j.encep.2024.02.011>
- Substance Abuse and Mental Health Services Administration (SAMHSA). Improving cultural competence: quick guide for clinicians. 2014. URL: <https://store.samhsa.gov/sites/default/files/d7/priv/sma16-4931.pdf> [accessed 2022-12-29]
- Sue DW, Sue D, Neville HA, Smith L. Counseling the culturally diverse: Theory and practice, 5th ed. (2008). *Counseling the Culturally Diverse: Theory and Practice*. <https://psycnet.apa.org/record/2007-13154-000>
- Sue, S., & Zane, N. (1987). The role of culture and cultural techniques in psychotherapy: A critique and reformulation. *The American Psychologist*, 42(1), 37–45.
- Sue, S., Fujino, D. C., Hu, L. T., Takeuchi, D. T., & Zane, N. W. (1991). Community mental health services for ethnic minority groups: a test of the cultural responsiveness

- hypothesis. *Journal of Consulting and Clinical Psychology*, 59(4), 533–540.
- Sun, H., Lin, Z., Zheng, C., Liu, S., & Huang, M. (2021). PsyQA: A Chinese Dataset for Generating Long Counseling Text for Mental Health Support. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 1489–1503). Association for Computational Linguistics.
- Tang, M. (2011). Can cultural values help explain the positive aspects of caregiving among Chinese American caregivers? *Journal of Gerontological Social Work*, 54(6), 551–569.
- Tao D, Awan-Scully R, Ash GI, et al. The effectiveness of dance movement interventions for older adults with mild cognitive impairment, Alzheimer’s disease, and dementia: A systematic scoping review and meta-analysis. *Ageing Res Rev* 2023; 92: 102120.
- Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), gae346.
- Thakkar A, Gupta A, De Sousa A. Artificial intelligence in positive mental health: a narrative review. *Front Digit Health*. 2024 Mar 18;6:1280235. doi: 10.3389/fdgth.2024.1280235. PMID: 38562663; PMCID: PMC10982476.
- The Recognize, Assist, Include, Support, and Engage (RAISE) Act Family Caregiving Advisory Council & The Advisory Council to Support Grandparents Raising Grandchildren. 2022 national strategy to support family caregivers.
- Thomas, L. (2016). *Psychotherapy dropout : the influence of ethnic identity and stigma on early termination* [University of Louisville]. <https://doi.org/10.18297/etd/2406>
- Tonja, A. L., Azime, I. A., Belay, T. D., Yigezu, M. G., Mehamed, M. A. A., Ayele, A. A., Jibril, E. C., Woldeyohannis, M. M., Kolesnikova, O., Slusallek, P., Klakow, D., & Yimam, S. M. (2024). EthioLLM: Multilingual Large Language Models for Ethiopian Languages

- with Task Evaluation. Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 6341–6352.
- Torous J, Benson NM, Myrick K, et al. Focusing on digital research priorities for advancing the access and quality of mental health. *JMIR Ment Health*. 2023;10:e47898.
- Tsai JL, Knutson B, Fung HH. Cultural variation in affect valuation. *J Pers Soc Psychol*. 2006;90:288–307.
- Tschuschke V, Koemeda-Lutz M, von Wyl A, et al. The Impact of Clients' and Therapists' Characteristics on Therapeutic Alliance and Outcome. *J Contemp Psychother* 2022; 52: 145–154.
- Tseng W-S. Culture and psychotherapy: Asian perspectives. *J Ment Health*. 2004;13:151–61.
- Ullrich HE. Culture, Empathy, and the Therapeutic Alliance. *Psychodyn Psychiatry* 2019; 47: 425–440.
- USAFacts. (2023, April 25). AAPI Demographics: Data on Asian American ethnicities, geography, income, and education. USAFacts.
<https://usafacts.org/articles/the-diverse-demographics-of-asian-americans/>
- Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can J Psychiatry*. 2019 Jul;64(7):456-464. doi: 10.1177/0706743719828977. Epub 2019 Mar 21. PMID: 30897957; PMCID: PMC6610568.
- Wallace, L. S., Rogers, E. S., Roskos, S. E., Holiday, D. B., & Weiss, B. D. (2006). Brief report: screening items to identify patients with limited health literacy skills. *Journal of General Internal Medicine*, 21(8), 874–877.

- Wampold BE. How important are the common factors in psychotherapy? An update. *World Psychiatry*. 2015;14:270–7.
- Wang, L., D., P., Carrington, D., S., M., Filienko, D., S., M., Jazmi, C. E., S., M., Xie, S. J., S., M., De Cock, M., D., P., Iribarren, S., D., P., Yuwen, W., & D, P. (2025). Large language model-powered conversational agent delivering Problem-Solving Therapy (PST) for family caregivers: Enhancing empathy and therapeutic alliance using in-context learning. In arXiv [cs.AI]. arXiv. <http://arxiv.org/abs/2506.11376>
- Wang, L., Smriti, D., Yuan, H., & Huh-Yoo, J. (2024). Artificial intelligence systems for supporting informal caregivers of people living with Alzheimer’s disease or related dementias: A systematic review. *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–11.
- Wang, Y., Liu, M., & Chi, I. (2021). An evaluation of self-care and caregiving training curriculum content for Chinese immigrant caregivers: A participatory design approach. *Journal of Applied Gerontology: The Official Journal of the Southern Gerontological Society*, 40(12), 1818–1827.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070.
- Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

- Xu, S., Lou, V. W. Q., Chi, I., Ng, W. C., Zhou, J., Huang, L.-K., Hok Ka Ma, C., & Jagasia, M. (2024). Validating interRAI Chinese self-reported carer needs (SCaN) assessment and predicting caregiving distress among informal Chinese caregivers of older adults. *BMC Geriatrics*, 24(1), 409.
- Yasukawa S, Tanaka T, Yamane K, et al. A chatbot to improve adherence to internet-based cognitive–behavioural therapy among workers with subthreshold depression: a randomised controlled trial. *BMJ Ment Health*; 27. Epub ahead of print 10 January 2024. DOI: 10.1136/bmjment-2023-300881.
- Yoo KM, Han J, In S, et al. HyperCLOVA X Technical Report. arXiv [cs.CL]. 2024.
- Yu, K., Jiang, H., Liu, M., Wu, S., Jordan-Marsh, M., & Chi, I. (2024). Chinese immigrant caregivers: Understanding their unmet needs and the co-design of an mHealth app. *La Revue Canadienne Du Vieillissement [Canadian Journal on Aging]*, 1–8.
- Zelvin, E., & Speyer, C. M. Online counseling skills, part I: Treatment strategies and skills for conducting counseling online. *Online counseling: A handbook for mental health professionals*. 2004;163–80.
- Zhao W, Mondal D, Tandon N, et al. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024:17696–706.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2306.05685>
- Zhou, L., Karidi, T., Liu, W., Garneau, N., Cao, Y., Chen, W., Li, H., & Hershcovich, D. (2025).

Does mapo tofu contain coffee? Probing LLMs for food-related cultural knowledge. In L. Chiruzzo, A. Ritter, & L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 9840–9867). Association for Computational Linguistics.

Appendices

Appendix 1. Interview Questions for Chapter 3

Appendix 1.1. Family Caregiver Interview Guide

Background Questions (Warm-up)

warm-up, get background information about the participant and the older family member he/she cared for.

- Who is the older family member you care for? What health conditions does [the family member] have?
- When did you start taking care of your [older family member]?
- Who else is involved in caring for your [older family member]? (Both informal and formal personnel.)

Part 1

Questions About Impact of Caregiving

- What is caregiving like for you? What are your primary responsibilities to care for your [older family member]?
- What do you do on a typical day? Out of these activities, which ones do you think are important/cannot miss/look forward to?
- How does caring for your [older family member] affect your typical day or week?
- Tell me about a time when you found it difficult to balance other important things in your life and your responsibility in caring for your [the older family member].
 - How did that make you feel?
 - What has been supportive or helpful in managing the tradeoff?
 - What support or tools would be helpful but were not available to you?

- How do you **feel** about caring for your [older family member]?
 - Can you describe a time when you felt good about caring for your [older family member]? How did you feel exactly? Why would you feel that?
 - Can you describe a time when you felt so good about caring for your [older family member]? How did you feel exactly? Why would you feel that?
 - If a person is there to console you, what can he/she say to make you feel understood?
 - What would be some ideal feelings for you?

Part 2

Questions About Self-care Practices

- What does self-care mean to you?
- When you are feeling [negative feeling] (sad, tired, stressed, upset, angry), what are some things you typically do to feel better?
- What is most important to your well-being?
- Tell me about a time when you found it difficult to balance the **things important to your well-being or self-care** and your responsibility in caring for your [the older family member].
 - How did that make you feel?
 - What has been supportive or helpful in managing the tradeoff?
 - What support or tools would be helpful but were not available to you?
- Have you thought about seeking mental health services when you are feeling stressed?
Why or why not?

Part 3

Questions About Digital Health and Chatbot Technology

- Do you use mobile apps in your daily life? Do you use it frequently?
- When you think of a mobile app to help you manage your self-care and improve your well-being, what would you picture the mobile app to be?

Explain the concept of 'Chatbot': Chatbot is a chat robot that acts like a human and can respond to you in real-time based on your input message.

- If there is a mobile app chatbot,
 - What kind of characteristics would you prefer a chatbot to have?
 - What could the chatbot do to better support your caregiving and self-care practices?
 - What would motivate you to use the chatbot regularly?

Appendix 1.2. Professional Staff Members Focus Group Interview Guide

Background Questions (Warm-up)

- What profession? What role are you in? (nurse? Professional caregiver? Staff member of a community organization?)
- What are your primary responsibilities in supporting families with Chinese American older adults?

Part 1

Questions about Caregiving

- What are some challenges that families of Chinese American older adults face?
- How is caregiving for Chinese Americans different from other American families?
- What kind of help or resources are you asked about most often? What tools or resources have you seen families use?
- How do values in Chinese culture contribute to the differences you see?
- How do existing tools or resources for family caregivers align or misalign with these Chinese values?

Part 2

Questions about Self-care

- What are some ways that Chinese American family caregivers take care of themselves?
- What kind of help or resources are you asked about most often?
- What tools or resources are there? Which tools or resources have you seen families use?
- How is family caregivers' self-care for Chinese Americans different from other American family caregivers?
- How do values in Chinese culture contribute to the differences you see?
- How do existing tools or resources for family caregivers' self-care align or misalign with these Chinese values?

Part 3**Digital Health Technology Brainstorm**

- When you picture a mobile app to help the families you work with manage their caregiving and self-care, what are some functions that the mobile app has?
- What would promote the usage of a mobile app for Chinese-American family caregivers?

Appendix 2. LLMs' System Prompts (Chapter 5)

Cultural Relevance Detection LLM:

Based on the user input below, determine whether it involves topics related to Chinese culture (such as communication with family members, language, family expectations, parents' feelings of loneliness, lack of understanding about health and social services, viewing self-care and relaxation as indulgent, lack of support from relatives or friends, values, family, immigration, etc.).

Context Relevance Validation LLM:

Please determine whether the following cultural context information is highly relevant to the user's three most recent inputs.

User Inputs: {chr(10).join([f"- {m}" for m in recent_user_inputs])}

Cultural Context Information: {retrieved_data}

Please respond with only one word: "Relevant" or "Not Relevant"

Therapist LLM with Cultural Prompting

Your responsibility is to guide the conversation with a caregiver ("user") to identify one significant symptom the caregiver is experiencing. **When responding to the caregiver, please fully consider their Chinese cultural background. Use language, values,**

and communication styles that align with Chinese culture to ensure your

message resonates with them. You will ask open-ended questions to identify and assess their challenges and stressors and improve their self-care. After you have identified one symptom that the caregiver can work on to improve their health, summarize the problem and understanding collaboratively in 2–3 sentences and check for accuracy. Avoid focusing on the care receiver. Remember, your job is to help the caregiver.

Use Motivational Interviewing (MI) techniques such as affirmation, reflection, emphasizing autonomy, giving information, normalizing, persuasion with permission, and seeking collaboration. Keeping in mind to cultivate change talk, soften sustain talk, strengthen partnerships, and show empathy. Ask only one question at a time, and do not question stack.

Here are some details:

During symptom assessment, here are some common dimensions to assess: timing (onset, duration, pattern), intensity/severity, location, distress, exacerbating/alleviating factors, and the degree of treatment relief. These dimensions help to characterize the nature and impact of symptoms on an individual's well-being and function. You can refer to the following elaborations on the dimensions when you conduct symptom assessment.

Timing:

This encompasses when the symptom first appeared (onset), how long it lasts (duration), and how it changes over time (pattern).

- When did the symptom first appear?

- How long has it lasted?
- Has it changed or varied over time?

Intensity/Severity:

This refers to the magnitude or strength of the symptom, often rated on a scale.

- How would you describe the intensity or impact of the symptom on a typical day?
- On a scale from 0 to 10, how bad does it feel at its worst?

Location or Type:

This describes where the symptom is felt, whether localized to a specific area or more widespread (diffuse or systemic).

- Where or how does the symptom manifest (physically, emotionally, cognitively)?
- Is it localized (e.g., headache, anxiety in social settings) or more generalized?

Distress: This refers to the emotional or psychological impact of the symptom, including the degree to which it bothers or disrupts the individual.

- How much does this symptom bother you or disrupt your daily life?
- What worries or concerns has it raised for you?

Symptom fluctuation:

Symptoms may fluctuate overtime or individuals may have tried certain coping skills or

treatments that changed the symptoms. Changes may include emotional change, behavioral change, cognitive changes, and physical symptom changes.

- is there a particular time during the day or week that the symptoms get better?
- What tends to make it worse or more noticeable?
- Are there things that help reduce it, even a little?
- Have you tried anything to manage it? (e.g., rest, therapy, medications, coping strategies)
- How well has that worked for you?

Impact on Life:

This refers to how the symptom affects an individual's daily activities, function, and quality of life.

- How is this affecting your daily activities, relationships, sleep, energy, or overall functioning?

Tone and Style Reminders:

- Use empathetic, nonjudgmental language.
- Reflect back key ideas (e.g., “It sounds like this has been really draining for you.”).
- Affirm strengths and autonomy (e.g., “You’ve been doing a lot just to keep things going.”).
- Normalize when appropriate (e.g., “Many caregivers feel overwhelmed—this isn’t just you.”).
- Ask for permission before offering guidance or suggestions.
- Foster collaboration and avoid giving direct advice unless invited.

Response Guidelines:

- Show a level of understanding that helps the client feel you truly grasp their situation
- Questions should be specific and easy to answer.
- Ask only one question at a time—do not ask multiple questions in a row.
- Always keep the focus on the client themselves, not on others (e.g., the care recipient).

Therapist LLM with Cultural Prompting **AND Cultural Context**

Therapist LLM with Cultural Prompting prompt (shown above) +

Response Guidelines:

- **Leverage the provided cultural context information to develop a deeper understanding of the client's cultural background and circumstances.**
- Show a level of understanding that helps the client feel you truly grasp their **cultural context** and situation.
- Questions should be specific and easy to answer.
- Ask only one question at a time—do not ask multiple questions in a row.
- Always keep the focus on the client themselves, not on others (e.g., the care recipient).

Cultural Context information: {Retrieved_data}

Appendix 3. Evaluation Questions (Chapter 6)

[Session Rating Scale (SRS)]

SRS - Relationship

0- I did not feel heard, understood, and respected.

10- I felt heard, understood, and respected.

SRS - Goals & Topics

0-We did not work on or talk about what I wanted to work on or talk about.

10 - We worked on and talked about what I wanted to work on or talk about.

SRS - Approach

0- The therapist's approach is not a good fit for me.

10 - The therapist's approach is a good fit for me.

SRS - Overall

0- There was something missing in this session.

10 - Overall, this session was right for me.

[User Satisfaction - Net Promoter Score (NPS)]

How likely are you to recommend this AI therapist to a family member or friend?

0 - Not at all likely

10 - Extremely likely

[Empathy]

Empathy- Emotional Reactions. To what extent did the therapist express warmth and care towards the user?

1-None

5-A great deal

Empathy- Interpretations. How accurately did the therapist reflect the user's feelings and experiences back to him/her?

1-Not at all

5-Extremely accurately

Empathy- Explorations. How well did the therapist help the user explore feelings the user hadn't initially expressed?

1-Not at all

5-Very well

[Cultural Competence Measure (CCM)]

CCM- Cultural Knowledge. How knowledgeable do you feel that the AI therapist is of Chinese culture?

1- Not at all

7-Knows a lot

CCM- Cultural Awareness. Do you feel as though the AI therapist is aware of the views that he or she may have towards clients with Chinese cultural background?

1-Not at all

7-Very aware

CCM- Well-equipped with Skills. Do you think that the AI therapist is well equipped to treat patients of Chinese cultural background?

1-Not at all

7-Very well equipped

CCM- Overall. Would you recommend the AI therapist to someone with Chinese cultural background?

1-Not at all

7-Highly recommend

[Cultural Relevance Questionnaire (CRQ)]

CRQ-1. The **people and cultural context** are reflected in the responses (e.g., social, political, economic, ethnic, historical).

1 - Not at all

5 - Very well

CRQ-2. The **goals** suggested in the response are tailored to work with the user from this cultural context (e.g., examples, personal stories).

1 - Not at all

5 - Very well

CRQ-3. The responses include written and oral communication that can be considered **dialects and jargon** relevant in this cultural context (e.g., saying, slang).

1 - Not at all

5 - Very well

CRQ-4. The responses involve **behavioral or emotional expressions** familiar to the cultural group being targeted.

1 - Not at all

5 - Very well

CRQ-5. The responses include **symbols and concepts** shared by the cultural group; for instance, cultural expressions of depression, ideas, or analogies about mental illness are included in the program.

1 - Not at all

5 - Very well